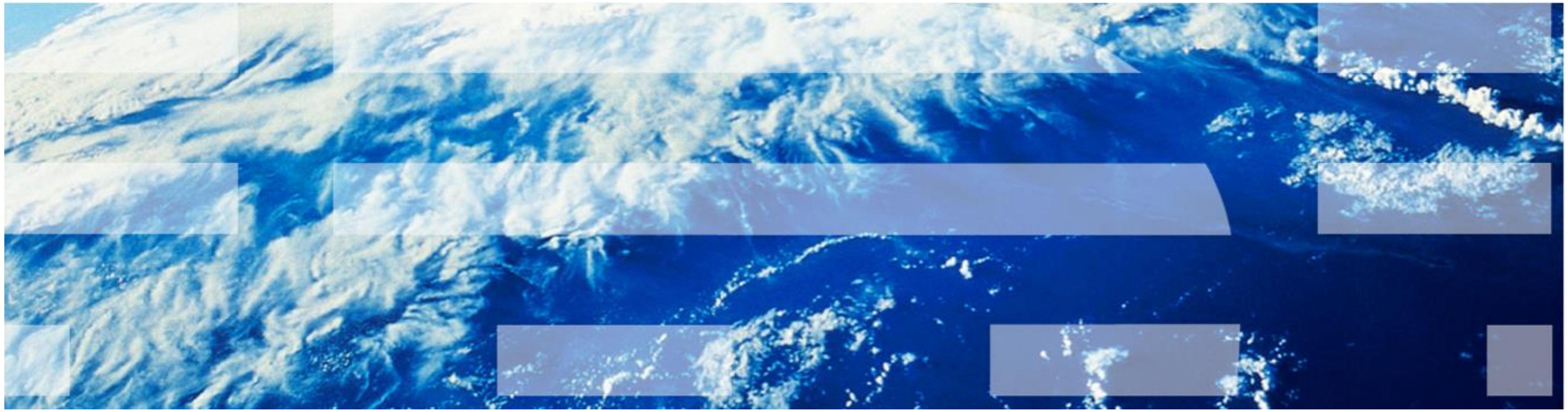
E6893 Big Data Analytics Project Proposal

Politics & Analytics

Sanjana Gopisetty - ssg2147
Saad Ahmed - sa3205
Jayni Chopda - jjc2253
Jivtesh Singh - jsc2226



November 19th, 2015

Motivation

Election season is coming around and our candidates are the hottest topics on social media. So, why not, use Big Data Technology to analyze election trends?

- Which candidate is being talked about the most?
- What are the sentiments for the candidates and the major parties?
- Where geographically are they being most talked about?
- Compare popularity and ranking in social media to polling data.
- Observe the change in trends over time.

Dataset, Algorithms and Tools

Dataset : Twitter API

Tracking certain keywords and collecting data like user information, geolocation and the text.

Algorithms:

1) Stream live tweets based on keywords using Twitter API and Node.js. We can also use Flume, Hive and HDFS if the data set is very large.

1) Determine the popularity of candidate/party based on the number of tweets.

1) Heat-Map based query based on keywords to display geolocation of candidates/ parties popularity in a particular area

1) Apply sentiment analysis on each tweet by computing the average sentiment score of each tweet and then compute the average sentiment score of all the tweets collected. Also do sentiment analysis on Internet data by using Alchemy API.

Tools: Alchemy API, Twitter API, Google Maps API, Node.js

Current Progress, Schedule and Expected Contributions

Progress:

- Framework discussion related to our proposed project
- Have Successfully used Twitter API to stream data

To-Do List	Expected Contributions	Schedule
Collect Tweet Data Sentiment Analysis	Sanjana and Saad	2 weeks
Heat Map Display Trend Analysis	Jivtesh and Jayni	2 weeks

References

1. https://dev.twitter.com/rest/public
1. https://developers.google.com/maps/documentation/javascript/examples/layer-heatmap
1. https://en.wikipedia.org/wiki/Sentiment_analysis
1. http://www.alchemyapi.com/api/sentiment/textc.html

THANK YOU

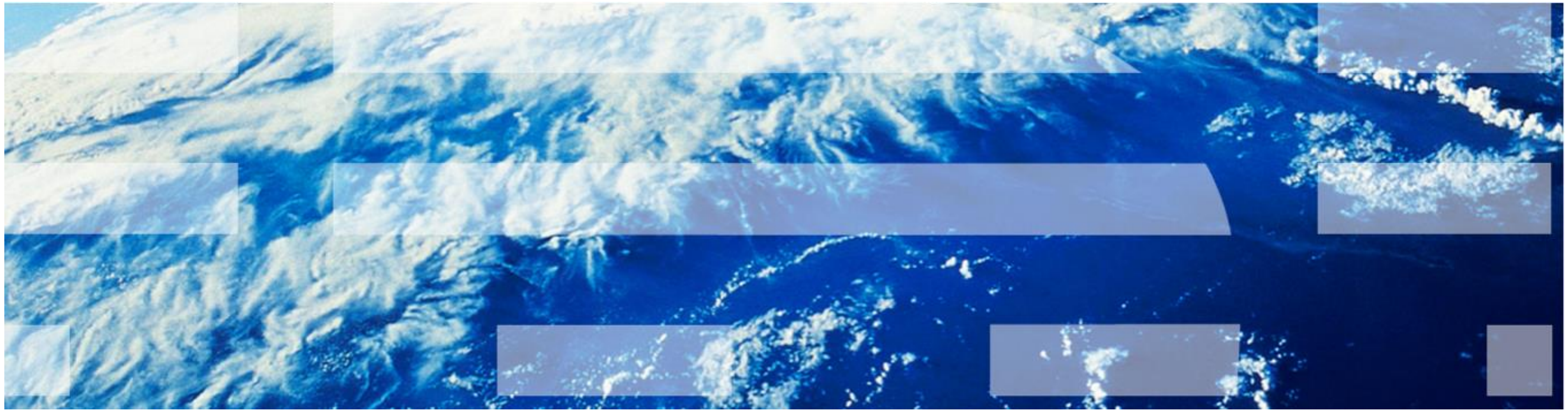
No questions Please ! :)

E6893 Big Data Analytics Project Proposal:

Uber *Max!*

Team: Munan Cheng, Lingqiu Jin, Chuwen Xu

UNI: mc4081, lj2379, cx2178



November 19th, 2015

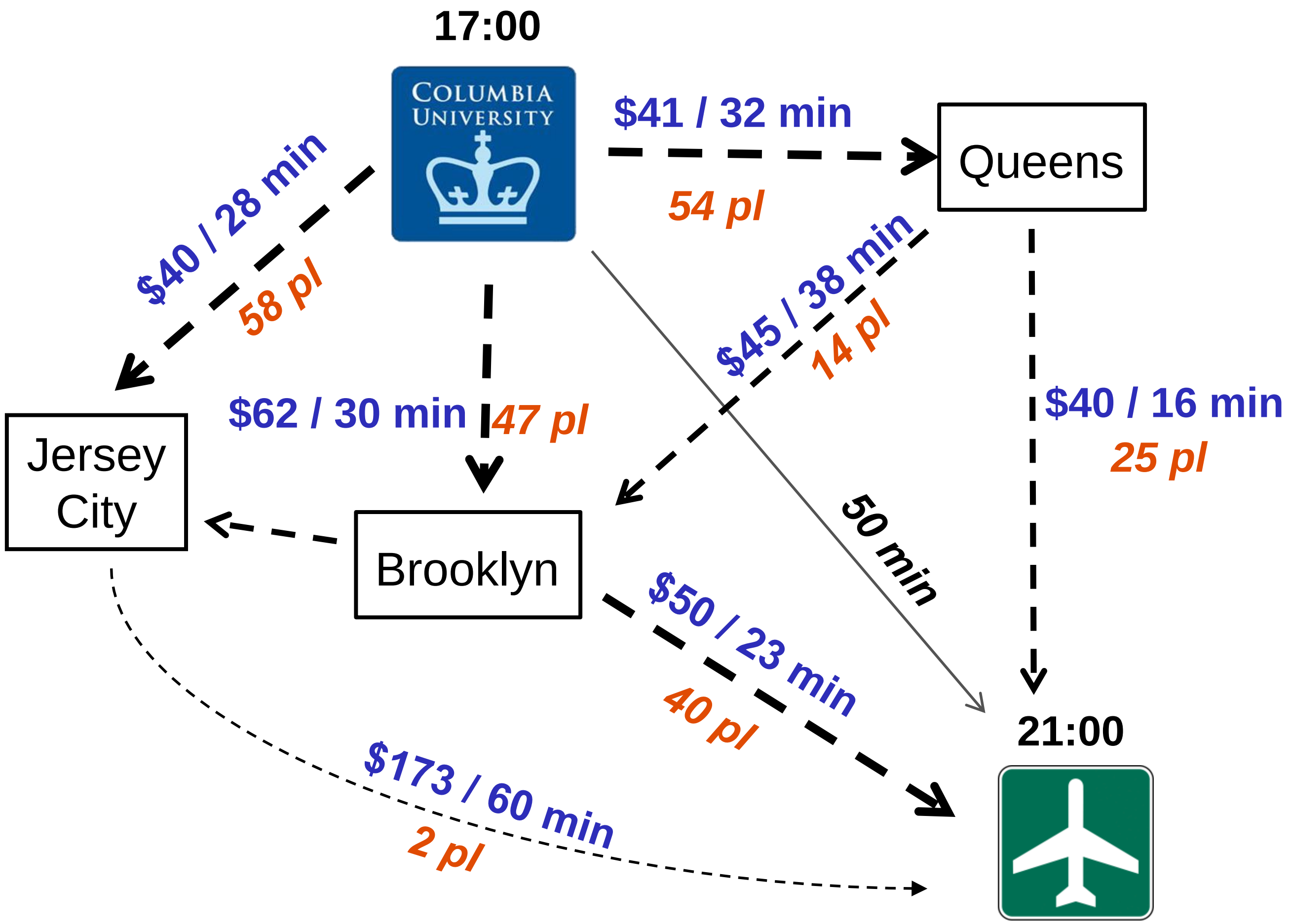
Motivation

Tom has just finished school at **5 p.m**. and has to pick his friend up at **airport** at **9 p.m.** In this period of 4 hours, he plans to **make some money** as an **Uber driver**. But now, **whom should he offer the ride to?**

? ***Trip time***

? ***Fares***

? ***Demands***



Dataset, Algorithms and Tools

Dataset:

- NYC Taxi Data
  - Dataset of taxi trips during last 7 years

Algorithms:

- (Pick-up & Drop-off) Estimation of
  - Demand, Fares, Trip Time
  - Time sensitive
- Route planning
  - Dynamic Programming

Tools:

- AWS
- Hadoop + Hive + Mahout
- Neo4J

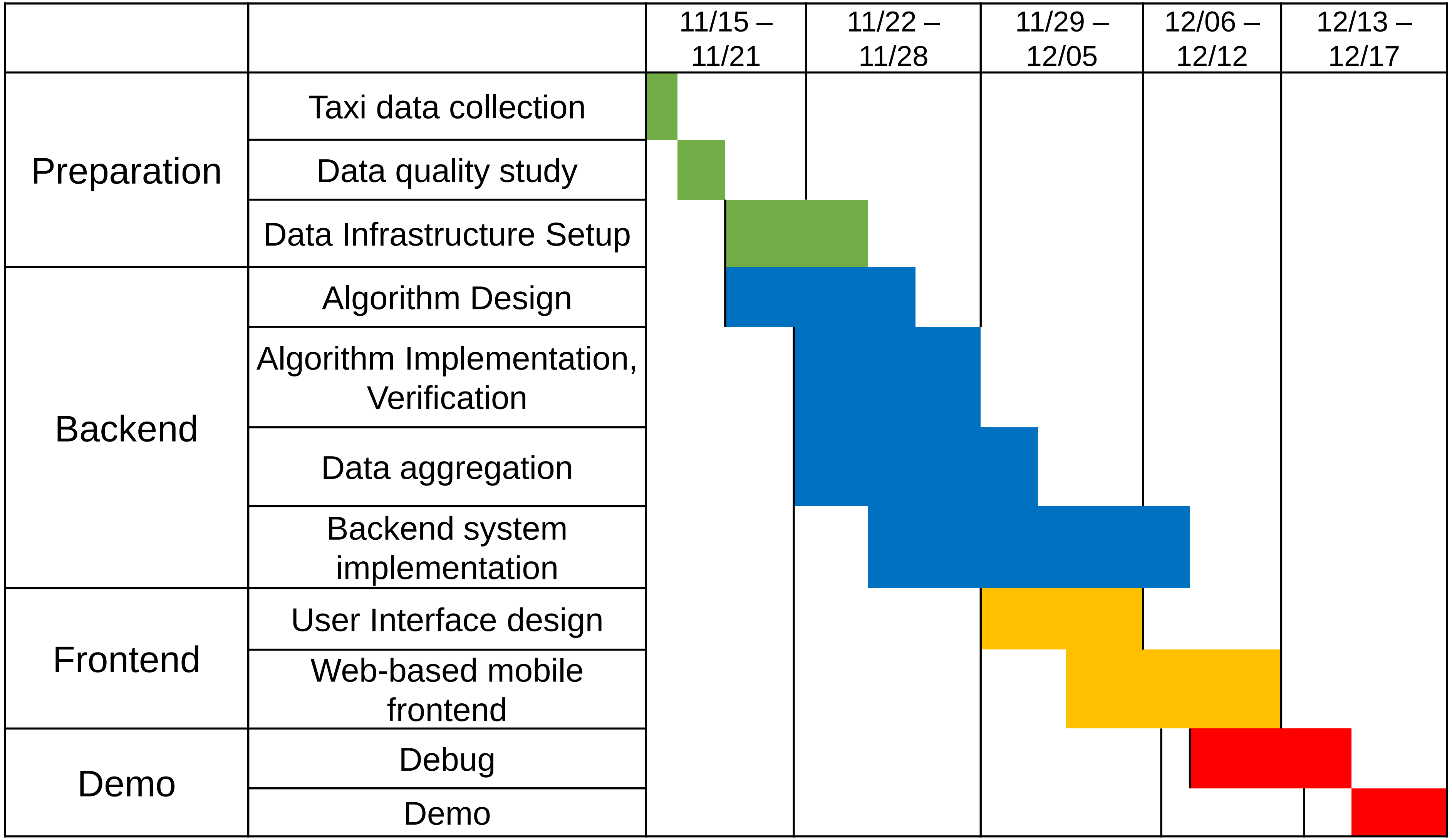
Input
Time constraints
Start, end location

Big Data
Fares
Trip time
Demand

Output
Preferred
next destination
that maximizes
the total profit!

http://www.nyc.gov/html/tlc/html/about/statistics.shtml

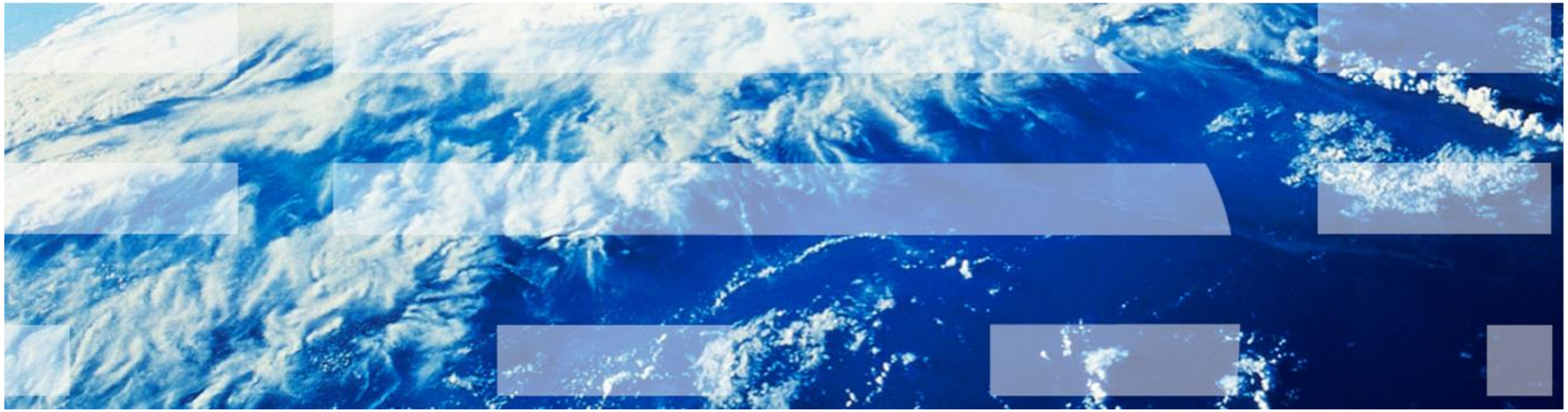
Current Progress, Schedule and Expected Contributions



E6893 Big Data Analytics Project Proposal:

Waste Management using Big Data

Hadeel Albahar, Shreya Yathish Kumar, Harnoor Singh Powar



November 19th, 2015

Motivation

- **Help New Yorkers achieve zero waste.**

  The average New Yorker throws out nearly 24 pounds of waste at home, at work, and at commercial establishments every week. [BigApps.NYC]

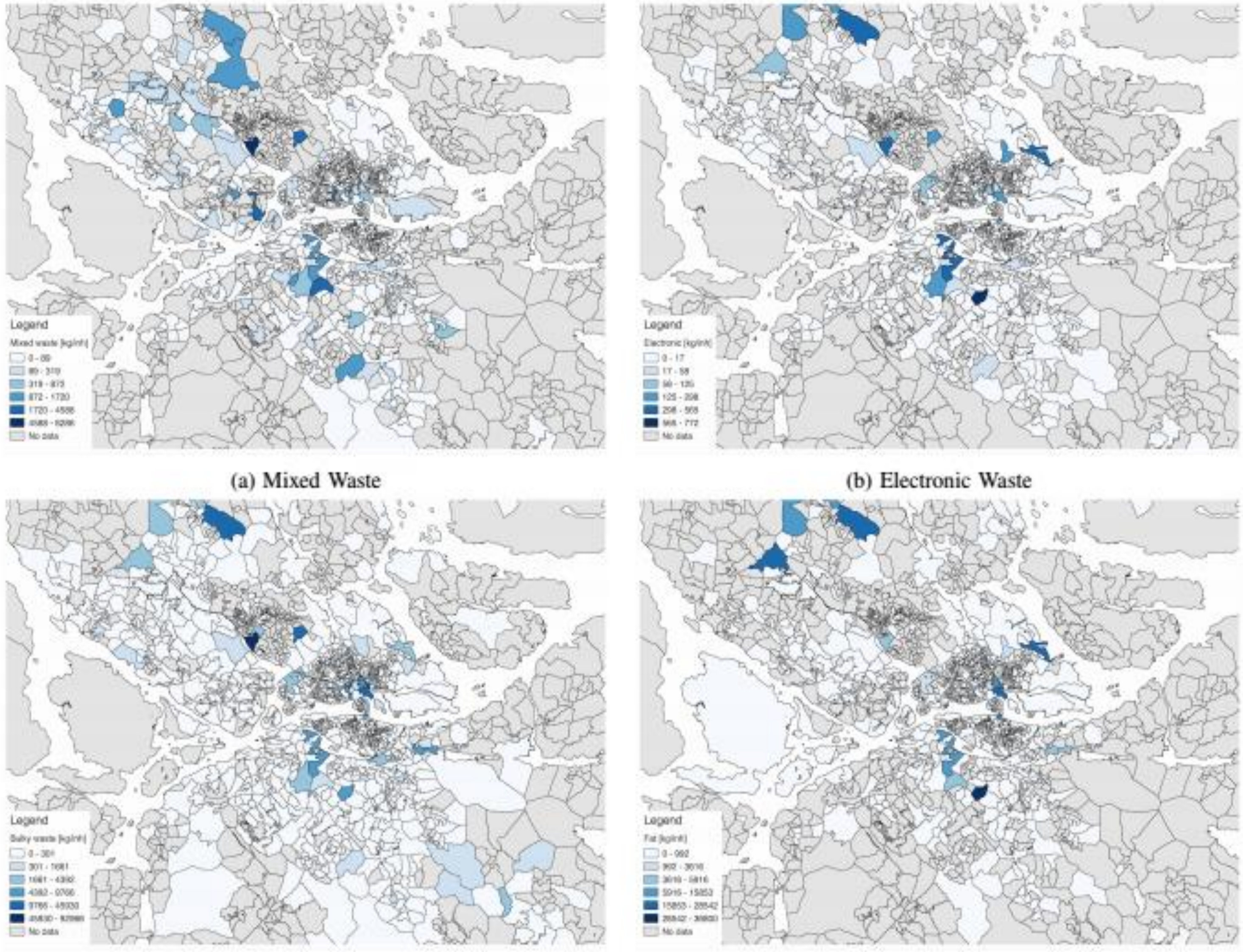
- **Provide an incentive to residents and businesses to audit their waste to green their home, neighborhood or workplace.**

  Provide a heat map of green neighborhoods.

- **Help New York City Department of Sanitation (DSNY) understand which parts of the city are saturated with recycling/compost bins, and which are underserved.**

- **Enable New Yorkers to find the "people's" route to bins locations as suggested by recommendation, using *PeopleMaps* which was implemented last year. (future work)**

Zero Waste Goal



(a) Mixed Waste

(b) Electronic Waste

(c) Bulky Waste

(d) Fat Waste

Dataset, Algorithms and Tools

- **Datasets:** (obtained from https://data.cityofnewyork.us)
  1. Locations of public recycling bins throughout NYC.
  2. DSNY's Refuse(waste) and Recycling Disposal Networks.
  3. Special Waste Drop-off Sites (batteries, motor oil, oil filters, car tires, …etc).
  4. Recycling Diversion and Capture Rates.

- **Algorithms:**

1. Filter the datasets as per the type of waste.
2. Suggest appropriate recommendations for the nearest drop-off location(Euclidean Distance recommendation).
3. Apply k-means clustering to visualize the capture rate data and recycling diversion rate on heat maps.
4. Enable the user to trace the route for the address suggested by recommendation using *PeopleMaps* (Implemented last year).

- **Tools:**

1. Apache Mahout
2. Apache Spark
3. Apache Hadoop(Pig)

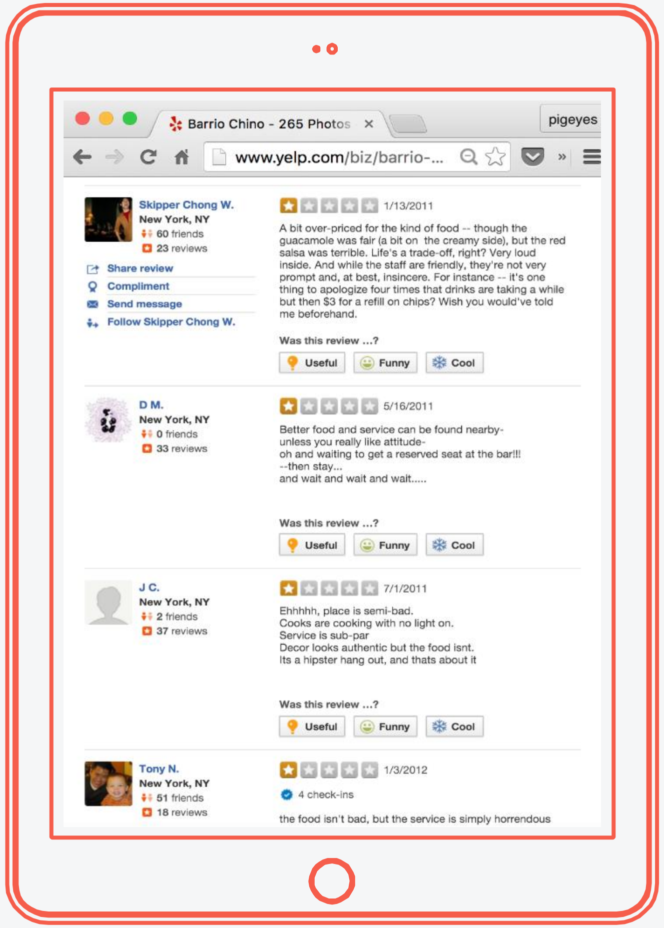
Current Progress, Schedule and Expected Contributions

- Collected and understood the datasets that will be used
- Prepared a flowchart for project implementation

Expected Contribution	Task name	Task Description (+issues, feasibility, …)	Target dates (start-finish)	S: started I: in progress, C: completed
All	Literature review		Nov 4 - Nov 17	C
All	Collecting datasets		Nov17 – Nov 20	S
Shreya , Harnoor	Starting Task 1	Recommendation :all types of waste	Nov 20-Dec 5	
Hadeel	Starting Task 2	Clustering: for the capture rate and recycling rate division	Nov 27 - Dec 5	
All	Integrate PeopleMaps		Dec 5 – Dec 10	
All	Write project report		Dec 7 - Dec 10	

Reverse-Recommendation on Yelp!

- find out what user really cares about from their low rating reviews.
- Send them a message to recommend local restaurants.



Dataset and Tools

Dataset

Yelp Dataset Challenge

- 1.6M reviews and 500K tips by 366K users for 61K businesses
- 481K business attributes, e.g., hours, parking availability, ambience.
- Social network of 366K users for a total of 2.9M social edges.
- Aggregated check-ins over time for each of the 61K businesses

Tools

- Server: AWS EC2, S3, Heroko
- Backend: Python-Flask
- Frontend: JS
- Yelp API
- NLP: Python NLTK
- Analysis tool: Spark - Milb-Cluster, Recommendation, MapReduce

Algorithms

Recommendation/ Clustering:

- SVM
- Sentiment Analysis
- Latent Dirichlet Allocation (LDA)

Project Timeline

Week-1
11/20-11/26

Categorize negative reviews

Week-2
11/27-12/03

Build Recommendation Model

Week-3
12/04-12/10

Web Application Buildup

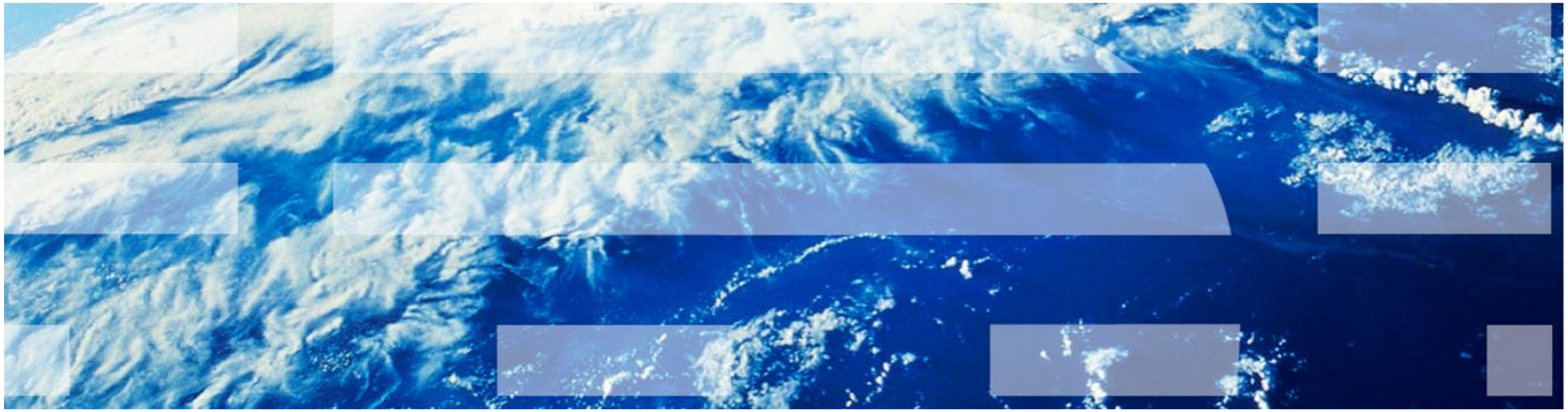
Week-4
12/11-12/17

Optimization

E6893 Big Data Analytics Project Proposal:

<MYOU : Music for You Recommender System>

<Yingtao Xu>



November 19th, 2015

Motivation

- Music fan

- Make what I have learned into practice

Dataset, Algorithms and Tools

✦Dataset: Yahoo music dataset

✦Algorithms:

- collaborative filtering algorithms (all included in Mahout)
- customized similarity metric(TF-IDF on lyrics)

✦Tools: Java, Mahout, Tomcat, Maven, Bootstrap

Current Progress:

- Preparation of the dataset
- Analysis of the project procedure

Schedule:

- Establish a server (GUI) for users to type in the songs they are interested in and return the recommendation results back to the users. (Java, Tomcat, Bootstrap)
- Build the recommender (Mahout)
- Do the recommendation testing

Expected Contributions:

- Users put the name of the songs they are interested in, the recommender will find the corresponding songs that match their interests.

Find your fit: University Edition

By
Ashwin Raghupathi ar3390
Senthil Krishna Mani sm3906

Motivation

- Wanted to work on a topic which was relevant to students
- Searched for a topic where we could have a lot of intuition on and see if the data proves or disproves our intuition
- Decided to work on UG colleges and student performance
- Goals:
  - Draw interesting relationships between available parameters
  - Build recommendation engine for students to identify best-fit UG institution

Dataset, Algorithm & Tools

- Dataset
  - Rich US gov dataset on UG colleges and future student performance
  - Exhaustive data for over 20 years
- Implementation
  - Spark SQL
  - Apache Mahout
- Algorithms
  - Recommendation algorithm
  - Clustering analysis

Data Metrics

Key considerations in trying to interpret data involves understanding importance of factors to identify best fit school (School of your dreams!)

- Earnings 10 years after Matriculation.
- SAT Score: Math, Written, Verbal
- Admission Rates: Highest, Lowest
- Percentage by Major: Engineering, Sciences, arts etc.
- School Type: Non Profit, For Profit, Public/ State, Private.
- Enrollment by number: Absolute and demographic split
- Price of Tuition and Overall Fees.

Current Progress & Schedule

- Current Progress
  - Data cleaning
  - Find relationships between different parameters
  - Identifying influencing factors for decision making
  - Plotting these relationships
- Schedule
  - Have to build the recommendation algorithm
  - Set to complete the project in the next 4 weeks

Appendix

- Query databases and libraries will be essential in this process of selection
- Each query and search result will be an amalgamation of factors important a specific user, i.e. some may find tuition as a limiter while others might find demographics both racial/cultural and gender based more important, this leads to different results being outputted
- Plots of overall displays and outputs reflect best on the list of “top” universities that help rank and guide users better and reflects better matches via the results

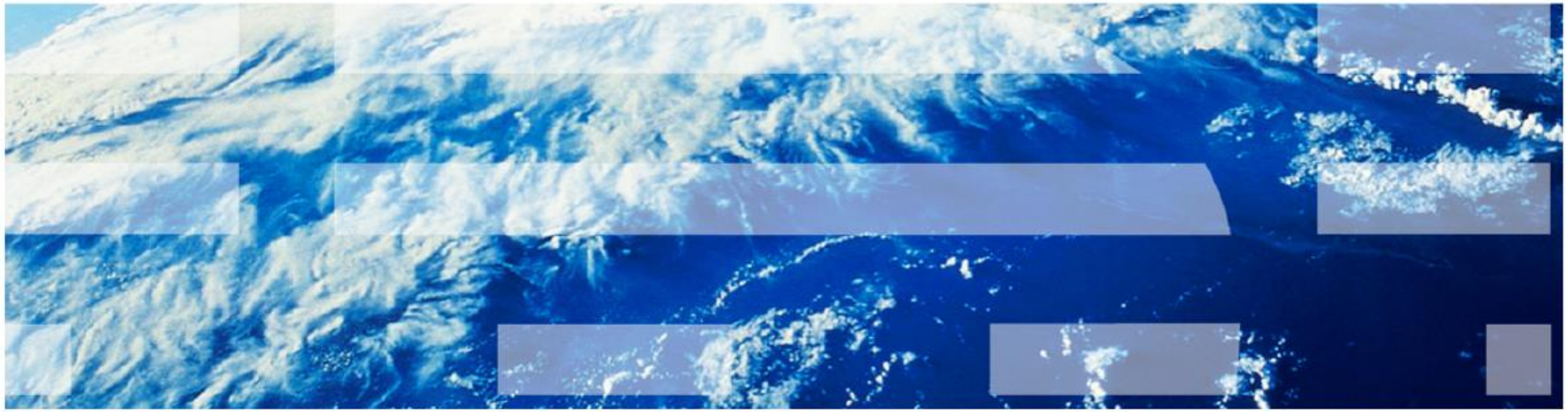
E6893 Big Data Analytics Project Proposal:

Sports Fandom

Mayank Mahajan - mm4399

Sheng Qian - sq2168

Brian Slakter - bjs2135



November 19th, 2015

Motivation

- The NBA is the world's fourth largest sports league by revenue, and popularity is on the rise both in America and around the world

League	Sport	Revenue
NFL	American football	$9 billion
MLB	baseball	$8 billion
English Premier League	soccer	$5 billion
NBA	basketball	$4.75 billion

- The league has an estimated 840 million fans on social media, including 17.9 million twitter followers

- Although fans often choose to root for teams they are closest to in geography, there are many others who will choose to focus on other teams in different parts of the country

- We would like to understand what teams people in different parts of the country root for and against, and help provide recommendations of what teams to follow for those interested in the sport

Dataset, Algorithms and Tools

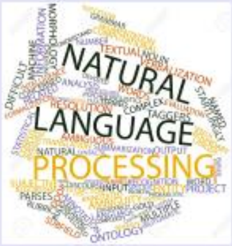
Dataset

- ~60k Tweets from users who follow @NBA
- Information collected for each tweet:
    - Tweet Text
    - Time of Tweet
    - User Location (Latitude, Longitude)

Algorithms and Tools

Gather NBA fan Tweets via Twitter API	Identify team mentions and perform sentiment analysis	Perform geographic clustering and provide team recommendations for each cluster

Twitter API → NLP → mahout



Current Progress, Schedule and Expected Contributions

Current Progress

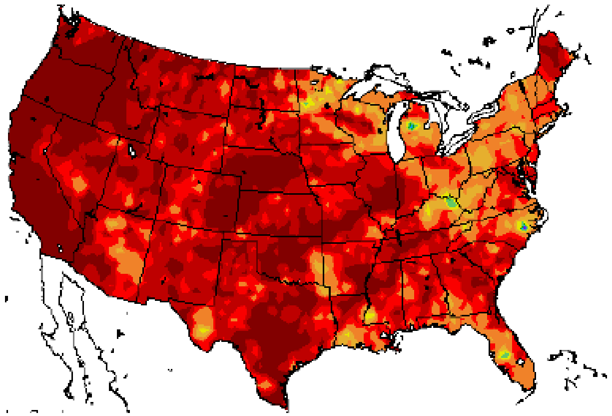
- Currently mining data from Twitter API
- Rate limit of 300 requests per 15 minutes

Schedule

- November 23: All Twitter data gathered
- December 4: Sentiment analysis finished
- December 11: Mahout clustering and recommendations complete

Expected Contributions

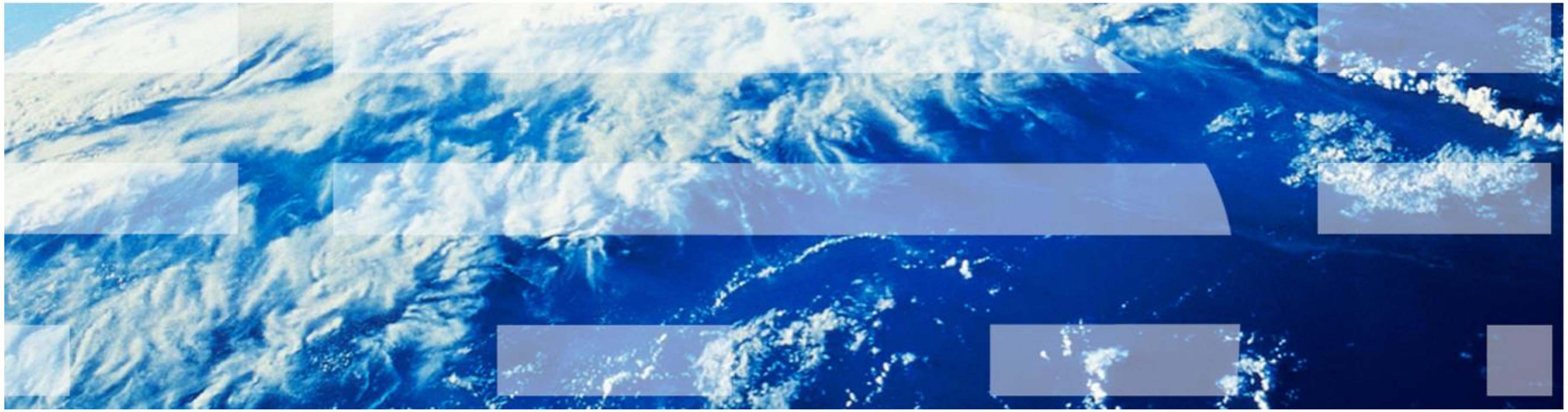
- Recommendations for teams to root for and against, by geographic cluster
- Heat maps that provide visualizations of clusters of fans who root for and against different teams



E6893 Big Data Analytics Project Proposal:

Regional Mood Assessment Application based on Tweets

Wenyu Zhang



November 19th, 2015

Motivation

✦**Proposal**:

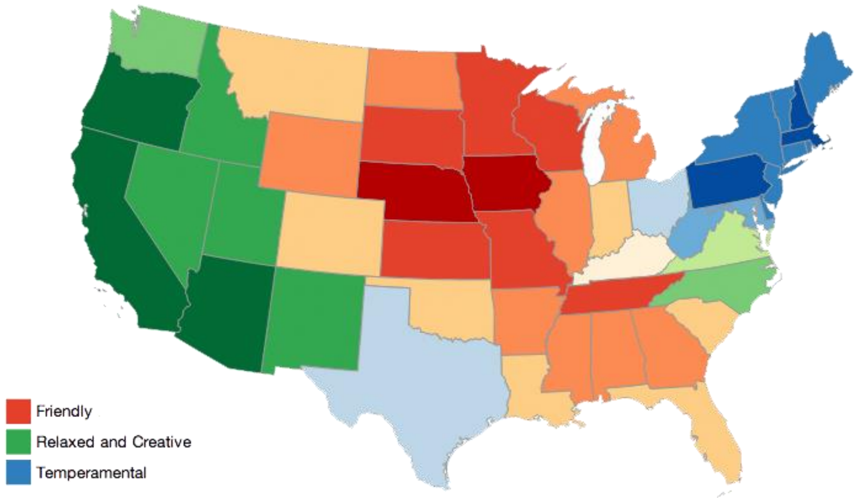
✦Use big data collected by Twitter to generate useful regional mood map

✦**Motivation**:

✦Help the government to get a sense of regional mood

✦Help people to decide where to move

✦Template application



Dataset, Algorithms and Tools

✦**Dataset**:

✦Twitter Streaming API (Using locations: -74,40,-73,41 to get NYC Tweets)

✦**Methods**:

✦Tokenizer: TextBlob (library for processing textual data)

✦Known lexical words: “Happy”, “Sad”, “Relaxed” …

✦Discover lexical words (big data): Compute the relative frequency of each term among all messages from within each area

✦(Optional) Score Mapping based on Dictionary

✦Map Mood (Score) on Application: Map Kit

✦**Tools and Languages**:

✦Xcode: Objective-C, Bash

✦PyCharm (Vim) : Python / Rstudio : R

Current Progress, Schedule and Expected Contributions

- **Progress**:
- Collected corresponding Twitter dataset
- Tested and analyzed Tweets sentiment

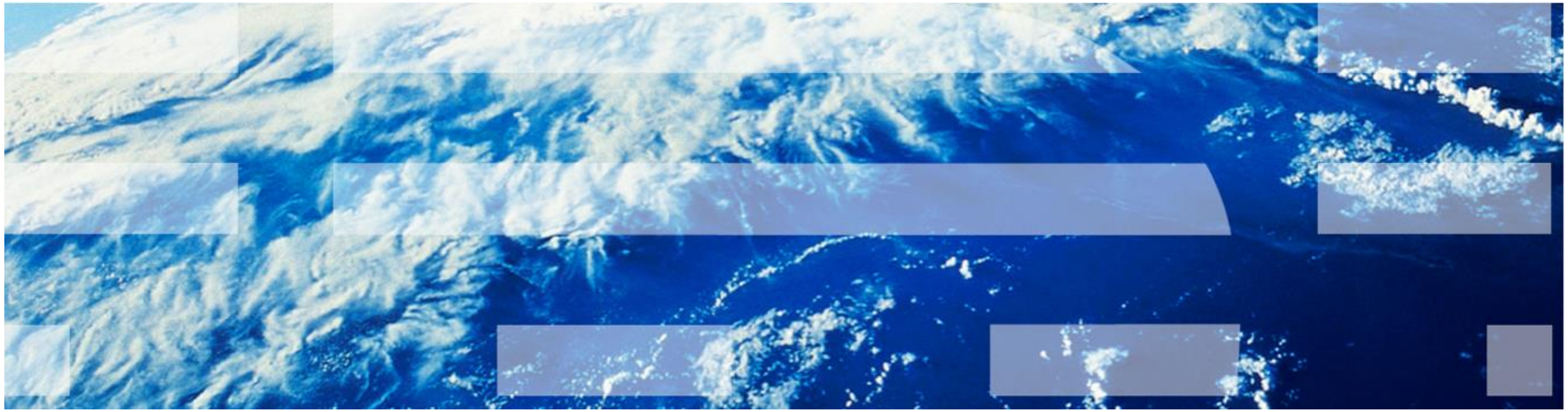
- **Schedule**:
- Week 12: Compute the relative frequency of each term among all messages from within each area
- Week 13: Mobile Application Implementation
- Week 14: System Integration and Analysis, Debug.
- Week 15: Tests and Final Presentation

- **Contributions**:
- Hope get some amazing data visualized maps!

E6893 Big Data Analytics Project Proposal:

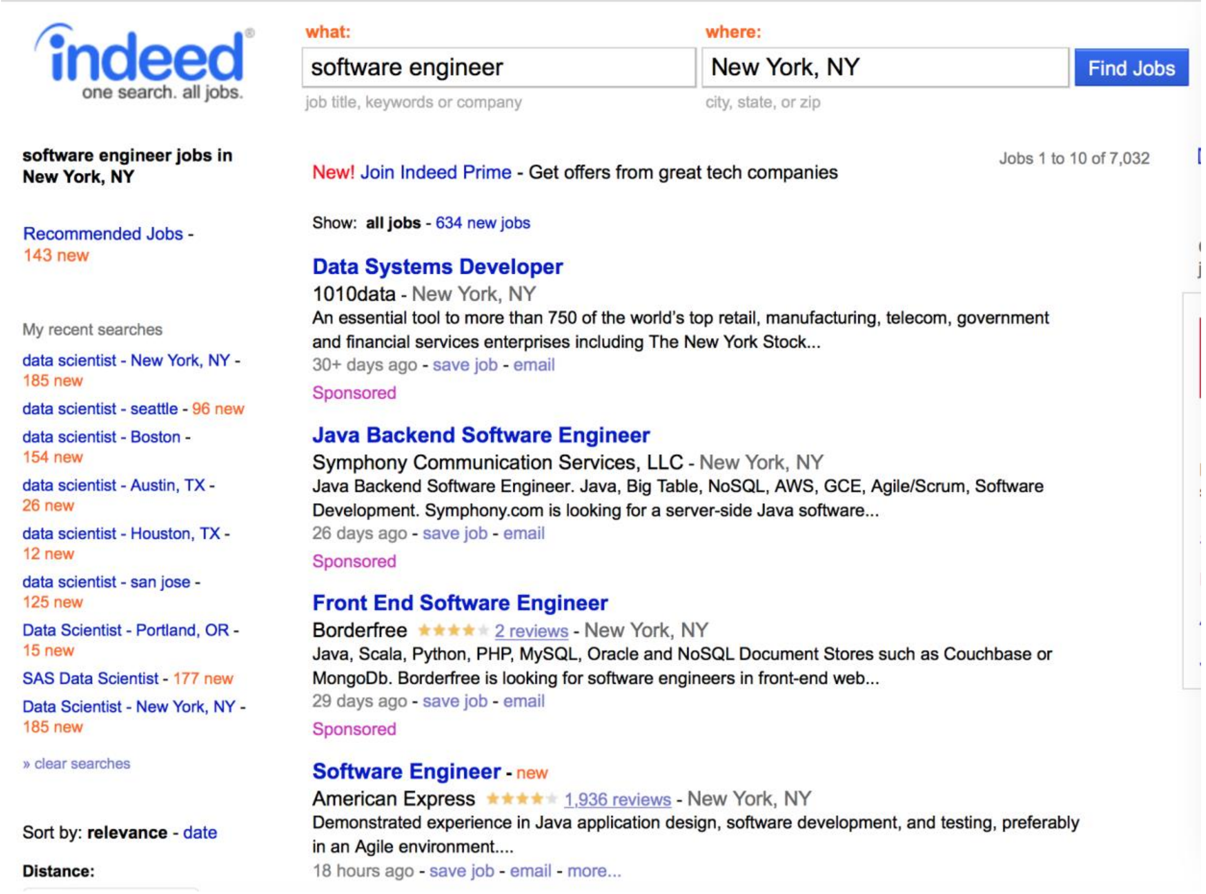
Job Recommendation System

Ke Shen



November 19th, 2015

- Most job search websites are using keyword searching to help find desired job
- Users are becoming more and more impatient when they visit websites and cannot find the desired jobs in an immediate way



Dataset, Algorithms and Tools

✦Dataset: a million of resumes and job descriptions scarped from job searching website

✦Algorithms: CRF, HMM, DP, LDA

✦Model: Combination of Collaborate filtering and content based filtering

✦Tools: Python, Spark, MongoDB

Current Progress, Schedule and Expected Contributions

Part I: web scraping (done)

Part II: NLP for JD (done)

Part III: Parsing for resume

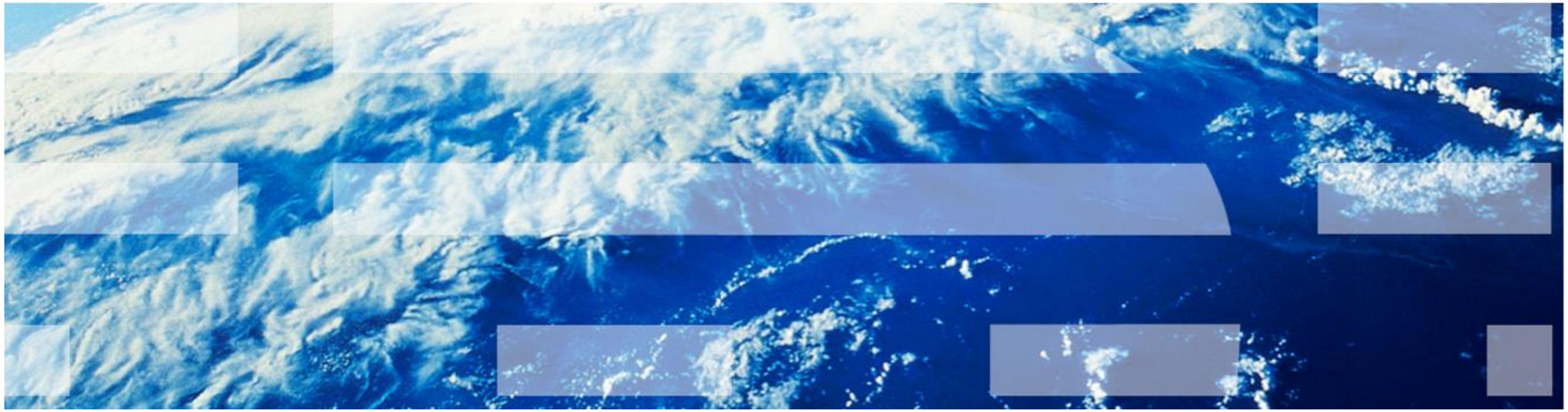
Part IV: Find topic words for each categories

Part V: Build model

E6893 Big Data Analytics Project Proposal:

Visual Analysis of Scholar Data

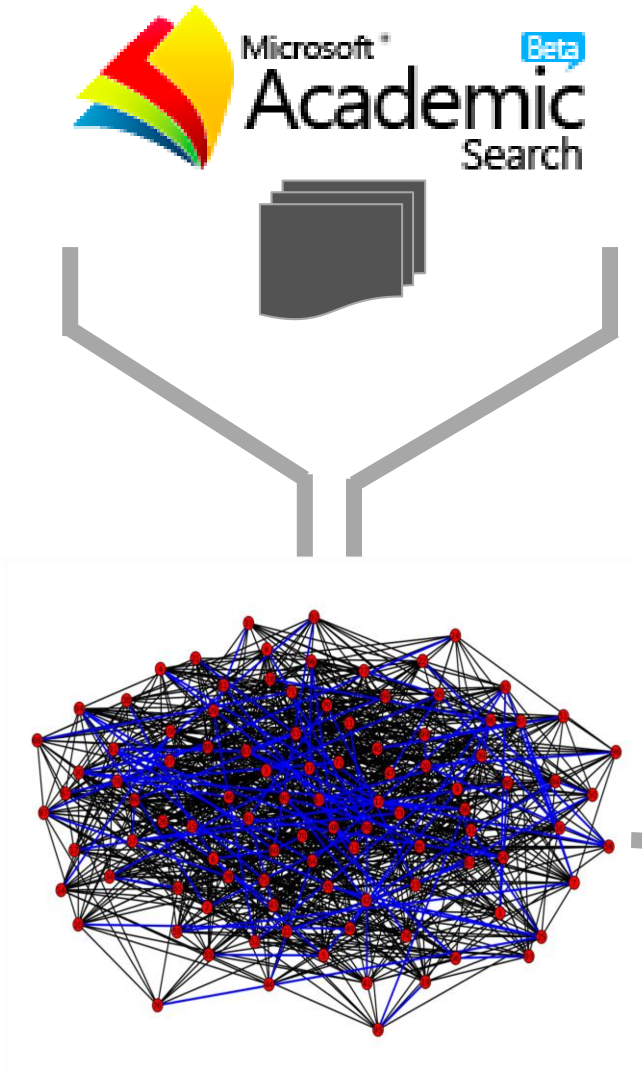
Michelle Tadmor, Miguel A. Yanez, YuHsuan Shih



November 19th, 2015

Columbia University

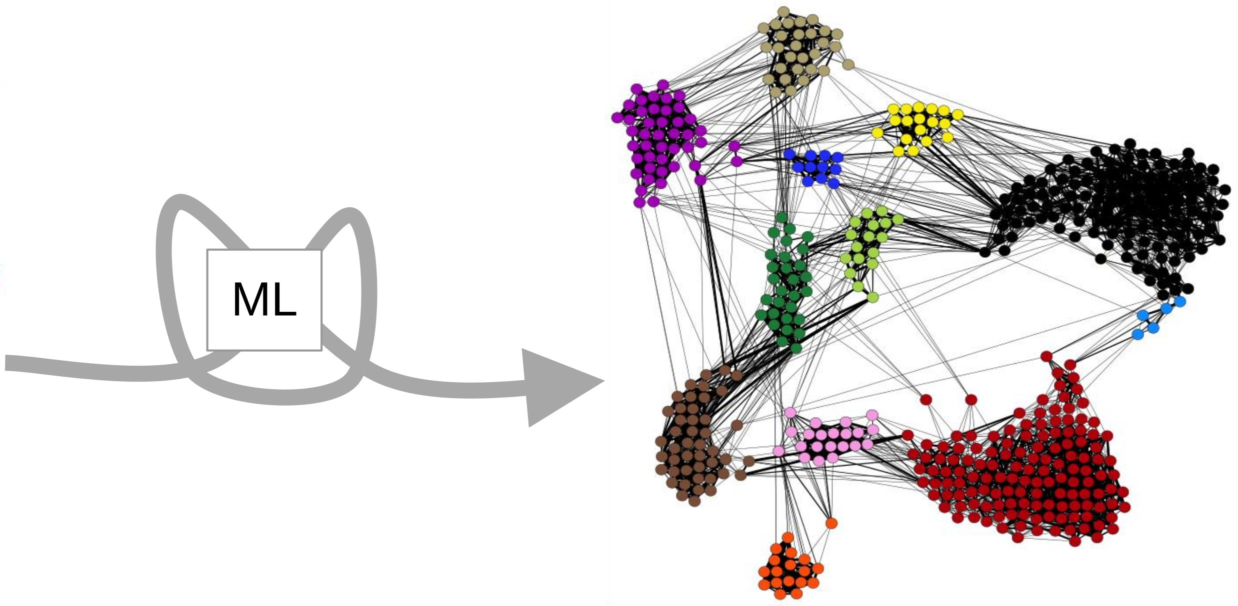
Organize, Visualize, and Analyze Scholar Publications



Automatic identification of trends

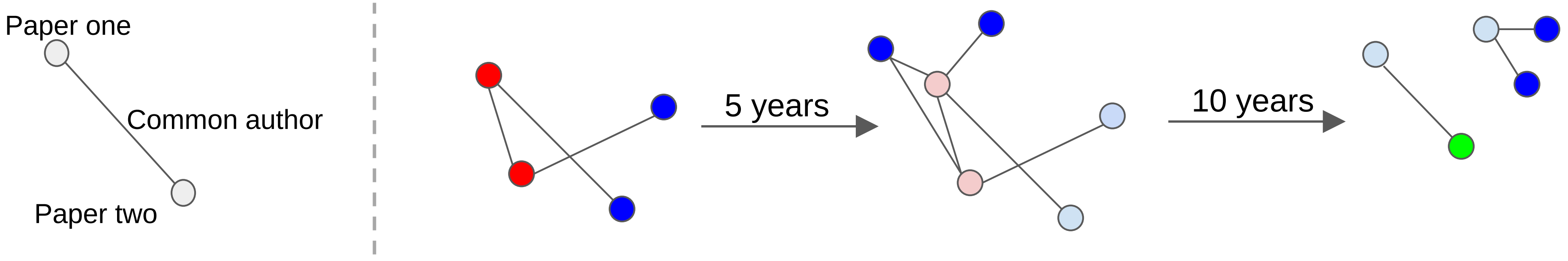
Highlight interdisciplinary publications

Visualize focus shifts as a function of time



Dataset, Algorithms and Tools

- **Dataset:** Newly released Microsoft Academic Graph. Part of an ongoing research project at Microsoft. Huge Dataset (29GB compressed). http://research.microsoft.com/en-us/projects/mag/
- **Tools**: IBM SystemG gShell Python API, D3js, Python Flask
- Topic grouping by Graph Based Clustering algorithm.
- Because of the scale of the data we will use a Cloud Instance on DigitalOcean to run our service.



Project

- Novel visualization of the Human Knowledge base.
- Analysis of trends and interdisciplinary relationships.

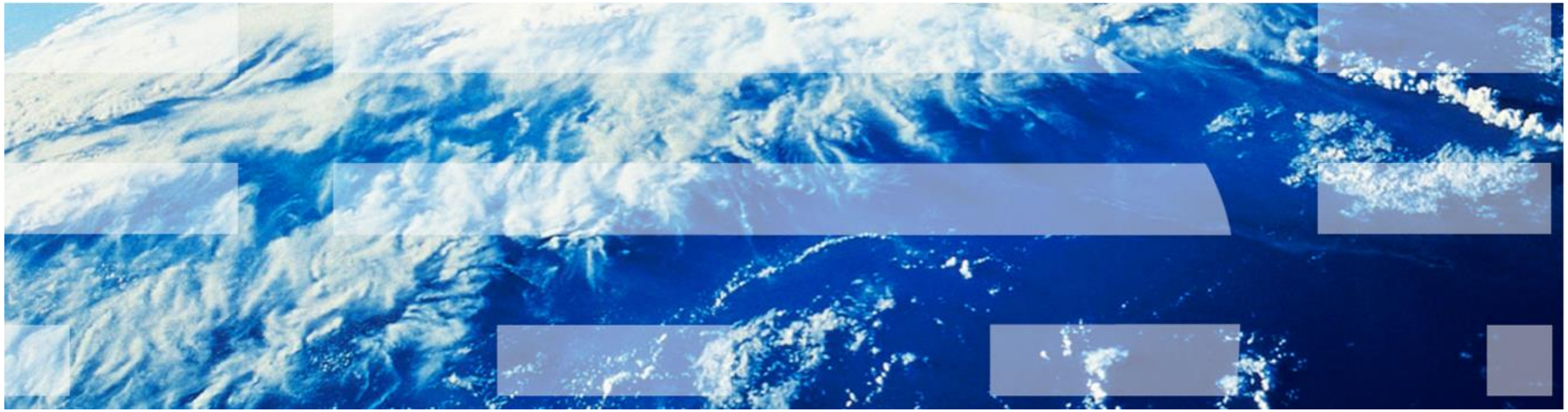
Individual

- **Miguel:** Cloud Infrastructure and Dataset preparation.
- **YuHsuan:** Visualization of the Dataset.
- **Michelle:** Computational identification of topics and interdisciplinary publications.

E6893 Big Data Analytics Project Proposal:

Product recommendation using customers' search or click behavior

By: Neha Gupta



November 19th, 2015

Motivation

Describe the motivation of your project:

•Before buying any product online, one must do intensive research on the product's reviews, ratings , number of people who rated them, ratings from recent users.
•Users' rating and reviews are very important factors that increase the chances of a product being sold online.

This leads to the question: Can we leverage on the users' buying behavior to recommend them more products that they would like to buy?

Dataset: The dataset is downloaded from Kaggle website. Click here or enter https://www.kaggle.com/c/acm-sf-chapter-hackathon-small to download the dataset. train.csv and test.csv contain informa- tion on what items users clicked on after making a search. Each line of train.csv describes a user's click on a single item. It contains the following fields:(user, sku, category, query, click_time, query_time). small_product_data.xml contains information about products like name, sku, release time, price and description. Only the description will be used in our content based filtering method.

https://www.kaggle.com/c/acm-sf-chapter-hackathon-big/data

Language: Python

Analytics: Recommendation System, TF-IDF, Multiclassification, SVC, Spell Check, Content Based Filtering, Collaborative Filtering

Current Progress, Schedule and Expected Contributions

Stage -1:

- Data pre-processing - In order to evaluate our algorithms, we randomly split train.csv into two parts: training part and test part. The proportion of training data and testing data is 9:1

Stage -2:

- Apply the recommendation algorithms

Stage -3:

- Evaluate correctness

Schedule:

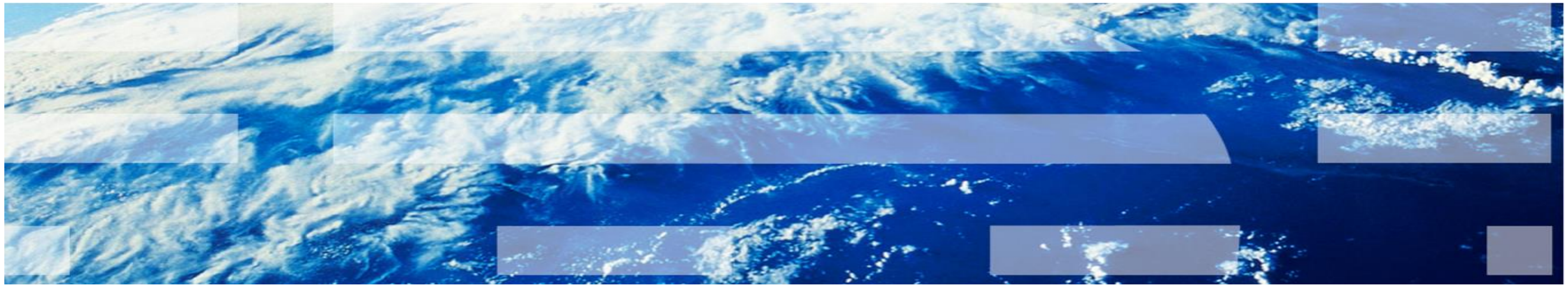
Total time frame:3 weeks

Current stage: Stage -1

E6893 Big Data Analytics Project Proposal

Clustering of Electricity Customers by Load Curves for Integration of Solar and Wind Energy Resources into the Grid

Akhilesh Ramakrishnan (ar3539)
Ankita Deshmukh (ad3293)
Kaustubh Upadhyay (ku2151)



November 19th, 2015

Motivation

- In the present energy market, energy vendors must commit to supplying a specific amount of energy in the next hour/day
- Utilities must be able to accurately predict the variation of the load and ensure that the supply matches the demand
- These factors make load pattern recognition and load forecasting essential to a reliable and efficient power grid
- Clustering and classification of electricity customers based on their hourly demand allows both utilities and energy providers to accurately predict the load they will have to meet
- Matching the daily demand patterns to renewable energy supply patterns will allow us to determine the optimal combination of solar and wind resources needed to meet the load for each type of customer throughout the day
- This will take into account the variation of supply and demand:
  - ➢ Hourly
  - ➢ Diurnally
  - ➢ Seasonally

Dataset, Algorithms and Tools

Data:

✦Demand data for customers
Electricity hourly demand data by zones - New York Independent System Operator website – http://nyiso.com/

✦National solar radiation database - http://rredc.nrel.gov/solar/old_data/nsrdb/

✦Wind resource data - http://www.nrel.gov/gis/data_wind.html

Algorithms & Tools:

✦Cluster customers based on the similarity of their demand curves through out a typical day by using cross correlation based kNN clustering

✦For a particular cluster/group of customers determine whether this group is best served by a solar energy source or a wind energy source or a combination of both

✦Apache Spark for clustering and python for data pre processing and general purpose scripting

Current Progress, Schedule and Expected Contributions

Current progress:

- Data preparation in progress
    - Research the different types of datasets available for the purpose
    - Preparing the data in usable formats
- Exploring the appropriate algorithms to be used

Schedule:

Week1: Finalize the data and algorithms

Week2: Run the clustering and demand/supply matching

Week3: Comparing the clustering results as available in literature

Week4: Refinements & documentation

Expected deliverables:

For each cluster of electricity customers based on demand curves obtain a combination of solar and wind supply to best serve the demand

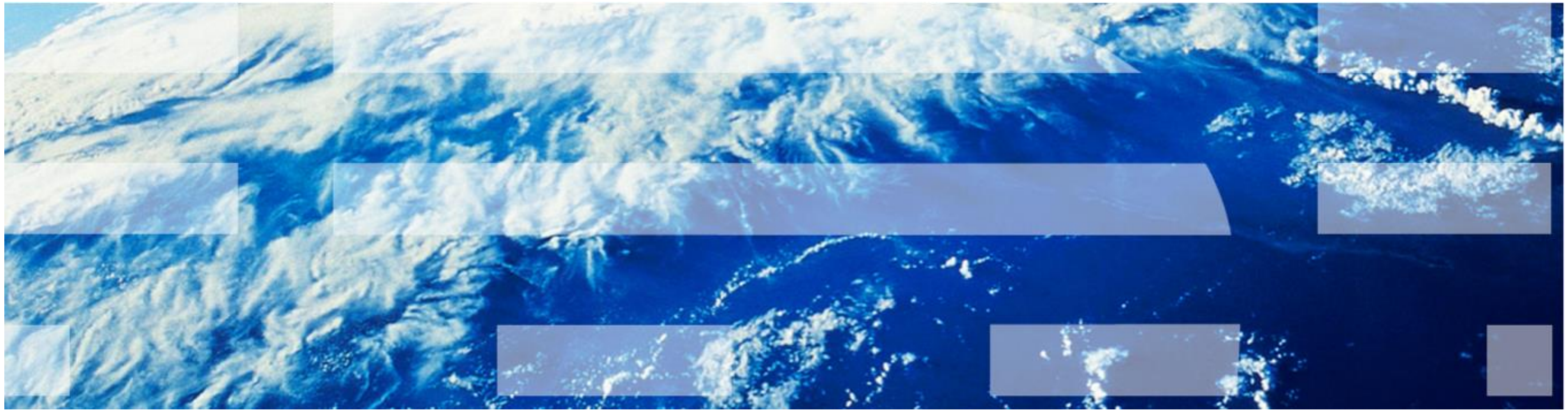
E6893 Big Data Analytics Project Proposal:

Image recognition with a huge dataset on iOS devices

Team members: Chang Chen (cc3757), Liang Wu (lw2589),

Changchang Wang (cw2826), Jialu Zhong (jz2612)

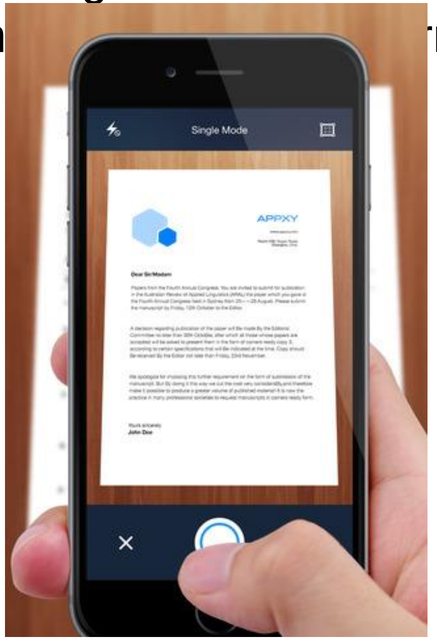
Supervisor: Larry Lai (IBM Watson Research Center)



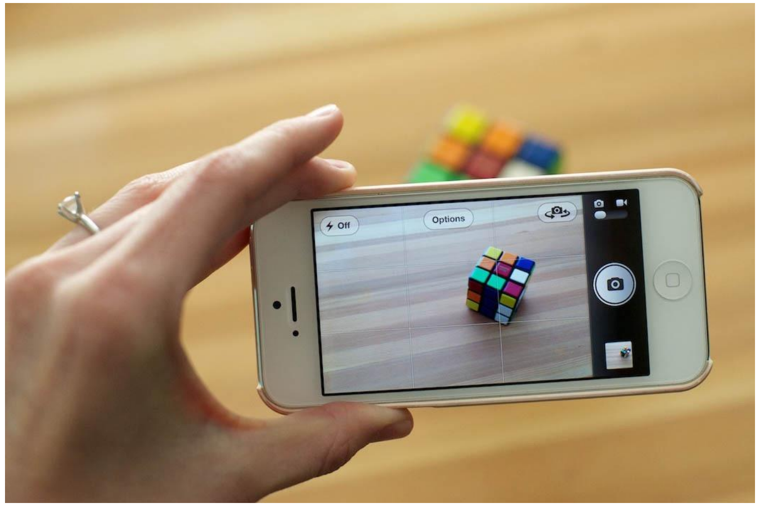
November 19th, 2015

- **Description**: Most companies do not allow their employees or visitors to take photos of any confidential materials, such as documents, white board, or screen shots, using their personal devices. So, the challenge here is how to help the companies immediately detect that the photo just taken contains confidential information.

- **Goal**: Design an APP and cloud service to detect if the user takes a picture contain rmation.



Confidential



Non-Confidential

- **Dataset**: A relatively large image dataset containing different content of confidential materials.
- **Algorithms**:
  - Text recognition to recognize the image content
  - Analysis on the image similarities
- **Tools**:
  - iOS developing (Swift)
  - OpenCV to calculate image features
  - CoreData on iOS to reduce the computation
  - Spark as the framework for big data processing

- **Current Progress**:
  - Skype meetings with Larry
  - Developed a small prototype of capture image using Swift and CoreData
  - Explored OpenCV with image recognition.

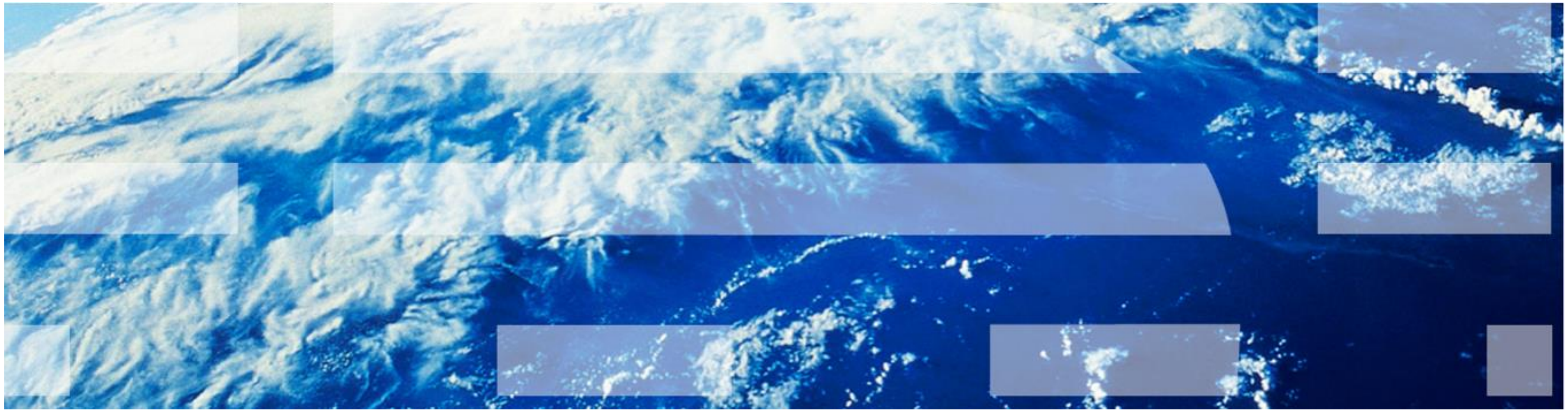
- **Schedule**:
  - Set up cloud server that allows our app to send images
  - Implement text recognition to detect the image content
  - Optimize our service to enable multi-processing at a time

- **Expected Results**:
  - An app with cloud service that has the functionality of camera, and the capability to immediately detect if the photo taken has confidential information

E6893 Big Data Analytics Project Proposal:

Social/Business network analysis for charitable fundraising

Janet Prumachuk, Sam Guleff, John Correa



November 19th, 2015

Motivation

Describe the motivation of your project

Analyze social networks to identify potential donors for charity causes

- Identify individuals with common interests with charity (college alma mater, political affiliation, previous donations, etc.)

- Use social/business network of those individuals to expand potential pool of donors

Dataset, Algorithms and Tools

Examples of datasets to be considered for use:

- Angel List: https://angel.co/
- LinkedIn: http://www.linkedin.com
- News article text mining for names and companies
- Board members for nonprofits and startups
- Federal Election Commission campaign finance records: http://www.fec.gov/finance/disclosure/disclosure_data_search.shtml

Tools:

- Neo4J, Cypher, Java, Python
- Apache Hadoop, Spark, Nutch
- JavaScript, D3

Algorithms:

- Entity extraction
- Term-document relevance
- Building a knowledge graph
- Shortest path (to find referrals)
- Clustering (to identify donor groups for different causes)

Current Progress, Schedule and Expected Contributions

Current Progress: Analysis phase.

- We have identified data sources and agreed on the concept.

Schedule:

Week 0: Learn Neo4J and Cypher

Week 0: Define graph node and edge model and properties

Week 1: Build web crawler, clean and transform data, load affiliations and properties.

Week 2: Inspect graphs and optimize data extraction and graph model

Week 3: Define queries, clusters, metrics and develop sample analysis results

Week 4: Build Web Interface, visualizations

Week 4: Develop presentation and assess lessons learned

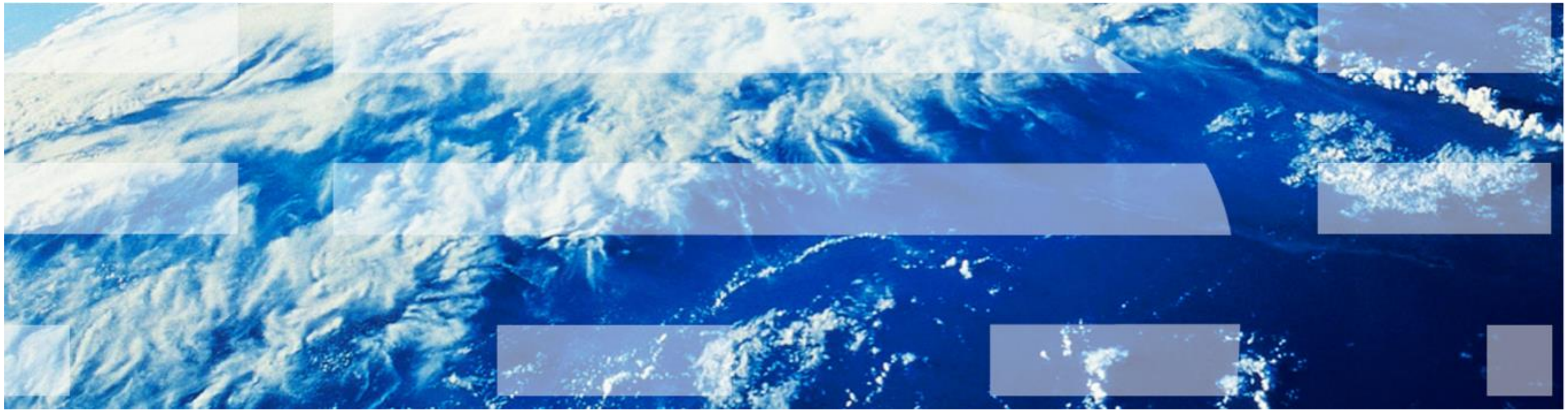
Expected Contributions:

- Janet Prumachuk: data sources, Sam Guleff: graph model, John Correa: software prototype
- Data Sources to be divided among team members for web crawling, data load
- Query/analysis/visualization to be divided among team members
- Develop presentation (team effort)

E6893 Big Data Analytics Project Proposal:

Predicting Dota2 game outcome

Li Qi(lq2156) Jiaqi Guo(jg3639) Xinyuan Hu(xh2251)



November 19th, 2015

Motivation

Dota2 is a free-to-play multiplayer online battle arena video game developed by Valve Corporation. Game is played in matches between two five-player teams, each of which occupies a stronghold in a corner of the playing field. A team wins by destroying the other side’s “ancient” building, located within the opposing stronghold.

The largest of the professional tournament in dota2 is known as The International. The 2015 edition of The International had the largest prize pool in eSports history, totaling over $18 million.

There are total 110 playable “Hero” characters in Dota2. So for each team, the hero selection can significantly influence the game outcome. Profession teams recognize the importance of this and in matches it usually takes up to 10 minutes for hero selections by both teams.

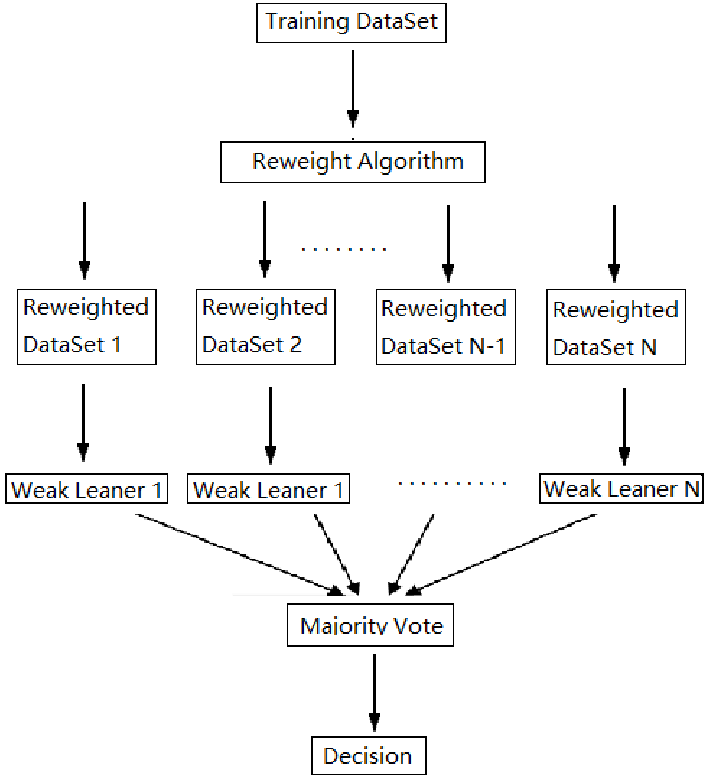
Our project’s goal is to predict the game outcome based on the hero selection.

Dataset, Algorithms and Tools

Dataset We use the Steam Web API for collecting dataset about public Dota2 matches. We will use Python script to record Dota2 matches periodically. And we use Mongo database to backup and restore our data.

Algorithms The prediction can be abstracted as a non-linear calssification problem with an ten dimensional input vector, namely the 10 selected heros, and a binary output.

ADABoosting is a kind of ensamble methods using independed dataset to train several weak learners to make majority vote. The key idea is the weak learner we choose and the algorithm to reweight the training dataset.



ADABoosting

Tools Python Machine Learning Flask Web Develop

Current Progress: Writing Python script and setting it up to record data from the 500 most recent public matches every 30 minutes.

Schedule:

11.20-11.25 Finishing python script and downloading dataset

11.26-12.10 Data analyzing. Compare the project result with the actual result.

12.11-12.15 Algorithm fixing up. Improve the accuracy.

Expected Contributions:

We will try to achieve about 70% accuracy for predicting match outcomes based on hero selection.

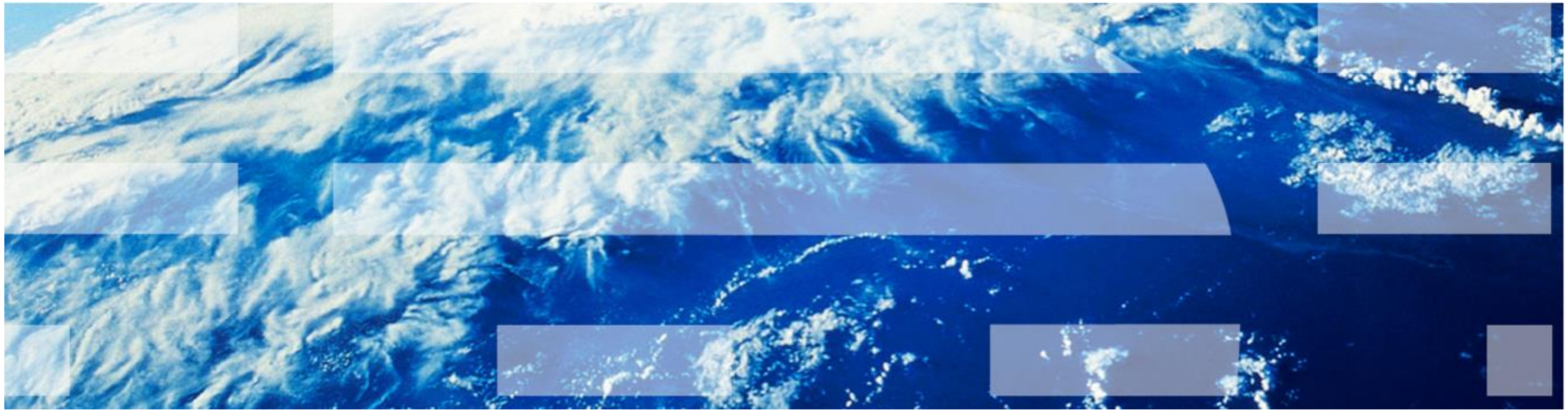
E6893 Big Data Analytics Project Proposal:

Item-based Event Recommendation Based on User's Preference

Shiwei Ren, MS EE

Yeran Zhang, MS EE

Yiqing Cui, MS CS



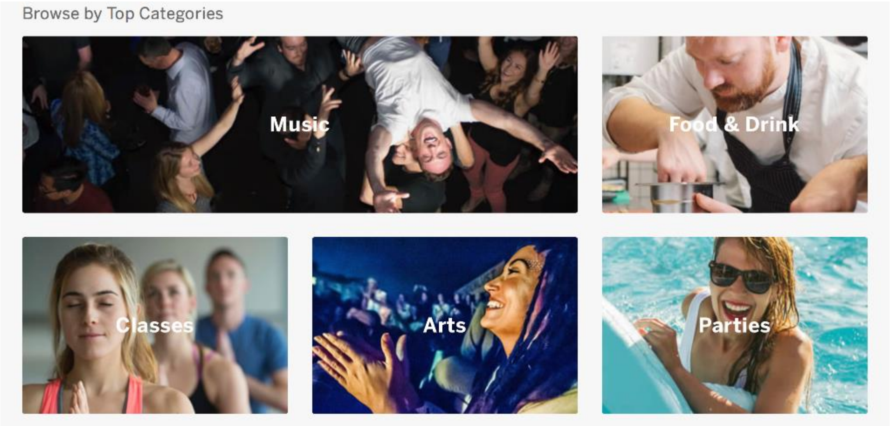
November 19th, 2015

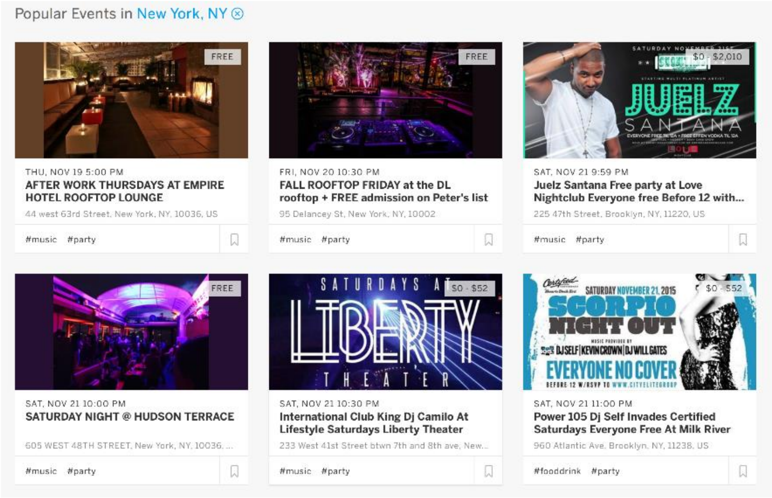
Motivation

Columbia University

- Events provided by Eventbrite can be overwhelming for users to make a choice.
- Even when users select some fields that they are interested in, there are still too much information for them.
- We are going to build a recommendation system which can predict suitable events for some specific users.

Eventbrite®





Dataset, Algorithms and Tools

- Dataset: Evenbrite
- Algorithms: Item-based Recommendation, KNN, MapReduce
- Tools: Python, Java, Spark, Hadoop

Current Progress, Schedule and Expected Contributions

- Current Progress: Data Fetching

- Schedule:
  11.20 - 11.23
  finish data fetching
  11.24 - 12.5
  implement and optimize the algorithms for recommendation
  12.5 - 12.10
  implement the front end to show the results
  12.10 - 12.12
  finish the technique report and presentation slides

- Excepted Contributions:
  data fetching: mainly Cui, Ren&Zhang assist
  algorithms: mainly Ren&Zhang, Cui assist
  front end: altogether

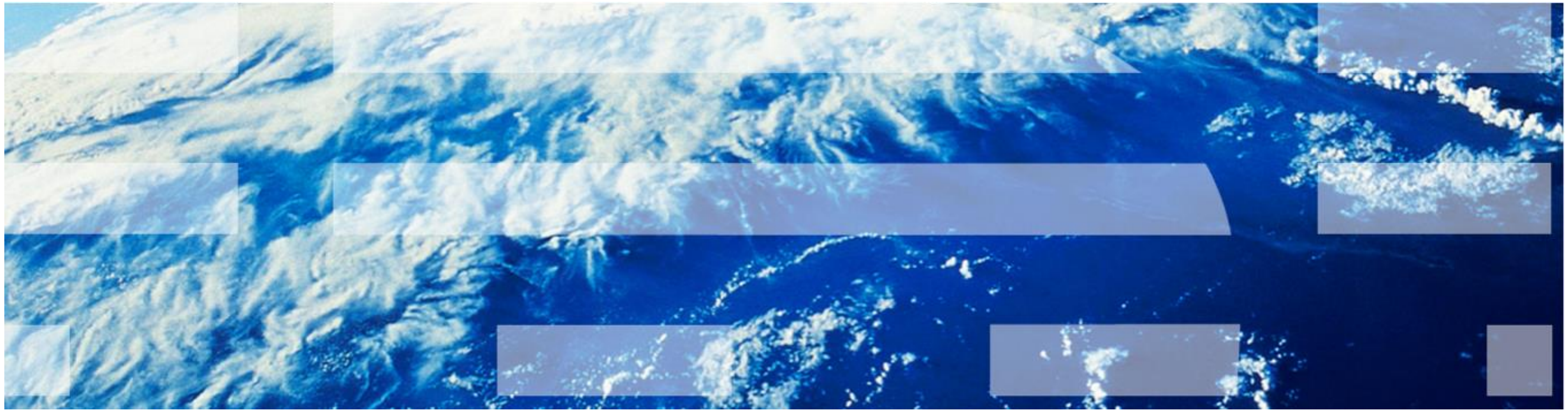
E6893 Big Data Analytics Project Proposal:

Reliable Reviews Recommendation

Chen Qian (cq2171)

Jiaqi Chen (jc4260)

Tianhe Shen (ts2957)



November 19th, 2015

Motivation

Opinionated social media are now widely for our decision making.

- Fake Reviews
- Spam Reviews
- Reviews not helpful

In our project, we would like to give each user the ‘best’ reviews based on different tastes and interests.

- **Dataset:** Yelp Dataset Challenge in 2016, including users, businesses, and reviews

- Spam reviews filtering: MapReduce, Sentiment Analysis;
- Reviews from similar taste users recommendation: Collaborative filtering, User Similarity Measurements;
- Reviews clustering: Naive Bayesian classification, TF-IDF.

Current Progress and Expected Contributions

Current Progress: Algorithms research and design

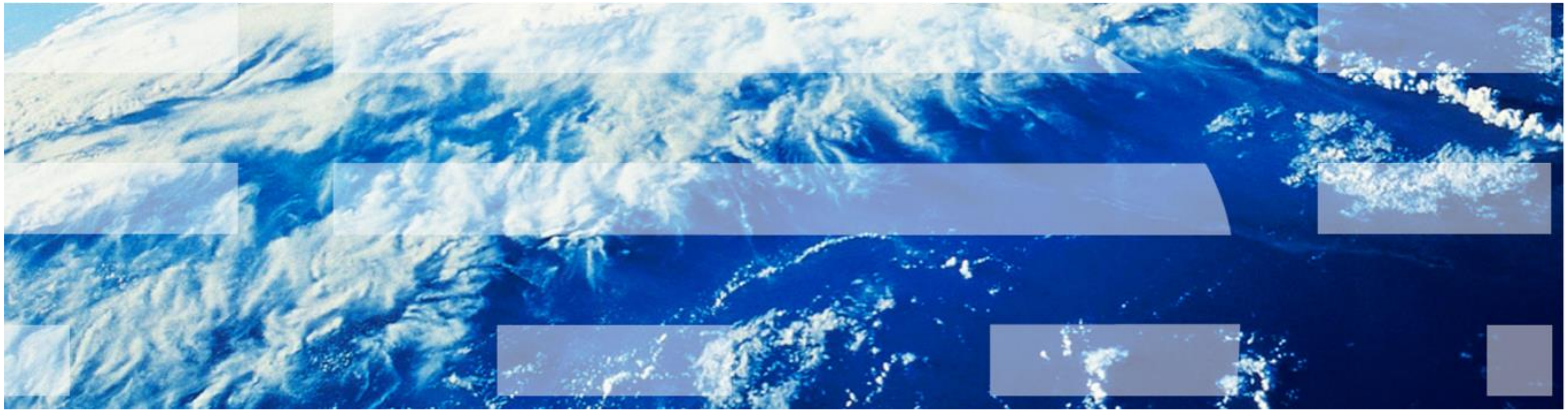
Expected Contributions:

1. Getting rid of spam reviews from the total review.
2. Recommend useful reviews to the user according to user's interest and taste.
3. Make reviews clustering by features such as environment, dishes, waiting time, service etc.

E6893 Big Data Analytics Project Proposal:

Plankton classification by Convolution neural network with Spark

Pan Li. pl2556
Ziheng Huang zh2220
Yifang Song ys2824



November 19th, 2015

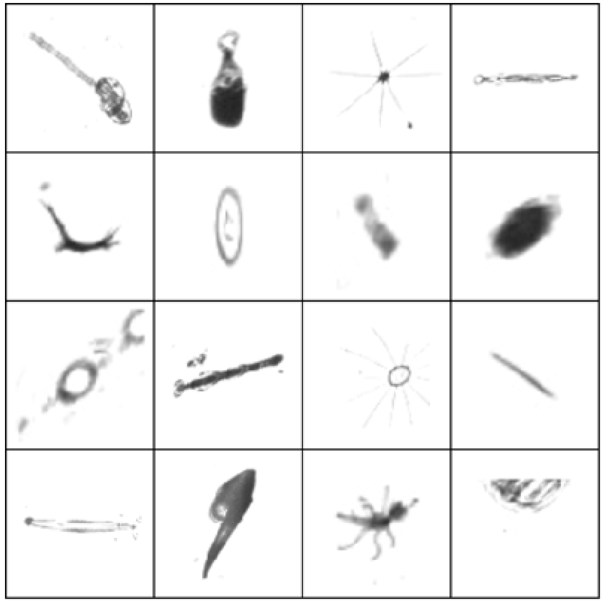
Motivation

- Deep learning methods have shown their power in recent years. Convolutional neural network is powerful deep network for image recognition. However, all deep networks suffer from extraordinary training time. In our project, we want to use the methods learned in this course to reduce that training time, and gain experience in both deep learning and big data.

Dataset, Algorithms and Tools

Dataset:

- We will use Plankton data set for this project. The data set contains more than 30000 labeled, 130000 unlabeled images of plankton. Our task is to classify them into 121 different kinds.



Tools:

- We will build this deep network in python, more specifically, by the python Theano package. After that, we will try to parallelize the training and prediction process in Spark.

Current Progress, Schedule and Expected Contributions

Current Progress and Schedule:

- By now, we have finished background reading and the preprocessing of data. By next week, we will finish the coding of first version model. We will use the time left to test our model and adjust the network architecture.

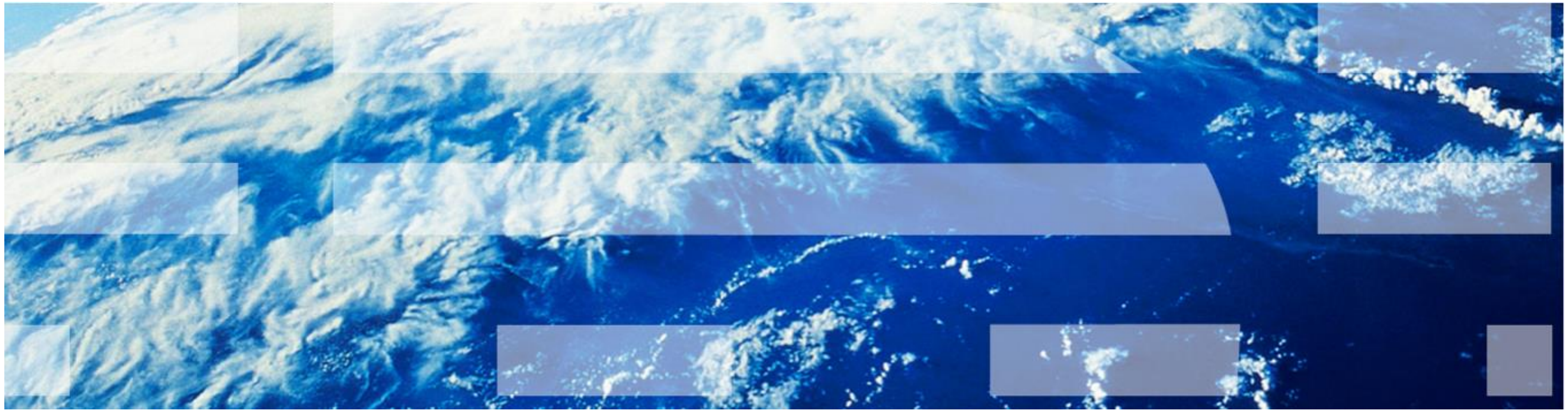
Expected Contribution:

- Ziheng and I will construct the neural network and adjust it. Yifang will see to how to parallelize the training process.

E6893 Big Data Analytics Project Proposal:

Movie Recommendation and Analytics

Tiancheng Jia
Xu Cao
Yanjing Chen



November 19th, 2015

Motivation

Why:

Many people love watching movies

Somewhat difficult to find new interesting movies after watching enough large number of them

What we will do:

Design a recommendation process to give people advice to watch new movies according to their taste

Analyze features among different genders and ages

- Yahoo! WEBSCOPE datasets
- MovieLens datasets
- Recommendation, Classification, Filtering
- Mahout, Hadoop, JAVA

Current Progress, Schedule and Expected Contributions

Current

Dataset downloaded

Pick up suitable algorithms

Schedule

This week: Analyze data tentatively

Next week: Apply different algorithms to the dataset

Expected Contributions

Recommend movies to individuals according to their rating records

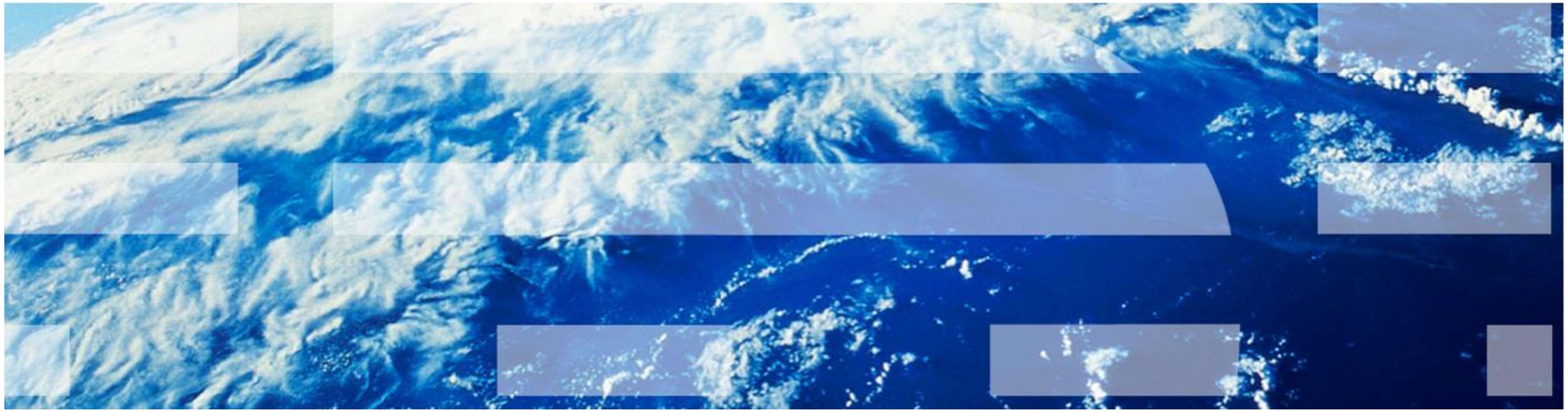
Compare characteristics in different groups of people

Optimize the recommendation algorithm by groups features

E6893 Big Data Analytics Project Proposal:

Restaurant recommendation based on Yelp data

Qianbo Wang, Yi Wu, Zuyi Wu



November 19th, 2015

Motivation

Many people rely on Yelp to explore new restaurants.

But Yelp always ‘surprises’ us with bad recommendations.

We found out that the traditional rating and recommendation has following limits:

1. No personalization.
2. Dummy users and fake review.
3. Extreme opinions affects too much.

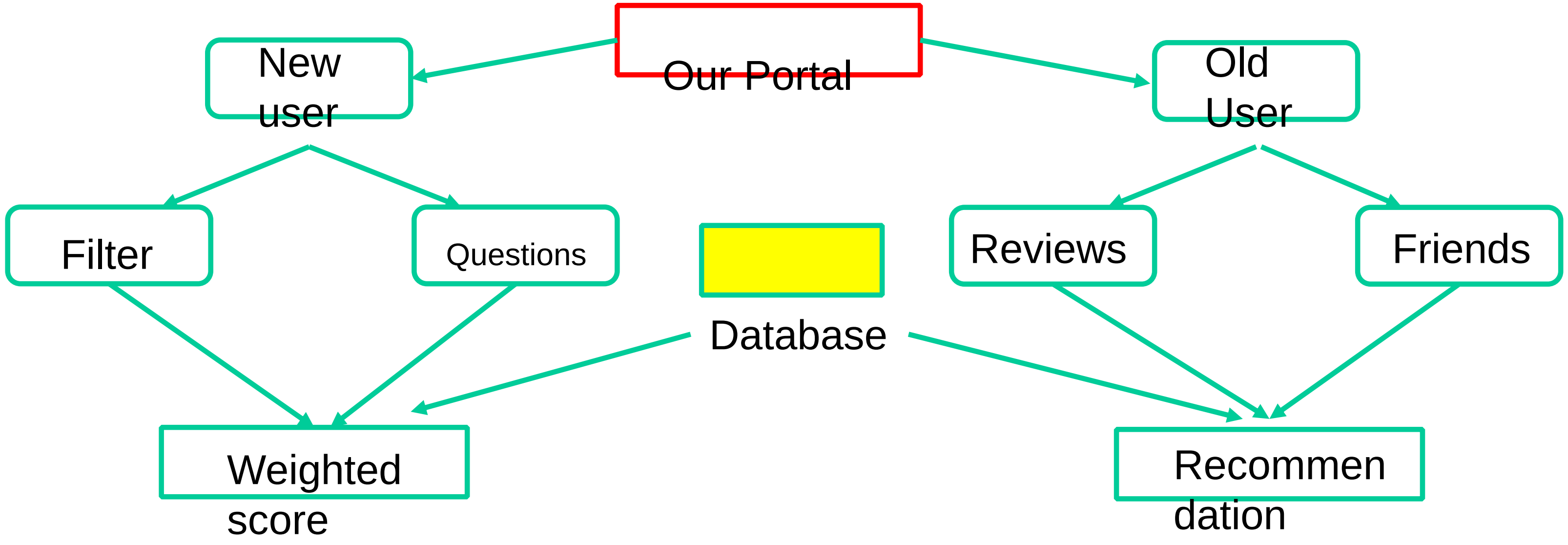
We decide to come up with a new rating and recommendation system that solves the problems above and gives more accurate advice.

✦Dataset: Yelp dataset of users, reviews, and restaurants from 10 different cities.

✦Algorithms: Weighted recommendation, natural language process, network analysis.

✦Tools:MySQL, Python, Spark, Objective C

Current Progress:
Cleaned dataset.
Adjusting key features in algorithms



Expected Contribution:
Come up with a new rating and recommendation system that solves the problems above and gives more accurate advice.

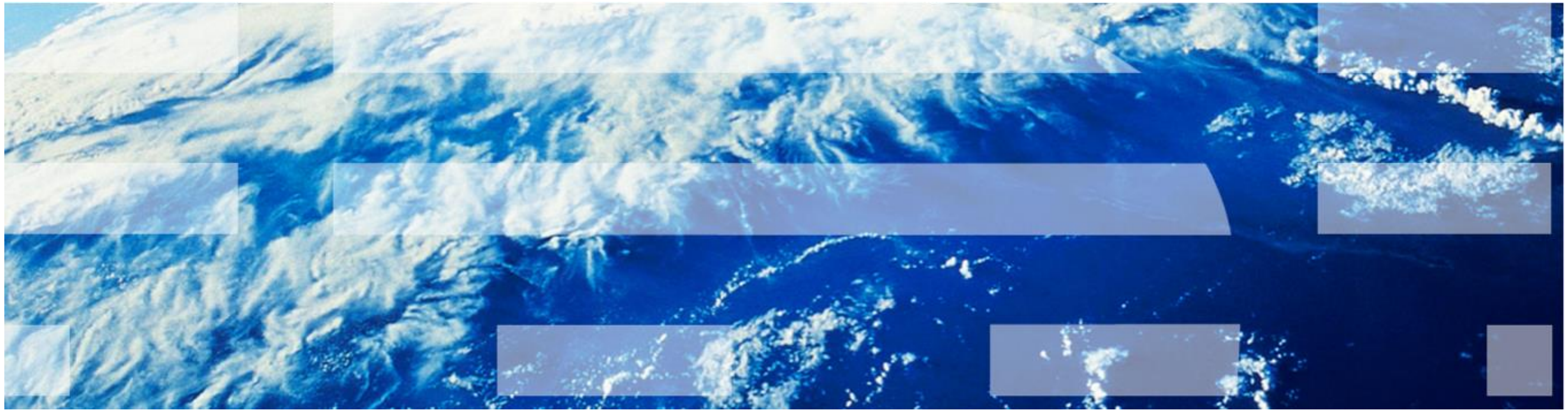
E6893 Big Data Analytics Project Proposal:

Map-based Restaurant Recommendation

Siyu Wang (sw3024)

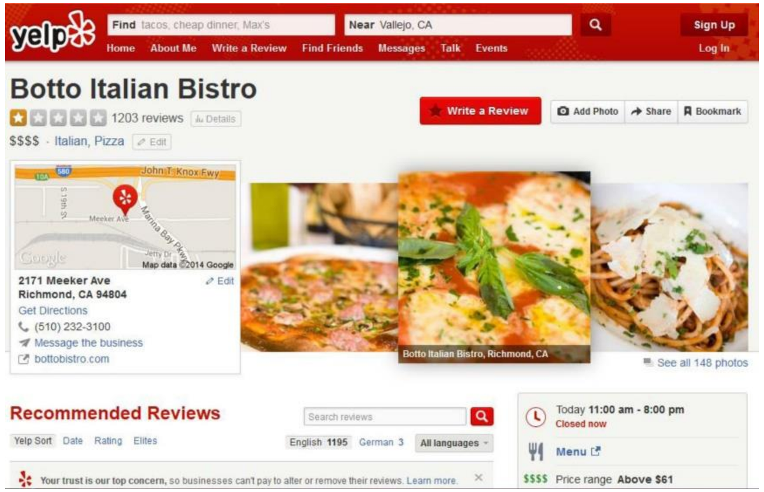
Ruoqi Wang (rw2612)

Yuyang Liu (yl3399)



November 19th, 2015

➢ Traditional recommendation system in Yelp is based on the rating simply, which is not aimed at specific customers. Our project is designed to give recommendations based on customers' own preferences.



➢ Our project visualizes the recommendation results in a map, using some map APIs to make it more clear and fascinating.

Dataset, Algorithms and Tools

- **Dataset**

  Apply the dataset from Yelp Dataset Challenge

- **Algorithms**

  Recommendation (user-based recommendation, etc.)

  Clustering

  Classification

- **Tools**

  Mahout, Hadoop, Eclipse, Google Map

Current Progress, Schedule and Expected Contributions

- **Current Progress**

  Collect the dataset

  Analyze the data which can be used in the future work

- **Schedule**

By Nov. 26th	finish analyzing data and recommendation
By Dec. 6th	apply the result in Google Map
By Dec. 16th	accomplish the whole project

- **Expected Contributions**

  Obtain and preprocess the dataset

  Analyze the data using mahout and Hadoop

  Visualize the result in the map

PHOTO: DARCY PETERKA

Analysis Between Economy and Student Studying Abroad

EECS6893 Final Project Proposal

By Presenter Weipeng Dang, Chuan Zhan

Situation

- Nowadays studying abroad is a global phenomenon.
- It has huge impact on the economy of a country.
- In 2013/2014, international students contributed over 27 billion dollars to the US economy.

There is more for analysis …

Goals

Initial Goal:

•Relationship between national economic growth and the number of students studying abroad.

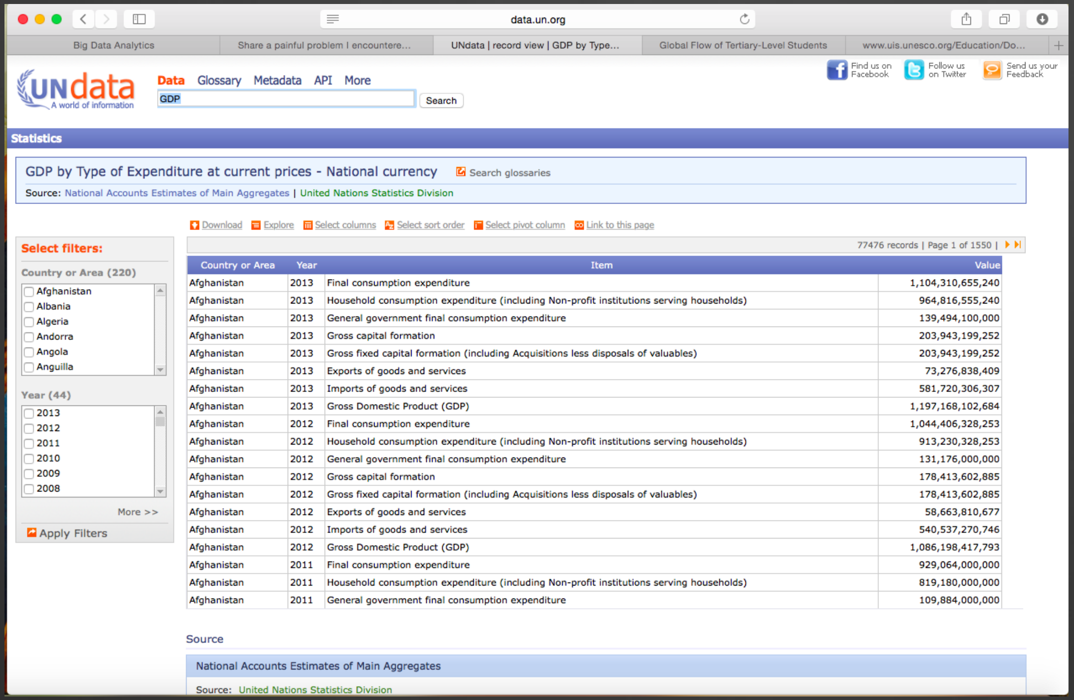
Final Goal:

•Predict national economic growth in upcoming years using the data of students studying aboard from previous years.

Sources - dataset

Data:

- http://data.un.org/Data.aspx?q=GDP&d=SNAAMA&f=grID%3a101%3bcurrID%3aNCU%3bpcFlag%3a0
- http://data.un.org/Data.aspx?q=student&d=UNESCO&f=series%3aED_FSOABS

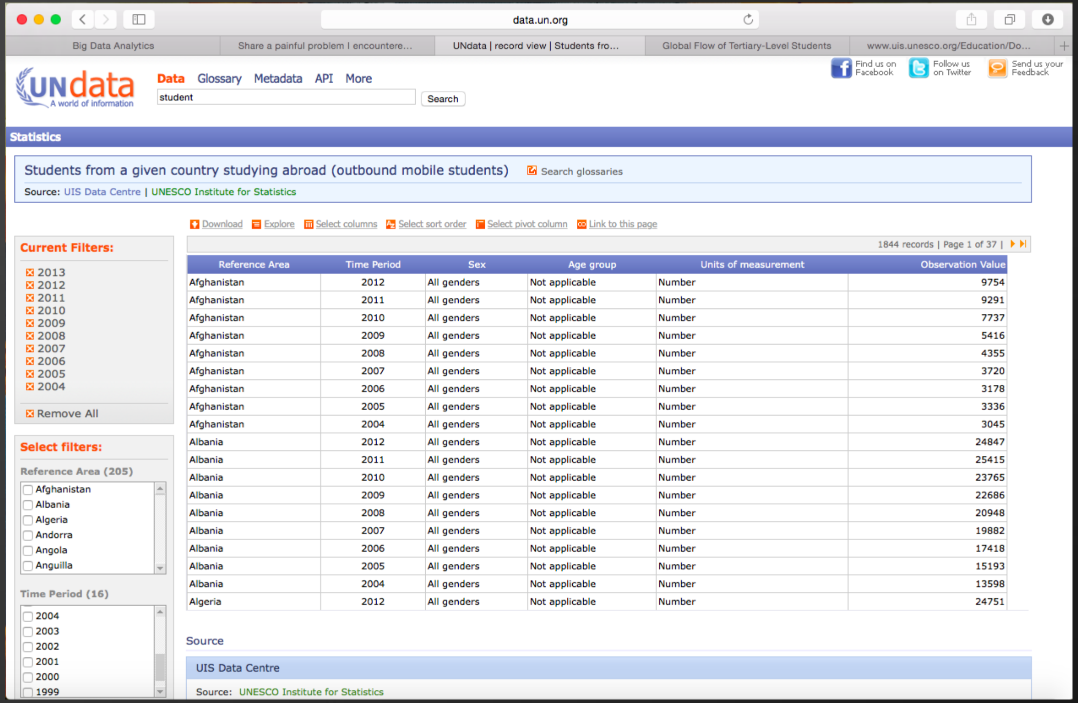


Country or Area	Year	Item	Value
Afghanistan	2013	Final consumption expenditure	1,104,310,655,240
Afghanistan	2013	Household consumption expenditure (including Non-profit institutions serving households)	964,816,555,240
Afghanistan	2013	General government final consumption expenditure	139,494,100,000
Afghanistan	2013	Gross capital formation	203,943,199,252
Afghanistan	2013	Gross fixed capital formation (including Acquisitions less disposals of valuables)	203,943,199,252
Afghanistan	2013	Exports of goods and services	73,276,838,409
Afghanistan	2013	Imports of goods and services	581,720,306,307
Afghanistan	2013	Gross Domestic Product (GDP)	1,197,168,102,684
Afghanistan	2012	Final consumption expenditure	1,044,406,328,253
Afghanistan	2012	Household consumption expenditure (including Non-profit institutions serving households)	913,230,328,253
Afghanistan	2012	General government final consumption expenditure	131,176,000,000
Afghanistan	2012	Gross capital formation	178,413,602,885
Afghanistan	2012	Gross fixed capital formation (including Acquisitions less disposals of valuables)	178,413,602,885
Afghanistan	2012	Exports of goods and services	58,663,810,677
Afghanistan	2012	Imports of goods and services	540,537,270,746
Afghanistan	2012	Gross Domestic Product (GDP)	1,086,198,417,793
Afghanistan	2011	Final consumption expenditure	929,064,000,000
Afghanistan	2011	Household consumption expenditure (including Non-profit institutions serving households)	819,180,000,000
Afghanistan	2011	General government final consumption expenditure	109,884,000,000

Source

National Accounts Estimates of Main Aggregates

Source: United Nations Statistics Division



Reference Area	Time Period	Sex	Age group	Units of measurement	Observation Value
Afghanistan	2012	All genders	Not applicable	Number	9754
Afghanistan	2011	All genders	Not applicable	Number	9291
Afghanistan	2010	All genders	Not applicable	Number	7737
Afghanistan	2009	All genders	Not applicable	Number	5416
Afghanistan	2008	All genders	Not applicable	Number	4355
Afghanistan	2007	All genders	Not applicable	Number	3720
Afghanistan	2006	All genders	Not applicable	Number	3178
Afghanistan	2005	All genders	Not applicable	Number	3336
Afghanistan	2004	All genders	Not applicable	Number	3045
Albania	2012	All genders	Not applicable	Number	24847
Albania	2011	All genders	Not applicable	Number	25415
Albania	2010	All genders	Not applicable	Number	23765
Albania	2009	All genders	Not applicable	Number	22686
Albania	2008	All genders	Not applicable	Number	20948
Albania	2007	All genders	Not applicable	Number	19882
Albania	2006	All genders	Not applicable	Number	17418
Albania	2005	All genders	Not applicable	Number	15193
Albania	2004	All genders	Not applicable	Number	13598
Algeria	2012	All genders	Not applicable	Number	24751

Source

UIS Data Centre

Source: UNESCO Institute for Statistics

Sources – algorithm & tools

Algorithm:

•Filtering

•Clustering

•Classfication

•Linear regression

Tools:

- Hadoop
- Pig
- Spark
- System G
- Excel

Schedule & Contribution

Week 11.20 – 11.26

•Weipeng: Analyzing economy data for various countries at some specific times (5-year period).

•Chuan: Analyzing student studying abroad data for various countries at corresponding times.

Week 11.27 - 12.3

•Weipeng: Track economy data trend for some specific countries in 15 years.

•Chuan: Track student studying abroad data for some specific countries in 15 years.

Week 12.4 – 12.10

•Computing and analyzing the relationship between economy growth and the number of students studying abroad.

Week 12.11 – 12.16

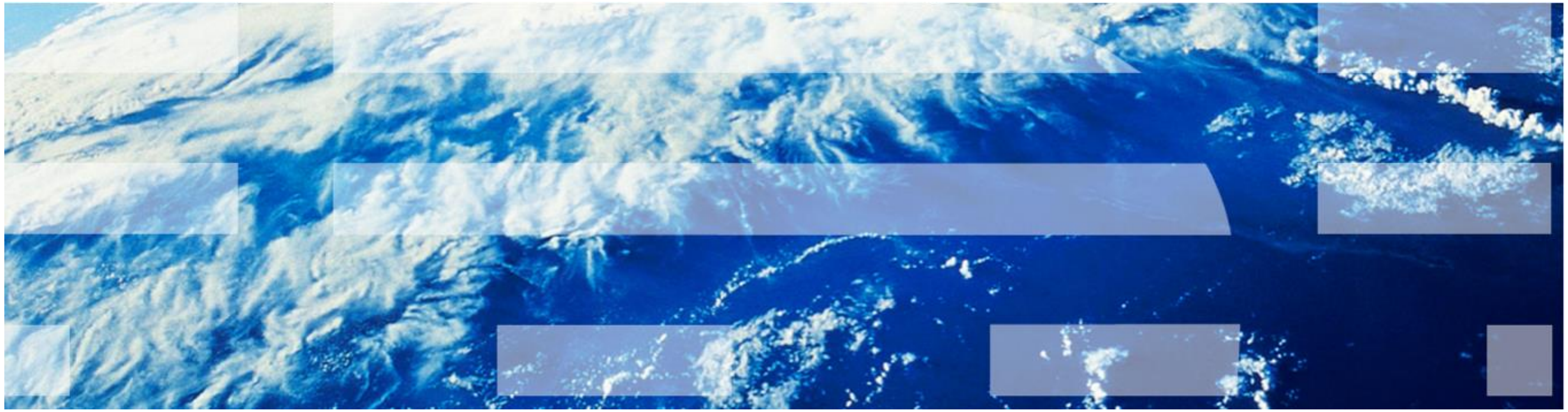
•Generating graphs and making presentation slides.

Thank you!

E6893 Big Data Analytics Project Proposal:

Cross-source Event Detection Through Social Media

Team Members: Cai, Zhuxi Wang, Sitian Shi, Yi



November 19th, 2015

Motivation

Columbia University

Simon Kuper @KuperSimon · Nov 13
I have no idea what those explosions outside the **Stade de France** were. May be benign but just heard police sirens.If anyone knows please say

RETWEETS 57 LIKES 19

12:24 PM - 13 Nov 2015 Details

local twitter

Huge Time Difference

CNN Breaking News @cnnbrk · Nov 13
Report: Several people killed, injured in Paris shooting. cnn.it/1MsFhuz

1:13 PM - 13 Nov 2015 · Details

Hide summary

Reuters Top News @Reuters · Nov 13
BREAKING: Deadly shooting in restaurant in central Paris: BSM TV

RETWEETS 523 LIKES 95

1:17 PM - 13 Nov 2015 · Details

Official Meida

Dataset:

2010-2015 target events local twitter data and major new service report data

Data sample:


```
{ "_id" : { "$oid" : "549c07837edd910e062851e8" }, "contributors" : null, "coordinates" :
null, "country" : "", "created_at" : { "$date" : "2014-12-25T12:48:03.000+0000" },
"created_date" : "2014-12-25", "curate_date" : "", "curate_flag" : "N", "embed_url" : "",
"entities" : { "user_mentions" : [], "symbols" : [], "trends" : [], "hashtags" : [], "urls" :
[] }, "favorite_count" : 0, "favorited" : false, "filter_level" : "medium", "geo" : null,
"id" : { "$numberLong" : "548097841668042752" }, "id_str" : "548097841668042752",
"in_reply_to_screen_name" : null, "in_reply_to_status_id" : null,
"in_reply_to_status_id_str" : null, "in_reply_to_user_id" : null, "in_reply_to_user_id_str" :
null, "lang" : "tr", "place" : null, "possibly_sensitive" : false, "rec_no" : 31787978,
"retweet_count" : 0, "retweeted" : false, "source" : "<a href=\"http://twitter.com\" rel=
\"nofollow\">Twitter Web Client</a>", "text" : "sevdiğin insanları kaybetmeye alıştığın zaman;
artık hayatı önemsememeye başlıyorsun..", "timestamp_ms" : "1419511683276", "truncated" :
false, "tweet_type" : "text", "user" : { "follow_request_sent" : null,
"profile_use_background_image" : true, "default_profile_image" : false, "id" : 824092764,
```


Algorithm:

Generalized Linear Model, NLP(majorly Sentiment Analysis)

Tools:

PySpark, D3, Twitter API

Current Progress, Schedule and Expected Contributions

Current Progress:

- Post comparison between twitter and famous global news service focused on Paris Terror Attack
- Sample twitter data extracted by Twitter API
- Hands-on experience in tools: PySpark, D3, Twitter API

Schedule:

- Until 11/19: Topic Selection, Data Collecting(part) and Presentation Slides
- 11/20 – 11/26: Finish Twitter data collecting and news data collecting on target events in year 2010-2015
- 11/27 – 12/03: Visualization of twitter data and news data on target events based on timeline and distribution in world map
- 12/04 – 12/10: Implement sentiment analysis in PySpark and build prediction model
- 12/11 – 12/17: Test model on Paris Terror Attack data

Expected Contributions:

- Discover the news value hidden in twitter comparing to official news agencies
- Build a model to predict the time, topic and content of coming global news based on earlier related twitter
- (Optional) Develop user interface to share news value of twitter

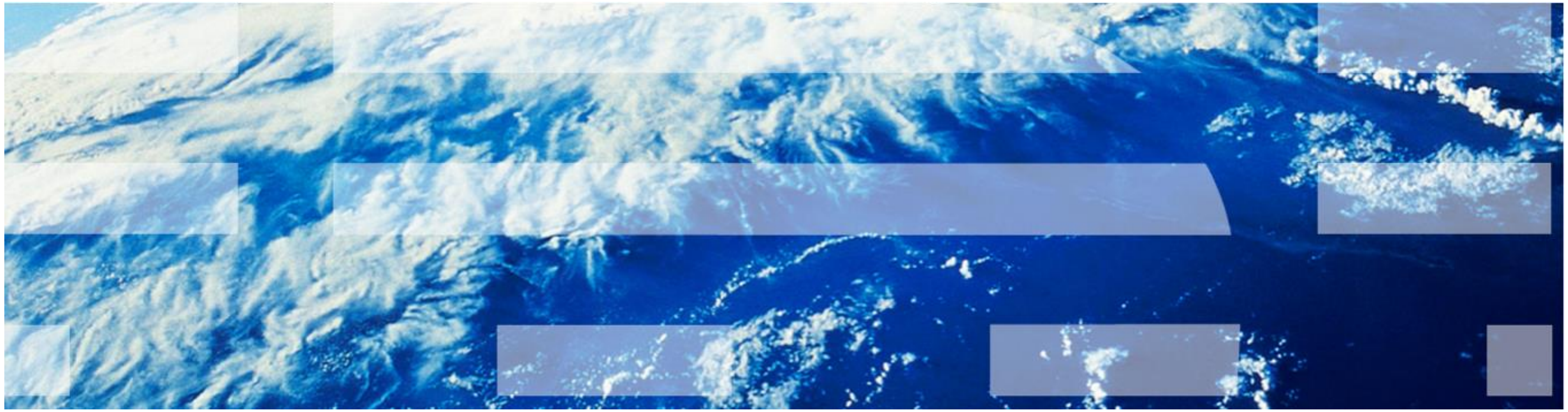
E6893 Big Data Analytics Project Proposal:

Face Detection

Justine Morgan

Stamatios Paterakis

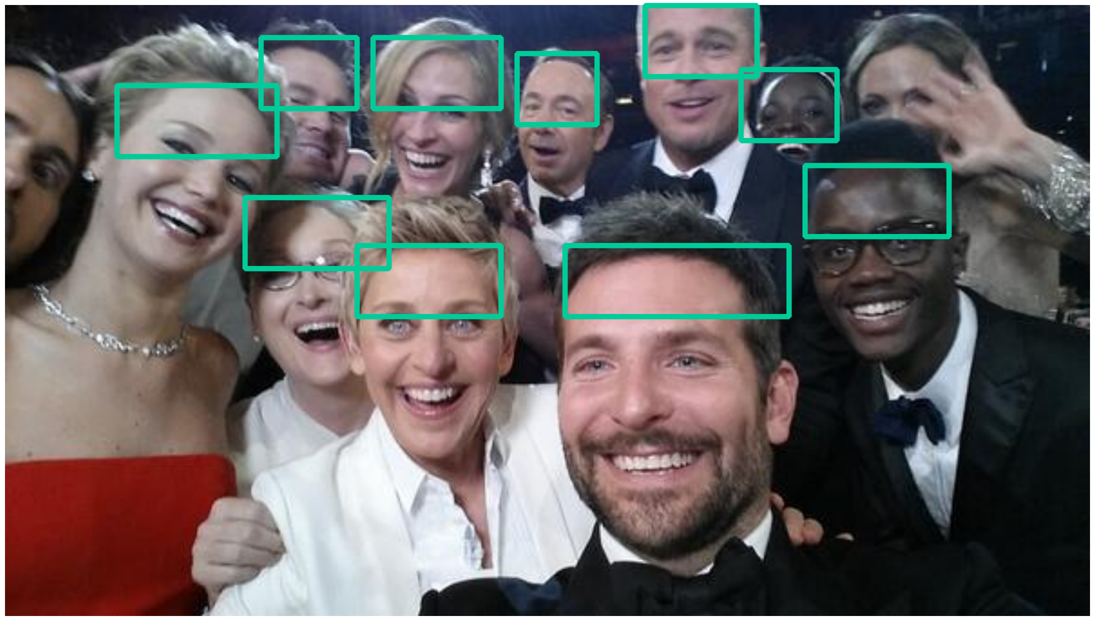
Lauren Valdivia



November 19th, 2015

Motivation

Goal:
Implement facial detection algorithms that are robust to lighting, angle, scale and background.



https://www.youtube.com/watch?v=aTErTqOIkss

Use Cases:
•Biometrics, often paired with facial recognition, for use in video surveillance, human computer interface, and image database management.
•Photography and Videography (autofocus, social media "tagging")

Dataset, Algorithms and Tools

Datasets

✦CMU/Vasc image database

✦FaceScrub celebrity photos

✦Feret image database

✦BioID face database

Algorithms

Preprocessing Phase

✦Edge Detection – Sobel Operator

✦Scaling / Window Extraction

✦Rotation Correction

Detection Phase

✦Neural Networks

✦PCA – Eigenface Decomposition

✦Viola-Jones (Adaboost with Cascades)

Tools

✦Apache Spark

✦Python

Current Progress, Schedule and Expected Contributions

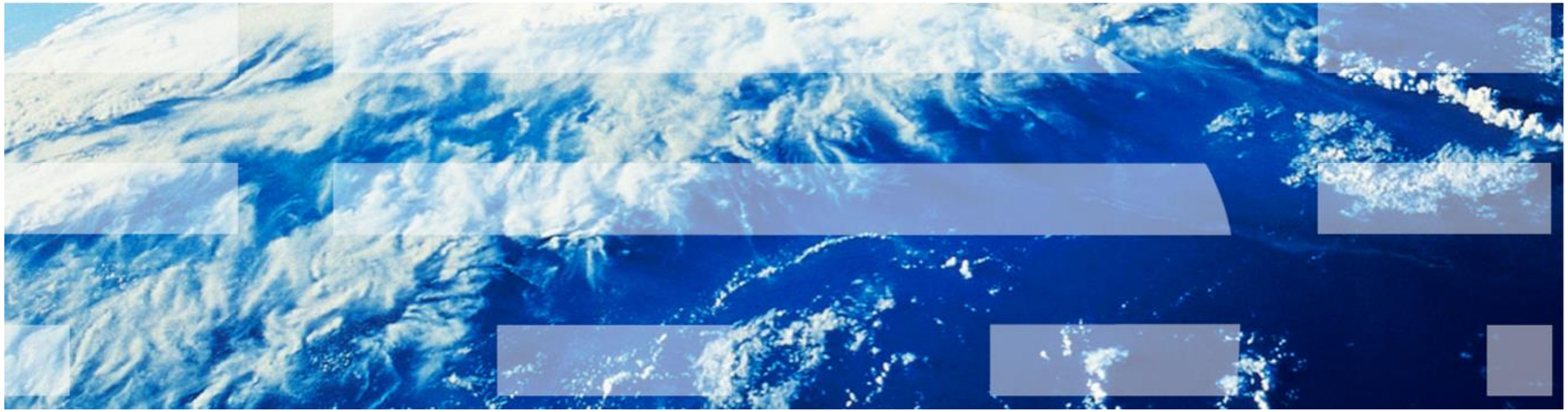
1. Research and choose topic -- Complete
2. Find datasets -- Complete
3. Compile research and choose algorithms to implement -- Complete
4. Clean and compile data -- Deadline: 11/23/15
5. Implement Algorithms – Deadline: 12/07/15
   - Neural Networks (Lead: Stamatios Paterakis)
   - PCA (Lead: Justine Morgan)
   - Viola-Jones (Lead: Lauren Valdivia)
6. Train and test algorithms -- Deadline: 12/14/15
7. Prepare Presentation -- Deadline: 12/17/15

Note: Each team member is expected to contribute equally in each step and fully understand each algorithm. Leads were assigned to the algorithms for organizational purposes.

E6893 Big Data Analytics Project Proposal:

Large Scale Video Search and Retrieval via CNN

Zheng Shou, Hongyi Liu, Weiye Hu



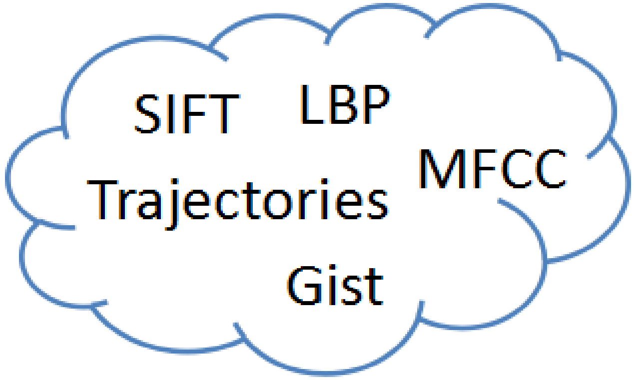
November 19th, 2015

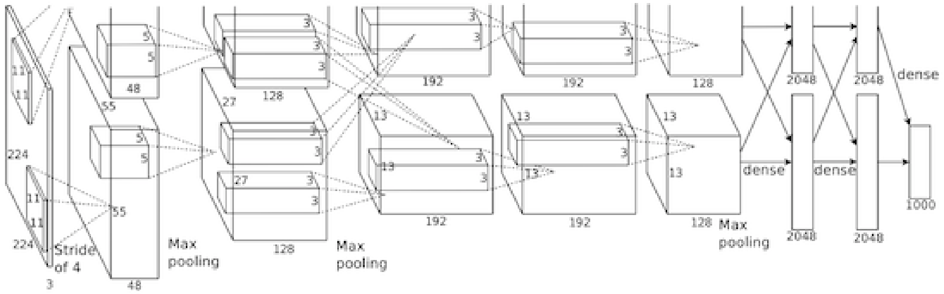
- **Motivation:**
  - benefit a lot of applications

Companies
Ads recommendation

Surveillance

- **Current Trends**





- CNN features: great success in many areas

Dataset, Algorithms and Tools

- Dataset:
  - UCF101:
  1. collected from YouTube
  2. 13320 videos from 101 action categories
  3. challenging: large variations in camera motion, object appearance and pose, object scale, viewpoint, illumination conditions, etc.
  - EventNet ?: a large scale structured concept library

- Algorithms:
  1. Extracting deep learning features. CNN model trained on ImageNet.
  2. Encode video into binary hash code.

- Tools:
  - Caffe: a deep learning framework
  - Matlab
  - Tomcat Apache, D3.js

Current Progress, Schedule and Expected Contributions

- Current progress:
    - Downloaded dataset
    - On-going: extracting CNNs features

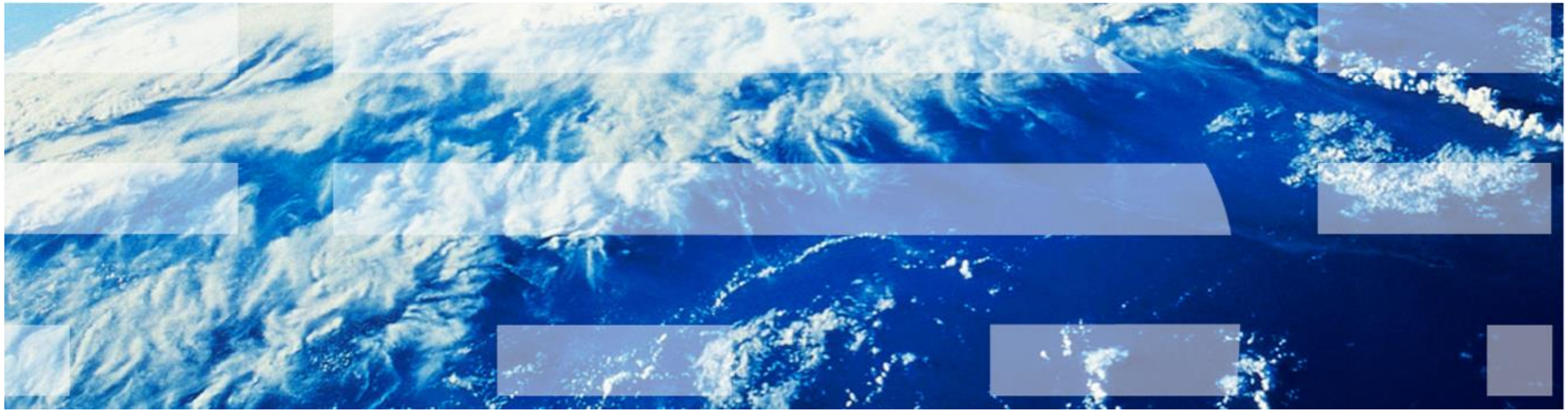
- Schedule:
    - Nov. 19 – Nov. 30. Feature Extraction, Generating Hashing Codes
    - Dec. 01 – Dec. 10. Development of Online Demo
    - Dec. 11 – Dec. 17. Technical Report and Presentation Preparation

- Expected contributions:
    - Fast retrieval of relevant videos
    - Demo: web interface
    - Technical Report

E6893 Big Data Analytics Project Proposal:

Product Review Helpfulness Prediction on Amazon Dataset

Chengcheng Du (cd2789), Qiurui Jin (qj2131), Jianhao Li (jl4350)



November 19th, 2015

Motivation

Columbia University

- Amazon is the largest internet-based retailer in the United States. High quality reviews are very important to help customer to make decisions when shopping at Amazon. However, some helpful reviews may be buried in overwhelming useless reviews. It would be helpful if we can dig them out
- Our project aims at predicting the helpfulness of reviews. So that we can know whether a review is helpful or not, even not many people see it. Then we can put helpful reviews in positions where customers can easily see and vote

Most helpful positive review

See all 104 positive reviews ›

173 of 177 people found the following review helpful

SONY 4K ULTRA HD CLARITY!!!

By JoeRod on May 20, 2015

So after having the X850C for about a week now I have to say we are very pleased. Out of the box the Sharpness needed to be dialed down some. I ended up between 45-50. I use VIVID and tweaked some of the picture settings. There is also another key setting to keep a lookout for after you hit the HOME button. It's under Picture and Display after you tap Settings. Go to Dynamic Range. Pause your image and toggle between LIMITED and FULL. Depending on your source one will blow the other out of the water. Image looked a little cloudy but after changing it blacks looked great and the image was full of POP! Also for best results turn the Light a Sensor off.

SETTINGS:

Read more

Dataset, Algorithms and Tools

Dataset

Product reviews and metadata from Amazon since May 1996 to July 2014. They include ratings, text helpfulness and votes

Number of reviews: 34,686,770

Number of products: 2,441,053

Algorithms

Logistic Regression, SVM, Naive Bayes, Gradient Boosted Decision Trees and Random Forest

Tools

NLTK: Tokenization, stemming, stop words removal, ngram generation

Pandas: Statistics and visualization of data

Scikit-learn: Machine Learning algorithms

Mahout: Distributed machine learning algorithms

Current Progress, Schedule and Expected Contributions

Current Progress:

Found interesting dataset
Finalized the goal of this project

Schedule:

11.16 - 11.24 Find proper data preprocessing tools and machine learning algorithms.
11.25 - 12.02 Prototype the whole pipeline
12.03 - 12.10 Iterate and adjust each components of the pipeline to get better performance
12.11 - 12.15 Summarize experiment results and prepare final project slides.
12.17 Final presentation.

Expected Contributions:

A system which can accept input of review data and output whether this review is helpful or not.

San Francisco Crime Rate

Chong Zhou
Chen Zheng

Objective

1. Where is the most high crime rate areas in SF?

2. The relationship between crime rate, crime type and location in SF.

3. The relationship between crime and time.

4. Which crimes and where are easy to solve?

5. How to distribute police manpower?

Data Source

- Date: time
- Category: crime type
- Descript: description of crime
- Resolution: resolve the crime or not
- Location
- Axis

Software and Algorithm

- Software: Python, Hive, Pig-Latin, D3.js, R.
- Platforms: Mac OSX
- Algorithm: Kmeans
    - Recommendation

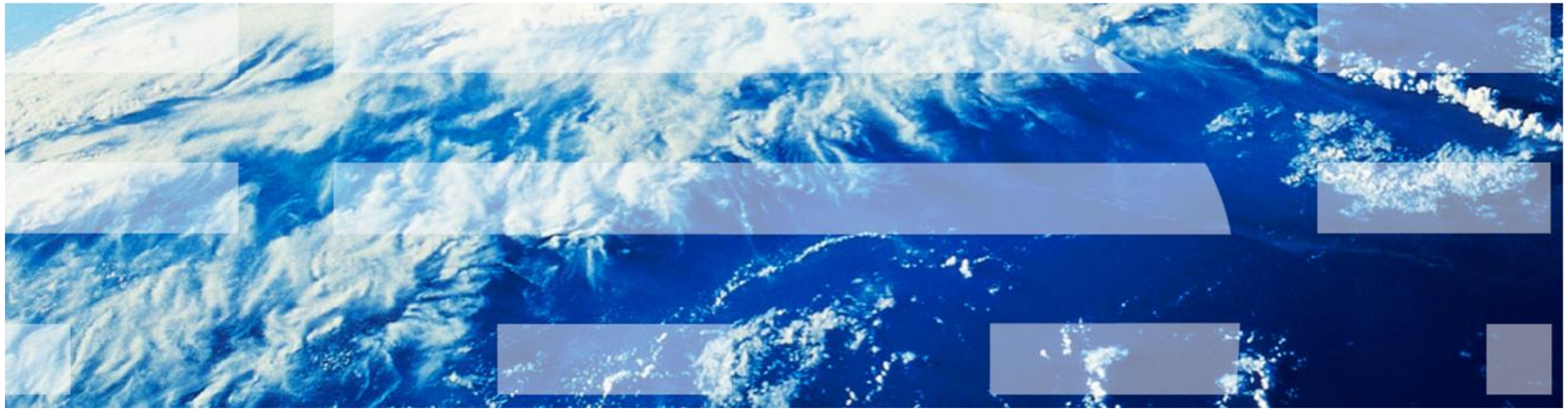
      Time Series

      Analysis

E6893 Big Data Analytics Project Proposal:

Visualization of Spatial Temporal Patterns of User Tweeting Behavior on Information Diffusion Process

Palash Sushil Matey (pm2824)
Sarat Chandra Vysyaraju (scv2114)
Shivam Choudhary (sc3973)



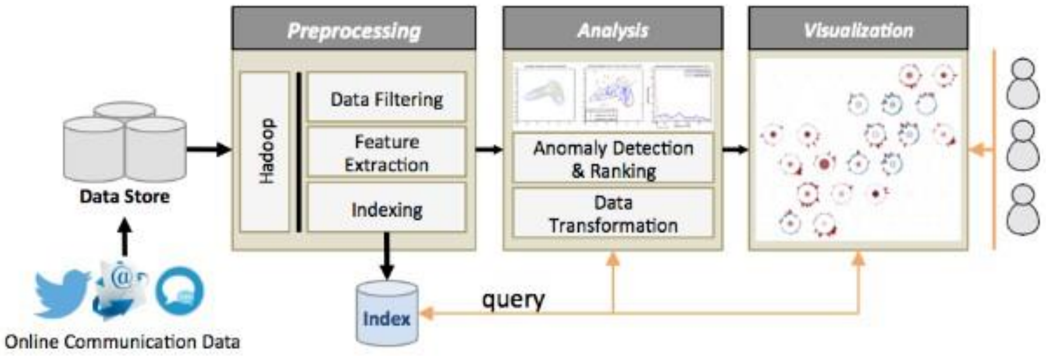
November 19th, 2015

Motivation

- Lots of data is available on social media websites like Twitter and Facebook which can provide valuable insight to researchers and practitioners in many application domains such as marketing.
- However, an important challenge is to discern the anomalous information behaviours leading to misinformation and rumours from the conventional patterns.
- These anomalous information trends may have a considerable impact and thus, It is very important to model and measure information diffusion patterns in social media.
- The complicated and highly dynamic nature of the data makes it important to involve human supervision in the analysis of anomalous information spreading.
- Thus we propose to develop an interactive visualisation platform to observe the spatio-temporal patterns on twitter data

Dataset, Algorithms and Tools

- We plan to use the Target Vue visualisation model, which is somewhat similar and improved version of the FluxFlow tool and then integrate it with the Whisper technique to incorporate the spatio-information of the tweets and re-tweets in the diffusion process.
- Target Vue uses machine learning algorithm based on the OCCRF (One-class conditional random fields) model because of the one-class nature of data (i.e., little knowledge about true anomalies) and highly time-dependent structures.
- Time-adaptive Local Outlier Factor (TLOF) Model -an unsupervised machine learning algorithm to score the users for ranking is used in the Target Vue system
- The dataset we are going to use will be in the gnip format, with a probable size of 7GB
  http://support.gnip.com/sources/twitter/data_format.html
- We are planning to implement the back-end analysis on Spark and Hadoop HDFS and then integrate the back-end analysis over to a visual interface



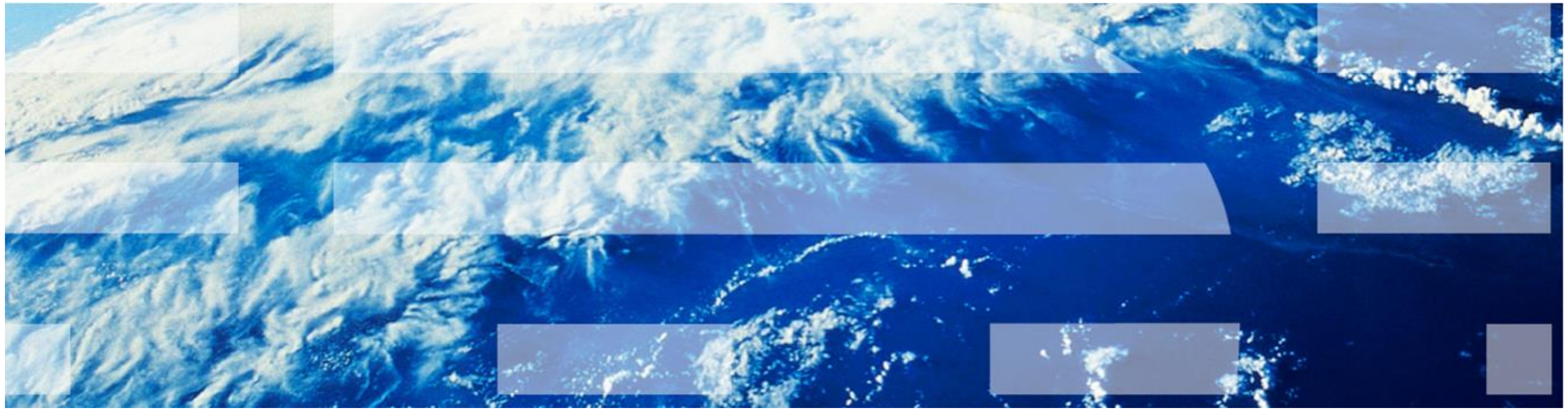
Current Progress, Schedule and Expected Contributions

Date	Description
16th Nov -22nd Nov	Background Research and Literature Review
22nd Nov -30th Nov	Implement the pre-processing steps
30th Nov -10th Dec	Implement Data Analysis Tasks (OCCRF and TLOF)
10th Dec -20th Dec	Integrate the processed data with Visual Interface

E6893 Big Data Analytics Project Proposal:

Structural Health Monitoring Of Bridges

Cihat Cagin Yakar cy2364
Karl Bayer ksb2153
Ziyue Jin zj2187

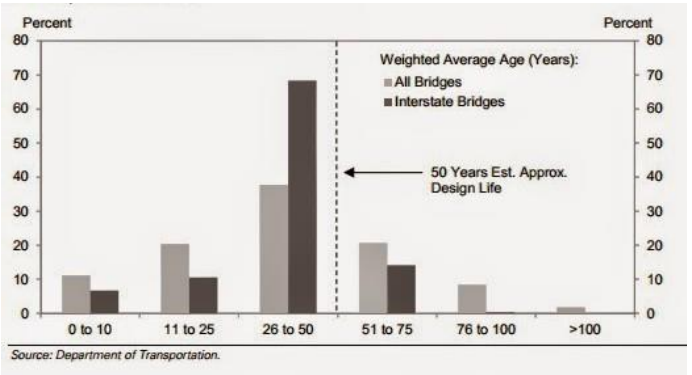


November 19th, 2015

Motivation

Describe the motivation of your project

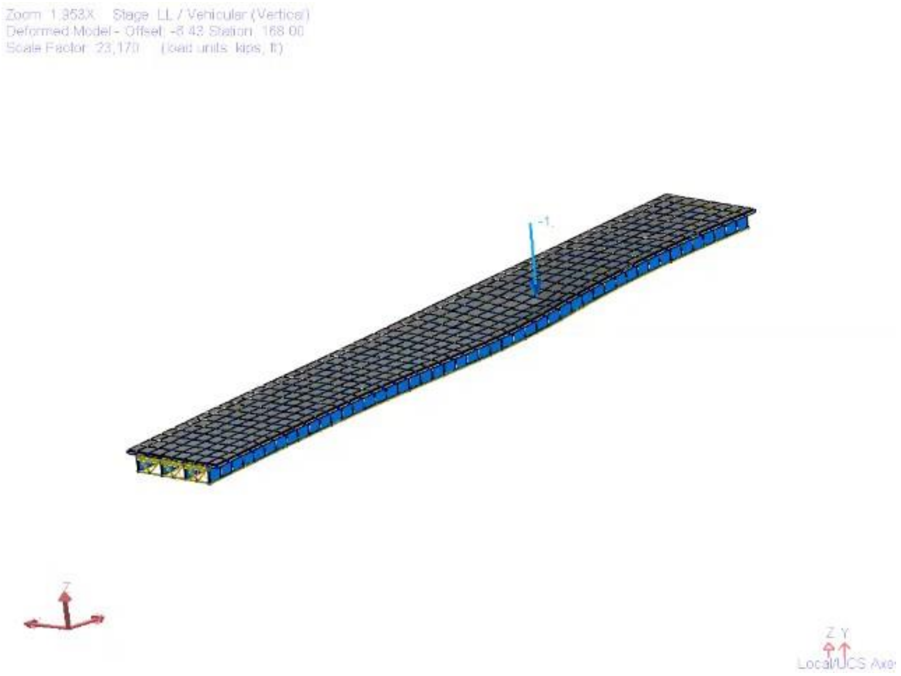
- Approximately 30% of bridges in the US are beyond their design lifetime.
- Given the aging infrastructure, how can we predict failure or know what to prioritize?



Andy Herrmann: “One of these arch bridges actually has a structure built under it to catch falling deck. See that structure underneath it? They actually built that to catch any of the falling concrete so it wouldn't hit traffic underneath it.”

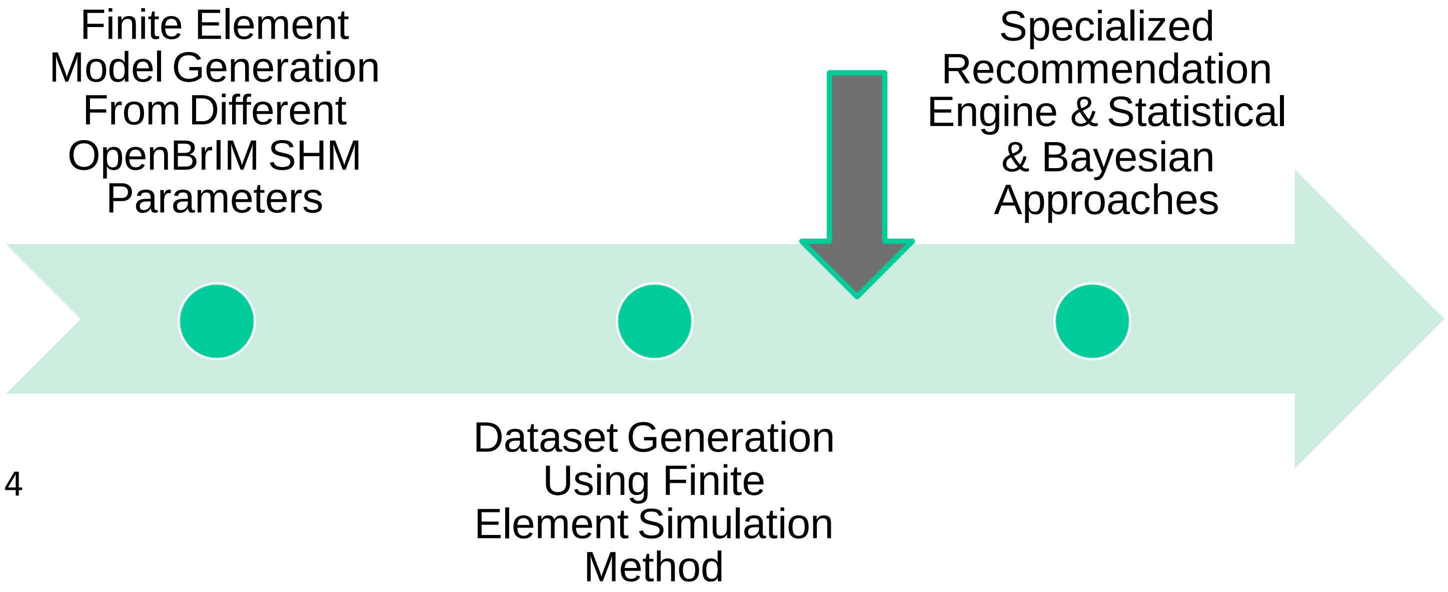
Dataset, Algorithms and Tools

- Dataset:
  - Simulation Results -Acceleration Response of a Bridge
- Algorithms:
  - Recommendation Algorithms
  - Finite Element Algorithms
- TOOLS:
  - LARSA 4D Finite Element Software
  - OPENBrIM
  - Mahout



Current Progress, Schedule and Expected Contributions

- Build finite element models to define different damaged state of the bridge
- Develop and train recommender to capture current state of bridge
- Damage location detection algorithm

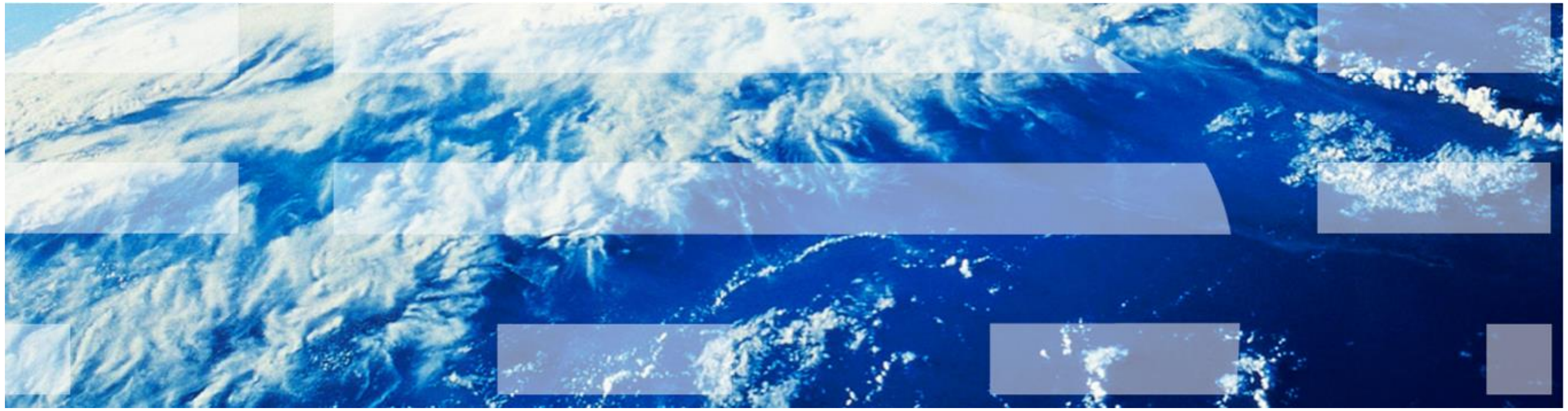


E6893 Big Data Analytics Project Proposal:

Cost and Return of College Education in the US

Lian Liu

ll2698



November 19th, 2015

College Scoreboard Project

US department of Education

https://collegescorecard.ed.gov/data/

Goal: *To provide more data than ever before to help students and families compare college costs and outcomes as they weigh the trade offs of different colleges, accounting for their own needs and educational goals*

Data Source: These data are provided through federal reporting from institutions, data on federal financial aid, and tax information.

Dataset

College Scorecard data from 1996 to 2013 for around 8000 schools

Category	Example
Root	Unique school ID, Location, etc.
School	Name, Type, Degree type, Religious Affiliation, etc.
Academics	Programs, etc.
Admission	Rate, SAT scores, etc.
Student	Number, Ethnicity, etc
Cost	Tuition and Fees, etc.
Aid	Loans, etc.
Repayment	Cohort default rate, etc.
Completion	Completion rates, retention rates, etc.
Earnings	Average and median earnings, etc.

Algorithms and Tools

Tools: Spark, Python, MongoDB, D3.js

Three main objectives:

1.Summary Statistics and interactive visualization (Spark, mongoDB, python, D3.js)
2.Prediction for a better decision: earnings, cost, completion rate, admission rate, debt? (Spark, Python)
3.School groups: How do schools compare with each other? (Spark, Python)

Automated Ticket Price Drop Reporting

Jake Dosoudil (JD3225)
Jake Wood (JGW2128)
Weiyi Zhou (WZ2333)

Job Automator

Problems with Oozie

- Environment dependent
- Can only run Hadoop jobs
- Compatibility issues
- Very complex

A Better Job Automator

- Environment independent
- Can run any job
- Easy to install/run

New Job Automator Features

- Asynchronous
- Time-based (cron)
- Dependency based
- Simple interface

Applying the Job Automator

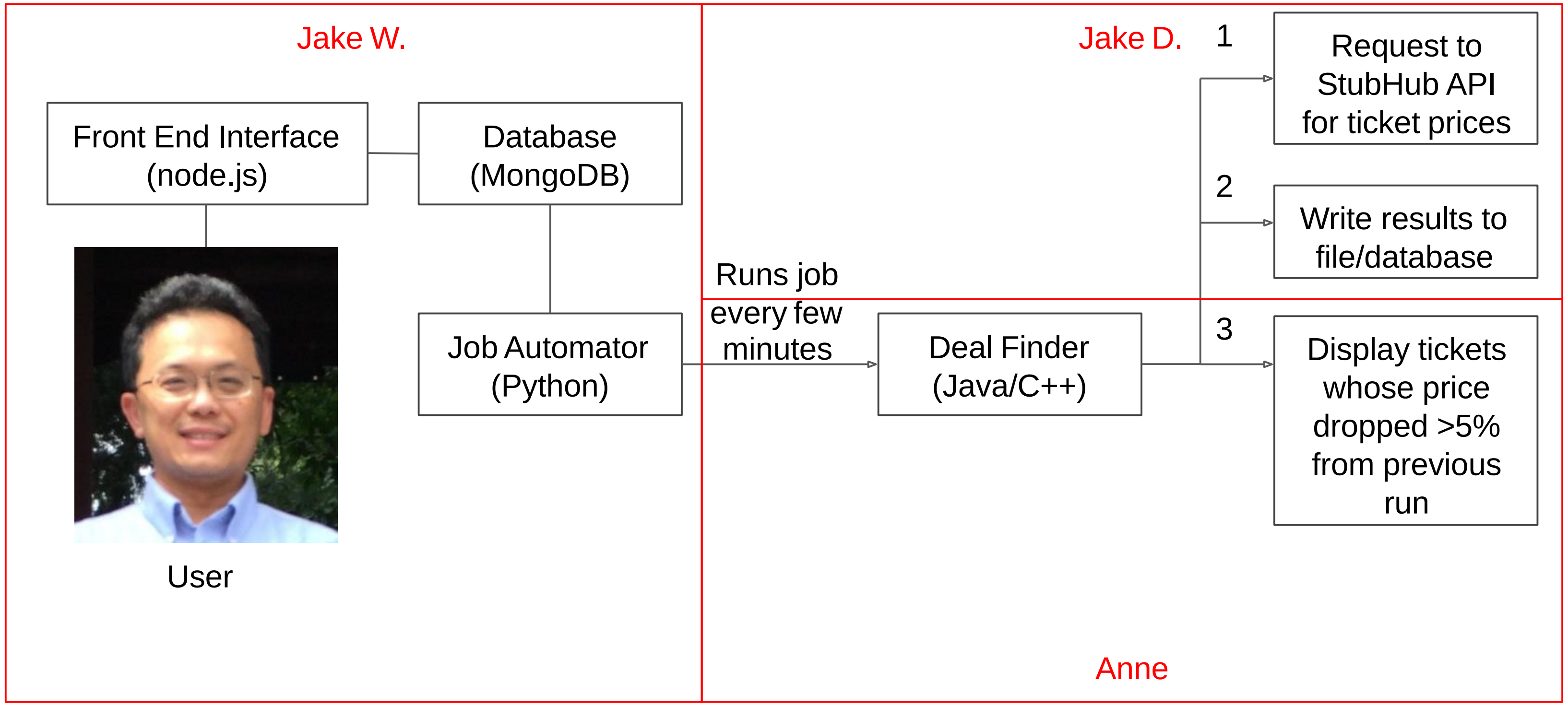
Goal:

- Create StubHub ticket price-drop reporter

Implementation:

- Display largest price drops in tickets (~5%)
    - using Job Automator on set time intervals
    - Organize data based on venue location, event

Block Diagram



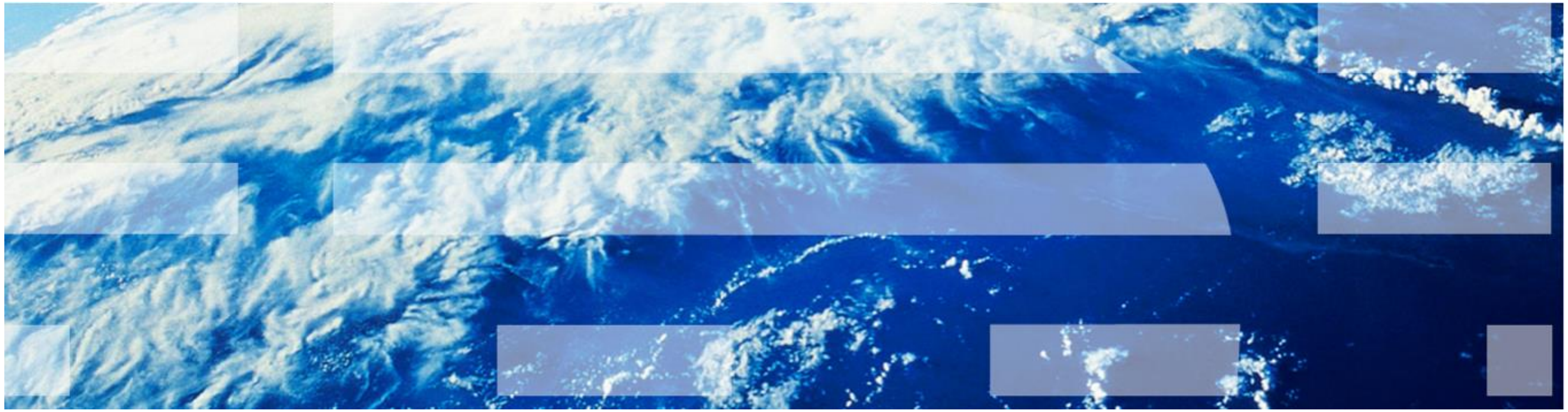
E6893 Big Data Analytics Project Proposal:

Decentralized Indoor Positioning Based on WLAN Fingerprints

Li Niu

Bin Wang

Chang Liu



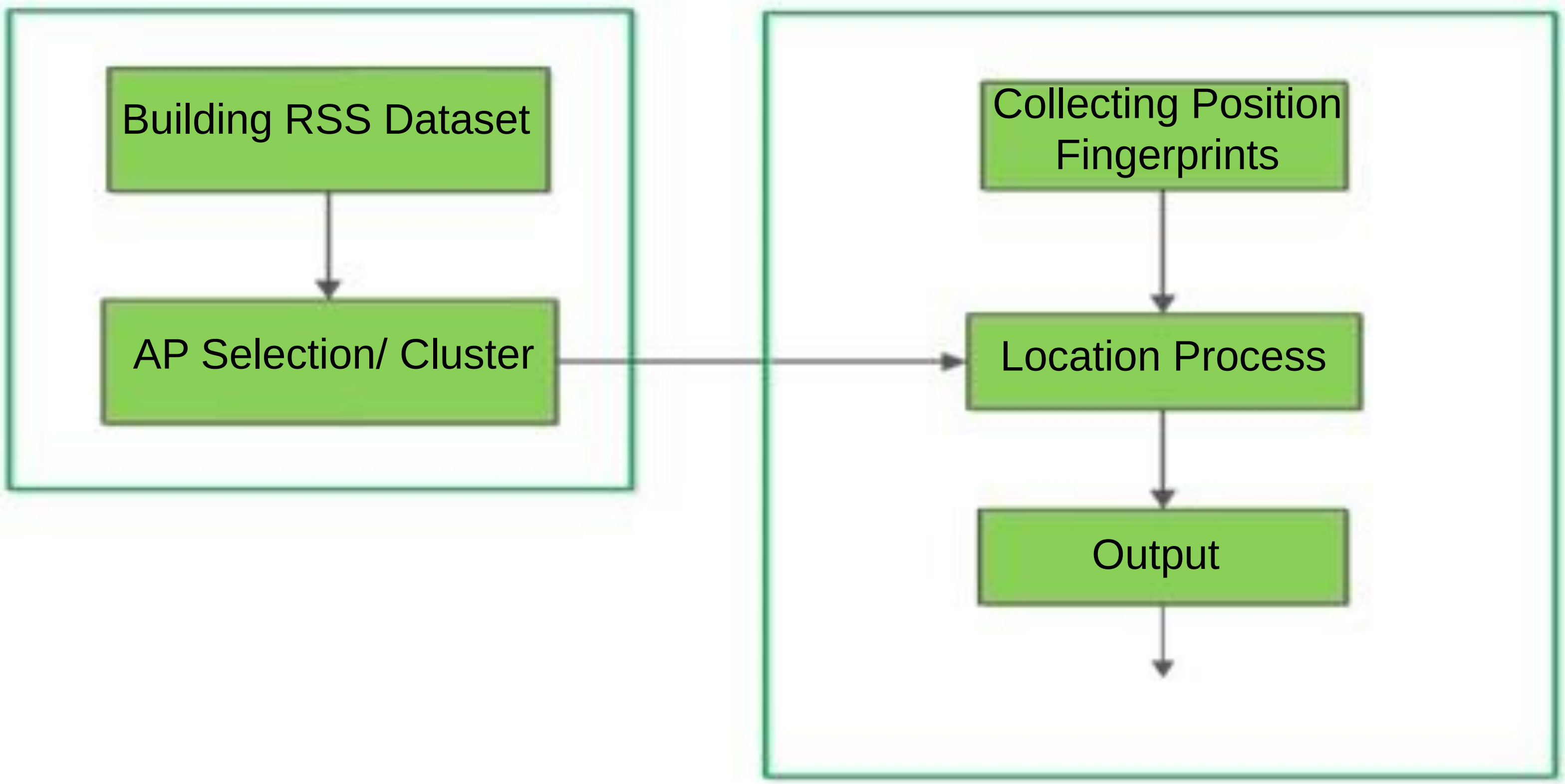
November 19th, 2015

Location Based Service:

Indoor vs Outdoor

Outdoor: GPS, Galileo satellite navigation system, BDS

Indoor: ZigBee, Bluetooth (iBeacon), WLAN



Dataset, Algorithms and Tools

Dataset: RSS Position Fingerprints

-*Training RSS is collected in advance*

-*Testing RSS is collected and applied*

Data Source:

WLAN RSS is obtained by self-made App

- *On-the-spot collecting*
- *Detected by the sensors of device*
- *Storing data in txt format*

Algorithms:

WLAN Positioning Algorithm

-*Fingerprint Indoor Positioning Algorithm*

Clustering Algorithm

-*k-Means Clustering*

Classification Algorithm

-*Naïve Bayes*

MapReduce

Toolset:

-*Java, Hadoop/Mahout, Android API*

-*Linux based Environment*

-*Google Nexus 7 Series Pad*

Current Progress, Schedule and Expected Contributions

Current progress: We did as following:

1. Investigated in current indoor positioning algorithms (e.g. fingerprint positioning algorithm) and compared them;
2. Came up with the project idea using big data to solve the positioning problem;

Schedule:

1. First week(11/19-11/25): collecting and training data, refining our algorithms;
2. Second week(11/26-12/2): analyzing data using proper clustering and classification algorithms, and completing the first version of our project;
3. Third week(12/3-12/9): refining our user interface;
4. Final week(12/10-12/16): refining other part of our project and preparing for the presentation.

Expected contributions:

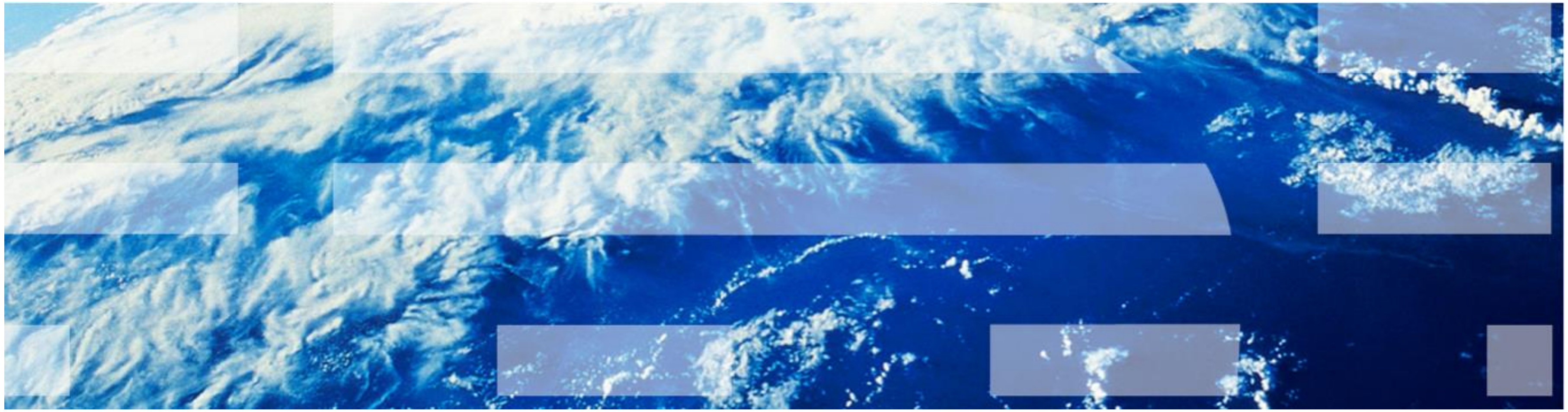
1. Developing an Android application to collect offline fingerprint dataset;
2. Processing the dataset with clustering and classification algorithms;
3. Developing an application to locate with the advanced offline fingerprints dataset.

E6893 Big Data Analytics Project Proposal:
Target your next READING : Book recommender

Shiyu Dong sd2810

Zewei Jiang zj2173

Zixuan Lu zl2348



November 19th, 2015

A system that knows your reading habit even better than yourself

We all like reading. Nowadays, there are just too many books and it might take a while to find a good one. Try a book for couple of days and then realize it is not really good for you is sad. We want to build a system that can recommend right books for the right person like you.

We want to build a book recommendation application. After login, user can see their top rated books based on their personal reading taste. If they don’t like any of these books, they can remove the book from recommendation. Our system will actively learn users’ preferences and keeps updating by time.

The system will support many more features beside the above basic one. It could recommend both on user-based and item-based algorithms. The final goal is a system that knows your reading habit better than yourself!

Dataset: Book-crossing
from *http://www2.informatik.uni-freiburg.de/~cziegler/BX/*
It contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.

Algorithms:
User based recommendation and Item based recommendation

Tools:
➔ Java, Spark, Hadoop and Mahout for the analytics
➔ Javascript, Python, and R for data gathering, web server, and visualization

Current Progress:

Data Collection
Preprocess and Analyse dataset
Come up with and select appropriate algorithms
Build user interface and software architecture

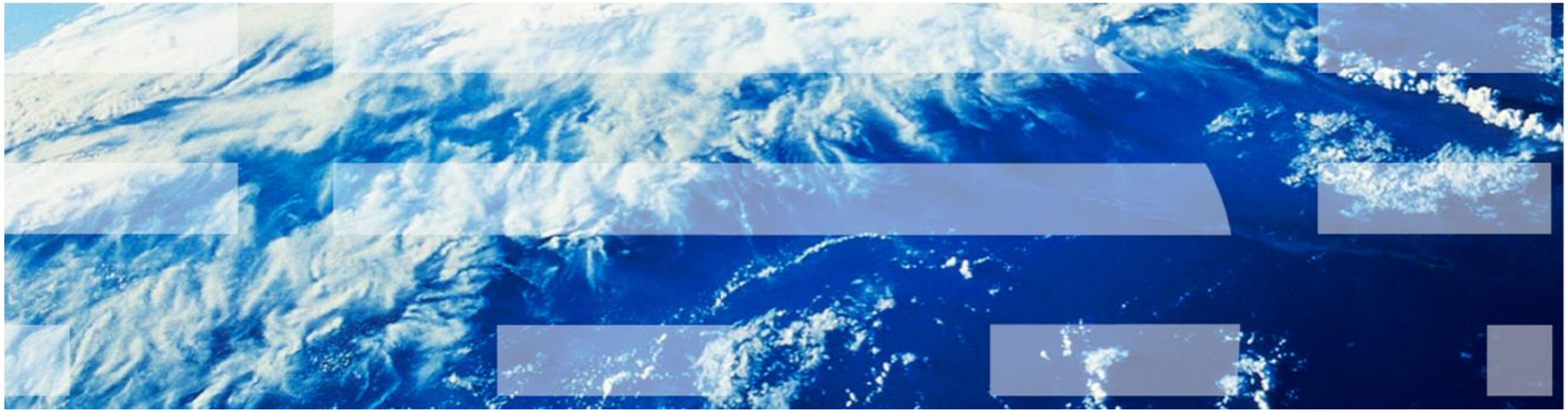
Expected contributions:

Enter a book you like and the site will analyse our huge database of all users' information to recommend books for you as your next suggested read.
Keep updating a reading list for each user.
User can specify reading category, author information, etc and get corresponding recommendations.

E6893 Big Data Analytics Project Proposal:

Data Visualization and Analytics of Columbia's Website Based on IBM System G

Chen Xu, Yue Yu, Zhongzhu Jiang



November 19th, 2015

Motivation

The IBM System G
is a comprehensive set of graph computing tools. Its key feature compared with the traditional analytic systems is that it is designed to deal with the data linked with each other. And the data on the Internet especially the links on a website fit this feature very well.

We think using a new tool to do some analytics on the university’s own website is fun. As there is few data analytics on that, we can show the construction of the university's website and find more about our university in this way.

Dataset, Algorithms and Tools

Dataset: The link information of the Columbia's website

Tools: IBM System G

Algorithms: PageRank, K-core decomposition, Degree Centrality etc.

Current Progress, Schedule and Expected Contributions

Install and configure IBM System G (finished)

Write a web crawler to obtain link information on Columbia's websites (ongoing)

Convert the link information into nodes and edges to construct a graph

Data visualization and analytics by using System G

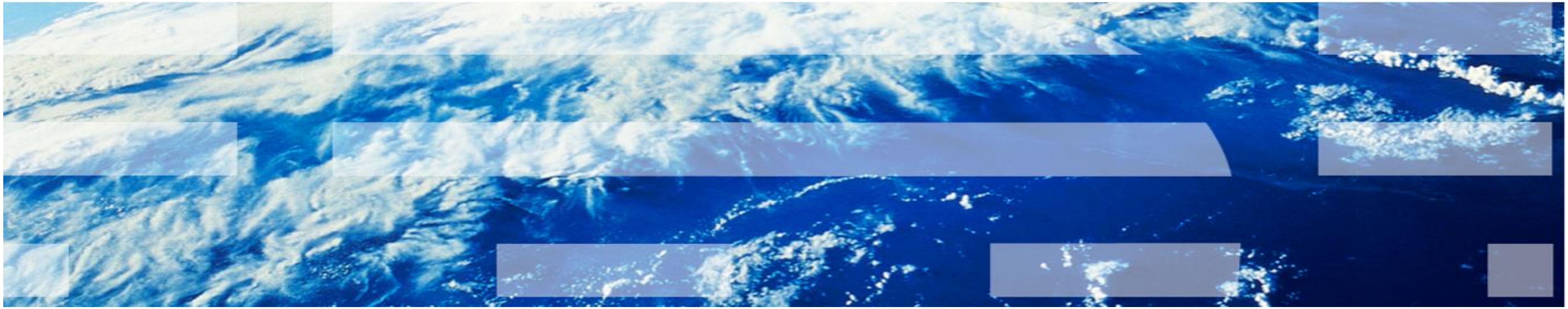
E6893 Big Data Analytics Project Proposal:

Twitter Based Movie Recommendation System

Jingmei Zhao UNI: jz2685

Xing Lan UNI: xl2523

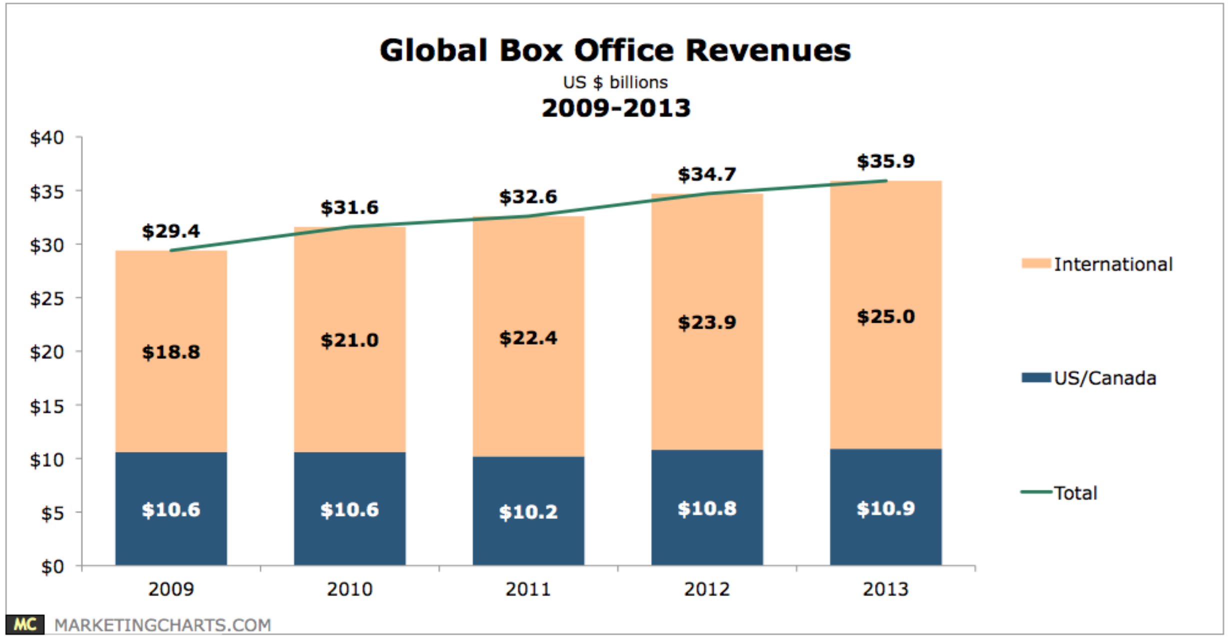
Yao Yang UNI: yy2641



November 19th, 2015

Project Motivation:

- Huge commercial market for specialized real time data based movie recommendation system targeting both business owner and consumer
- Recommend movie lovers with the recent trending movies in their geo location



Source: Motion Picture Association of America

Dataset, Algorithms and Tools

Train Data → Data preprocessing → Feature Extraction → Build Linear SVC Model → Find Sentiment using the built Linear SVC Model

Twiiter Data → Data preprocessing → Feature Extraction → Find Sentiment using the built Linear SVC Model

Find Sentiment using the built Linear SVC Model → Calculate the rating from the average of sentiment score

- Training data:
  Standford AI Lab Large Movie Review Dataset
  http://ai.stanford.edu/~amaas/data/sentiment/

- Data visualization:
  D3.JS
  http://d3js.org

scikit learn

PyEnchant
a spellchecking library for Python

NLTK

Twitter API

IMDbPY

Current Progress, Schedule and Expected Contributions

Current progress:

Project goal finalized
Twitter data structure research initiate

Schedule:

25/11/2015 Collect and clean up raw data
30/11/2015 Build up data training model
06/12/2015 Refine data training model
13/12/2015 Visualize result and prepare for the presentation
17/12/2015 Final presentation

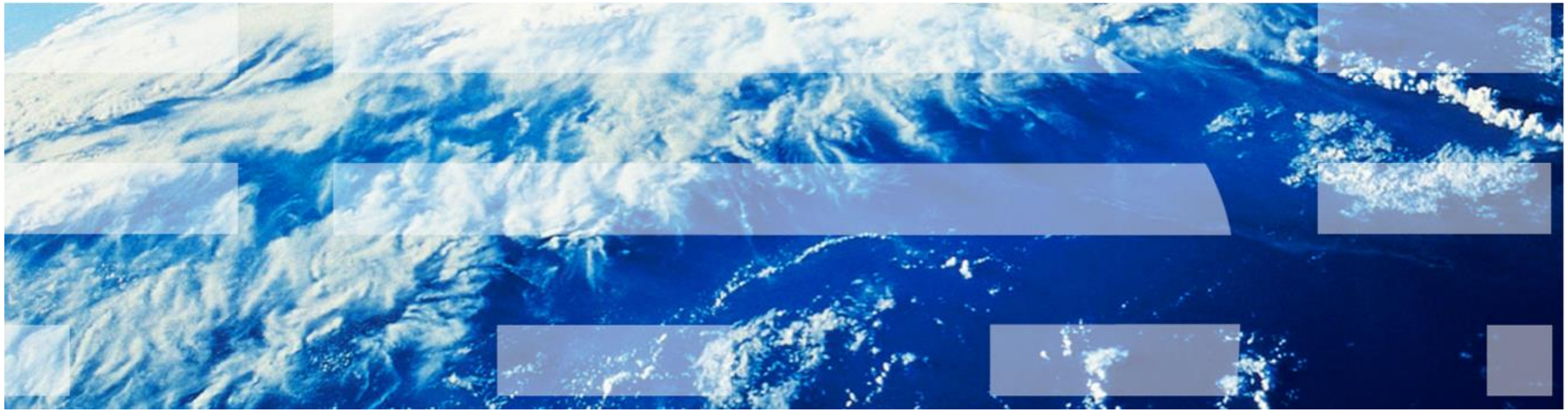
Expected Contributions: *of course, we work together on difficult issues

- Jingmei Zhao: focus on twitter data retrieval consolidation and work on modelling
- Xing Lan: primarily looking at training Model with geo tag
- Yao Yang: concentrate on visualization of regionalized recommendation

E6893 Big Data Analytics Project Proposal:

Yelp dataset analysis

Name:	UNI:
Yaxin Wang	yw2770
Zhibo Wan	zw2327
Jingtao Zhu	jz2664



November 19th, 2015

Motivation

- Provide more accurate interest recommendation for customers.

- Some comments may useless and we try to extract “bad” comments out.

Dataset, Algorithms and Tools

About the dataset:

- **1.6M** reviews and **500K** tips by **366K** users for **61K** businesses.
- **481K** business attributes, e.g., hours, parking availability, ambience.
- Social network of **366K** users for a total of **2.9M** social edges.
- Aggregated check-ins over time for each of the **61K** businesses.

Algorithms:

- Clustering, Classification, Recommendation.

Tools:

- Mahout and Spark.
- Java, HTML.

Current Progress, Schedule and Expected Contributions

Current Progress:

- Download yelp dataset.
- Learn to use tools and algorithms.

Schedule:

- By 11/30: finish clustering part.
- By 12/5: finish Classification part.
- By 12/9: finish Recommendation part.
- By 12/11: finish extra analysis of dataset.
- By 12/16: finish final report and presentation preparation.

Expected Contributions

- Help to identify customers` interests.
- Identify “good” and “bad” commends and figure out the accuracy.

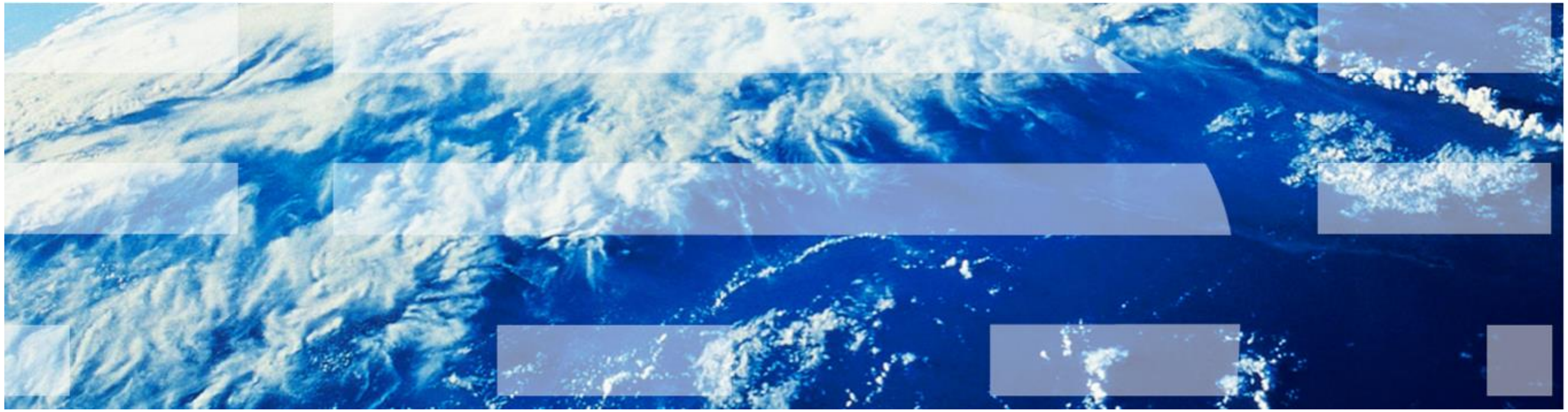
E6893 Big Data Analytics Project Proposal:

Accident Prediction System

Abhijit Roy

Juan Pablo Colomer

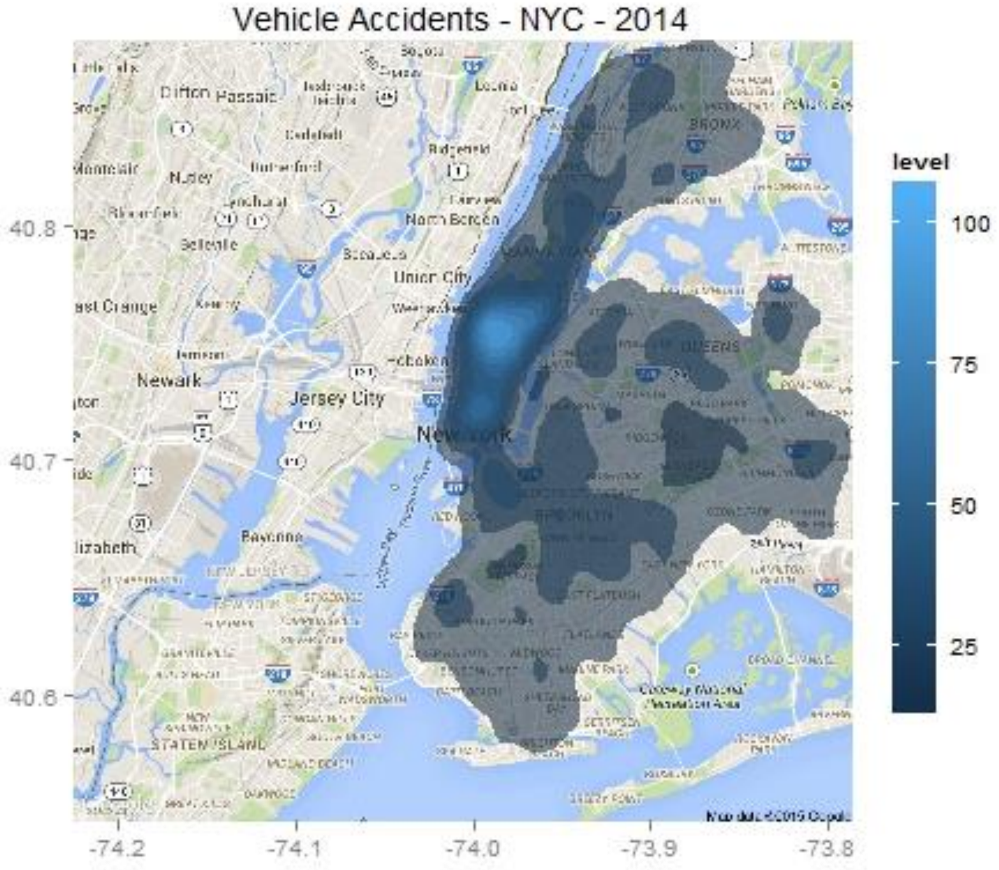
Pedro Perez Sanchez



November 19th, 2015

Motivation

- Weather condition is one of the causes of traffic accidents
- Portions of the city are more prone to accidents for a particular weather condition
- Highlight the areas of the city one should avoid for today's weather conditions



Dataset, Algorithms and Tools

Datasets:

•National Climatic Data Center, NOAA

•NYC OPEN DATA: NYPD Motor Vehicle Collisions

Algorithms:

•Classification: TBD (SVMs, Perceptron, Naïve Bayes)

Tools:

•AWS

•Mahout and/or Mlib

•Hadoop

•Spark

•CartoDB or D3.js

Current Progress:

- Inception Phase - Complete
- Data Gathering - Complete
- Design Phase - In Progress

Schedule:

- Design Phase to be completed by November 24th, 2015.
- Implementation to be completed by December 10th, 2015.
- Testing to be completed by December 14th, 2015.
- Final slides and Project video demo to be completed by December 16th, 2015

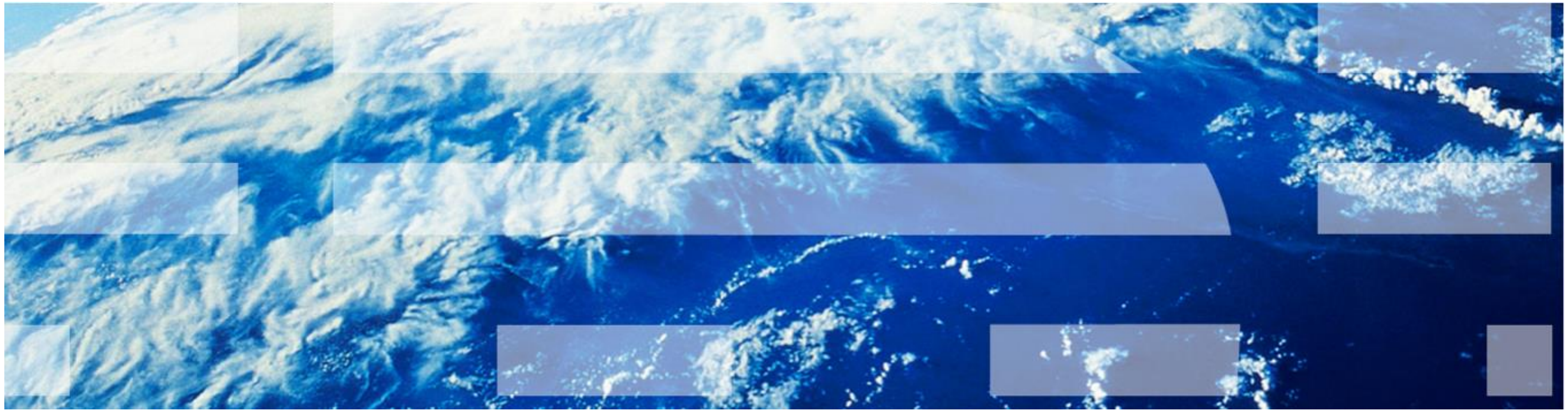
Expected Contributions:

- AWS: Juan Pablo Colomer and Abhijit Roy
- Environment setup: Juan Pablo Colomer, Pedro Perez Sanchez
- Implement ML algorithm: Juan Pablo Colomer, Pedro Perez Sanchez, Abhijit Roy
- UI implementation: Pedro Perez Sanchez, Abhijit Roy

E6893 Big Data Analytics Project Proposal:

Yet Another Evaluation System

Shengtong Zhang(sz2539), Tiezheng Li(tl2693), Ruiqi Duan(rd2704)



November 19th, 2015

We focus on

implementing the integration of the information of a specific product

and

making an overall evaluation.

Dataset, Algorithms and Tools

Potential Dataset:

Algorithm:

Build up the dictionary of (picture -> item name & id)

For every input item:

Find all the similar items to the input item using Near Neighbors Algorithm

Extract the list of matching items

Search the Best Buy products API to find all products matching the item and retrieve product ratings and customer review texts

Search Twitter Developer API for Tweets matching the search product, and calculate sentiment

Search NYTimes Articles API for news articles matching the search product.

Calculate the sentiment value of the review text using NLP tools

Return the result to user

Potential Tools: Hadoop, Mahout, Matlab, R, python

Current Progress

Schedule

Design
- System Architecture
- Core Algorithm

Development
- Data Crawling
- Data Filtration
- Algorithm Implementation

Test
- Function test
- Performance test

Expected Contributions

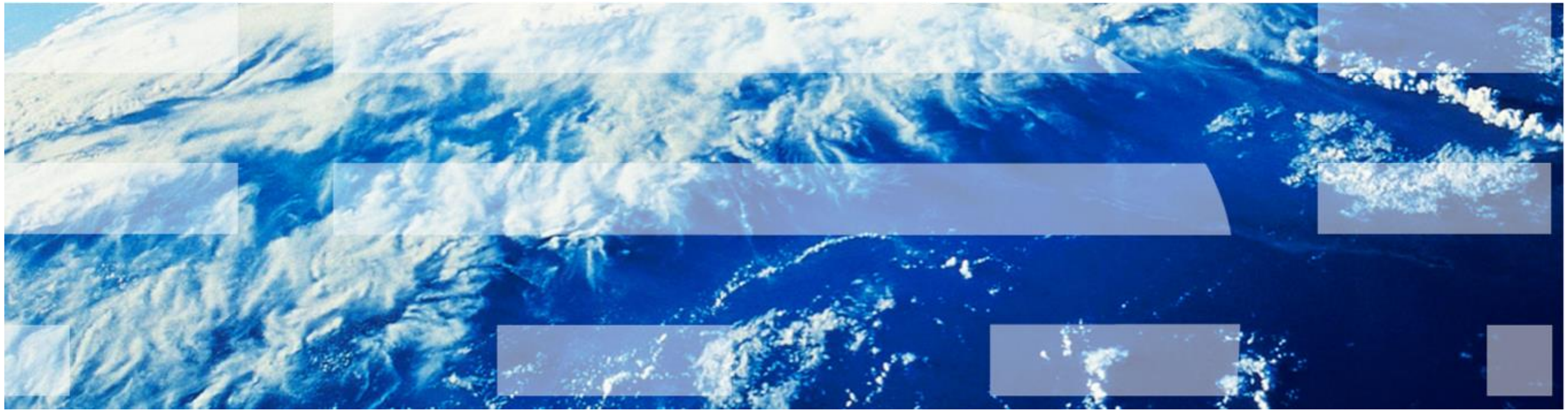
Shengtong Zhang: Product Manager, Core Algorithm Designer
Tiezheng Li: System Architect
Ruiqi Duan: Data Engineer, Test Engineer

E6893 Big Data Analytics Project Proposal:

Data analysis on soccer team performance

RUI WANG

SHUAIYU HAN



November 19th, 2015

Motivation

Inspired by the homework we have done, we designed this project to analyze the data of a soccer team. My partner and I are both fond of soccer game. We spent lots of time on watching soccer event, visiting media web sites (e.g. Sina, Yahoo) to see comments about the performance of each team. It is common to find out that some critics are prejudiced about some teams. Thus, we generated an idea that we can analyze a teams' performance by ourselves. We can share the results with our friends who have same interests. If it turns out that some critic has misjuged some team, we can put our results on social web site to refute them or even help our favorite team.

Dataset, Algorithms and Tools

•Data set: The data can be downloaded from the web site below: http://www.goalzz.com/.

•Algorithm:

1. Classification

Check the relationship between each term and the result of game. (correlation)

Plot the relationship

Combine the chosen terms together to see their effects on the result of game(train model: LDA, Random Forest & Classification Tree)

2. Regresion

Change the traget from game results (win,lose or tie) to the goal difference. Change the model from classification to regresion.

•Tool:R programming, Hadoop(Mapreduce)

Current Progress, Schedule and Expected Contributions

Progress: We have downloaded the data set and made some fundamental tests on the data with Hadoop.

Schedule:

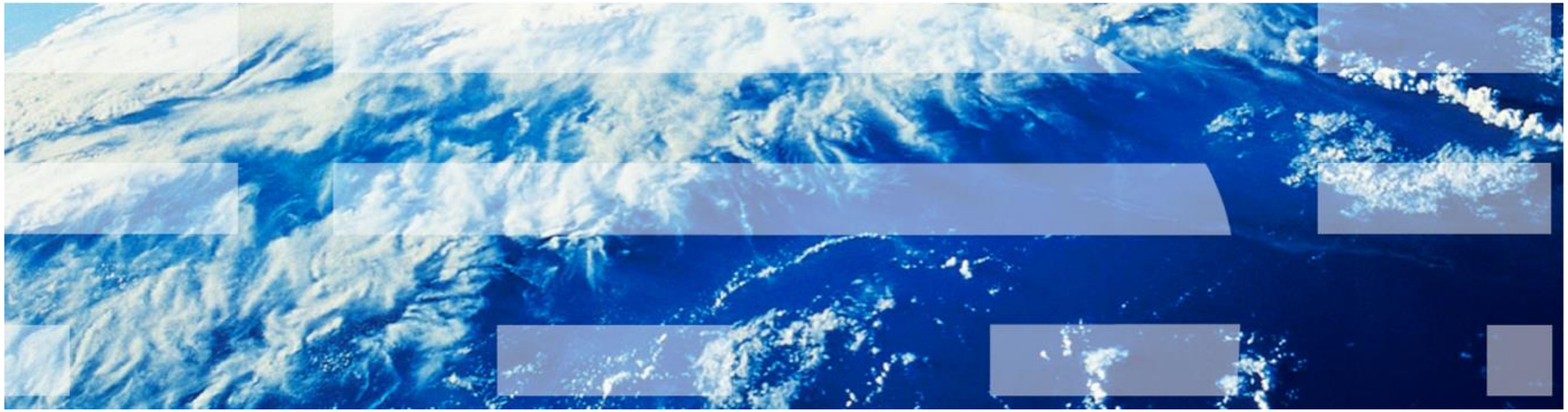
Nov.14 --- Nov. 18	**Download data and analyze**
Nov.19 --- Dec.10	**Classification**
Dec.11 --- Dec.15	**Regresion**
Dec.16 --- Dec. 17	**Double check, Presentaion**

Expected Contributions: We want to compare our results with the professional analysis of the website. The If the we have the same analysis of a soccer team, we can conclude that our project is successful. If not, we can put our results on the forum to discuss with soccer fans.

E6893 Big Data Analytics Project Proposal:

Large-Scale Visual Search

Moning Zhang(mz2499), Hongyi Jin(hj2405), Yang Liu(yl3318)



November 19th, 2015

Motivation

Search by image (Already Exist)

Video is more expressive than image, how can we extend image search to video search?

Dataset, Algorithms and Tools

✦**Dataset:** More than 1TB videos, cover various categories

✦**Algorithms:**

1) Video as “bag of images” (similar to the notion of “bag of words” in document modeling problem)

2) Labeling each image automatically by clustering and image retrieval

3) Using topic model (LDA) to infer the topic distribution for each video

✦**Tools:**

C++

CDVS (image retrieval tool)

Hadoop (dataset is quite large)

Schedule and Expected Contributions

By **Nov. 22:** Running CDVS (Yang Liu)

By **Nov. 28:** Set hadoop environment to run HDFS on AWS (Hongyi Jin)

By **Dec. 10:** Training model by dataset (All members)

By **Dec. 20:** Adjusting parameters (Moning Zhang)

Find People just Like You!

A social application clustering similar users

Haowen Pan
Kun Chen
Xuran Li

A new kind of social app that let you meet people who:

- Have similar education background!
- Have similar social status!
- Have similar professions!
- Have similar hobbies!
- Have similar favorite stars!

How this project works

Datasets:

- The names, occupations and schools downloaded by API via LinkedIn
- The names, usernames, tweets with keywords downloaded by API via Twitter
- Identify and merge the two datasets of common users by the names as keys.

Analysis:

- Cluster the users in various fields
- Use system G to present the outcome of adjacent groups

Processes and contributions

Now:

- 3,000 tweets for each user and most recent Tweets (Haowen & Kun)
- Followings of each accounts (Haowen & Kun)
- Presentation (Xuran)

Are downloaded from API

Schedule:

- Nov. 26: Fetch similar data from LinkedIn (Xuran)
- Nov. 30: Merge two sets of data (Haowen)
- Dec. 03: Complete the clustering and mapping of data (Kun and Xuran)
- Dec. 16: Format the final report and presentation (Haowen, Kun & Xuran)

Thank you!

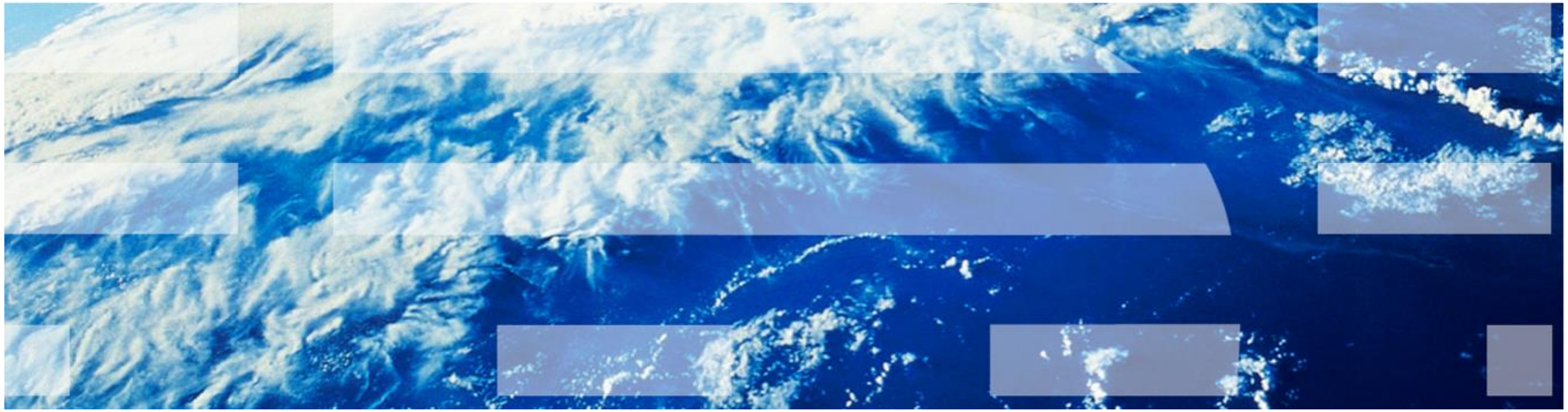
E6893 Big Data Analytics Project Proposal:

Otto Group Product Classification

Qiuyang Shen qs2147

Peng Song ps2839

Yun Sun ys2816



November 19th, 2015

The Otto Group is one of the world's biggest e-commerce companies and sells millions of products worldwide. Everyday there are several thousand products needing to be added to the product line.

- Consistent analysis of the performance of the products is crucial.
- Due to the diverse global infrastructure, many identical products get classified differently.
- The quality of the product analysis depends heavily on the ability to accurately cluster similar products.

Dataset

A dataset with 93 features for more than 200,000 products.

Download Link:

Https://www.kaggle.com/c/otto-group-product-classification-challenge/data

Algorithms

Neural Networks, Random Forest, SVM, Linear Model, XGBoost,

Regularized Greedy Forest…

Tools

Programming Language: Python

Packages: Numpy, Pandas, Scikit-learn, pylearn, scipy…

Current Progress, Schedule and Expected Contributions

Schedule

- Week 1: do a survey on various classification models and algorithms and have a deep understanding of the dataset.
- Week 2: develop models and algorithms to classify products according to their features.
- Week 3: refine models and algorithms and compare them.
- Week 4: visualize the result and prepare for the demo.

Current Progress

Understood the dataset.
Did a survey of various classification models and algorithms.
Work division between team members.

Expected Contributions

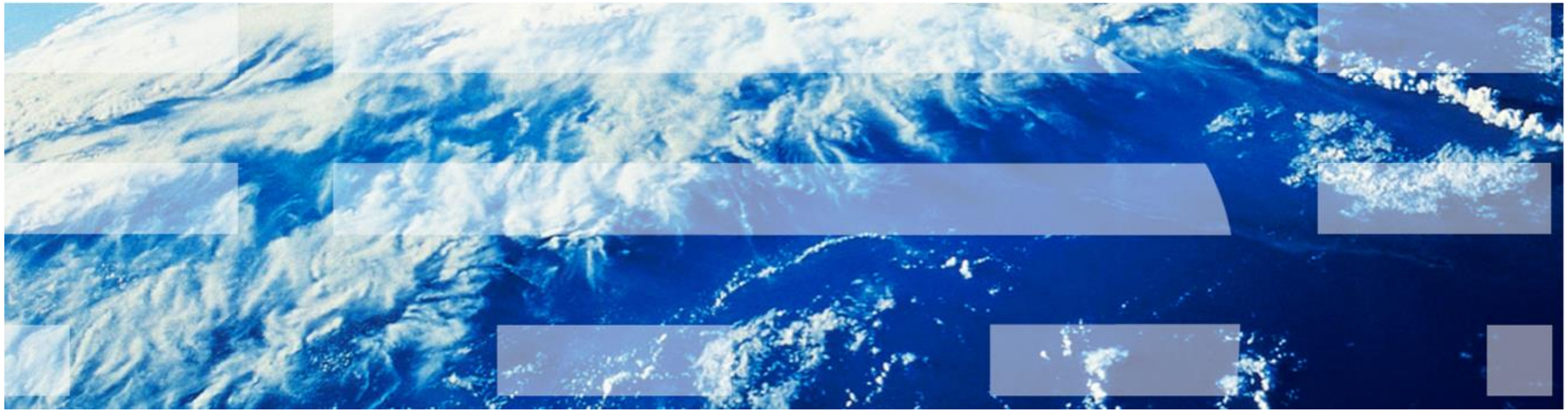
- High-accuracy classification results.
- Comparison and analysis between different algorithms.
- Visualization that shows the info of categories and products.

E6893 Big Data Analytics Project Proposal:

Image Quality Assessment with Different Resolution

Youjia Zhang, UNI:yz2797

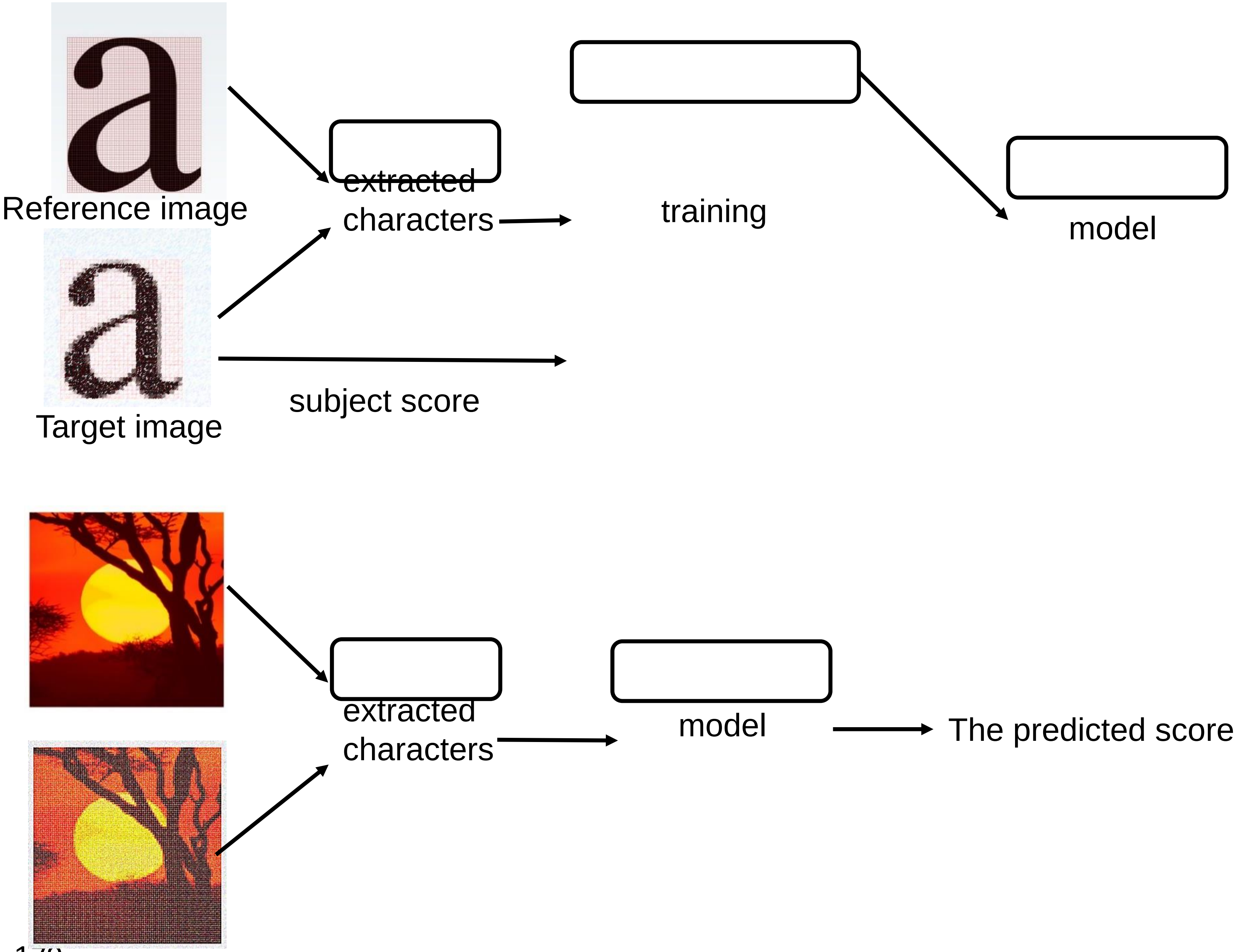
Zhili Zhang, UNI:zz2361



November 19th, 2015

The same image on the screens with different resolutions can derive different subject visual experience.

Images with different degrees of detailed information derive different subject visual experience on the same screen


Reference image
Target image
extracted characters
training
model
subject score
extracted characters
model
The predicted score

Current Progress, Schedule and Expected Contributions

Current progress: defined the general idea of algorithm, collecting appropriate images for subject assessment

Schedule: Grade the image pairs in one week and then figure out the detailed algorithm

Expected contributions: provide information for multimedia providers to improve their service; offer performance evaluation for future compression and coding method in image processing

San Francisco Crime Classification

Final Project – Big Data Analytics

By Sirui Tan, Guihao Liang, Haoyue Bai

BACKGROUND

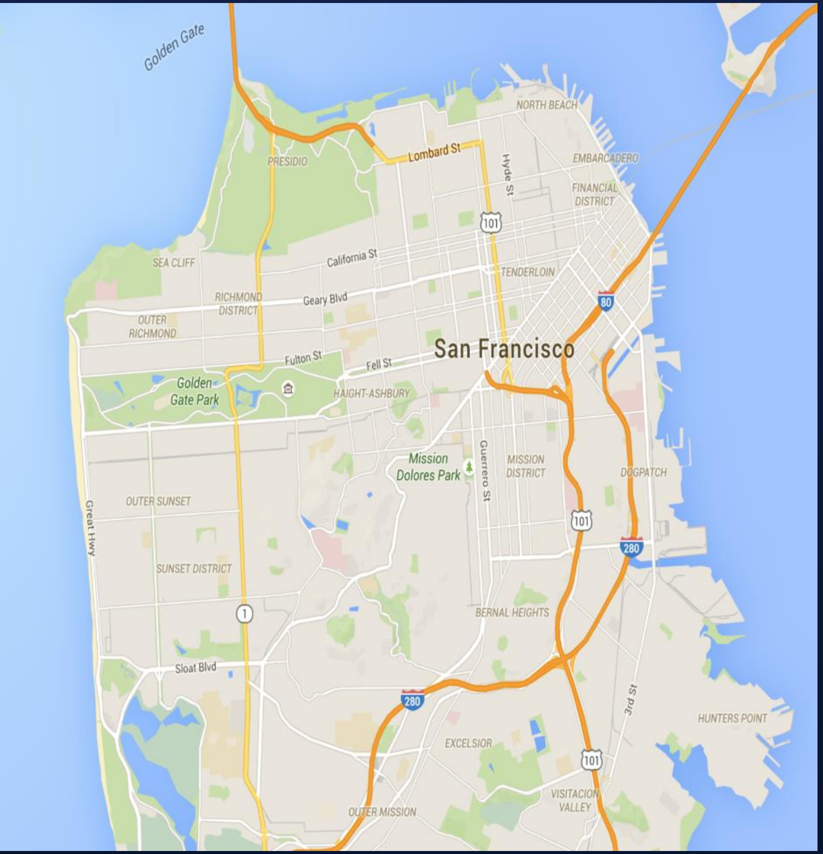


Fig.1 San Francisco Map

- Predict the Category of Crimes

- Visualize Dataset to A Crime Map

DATASETS

2003-01-07 07:52:00	WARRANTS	WARRANT ARREST	Tuesday	SOUTHERN	ARREST, BOOKED	5TH ST / SHIPLEY ST	-122.402843	37.779829
2003-01-07 04:49:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Tuesday	TENDERLOIN	ARREST, BOOKED	CYRIL MAGNIN STORTH ST / EDDY ST	-122.408495	37.784452
2003-01-07 03:52:00	WARRANTS	WARRANT ARREST	Tuesday	NORTHERN	ARREST, BOOKED	OFARRELL ST / LARKIN ST	-122.417904	37.785167
2003-01-07 03:34:00	WARRANTS	WARRANT ARREST	Tuesday	NORTHERN	ARREST, BOOKED	DIVISADERO ST / LOMBARD ST	-122.442650	37.798999
2003-01-07 01:22:00	WARRANTS	WARRANT ARREST	Tuesday	SOUTHERN	ARREST, BOOKED	900 Block of MARKET ST	-122.409537	37.782691
2003-01-06 23:30:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	BAYVIEW	ARREST, BOOKED	REVERE AV / INGALLS ST	-122.384557	37.728487
2003-01-06 23:14:00	WARRANTS	WARRANT ARREST	Monday	CENTRAL	ARREST, BOOKED	BUSH ST / HYDE ST	-122.417019	37.789110
2003-01-06 22:45:00	WARRANTS	WARRANT ARREST	Monday	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421
2003-01-06 22:45:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421
2003-01-06 22:19:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	NORTHERN	ARREST, BOOKED	GEARY ST / POLK ST	-122.419740	37.785893
2003-01-06 21:54:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	NORTHERN	ARREST, BOOKED	SUTTER ST / POLK ST	-122.420120	37.787757

Fig.2 An Example of Dataset

- Incidents derived from SFPD Crime Incident Reporting systems
- Ranges from 1/1/2003 to 5/13/2015.
- Data fields
  - Dates
  - Category
  - Descript
  - Day Of Week
  - Pd District
  - Resolution
  - Address
  - X - Longitude
  - Y - Latitude

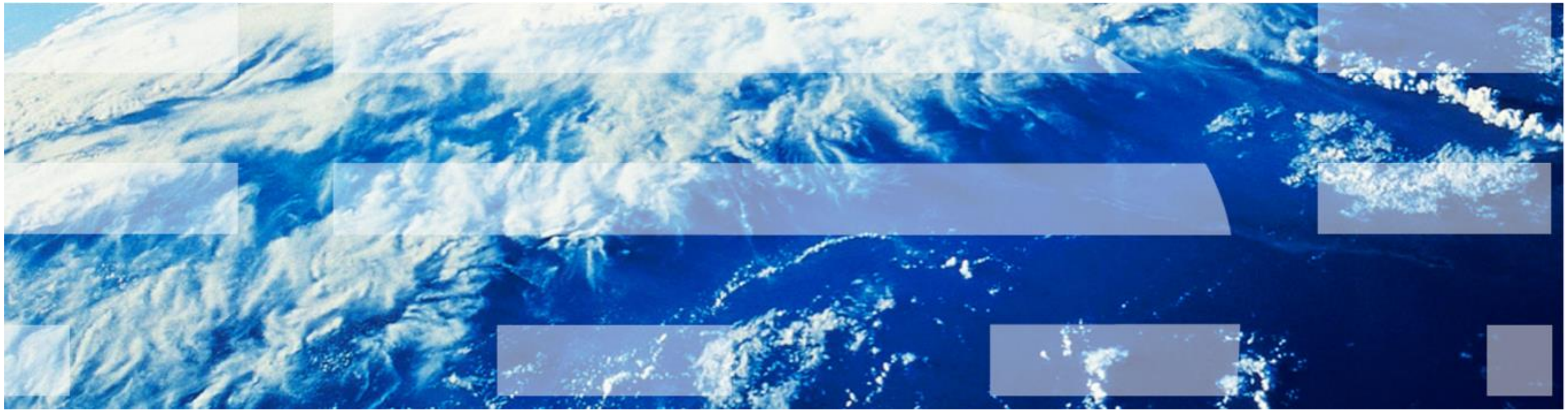
METHODS & ALOGRITHMS

- Machine Learning Algorithms, e.g. SVM, HMM, ANN
- Spark
- Python-based Machine Learning and Statistics Libraries: pandas, scipy, scikit-learn, matplotlib
- Visualization - making a crime distribution graph based on the map of San Francisco, and a generic table based on crime category.

E6893 Big Data Analytics Project Proposal:

Yelp Recommendation

Yufei Ou(yo2265), Ke Li(kl2831), Ye Cao(c3113)



November 19th, 2015

Motivation

Recommendation is more and more important in modern society.

Review analysis has become a critical reference in recommendation and business strategies nowadays. Exploration into the feedbacks of the users can grant us incredible insights.

Given such untapped treasure of resources, we aim at harnessing the fusion of the review analysis and recommendation, and try to extract valuable advice for business management

Dataset, Algorithms and Tools

- Dataset:
- Yelp Dataset Challenge, including users, businesses, and reviews

- Algorithms:
- KNN, LDA, BP

- Tools:
- Hadoop, Spark, Mahout

Current Progress, Schedule and Expected Contributions

Current Progress:
get dataset, design algorithm

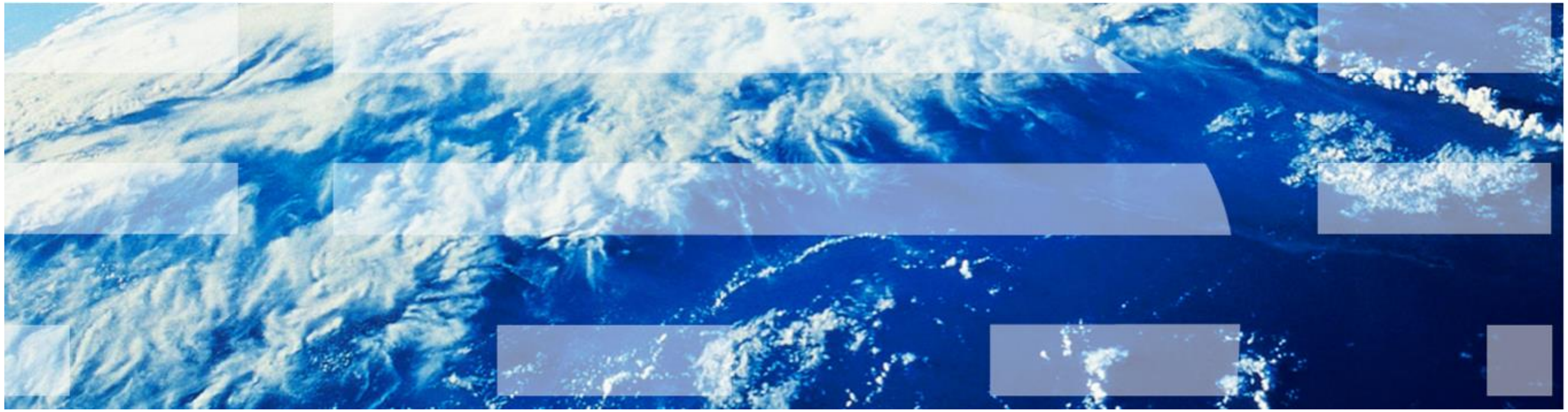
Schedule:
Nov.20 - Nov.27 : dataset extraction
Nov. 28 - Dec.5 : algorithm implementation
Dec.6 - Dec.16 : result analysis, prepare for final report

Expected Contributions:
Get recommendation result with high accuracy

E6893 Big Data Analytics Project Proposal:

Earnings Predictor: A system that predicts whether a company will beat consensus earnings estimate

Roberto Martin, Kedar Patil



November 19th, 2015

Motivation

Many analysts are paid to give estimates of earnings for different companies. The consensus earnings estimate is an average of these estimates. This consensus is correct approximately 60% of the time. There are a number of reasons why analysts incorrectly predict earnings for companies:

- Conflicting Interests
- Various Biases
- Manipulation

We think that building a model that is based on past prices and other company data will correct for these shortcomings. The model will not attempt to predict a company's earnings directly, instead, it will predict whether earnings will beat analysts estimations.

Dataset, Algorithms and Tools

Dataset

- 10 years of daily stock prices from all the stocks in the technology sector (1172 at last count). The data will consists of the following columns: **Open**, **High**, **Low**, **Close**, **Adjusted Close**, **Volume**.
- Possibly also use sentiment data from twitter as an additional feature.

Algorithms

- We will create a model using decision trees and svm and score each to see which performs better. We will use Spark's MLLib to generate the model.

Tools

- Python
- Spark (PySpark)
- MongoD
- Openscoring

Current Progress

We are in the data gathering phase at this point. OHLC stock data is being downloaded from Yahoo with a script that runs nightly.

Consensus estimates is very hard to come by. We use the Zacks dataset on **Quantdl** for this. **Zacks** (https:www.quandl.com/data/ZEEH) and **Quantdl** was kind enough to give us access to this for free (It costs $1800/year for the cheapest license)

Schedule

1. Finish gathering data – 11/20
2. Aggregate/summarize data – 11/27
3. Setup Big Data Pipeline – 11/24
4. Build model and iterate – 12/11

Expected Contribution

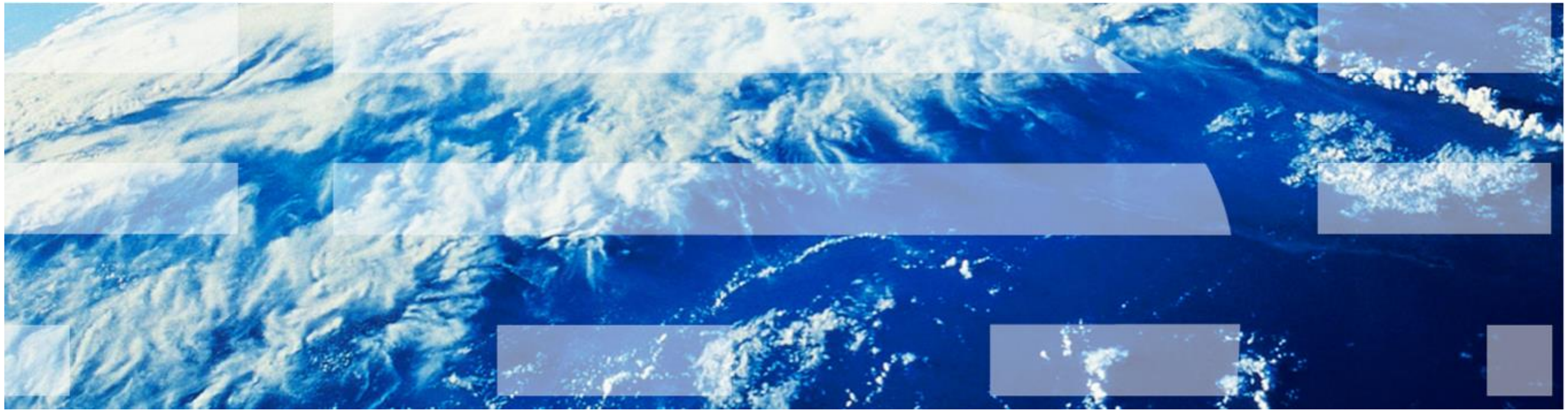
Roberto: Gather Data, Aggregate/Summarize, Setup Pipeline, Build Model

Kedar: Aggregate/Summarize, Pipeline Setup, Build Model

E6893 Big Data Analytics Project Proposal:

Predicting Optimal Daily Fantasy Basketball Rosters

Michael Raimi (mar2260)
Justin Pugliese (jp3571)



November 19th, 2015

Motivation

Daily fantasy, the newest incarnation of fantasy sports, is the fastest growing segment and is currently an over $3 billion industry.

Daily fantasy has been featured heavily in the news over the last few weeks. The Attorney General is currently close to banning it outright in the state of New York following a similar ban in Nevada. The rationale is that daily fantasy is purely luck and is thus illegal gambling. Hopefully we can make an informed evaluation of that statement after performing some research in the daily fantasy sector.

Luckily for us the format of daily fantasy lends itself rather well to certain kinds of optimization. Considering that it's daily, it also has tons of data waiting to be integrated into predictive models.

Bloomberg

Dataset: We intend to scrape http://rotoguru.net/ which has a few years worth of daily fantasy records. At 82 games a year, across 30 teams and hundreds of players we we will have enough data for meaningful predictions.

Algorithms: We want to offer several forms of predictions through clustering (k-means and Gaussian mixtures), optimization (stochastic gradient descent), and recommendation (Euclidean, cosine, Pearson, etc.)

Tools: We have settled on Spark for our Machine Learning algorithms and python for scraping the web. We would like to use HDFS for storage.

Current progress

1. Git Repository setup
2. Data collection
   a. Evaluate
   b. Parse
   c. Store
3. Algorithm Architecture

Schedule

1. Parse and sanitize to HDFS (1 week)
2. Build prediction pipeline (1 week)
3. Build clustering pipeline (1 week)
4. Build recommendation pipeline (1 week)

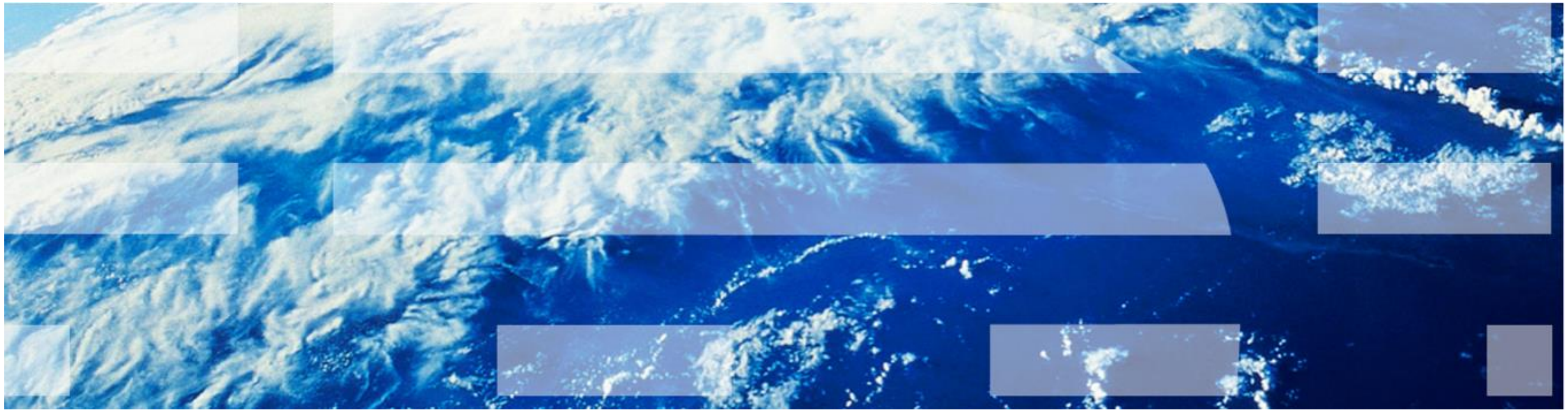
Expected Contributions

We plan to collaborate on all aspects of the project including: development, evaluating results, and creating deliverables.

E6893 Big Data Analytics Project Proposal:

Passenger-and-Driver-Based Analytics of NYC Taxi Database

Yunzhe Li UNI: yl3390
Changtai Liu UNI: cl3391
Xiaonan Duan UNI:xd2169



November 19th, 2015

NYC is a highly trafficked city where people valued time and efficiency more.

➢**For passengers**
• Request for special service
• Difficulty in taking taxi
• Confusion about estimated cost & tips

➢ **For driver:**
• Low passenger load factor
• Too busy to remember license expiration

Good luck, lady. Photo: Jacobs Stock Photography/Getty Images

With more suggestions and analysis, which hopefully will be provided by our project, both passengers and drivers can make their trip more efficient and convenient.

Dataset, Algorithms and Tools

Dataset

- **Green Taxi Trip and Yellow Taxi Trip Data**
- **Assistance_Trained_Data**
  Pick Up and Drop Off Data(date, longitude and latitude)
  Trip Distance, Total Amount and Tip amount
  Drivers License Information and Special Assistance

Algorithms

- **Recommendation:** item-based, user-based similarity measurement
- **Filter:** collaborative filtering
- **Clustering:** k-means

Tools

- **Hadoop, Mahout, Eclipse, Pig, Matlab……**
- **Languages: Java, Pig Latin, Matlab……**

Current Progress, Schedule and Expected Contributions

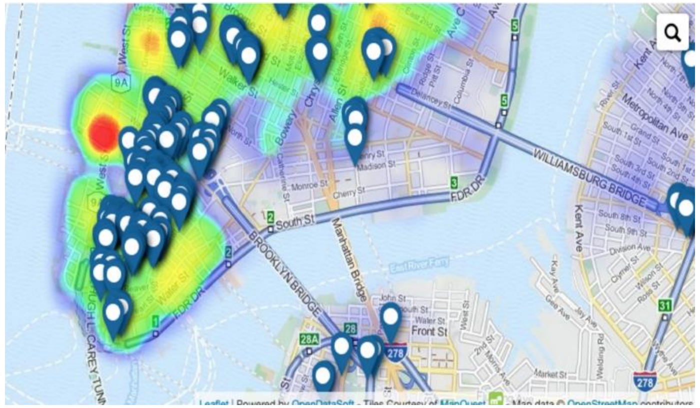
Progress:

- Acquired NYC Yellow and Green Taxi database and begun initial testing
- Setup an environment for Java Recommendation and Hadoop distributed system
- Selected algorithms and Tools to complement recommendation, clustering, and filter.

Expected Contributions:

- **Driver-Based**
  Driver license with Disabled Service expiration date reminder
  Time based popular pick up location recommendation

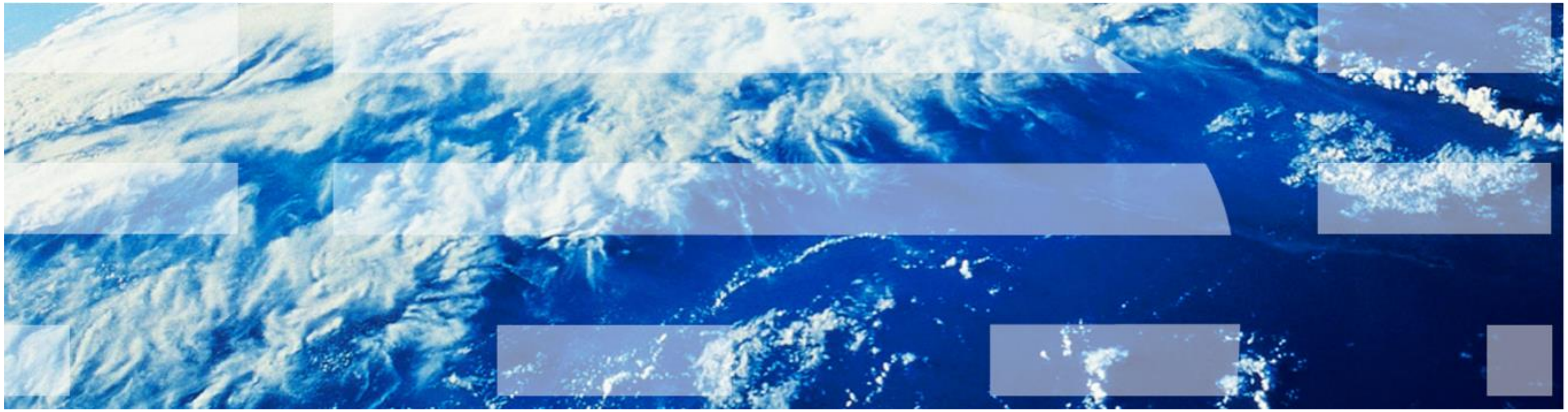
- **Passenger-Based**
  Trip fare and time estimation
  Tip amount recommendation
  Popular boarding location recommendation
  Disabled Service request



E6893 Big Data Analytics Project Proposal:

Analyzing the Yelp Review Dataset with Topic Modeling

Jon Adelson, Kyle DeRosa & Karthik Jayaraman



November 19th, 2015

Motivation

- Use cutting-edge topic modeling techniques to analyze the Yelp Reviews dataset.

- Find a list of differences in topics between high-rating and low-rating reviews for the same business or class of businesses

- Examine the feasibility of reproducing the Yelp category hierarchy purely by analyzing the review text.

- Use LDA as a baseline and then, as time permits, see what improvements can be achieved by using more cutting edge techniques such as Hierarchical Dirichlet Process Topic Model or Collaborative Topic Models.

Dataset, Algorithms and Tools

COLUMBIA UNIVERSITY

- ✦Dataset – Yelp Academic Challenge Dataset from http://www.yelp.com/dataset_challenge
- ✦Tools
- ✦Hadoop + HDFS – distributed document store
- ✦NLTK – for lemmatization, named entity recognition and other text preprocessing prior to LDA
- ✦Mahout – For baseline topic modeling using LDA
- ✦HDP(Hierarchical Dirichlet Process), CTM(Collaborative Topic Modeling), DEF(Deep Exponential Families) – Open source libraries for topic modeling from various academic research groups.
- ✦Amazon Web Services Elastic MapReduce + S3 – S3 for document storage and EMR to speed up processing by using a cluster

Current Progress, Schedule and Expected Contributions

- Current Status

  Selected dataset for analysis, performed preprocessing to convert it into input format for input to topic modeling tools

- Schedule

  - Before Nov 30th – finish initial preprocessing, run numerous iterations of LDA to find optimal parameters for our dataset

  - Nov 30th – Dec 10th – experiment with more cutting-edge topic modeling techniques such as Deep Exponential Families or Collaborative Topic Modeling

  - Dec 10th – Dec 17th – continue experimentation, create simple web application to act as a front-end to display results, analyze and summarize results for presentation

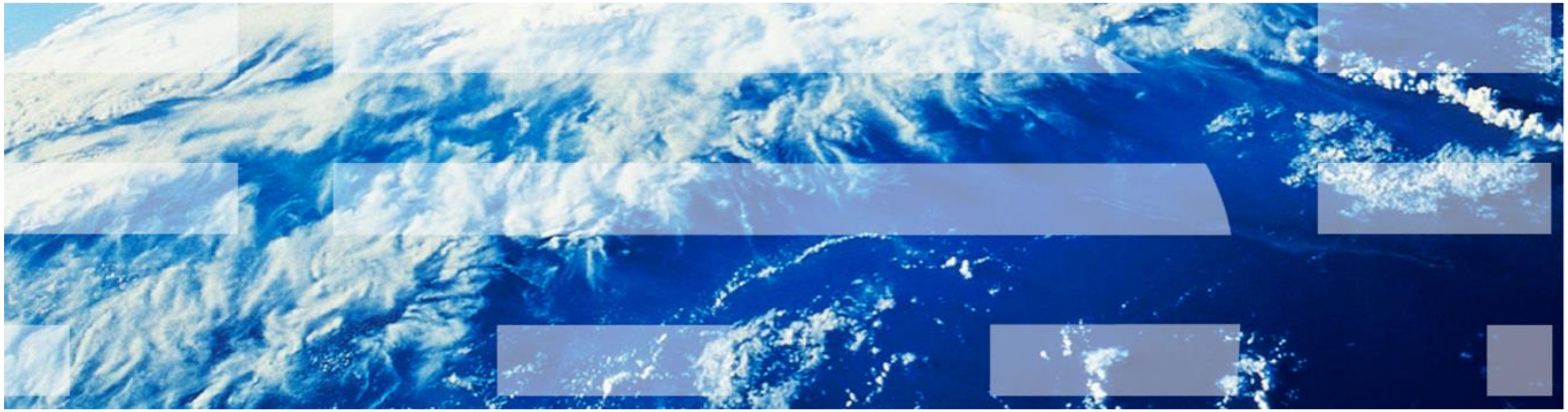
- Expected Contributions

  - We expect to split the work pretty evenly across all three participants with Jon Adelson playing a slightly greater role in experimenting with newer topic modeling frameworks and Karthik and Kyle playing a slightly greater role in running iterations of LDA to find the optimal parameters, setting up our tools to work on AWS if needed and putting together presentations.

E6893 Big Data Analytics Project Proposal:

New York City Taxi Trips

Kevin Graney



November 19th, 2015

Motivation

The NYC Taxi & Limousine commission provides a dataset containing detailed information about every taxi trip taken in the city. We plan to use this information to gain insights into how New Yorkers use taxis.

Dataset, Algorithms and Tools

Dataset
NYC Taxi & Limousine Commission's Trip Record Data (2009 through 2015)
http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

A database of NYC attraction locations (e.g. theaters, airports, hotels, etc.)
A database of NYC neighborhood boundaries

Algorithms
- Clustering algorithms (e.g. K-means) applied to
  - Geographic locations (i.e. trip start and end points as well as attraction locations)
  - Trip data (i.e. pairs of start and end points, and possibly duration, fare, etc.)
- Possibly some statistical testing around fares for different trips

Tools
- HDFS for storing the dataset CSV files
- Spark for fast iterative analysis of the dataset and use of its built-in algorithms

Current progress

2009-2015 TLC dataset is fully downloaded to HDFS

A 16-node Hadoop/Spark cluster with plenty of RAM (200GB/node) is configured and ready for use

Schedule

This project will be broken down into several distinct phases

<u>Phase 1</u>: Clustering start and end points

We will start by clustering start and end locations of trips. This will be done geographically using Euclidean distance.

<u>Phase 2</u>: Giving clusters an identity

Each cluster will be given an identity based on its geographic location. This identity might be an attraction (e.g. Lincoln Center) or a more general term that applies to the area (e.g. Residential if the point is on a primarily residential block).

<u>Phase 3</u>: Clustering trips

Within each identity we will cluster individual trips together. This should help us identify patterns (e.g. Residental UES to Financial District, or Midtown to JFK) that occur in the trips. We may cluster multiple identities together for the purpose of this analysis.

Expected Contributions

Kevin will contribute the entire project

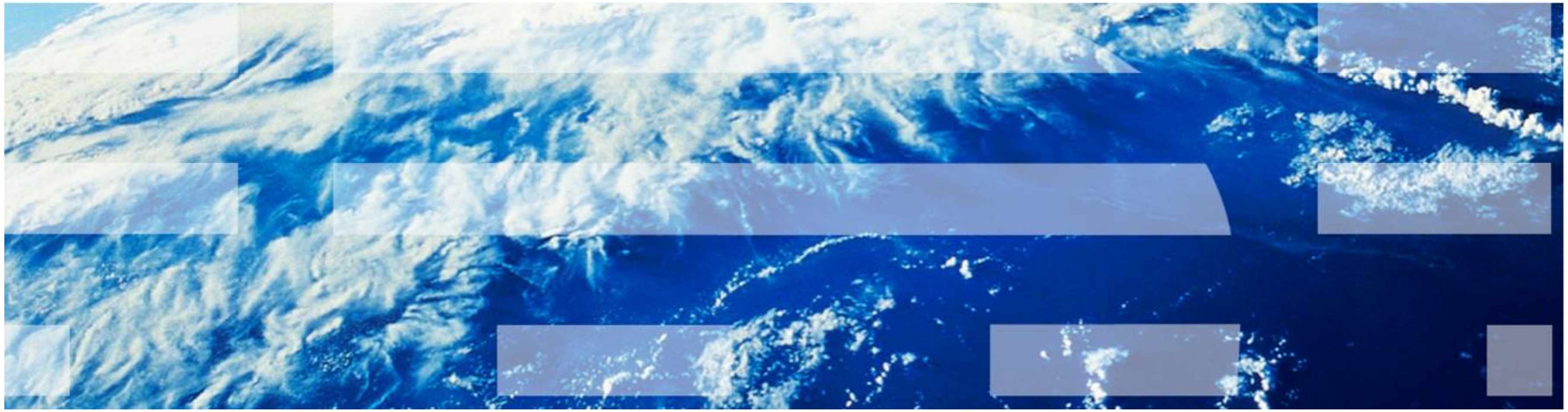
E6893 Big Data Analytics Project Proposal:

Visualization of Machine Learning Algorithms in MapReduce

Yubin Shen

Ziyu He

Jie Yuan



November 19th, 2015

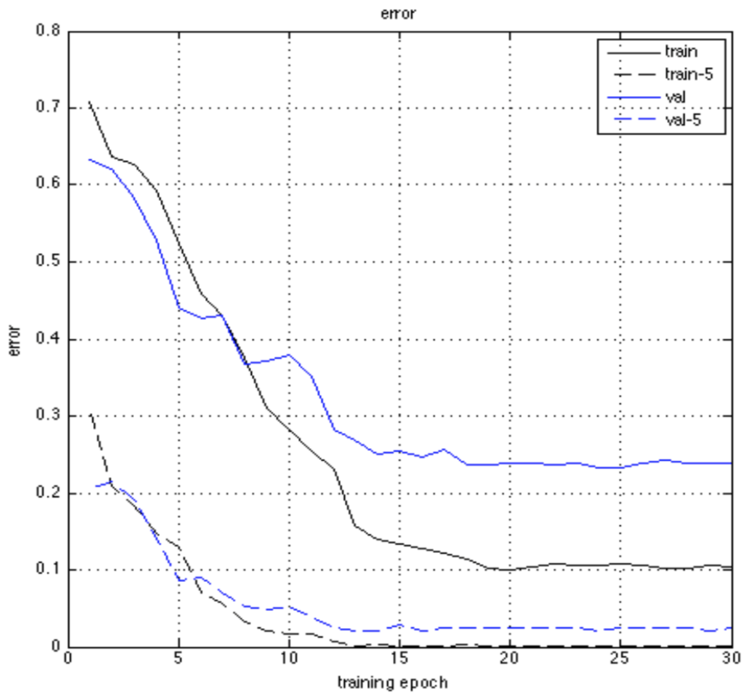
Currently, we treat machine learning packages such as Mahout as black boxes – we would like to make an ML package that is more transparent to the user

•Implement ML algorithms in MapReduce on handwritten data

Random Forest
Neural Network

- Visualize summary of inputs and outputs into mappers and reducers as a graph (in real time if possible).

- Visualize performance metrics related to the particular algorithm (in real time if possible)



https://courses.cs.ut.ee/MTAT.03.291/2015_spring/uploads/Main/Presentation%20-%20Introduction%20to%20Computational%20Neuroscience%20Project%20Classification%20of%20LFW%20Dataset%20Using%20MatConvNet.pdf

Dataset

✦MNIST: vectors of pixel intensity for handwritten numbers

Algorithms

✦Machine Learning algorithms: Random Forest, Neural Network

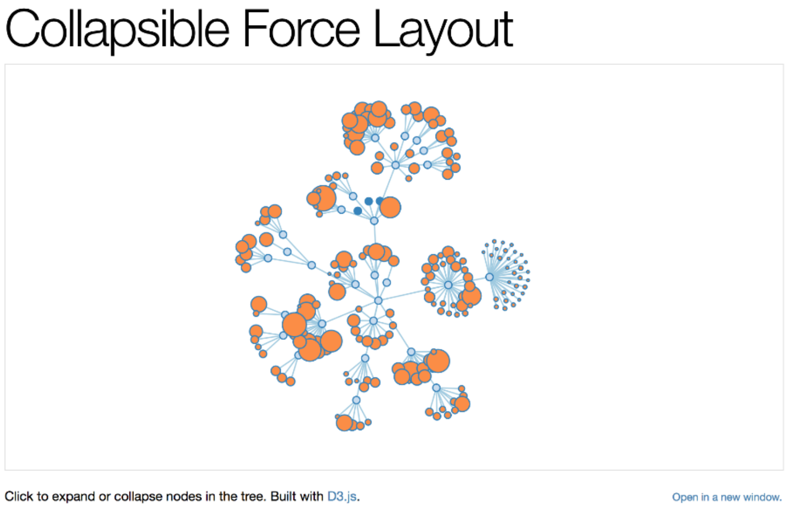
✦Neural Network: test the performance of dropout with different drop probabilities

Tools:

✦MapReduce: Hadoop, Java

✦Parsing Log output of Hadoop: Python

✦Visualization: D3.js, javascript



http://bl.ocks.org/mbostock/1062288

Current Progress, Schedule and Expected Contributions

Current Progress

- Exploring usage of D3.js
- Investigating creation of MapReduce jobs

Schedule

- Parsing of text dataset into appropriate input format
- Implement Random Forest/Neural Network in MapReduce
- Design summary visualizations for algorithm output (e.g. graph showing Mappers and Reducers)
- Try to get the visualizations to update in real time, as the MapReduce job runs

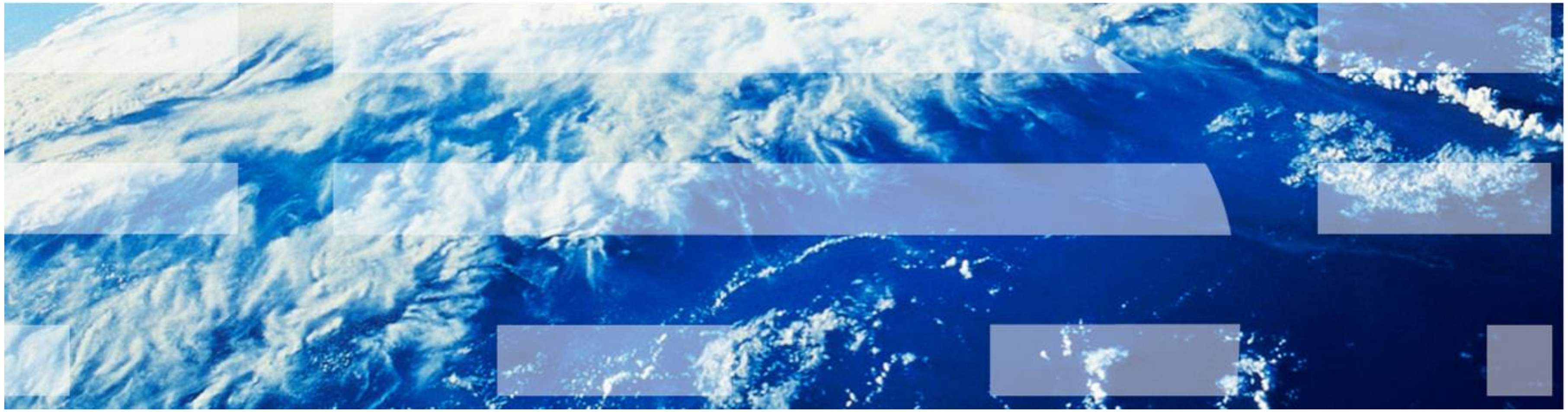
Expected Contributions

- MapReduce implementation: Ziyu, Yubin, Jie
- Transferring data to log files and to front-end: Ziyu, Yubin
- Front end (D3.js): Jie Yuan

E6893 Big Data Analytics Project Proposal:

Visualization and Analysis based on NYC Taxi Trip Data

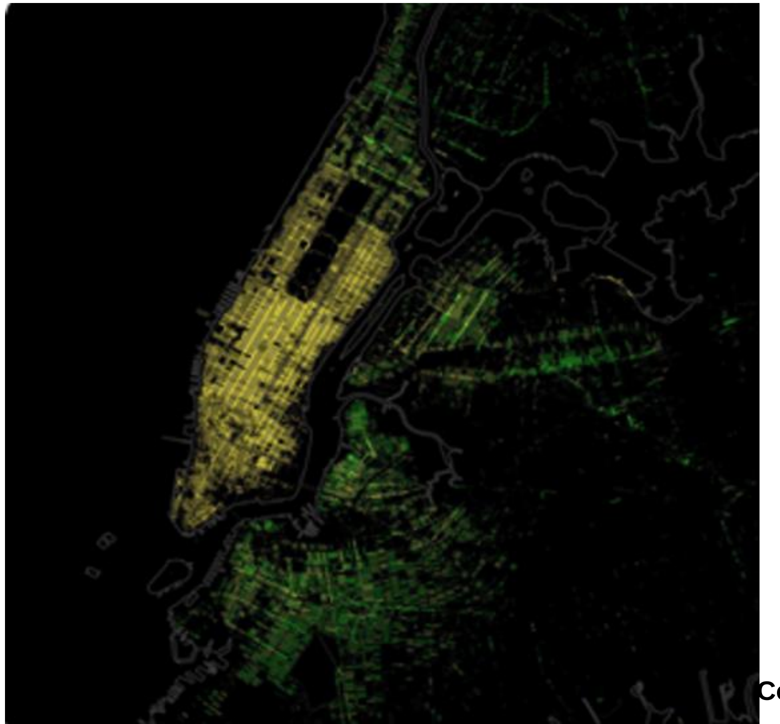
Xianglu Kong, Guochen Jing, Junfei Shen



November 19th, 2015

Motivation

- Identify popular taxi pick-up & drop-off locations at different time of day
- Help people get taxis more efficiently
- Suggest locations good for taxi drivers to pick up potential passengers
- Present NYC taxi trips information in an intuitive way - visualization



Dataset, Algorithms and Tools

- Dataset: NYC Taxi Trip Records
- http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
- Date time, longitude and latitude for pick up and drop off
- Passenger count, trip distance, fare amount etc.

- Methods
- MapReduce: Count taxi trips in each time period
- Cluster: Find out geographical clusters i.e. popular locations
- Visualize

- Tools
- Hadoop, Spark, JavaScript etc.

Current Progress, Schedule and Expected Contributions

- Schedule:
- Step 1: Count taxi trips ←
- Step 2: Draw maps
- Step 3: Find clusters

- Expected Contributions:
- A clear mind of how taxis are distributed and moving in NYC
- Make finding a taxi in NYC easier

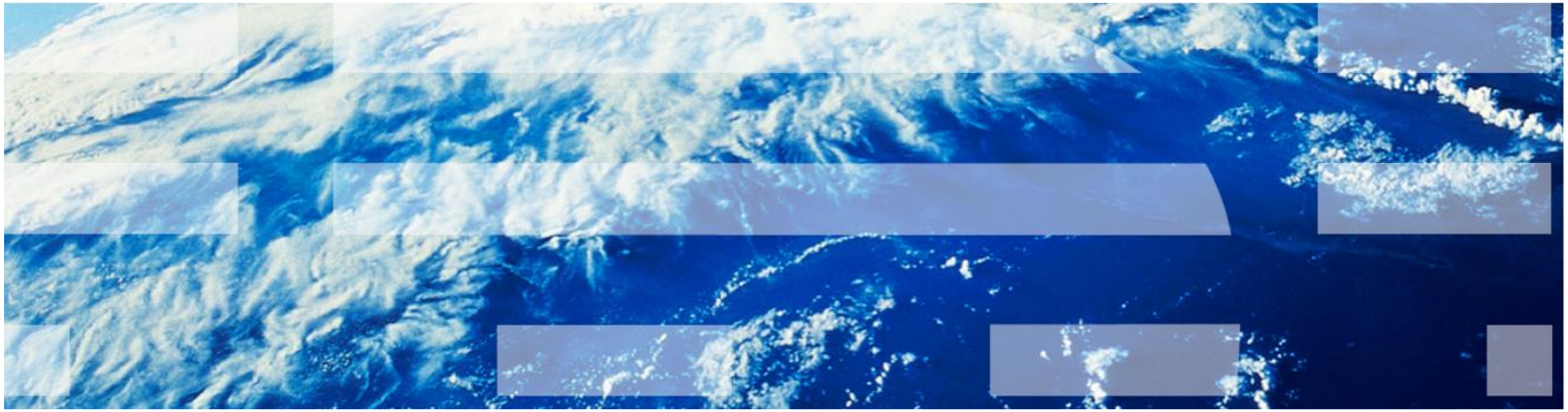
E6893 Big Data Analytics Project Proposal:

Pedestrian Tracking for ATC Shopping Mall

Yan Lu (yl3406)

Mengzhuo Lu (ml3806)

Dingyu Yao (dy2307)



November 19th, 2015

Motivation

Asian Pacific Trade Center (ATC), located in Osaka Japan, is the largest international mall complex in Kansai.

We analyze its pedestrian flow information to help ATC distinguish target client and come up with store deploy strategy.

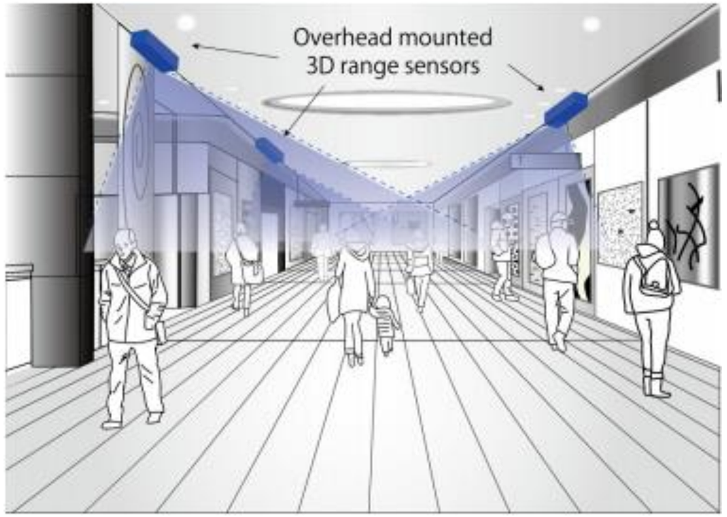
1. People behavior in ATC shopping mall

2. Difference between single, couple, group shopper.
2. Children ratio among all shoppers
3. Popular area in the shopping mall

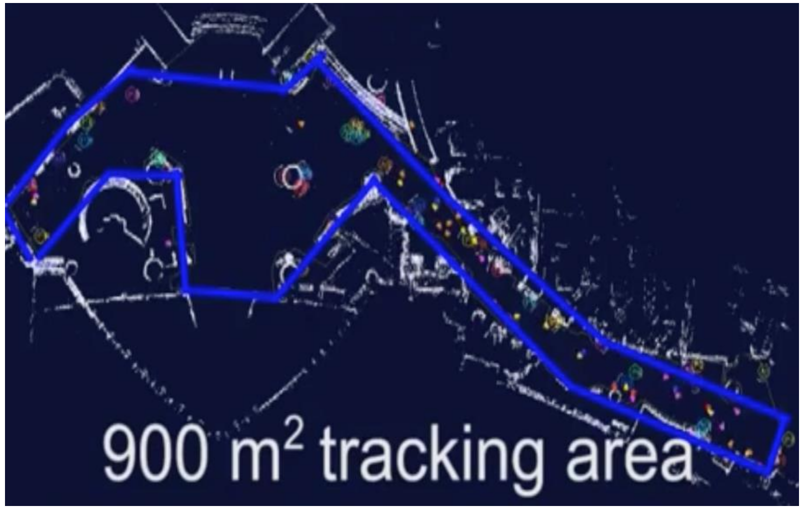
[1] www.timberwyck.org [2] www.visituzbekistan.travel [3] fotomen.cn [4] www.timberwyck.org [5] simpleclassroompsychology.edublogs.org

Dataset, Algorithms and Tools

✦A censoring system was set up in ATC shopping mall, which contains multiple 3D range sensors



D. Brscic, T. Kanda, T. Ikeda, T. Miyashita, "Person position and body direction tracking in large public spaces using 3D range sensors", IEEE Transactions on Human-Machine Systems, Vol. 43, No. 6, pp. 522-534, 2013



http://www.irc.atr.jp/crest2010_HRI/ATC_dataset/

✦The dataset was collected between October 24, 2012 and November 29, 2013, Wednesday and Sunday, 9:40-20:20. It contains 92 days in total.

✦It tracks people with their height, coordinate, velocity and group interaction, etc.

✦Pig, Mahout, spark, and so on…

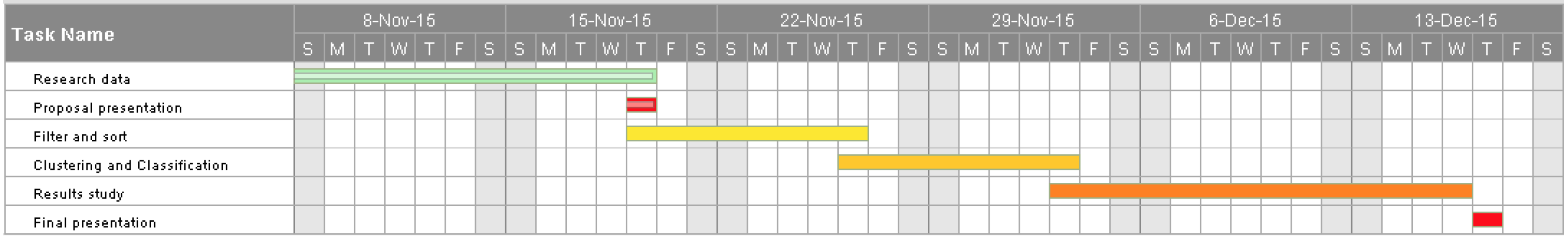
✦Filter and sort pedestrians by their behavior (Pig),clustering them by coordinates (Mahout), analysis their behavior and give classification for new coming people and groups (spark).

Current Progress, Schedule and Expected Contributions

Current progress:

- Gathered ATC pedestrian tracking dataset
- Researched data and environment background
- Developed potential big data analytical methodologies

Schedule:



Expected delivery:

- Find patterns for pedestrian behavior
- Make suggestions to new building construction

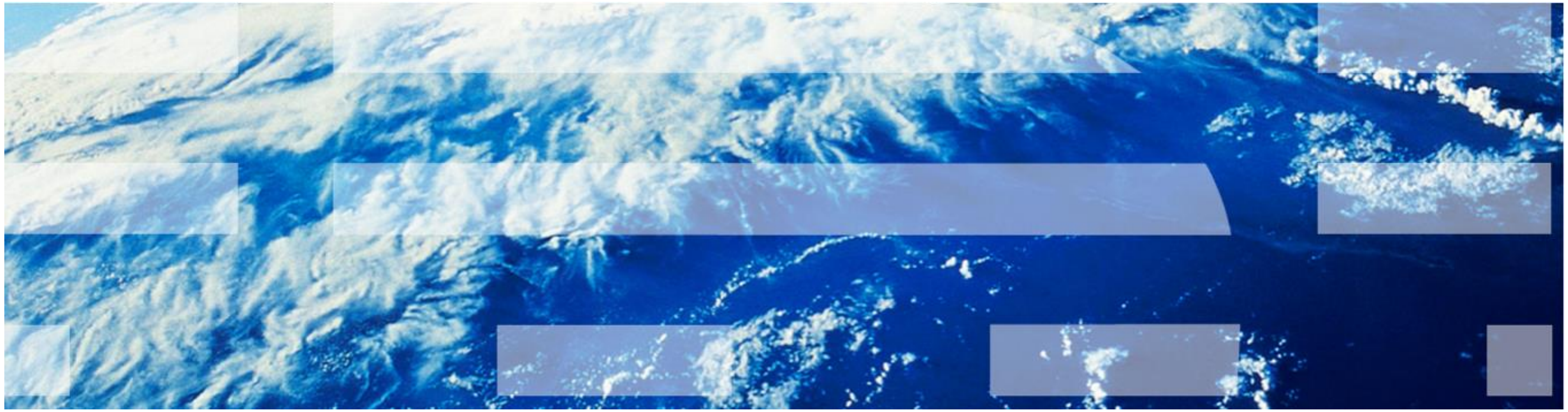
E6893 Big Data Analytics Project Proposal:

Analysis of Traffic Accidents in NYC

Shuo Chang (sc3919)

Sheng Qie (sq2179)

Baochan Zheng (bz2269)



November 19th, 2015

Motivation

- New York City ranks number five in the top 10 worst traffic cities in the U.S. by INRIX Traffic Scorecard[1]

•According to Forbes, NYC is 41.1% greater-than-average accident frequency in the U.S.[2]

•Therefore, certain solutions need to be designed, from the analysis of associated dataset, in order to reduce the rate of traffic accidents in NYC

1 http://inrix.com/new-york-city-ranks-5-in-the-top-10-worst-traffic-cities/

2 http://www.forbes.com/sites/jimgorzelany/2012/08/28/cities-with-the-worst-drivers-2012/

Dataset, Algorithms and Tools

Dataset

- The dataset of traffic accidents in NYC (2012 - 2015) can be downloaded from the following link:

  http://www.wnyc.org/story/nyc-opens-traffic-crash-data-finally/

- The accidents information is compiled in the format of DATE, TIME, BOROUGH, ZIP CODE, LATITUDE, LONGITUDE, CONTRIBUTING FACTORS, VEHICLE TYPE etc.

DATE	TIME	BOROUGH	ZIP CODE	LATITUDE	LONGITUDE	LOCATION	ON STREET NAME	CROSS STREET NAME
11/09/2015	19:00	QUEENS	11419	40.6881124	-73.8193308	(40.6881124, -73.8193308)	LIBERTY AVENUE	125 STREET
11/09/2015	19:00	QUEENS	11365	40.740876	-73.7877345	(40.740876, -73.7877345)	187 STREET	HORACE HARDING EXPRESSWAY
11/09/2015	19:00	MANHATTAN	10013	40.7179306	-74.0009884	(40.7179306, -74.0009884)	LAFAYETTE STREET	WALKER STREET

Algorithms

- In order to determine the optimal algorithm, we plan to test the performances of various classification and clustering algorithms
- Naïve Bayesian Classifier as a starting point

Tools

- Mahout and Spark

Current Progress

-

Title	Duration (hours)	Key	Week 8	Week 9	Week 10	Week 11
Team Formation	1	1				
Topic Selection	3	2				
Background Research	8	3				
Dataset Search	1	4				
Algorithms Search	3	5				
Implementation of Naïve Bayesain Classifer	5	6				

Schedule

-

Title	Duration (hours)	Key	Week 12	Week 13	Week 14
Futher Research on Classification and Clustering Algorithms	10	1			
Futher Implemmentation of Classification and Clustering Algorithms	20	2			
Performance Verification	1	3			
Performance Enhancement	5	4			
Final Report Writing	10	5			
Presentation	0.5	6			

Expected Contributions

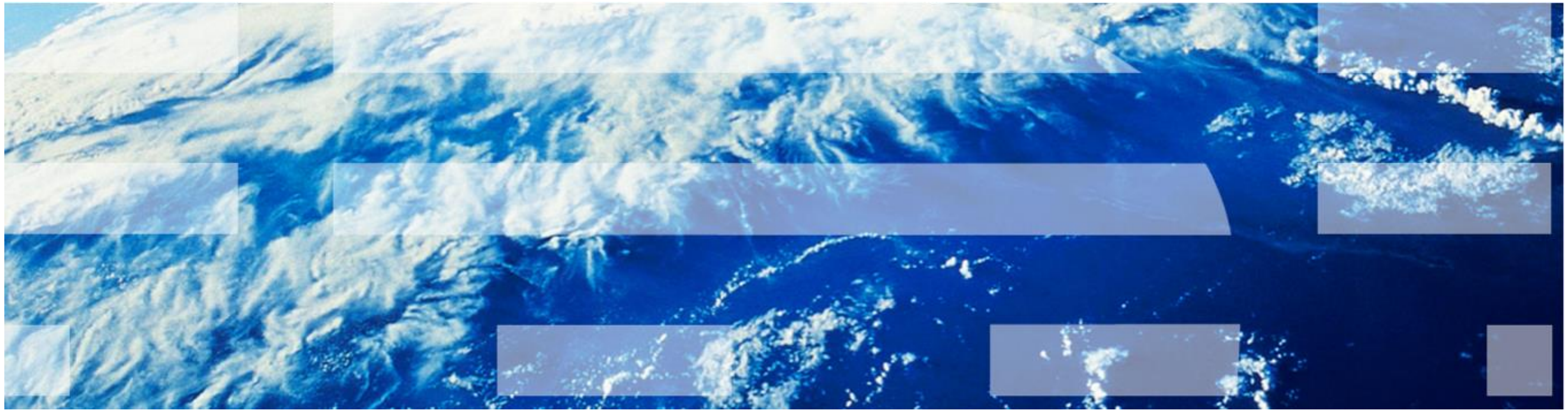
•Determine the contributing factor and vehicle type involved in the traffic accidents with the highest frequency, at given time, location etc.

•Propose solutions to reduce the rate of traffic accidents

E6893 Big Data Analytics Project Proposal:

<Twitter Based Youtube Video Recommender>

<Hanyi Du, Baokun Cheng, Zhe Li>



November 19th, 2015

Why:

1. Information explosion.
2. People are busy, time is money.
3. Profit.

How:

1. Using Twitter API to acquire user's tweets.
2. Analyzing these data to get his/her interest.
3. Recommend related video from Youtube.

- **Datasets:**
- **Tweets from Twitter API**
- **Videos from Youtube API**

- **Algorithms:**
- **Feature Selection**
- **Classification: Decision Trees, Clustering, Naive Bayes, TF-IDF**

- **Tools:**
- **language: Python(sklearn for NLP and ML algorithms), Scala**
- **Tool: Spark**

Current Progress:

1. Got familiar with hadoop,mahout and some related algorithms to deal with and analyze large dataset.
2. Getting familiar with twitter and youtube APIs.

Schedule:

1. First week : parsing twitter dataset part.
2. Second week : Finding youtube videos part.
3. Third week: recommend video to twitter users.

Team Contributions:

1. Hanyi Du : presentation+data&algorithm analysis
2. Baokun Cheng: programming
3. Zhe Li : algorithms choosing and some programming

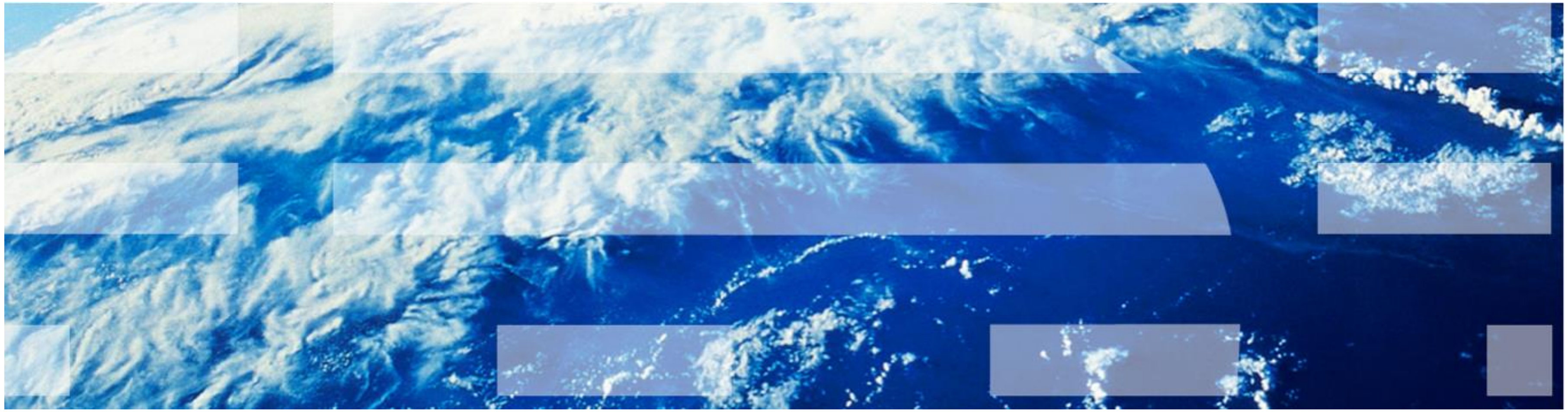
Expected Result: recommend videos that interest twitter users.

E6893 Big Data Analytics Project Proposal:

Yelp Dataset Visualization and Customized Recommender System

Wendan Kang

Jing Hu



November 19th, 2015

Yelp offers a platform for consumers to find restaurants especially through reviews and ratings. A typical search on Yelp displays the best match of the keywords. However, the same keywords will give same search results to different customers so that each customer still has to go though many reviews and ratings before making a choice.

Our project is designed to analyze the Yelp open database and provide customized recommendation based on users' preference. The database analysis part will include implementation of big data analytical tools such as Hadoop, Hive on AWS EC2 and certain visualization tool. The customized recommender system will include implementation of Yelp API, recommendation algorithms and UI on an android app.

Dataset, Algorithms and Tools

- Dataset
  - Yelp Open Dataset
- Algorithm
  - Collaborative Filtering Recommender Algorithm
- Tools
  - Hadoop
  - Hive
  - AWS
  - Yelp API
  - Tableau (Data Visualization Tool)
  - Java

Current Progress, Schedule and Expected Contributions

- ✦Current Progress
  - ✦Dived into the Yelp Academic Dataset
  - ✦Start using Yelp API
- ✦Schedule
  - ✦11/19 — 11/30: Data Pre-Processing and Analysis
  - ✦12/01 — 12/08: Recommender System
  - ✦12/09 — 12/16: UI
- ✦Expected Contribution
  - ✦Data Pre-Processing and Analysis: Jing Hu
  - ✦Recommender System: Wendan Kang
  - ✦UI: Jing Hu & Wendan Kang

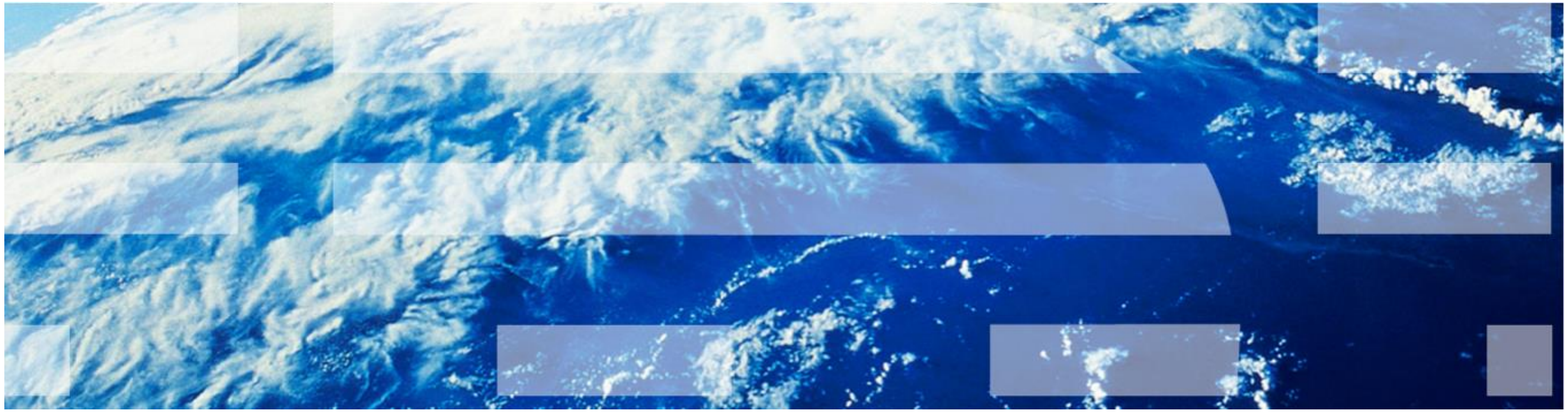
E6893 Big Data Analytics Project Proposal:

Auction Recommendation for Advertiser

Qi Xu (qx2155)

Chen Chen (cc3701)

Xiaowen Li (xl2519)



November 19th, 2015

Motivation

As many recommendations aim at the users based on their phrase searching and clicking. We want to design the recommendation for another kind of users, that is, the advertisers. According to the keyword phrases they bid on, we hope to recommend several appropriate keyword phrases for each advertiser.

Dataset, Algorithms and Tools

✦**Dataset**

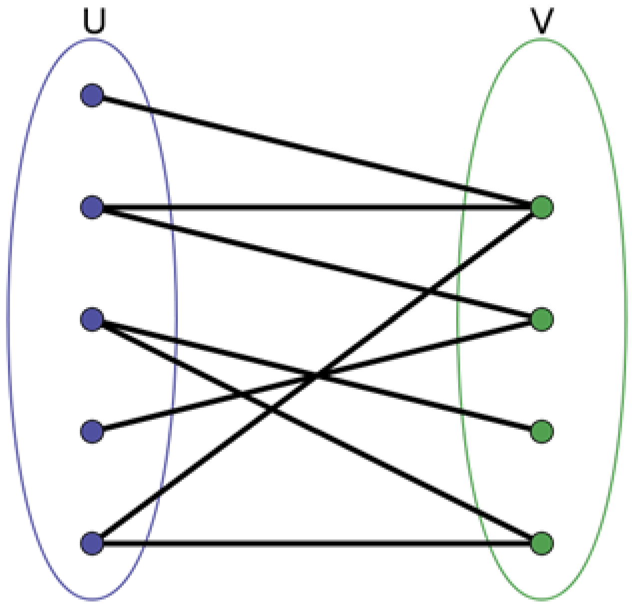
Search Marketing Advertiser-phrase Bipartite Graph (14MB)

Anonymized graph reflecting the pattern of connectivity between advertisers and some of the search keyword phrases they bid on.

Total nodes: 653,260459,678：

✓anonymous phrases ids193,582

✓anonymous advertiser ids2,278,448 edges, representing the act of an advertiser bidding on a phrase.



✦**Algorithms**

Shortest path problem

Maximum weight matching problem Tools

✦**Tools**

Python: Pre-process dataset

System G: Graph visualization

Spark: Process large-scale data

Current Progress, Schedule and Expected Contributions

- **Current progress:**

  We've accomplished the first stage of data analysis, converting the raw data into node and edges.

- **Schedule:**

  Approximately 4 weeks:

  Week 1 : Organize the raw data

  Week 2 : Working on making improvement based on existing algorithm

  Week 3 : Utilizing the algorithm on our data and evaluate it

  Week 4 : Organize the result and write a final report

- Expected contribution:

  Qi Xu: Data analysis and algorithm implementation

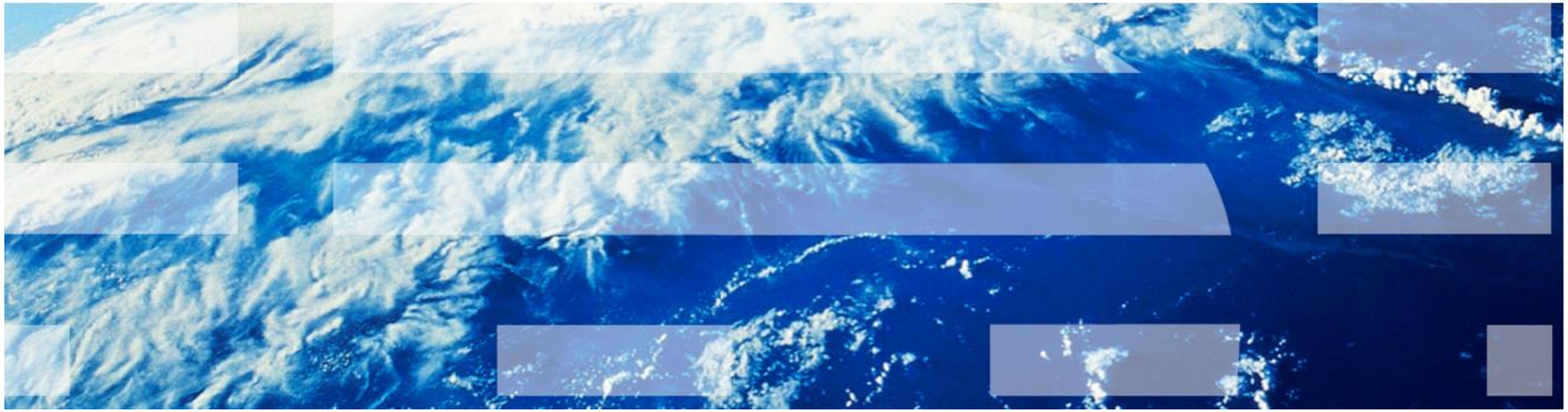
  Chen Chen: Algorithm design and implementation

  Xiaowen Li: UI design and implementation

E6893 Big Data Analytics Project Proposal:

RelEx: Relationship Explainer Using Knowledge Bases

Wangda Zhang (wz2295)

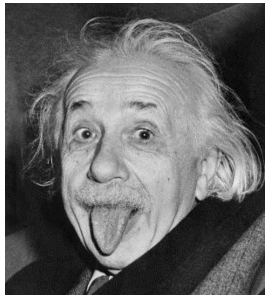


November 19th, 2015

Motivation

Given two objects, what is their relationship?

- <person> lives in <country>
- <person> works at <institution> located in <country>
- <person> married to <person> born in <country>
- ….



Novelty: objects from different domains; relationships are more complex

- Webpages: pagerank
- Facebook friends: common friends, friends of friends

Use **knowledge bases** for objects from general concepts

Dataset, Algorithms and Tools

Dataset: Yago, DBPedia (knowledge bases extracted from Wikipedia)
Storage: property graphs in graph databases (e.g. Neo4j)

Build a system for explaining relationships:

- Online traversal from both objects
    - May be slow for longer path
    - Which relationship is more important?
- Offline learning:
    - Use object class information (hierarchical classes)
    - Discover path patterns: e.g. random walk
    - Rank path patterns: e.g. logistic regression

Tools: Neo4j for storage, Spark MLlib for learning, Alchemy.js for visualization

Current Progress, Schedule and Expected Contributions

Current Progress: data preparation

- Load DBPedia into Neo4j using open source importers
- Implementing online traversal for query processing

Schedule:

1) Finish online query framework
2) Perform path pattern learning
3) Build visualization module
4) Integrate entire explainer system

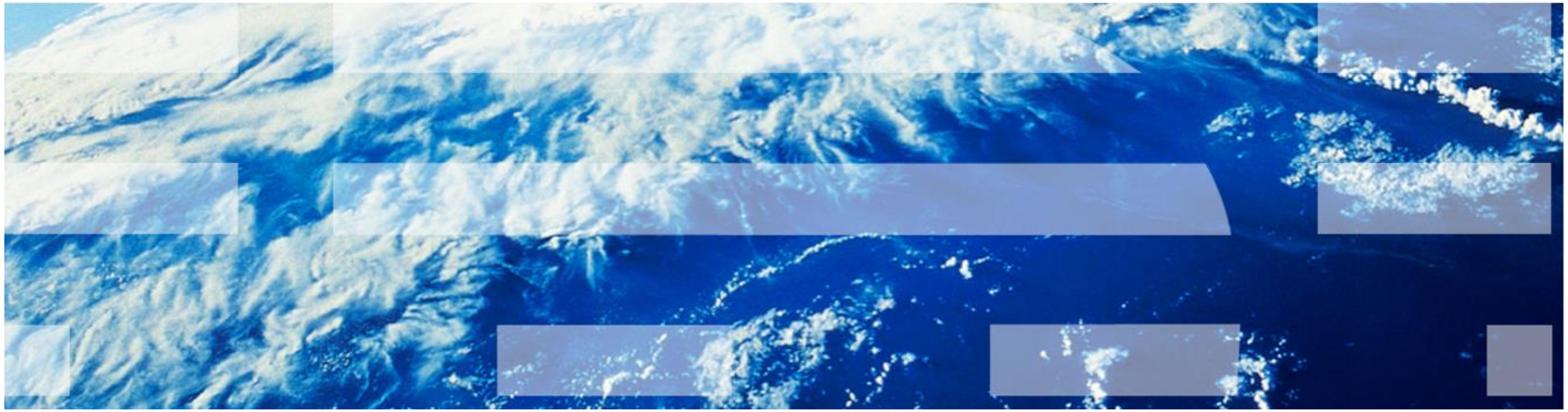
Expected Contributions:

- A prototype system for explaining relationships between general objects

E6893 Big Data Analytics Project Proposal:
Delving into the Q&A network – graph analysis and text mining

Zhen Liang, Xinli Wang

Zl2406, xw2341



November 19th, 2015

Q&A platform is increasingly important for students, engineerers and scientists sharing their knowledge and get their questions answered. Piazza, Stack Exchange are two of popular forums for us.

As users, we are inrerested in:

•What are heated discussed topics

•How easily they get their problems solved using such platforms

As developers, we are interested in:

•The problems users are facing and how they can take such information to improve their products and documentation.

Our project addresses such problems by

•Extracting the key statistics out of large amount of users' data

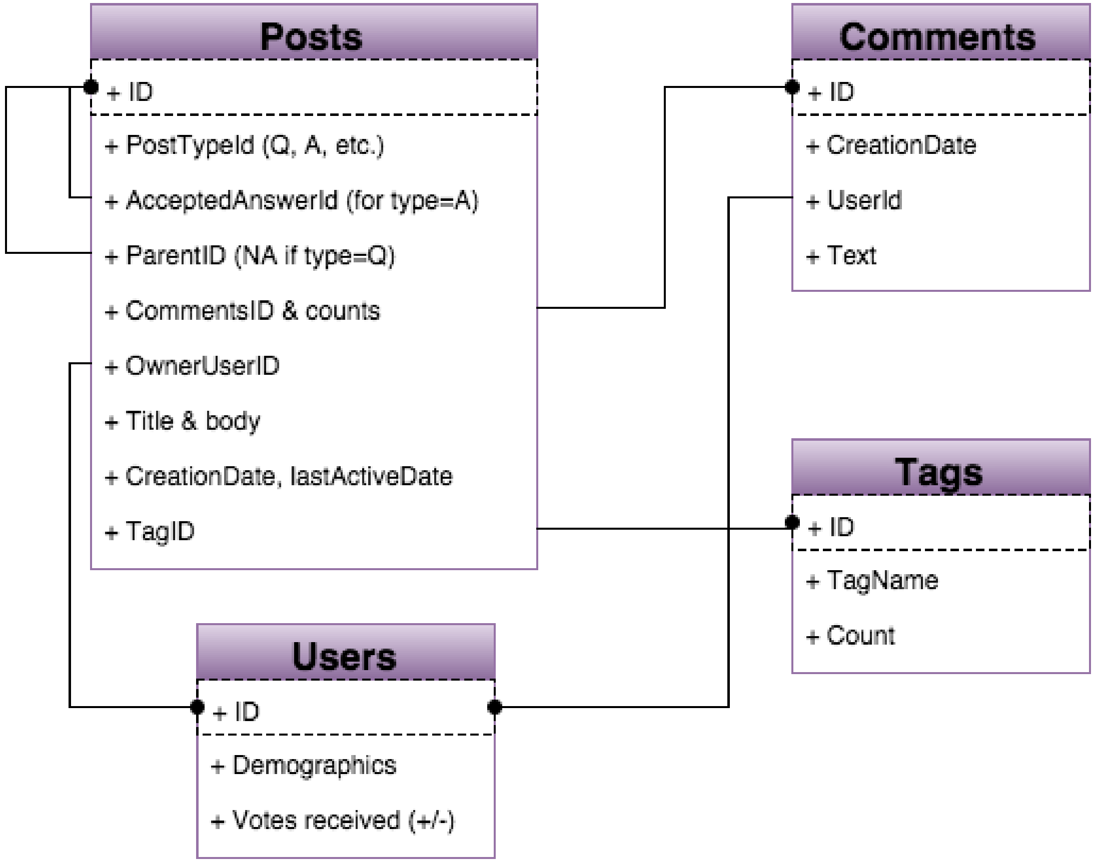
•Merging similar information to reduce information duplicates.

•Visualizing the "network" of questions, to know what's the trends and relationships among discussed topics

- Dataset: **Stack Exchange Data Explorer** (SEDE)

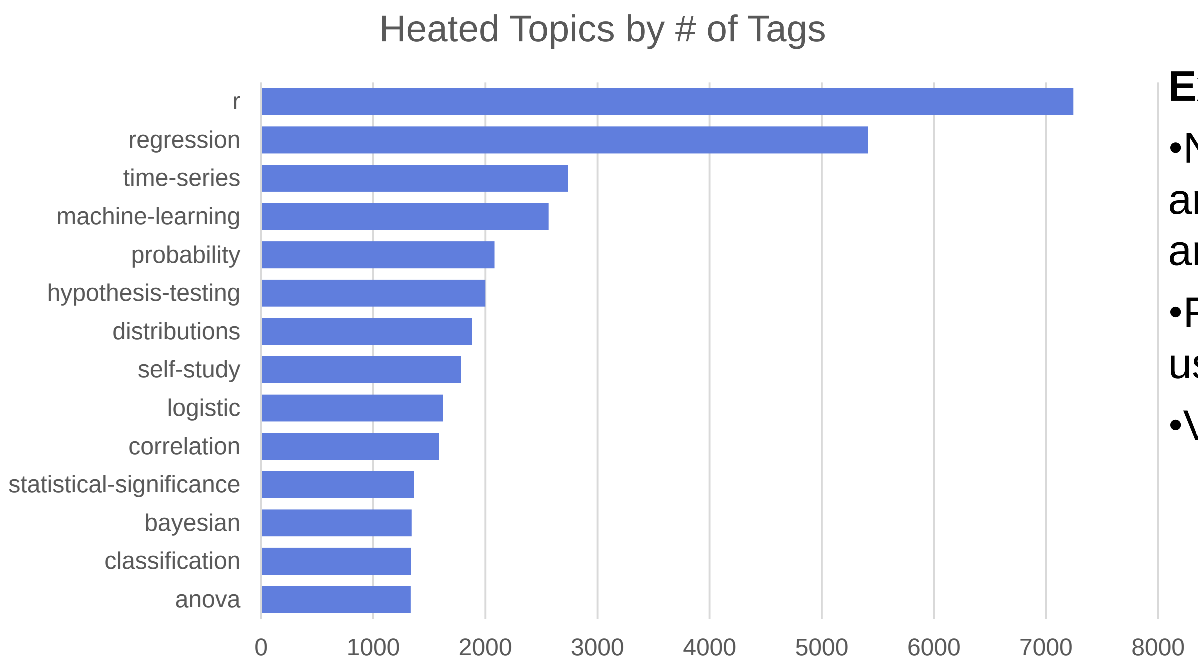
✦Algorithm & Tools:

- ➢ Python(getting and cleaning data, topic modeling)
- ➢ Spark(clustering, sentiment analysis)
- ➢ SystemG, d3.js(visualization)



Source: Stack Exchange website. http://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede

Current Progress, Schedule and Expected Contributions



Expected Contribution:

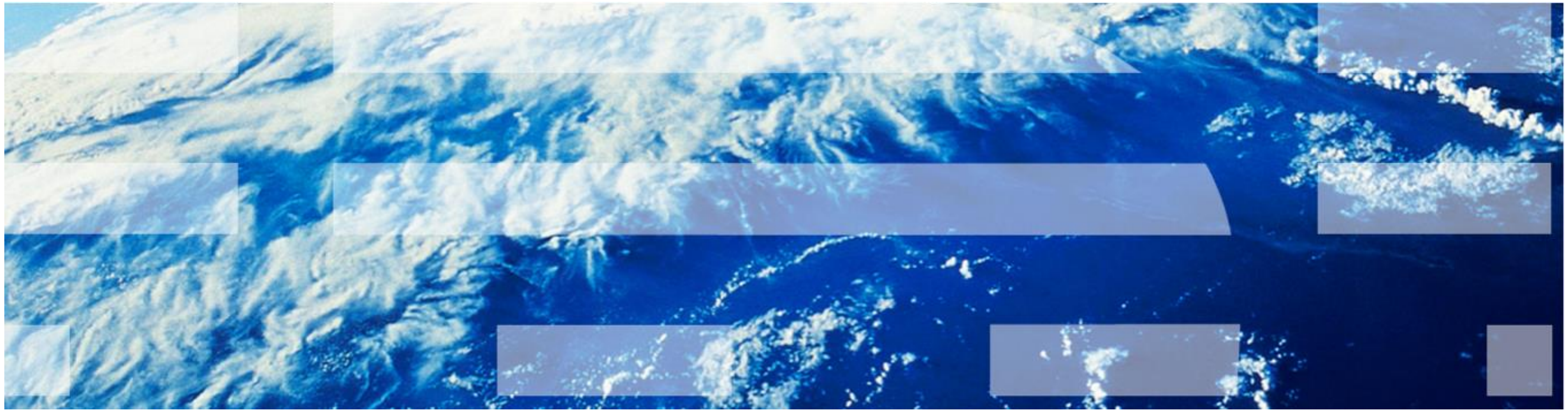
- Novel application of graph analysis in text and users analysis
- Finding trends in topics and user behaviors.
- Visualization dashboard.

	Achievement	Time
Current Progress	Got and cleaned data and performed basic analysis	Nov. 19
Schedule	Text Mining (topic modeling, sentiment analysis)	Nov.20 – Dec. 4
	Graph Analysis	Dec. 4 - 11
	Visualization	Dec 12 - 16

E6893 Big Data Analytics Project Proposal:

<Analysis of Motor Vehicle Accident in NYC>

Team Member Names：Jimin Ge, Xiaowen Zhang, Peiran Zhou



November 19th, 2015

Motivation

New York City has one of the most extensive and oldest transportation infrastructures across the country. However, NYC is infamous for its world's most notorious traffic condition for its high rate of motor vehicle accidents. Today, the city is renowned for its commercial and prosperous scene. With the large population of motor vehicle holders in NYC, we noticed that we could use data science tools to probe into this phenomenon.

To achieve this goal:

•Grab dataset of motor vehicle accident reports for NYC
•Analyzing the historical relationship between accidents and time / location
•Build category by descriptions using topic modeling
•Build and train classifier to classify different category, and predict classification result given time and location information

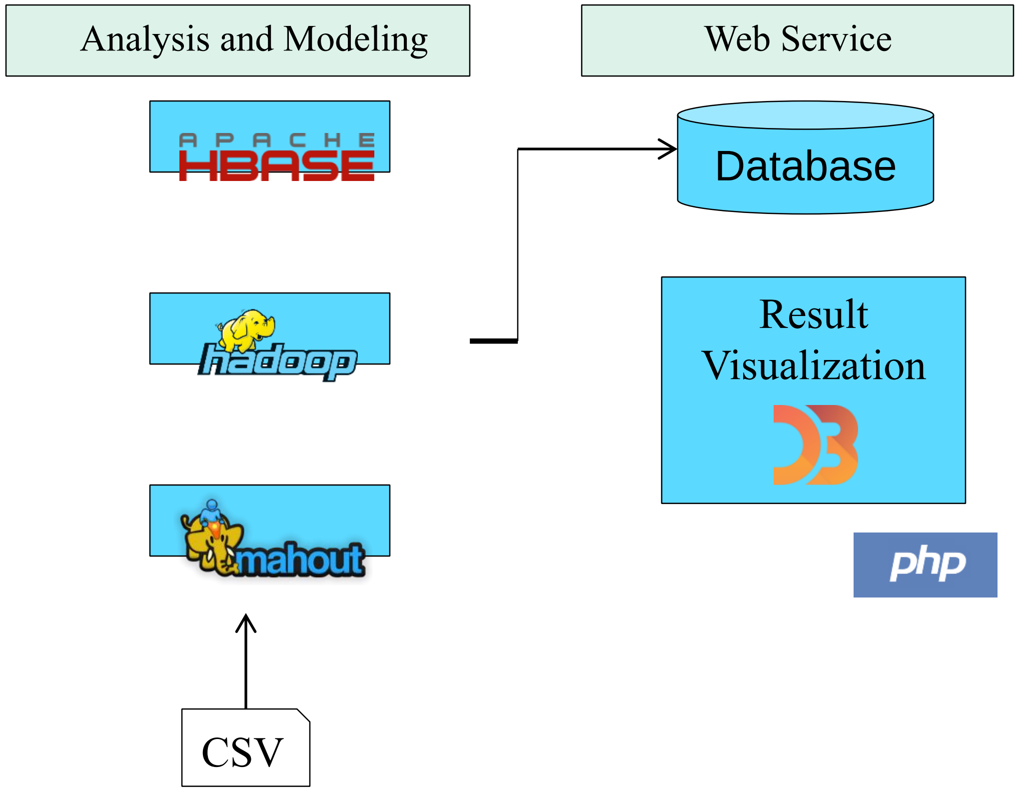
Dataset, Algorithms and Tools

✦Dataset

NYPD_Motor_Vehicle_Accidents.csv; *(https://data.cityofnewyork.us)*

✦Algorithms

1. Naive Bayes Classification
2. K-Means Clustering
3. Latent Dirichlet allocation



✦Tools

1. Hadoop, Mahout, Hbase

2. R, Python

3. PHP, HTML, JavaScript

✦Current Progress

1. Downloaded the data from data.cityofnewyork.us, and transformed the data set into the common CSV format. Also, we have created the train.csv and the test.csv files on the basis of Naive Bayes.

2. Designed the front-end of interactive visualization module.

✦Schedule

1. Realizing Classification Engine before December.

2. Implementing interactive visualization module about December 10.

3. Test and debug the system, analyzing the statistic result, preparing the final presentation.

✦Expected Contributions

1. train.csv; test.csv - Naive Bayes Classification / Latent Dirichlet allocation; *(Xiaowen Zhang)*

2. Motor_Vehicle_Accident-Based Classification Engine; *(Jimin Ge, Xiaowen Zhang)*

3. D3.js-based visualization for Bayesian networks; *(Peiran Zhou, Jimin Ge)*

4. PHP-Based Interactive Visualization; *(Peiran*

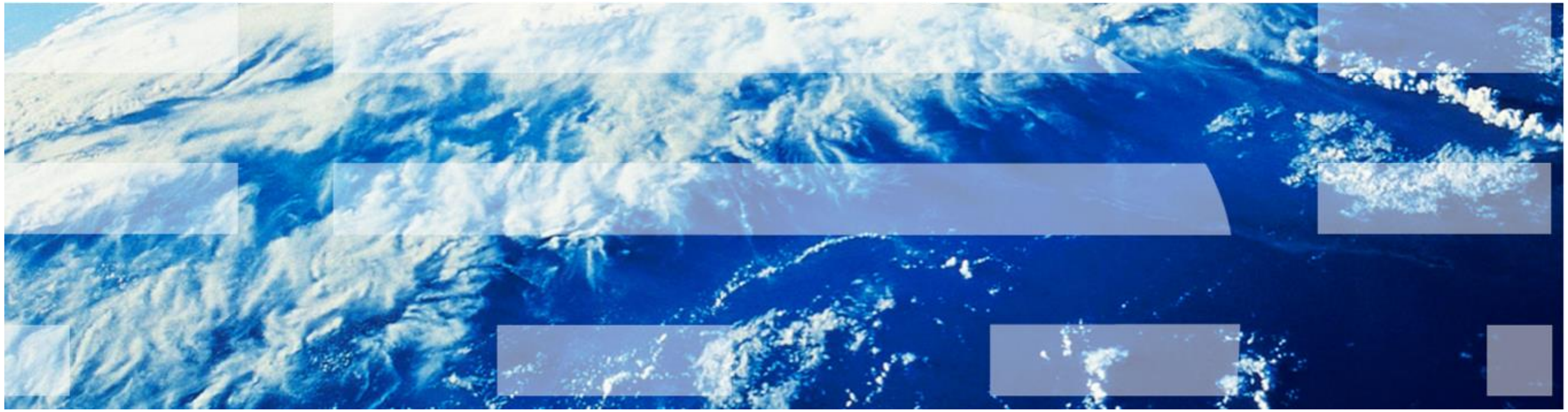
E6893 Big Data Analytics Project Proposal:

Hospital Charge Data Analysis

Anubha Bhargava

Caleb Perry

Turab Ali



November 19th, 2015

Motivation

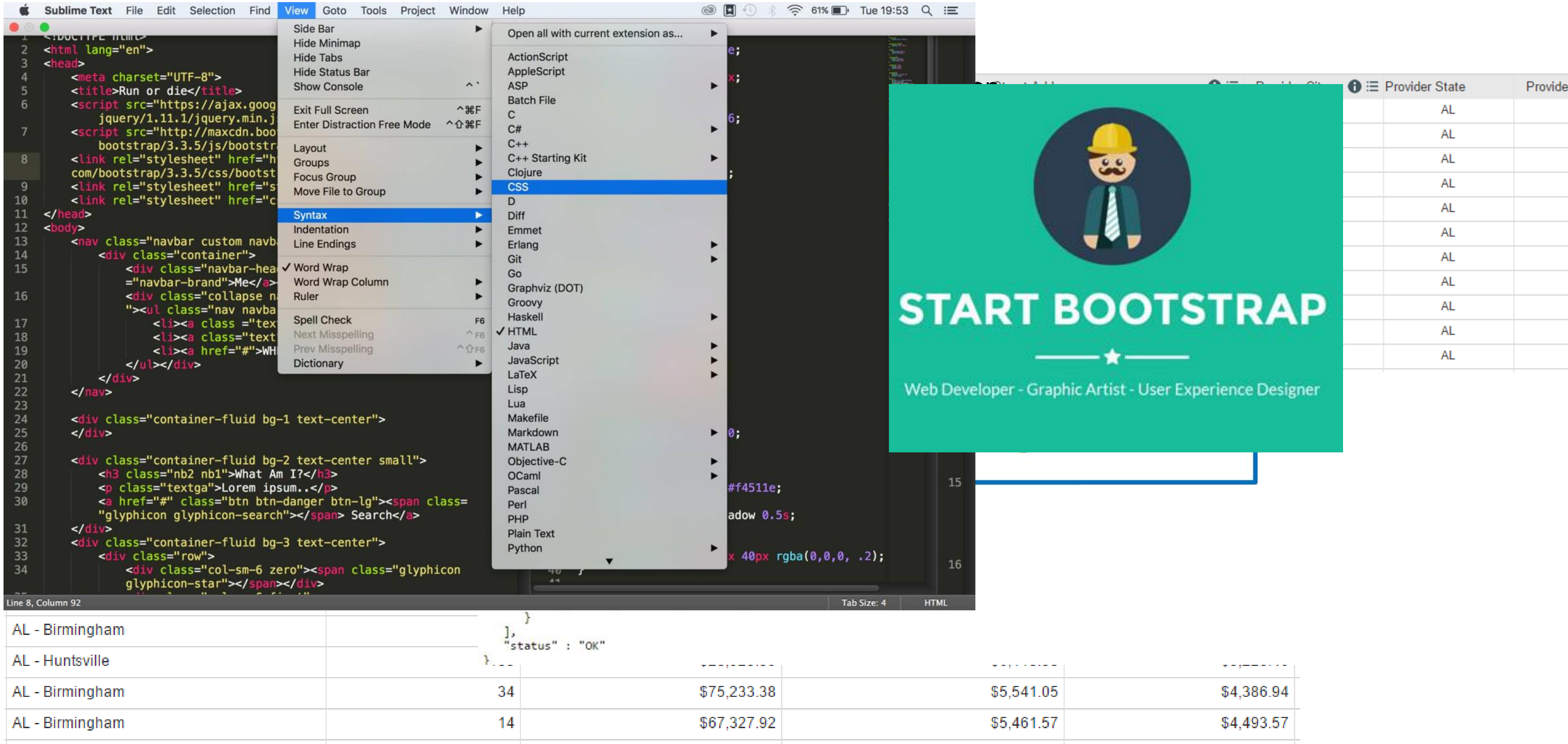
- We want to create a useful, problem-solving tool.
- Patients in hospitals want to know the medical expenses prior to receiving care.

We will create a webpage that will:

1) Allow patients to identify which hospitals offer lower prices
2) Focus a user's search on the hospitals closest to them
3) Give patients an idea of how much their care may cost

Dataset, Algorithms and Tools

- Our primary dataset is the “Inpatient Prospective Payment System (IPPS) Provider Summary for the Top 100 Diagnosis-Related Groups (DRG) – FY2011”



Current Progress, Schedule and Expected Contributions

Current Progress:

Deadlines identified

Project goals and definition

Division of labor

Schedule:

11/19/2015 Proposal presentation – begin coding upon project approval

12/7/2015 Share code, slides, and report contributions for integration

12/10/2015 Team members test, peer review, and update each other's work

12/17/2015 Final project submission

Expected Contributions:

Anubha Bhargava

1) Geocoding addresses and sorting by distance
2) Integrating the final report

Caleb Perry

1) DRG lookup and price range
2) Integrate final presentation

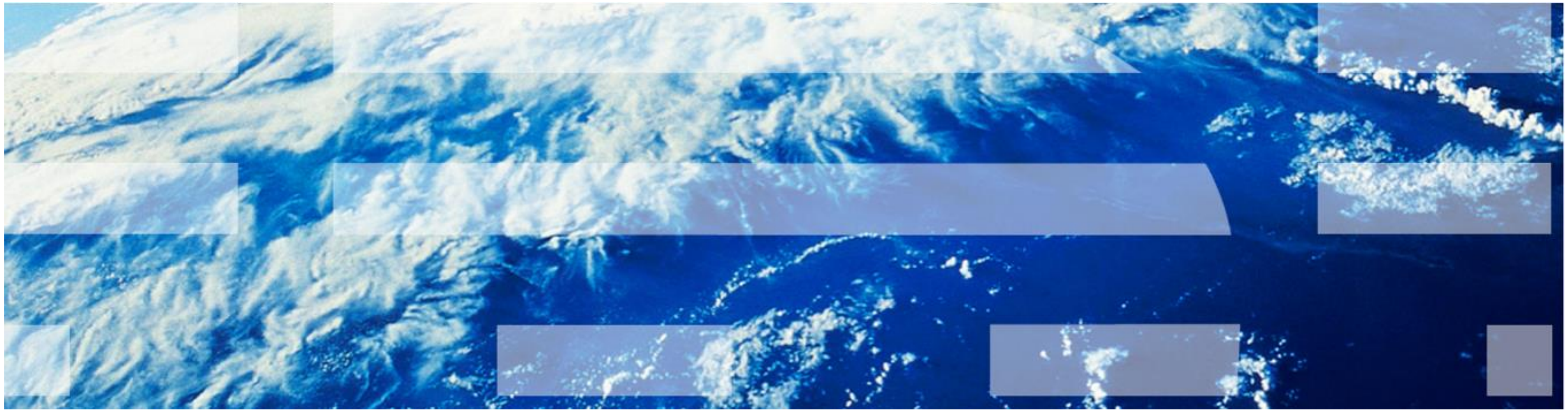
Turab Ali

1) Sorting hospitals by relative cost
2) Website user interface

E6893 Big Data Analytics Project Proposal:

<Twitty-Foodie : Twitter-Based Food Recommendation>

<Tianlong Li, Mei Mei, Shanqing Tan>



November 19th, 2015

Motivation

Twitter users offer a variety of insight about restaurants that is largely missed in various approaches to make best dining choice. We plan to gather such data through Twitter's streaming API, targeted at tweets about restaurants near our campus, generate useful results in terms of quality and popularity of restaurants for students and residents around campus. To better serve our goal, a visualization front end will also be implemented, aggregating data and curating an appropriate view based on use case and user input.

Dataset, Algorithms and Tools

✦Dataset: Twitter Streaming API, both real-time and stored results.

✦Algorithms: Geospatial and textual analysis related algorithms.

✦Tools: Tweepy as Twitter streaming library, Django or Flask for frontend website, Node.js for backend server, Amazon EC2 & S3 & SimpleDB or DynamoDB for hosting visualization website and storing raw as well as parsed data, Amzon SNS&SQS or Kafka or RabbitMQ for message queue services, D3.js and Google Map API or Leaflet for visualization of data, Alchemy API and NLTK for natural language and sentimental analysis, Mahout or Spark for generating recommendation, Pig for batch extracting and processing tweets from raw data, contingent upon further implementation and designing.

Project Strategy

1, Gather Twitter and instagram data on restaurants near Columbia University.

2, Identify tweets related to restaurants.

3, Create a web-based visualization of the gathered data that shows a map of the restaurants and tweets related to them.

4, Analyze the data to provide users with useful insights about these data, based on specific factors such as number of retweets.

5, Produce a rating system based on the collected data from Twitter

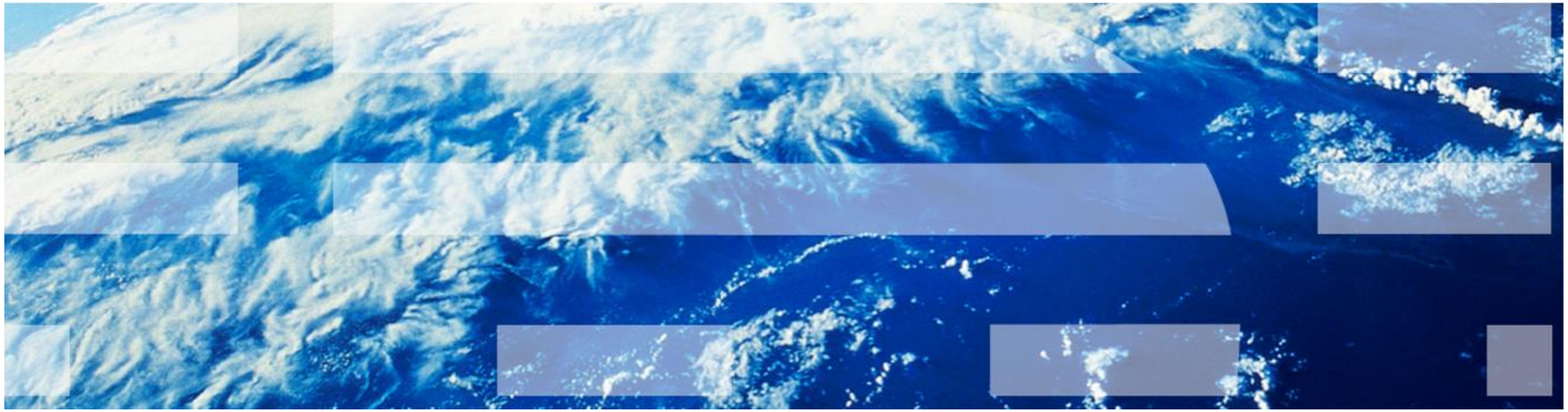
Current Progress, Schedule and Expected Contributions

- **Current Progress**
  - Deciding on exact algorithms and tools for restaurant recommending.
  - Learning to use unfamiliar tools and frameworks for the project.
- **Project Timeline**
  - Week of 11/16: Gather experimental tweets for processing and trying algorithms.
  - Weeks of 11/23 &11/30: Build up frontend visualization website and backend server logic using collected raw data for testing.
  - Weeks of 12/07 & 12/13: Deploy all components to AWS, beta testing and initial write-up of reports as well as slides.
  - Last week: Finalize write-up and demo videos.
- **Expected Contributions**
  - Dynamic adjustment of contributions and responsibilities is decided since this is covering a wide range of stacks and they are closely interconnected.

E6893 Big Data Analytics Project Proposal:

Photo Similarity & Recommendation for Journey

Team Members: *Zhengrong Li (zl2438)*
Xingying Liu (xl2493)
Sen Lin (sl3773)



November 19th, 2015

Motivation

The information of photo album is useful. It reveals user’s travel reference

Classic, Modern, Natural, Arts …

We want to implement an App which could recommend some places to visit based on user’s photo album

Dataset

Google Street View, Local photo album

Algorithm

Extract feature:

Histogram, SURF, SIFT, Shape Detection

Similarity Match:

Euclidean Distance Similarity,

Cosine Measure Similarity,

Nearest-neighbor search

Speed up process:

Locality Sensitive Hashing (LSH)

Trade-off:

Accuracy, Speed VS. Reliability, Scalability

Tools

Hadoop, Pig, OpenCV

Current Progress, Schedule and Expected Contributions

Current Progress
1.Architecture design
2.Research on related algorithms

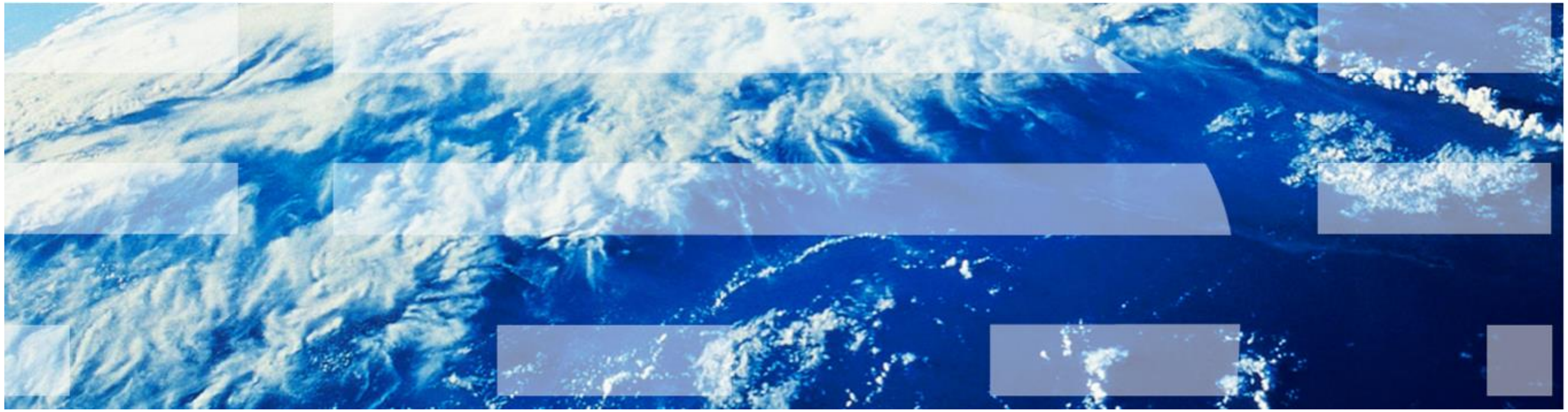
Schedule:
1.Feature extration & Similarity computation by 25 Nov
2.Front end by 30 Nov
3.Report by 5 Dec

Contributions
1.Find keypoint features of local image
2.Compute similarity of features of local image and dataset
3.Determine quality of match, find and record five pictures with highest quality
4.Repeat step 1 to 3 until all local images are processed
5.Compute the appearance time of cities in step 3. Find the cities with the highest frequency

E6893 Big Data Analytics:

Yelp Review Analysis and Recommendation

Team Members: ***Lan Yang, HongYang Bai, YaZhuo Nan***



November 19th, 2015

Describe Our Topic: Cultural Trends

Lots of aspects,...

- What is the usual lunch time at different area? Do Americans eat late than English or Chinese?
- What is the food preference at different area about a particular food? Is sushi very popular in all states of USA?
- What is the trending of giving tips at different area?

Commercial Contribution

restaurant owners..

customers..
motivation of your project

5 json files contained inside

- yelp_academic_dataset_business.json
  - keep track of the information of business(address, open/close hours, categories, city, name, stars, delivery,)
- yelp_academic_dataset_review.json
  - users' reviews(user id, comment date, stars, text)
- yelp_academic_dataset_tip.json
  - users' tip for a business store(user id, tip text, business id)
- yelp_academic_dataset_user.json
  - users' information summary(user id, name, helpful, cool, average stars, review count, friends' id)
- yelp_academic_dataset_checkin.json
  - Users' checking information

related to cultural trends:

- **business table:**
  - delivery attribute
  - parking attribute
  - accept credit card, wifi free or not, price range

- **tip table:**
  - tips for a user on a typical business store

- **user table:**
  - a user's review on some store may also be prefered by his/her friends.

Method & Language

PIG: Manage the yelp database using the PIG platform for analyzing.

Clustering: K-mean clustering method using Mahout or Spark to cluster users according to their ratings on various venues.

Recommendation: A typical user's like/dislike on a business store may influence his/her friends' taste. Create maven project in eclipse, use java to gain recommendation information for a user's friends.

Visualization: Use IBM System G to visualize the cultural trend disseminating among the Yelp users.

Challenges

- There are might be duplicate data(users, tips, ratings)
- There are might be corrupted data(outliers)
- Association between different datasets may not be quite obvious, various methods might be needed to improve performance
- Datasets may be too large for a single PC to process

Expectation

- Specific data analysis result on a list of cities.

- Visualized presentation of data analysis result

Questions and comments

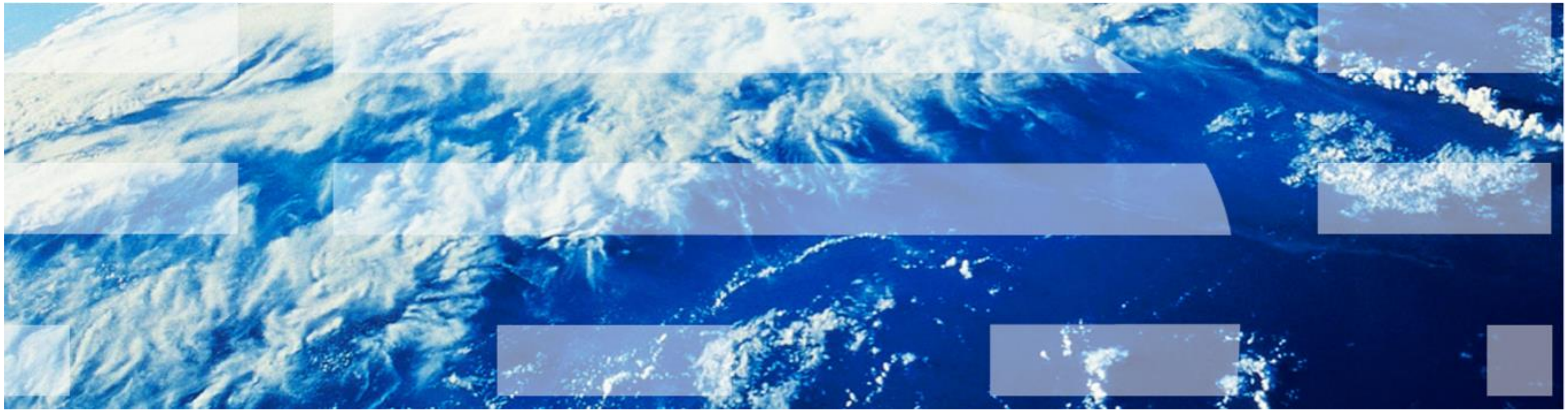
E6893 Big Data Analytics Project Proposal:

Hot Issue Extractor

Guangshi Chen

Haitian Sun

Sihan Zhao



November 19th, 2015

Motivation

Online Social Media becomes popular now. People leave different comments about the current hot issues on the internet.

We will use big data strategy to analyze all users' online comments and try to find the current hot issue.

Dataset, Algorithms and Tools

Dataset:
Comments from the users of microsoft forum

Preprocessing:
Tokenize word and normalize sentences

Algorithm:
Pagerank(popular sentences)
Cosine similarity(common words)
Square root is to reduce the effect of the long-sentence to the whole distribution.

Tools:
Spark(python), Mysql

Current Progress, Schedule and Expected Contributions

Schedule:

Now - 11.25: Collecting data and Pre-processing
11.26 - 12.2: Calculation of Similarity
12.3 - 12.6: Extraction of hot issues with pagerank
12.7 - 12.13: Optimization of execution time and more advancement
12.14 - 12.17: Preparing for the final project report

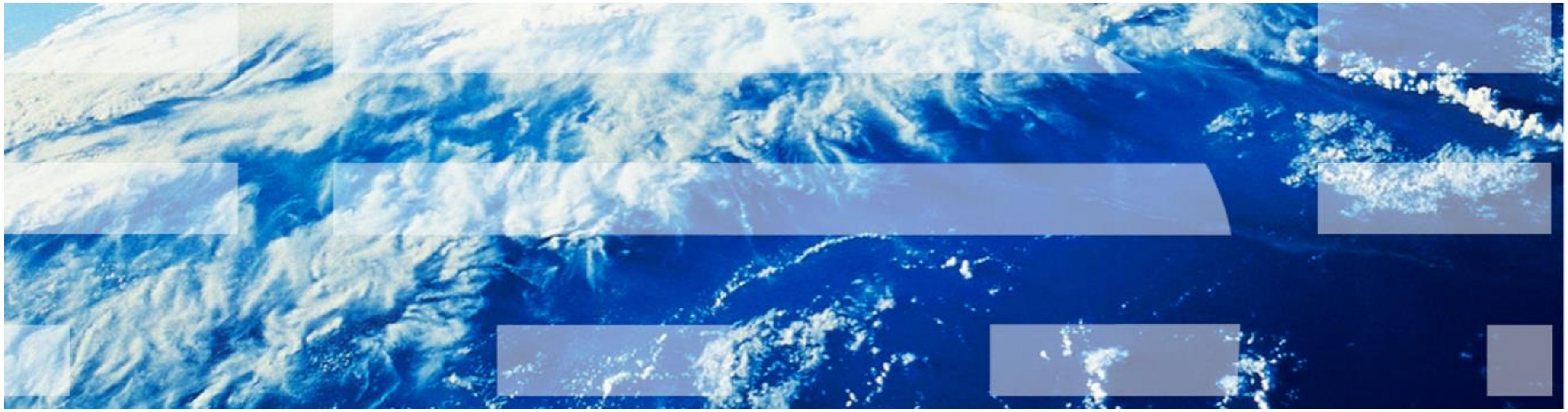
Expected Contributions: to construct an extractor for hot issues from the Internet based on big data analytics.

Writing a python script and deal with the database together

E6893 Big Data Analytics Project Proposal:

Factors Lead to Win NBA Games

Team Members:
Xuhui Wang, Yuantuo Yu, Jiadong Yan



November 19th, 2015

- Explore individual sporting interests

Find out the critical factors lead to win NBA games

Rebound, Assist, Points, Block and etc.

Help coach by using the results to strong his team

Dataset, Algorithms and Tools

Data retrieved from http://www.databasebasketball.com/
Covers all NBA basketball stats such as rebounds and assists for every NBA teams from season 1976 to season 2009

1	team	year	leag	o_fgm	o_fga	o_ftm	o_fta	o_oreb	o_dreb	o_reb	o_asts	o_pf	o_stl	o_to	o_blk	o_3pm	o_3pa	o_pts	d_fgm	d_fga	d_ftm	d_fta	d_oreb	d_dreb
2	ATL	1976	N	3279	7176	1836	2451	1244	2512	3756	1882	2302	733	1779	330	0	0	8394	3409	7137	1909	2527	1121	2533
3	BOS	1976	N	3462	7775	1648	2181	1241	2966	4207	2010	2039	506	1673	263	0	0	8572	3559	7904	1616	2180	1110	2753
4	BUF	1976	N	3366	7475	1880	2492	1213	2623	3836	1883	1842	683	1699	392	0	0	8612	3786	7917	1404	1859	1268	2721
5	CHI	1976	N	3249	7186	1613	2159	1292	2705	3997	1989	1871	699	1552	364	0	0	8111	3306	7095	1425	1907	1055	2559
6	CLE	1976	N	3451	7688	1468	1993	1312	2563	3875	1845	1951	579	1356	472	0	0	8370	3265	7268	1748	2325	1202	2711
7	DEN	1976	N	3590	7471	2053	2783	1288	2700	3988	2262	2142	953	2011	471	0	0	9233	3585	7743	1635	2231	1269	2481
8	DET	1976	N	3764	7792	1442	1960	1169	2495	3664	2004	2200	877	1718	459	0	0	8970	3561	7539	1933	2543	1317	2637
9	GSW	1976	N	3724	7832	1649	2172	1300	2639	3939	2120	2058	904	1624	432	0	0	9097	3567	7584	1699	2282	1256	2640
10	HOU	1976	N	3535	7325	1656	2103	1254	2632	3886	1913	2132	616	1600	411	0	0	8726	3424	7356	1746	2252	1121	2232
11	IND	1976	N	3522	7840	1714	2297	1409	2584	3993	2009	2030	924	1609	458	0	0	8758	3599	7629	1705	2252	1378	2770
12	KCK	1976	N	3561	7733	1706	2140	1222	2593	3815	1982	2173	849	1576	386	0	0	8828	3422	7244	1912	2513	1097	2739
13	LAL	1976	N	3663	7657	1437	1941	1177	2628	3805	2057	1867	801	1538	445	0	0	8763	3515	7781	1510	1990	1348	2625
14	MIL	1976	N	3668	7840	1553	2072	1220	2519	3739	1970	2094	790	1648	342	0	0	8889	3712	7753	1721	2330	1265	2613
15	NOJ	1976	N	3443	7602	1688	2183	1249	2828	4077	1854	2099	613	1706	357	0	0	8574	3486	7712	1833	2448	1318	2781
16	NYK	1976	N	3659	7530	1587	2078	974	2680	3654	1956	2007	714	1680	304	0	0	8905	3577	7610	1752	2327	1163	2716
17	NYN	1976	N	3096	7222	1673	2274	1157	2547	3704	1422	2178	802	1630	435	0	0	7865	3279	7074	1863	2488	1149	2937
18	PHI	1976	N	3511	7322	2012	2732	1293	2752	4045	1966	2074	814	1915	561	0	0	9034	3575	7920	1561	2074	1416	2448
19	PHO	1976	N	3406	7249	1791	2345	1059	2493	3552	2100	2089	750	1830	346	0	0	8603	3320	7192	1903	2525	1180	2594
20	POR	1976	N	3623	7537	1917	2515	1260	2703	3963	1990	2220	868	1757	492	0	0	9163	3408	7404	1889	2514	1197	2510

Using pig to list out won rates vs. every single specific stat

Map-reduce method is deployed to do the job
- Map stage to pick out a specific stat with won and lost from dataset
- Reduce phase to combine and deal with the result from map operation
- Map-reduce has optional finalized stage to make optimization to the results

Current Progress, Schedule and Expected Contributions

Currently Progress :

We have found a dataset including rebounds, assists, steals, won rates, and many other basketball stats for every NBA teams from 1976 to 2009.

Schedule:

Week 10: analyze the dataset, and list all won rates respected to every single specific stat by using PIG, for example, every single won rate at total rebound number from 30 to 40

Week 11: after having all generated data, we are going to build a diagram to describe the relation between won rate and total rebound number for instance

Week 12: from all the diagram we build, we are going to find out which of those stats contribute relatively more to a win of NBA basketball game

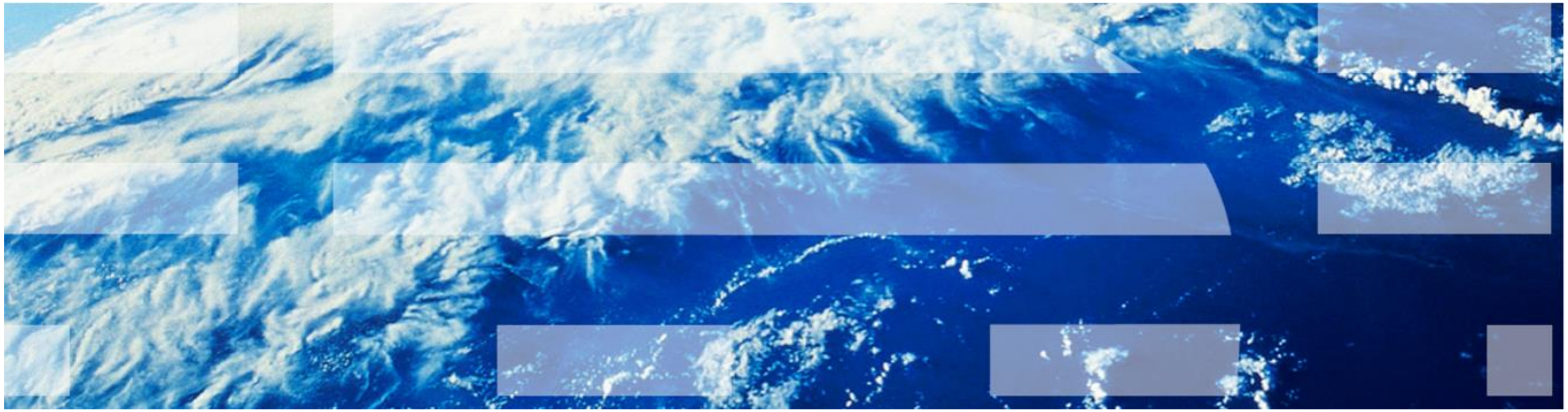
Expected Contributions:

- help analyze critical factors lead to a NBA basketball win
- help team build with different types of players who are able to contribute those critical factors
- help a team know which part they need to emphasize in order to get a better won rate

E6893 Big Data Analytics Project Proposal:

Map-Reduce for Algorithmic Trading

Akshaan Kakar, Alice Berard



November 19th, 2015

Motivation

- Algorithmic Trading
  - Algorithmic trading is a highly competitive sector of global financial markets.
    - -Marginal profit per trade is low
    - -Potential for profit is high
  - Trading algorithms use metrics and heuristics to generate trading signals.

- Why Big Data?
  - Testing is the most crucial aspect of algorithm development.
  - Involves testing strategies on stock tick data
    - -Backtesting
    - -Live testing
  - Historical stock tick data (minute/day resolution, multiple symbols)
  - Live stock ticker data : data store keeps expanding with time

- Our Goal
  - To build a trading algorithm testing engine with result visualization

Dataset, Algorithms and Tools

- Dataset
  - Platform is data source agnostic
  - We will use historical data for S&P 500 symbols from QuantQuote
  - Live data stream from Yahoo! finance API

- Algorithms
  - Map-Reduce paradigm to retrieve time slices of large time series
  - Custom algorithm to apply trading algorithm rules to data
  - Numerical algorithms to computer moving averages, Sharpe Ratio etc.

- Tools
  - Hadoop Distributed File System
  - Daemons to update HDFS store in batches
  - Spark atop Hadoop to run trading algorithms
  - Spark visualization modules to depict algorithm performance

Current Progress, Schedule and Expected Contributions

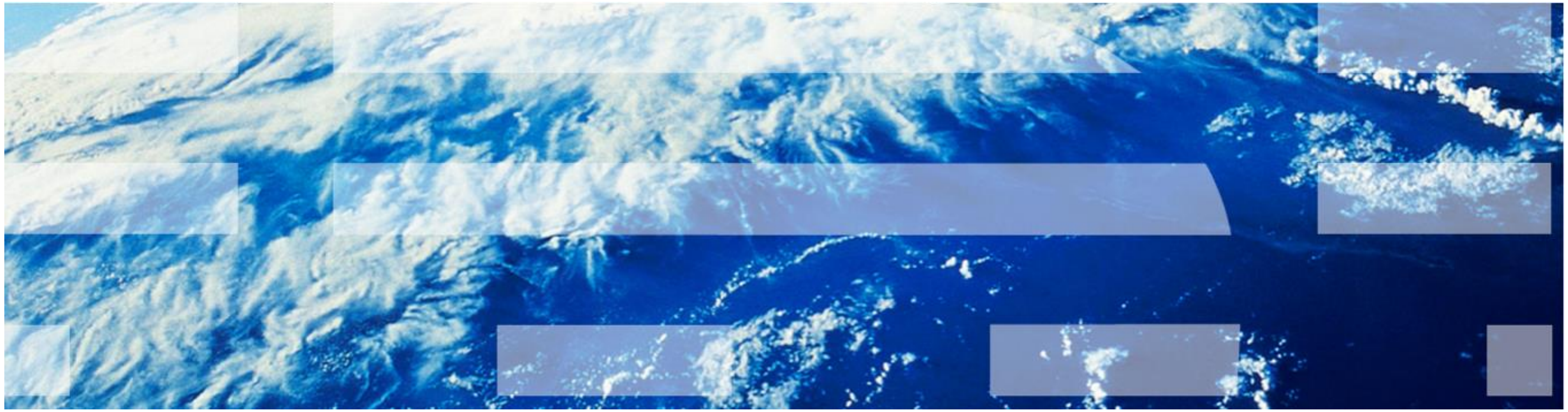
- Progress & Schedule
  - The high-level layout has been confirmed
  - The required data sources have been explored
  - The scope of trading algorithm features supported is to be decided
  - Next step is to implement execution of trading rules and performance viz.

- Expected Contributions
  - We expect to deliver an easy-to-use, inherently distributed, algorithmic trading engine with the following features
    - -Extensive backtesting capability
    - -Live testing features
    - -Performance metric computation
    - -Performance visualization

E6893 Big Data Analytics Project Proposal:

Movie clustering and recommendation based on Netflix movie rating data

Tianchun Yang, Ziyi Luo, Pengyuan Zhao



November 19th, 2015

Motivation

Based on the Netflix movie rating data over 17 thousands movies from 480 thousand customers, our group propose to make the following analytics:

1. Movie clustering based on the movie date, rating, number of rating;
2. Movie recommendation for certain customer based on the customer rating record.
3. Based on the clustering result of movie clustering, check whether the recommendation for customers are reasonable.

Dataset, Algorithms and Tools

Dataset information:

✦Netflix Prize Data Set (Data size: 2 GB)

✦Available online: http://academictorrents.com/details/9b13183dc4d60676b773c9e2cd6de5e5542cee9a

✦Note: The dataset is used as a data analysis competition (i.e., rating prediction). Here we use the dataset for different analysis.

Algorithms:

✦Clustering algorithm: K-means

✦Recommendation algorithm: Knn Item-based recommendation with log likelihood similarity.

Tools:

✦Hadoop & Mahout

✦AWS amazon cloud computing platform

✦Eclipse

Current Progress, Schedule and Expected Contributions

Current Progress:

- The Netflix dataset has been rearranged for clustering and recommendation respectively.
- AWS platform is already available for data analysis.

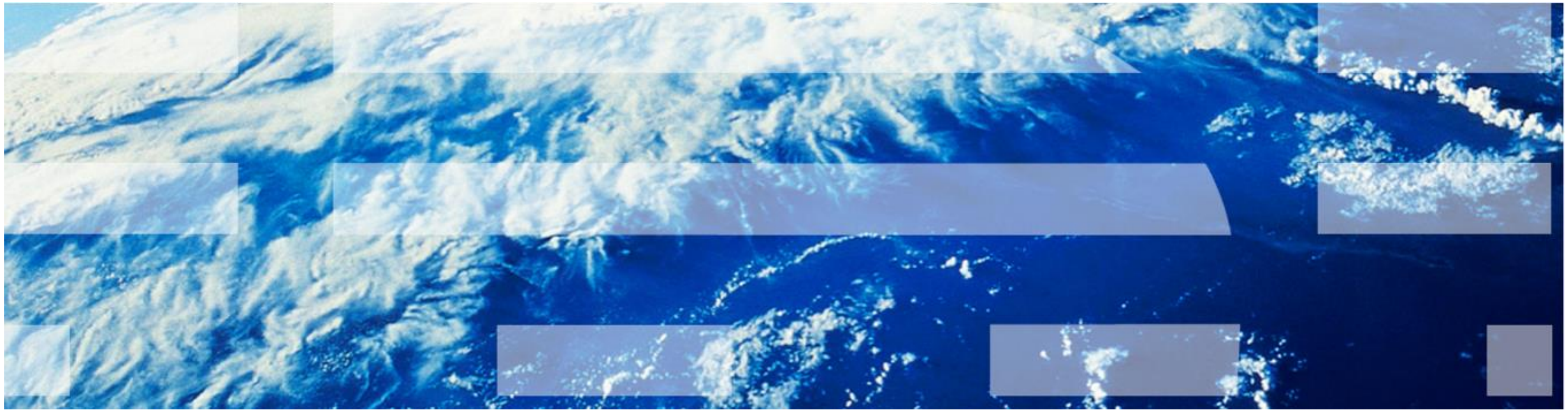
Schedule:

- Extract files can convert into Hadoop file respectively.
- Launching clustering and recommendation jobs on AWS
- Analysis the results and do technical report

E6893 Big Data Analytics Project Proposal:

League of Legends team builder

Chenli Yuan (cy2403)



November 19th, 2015

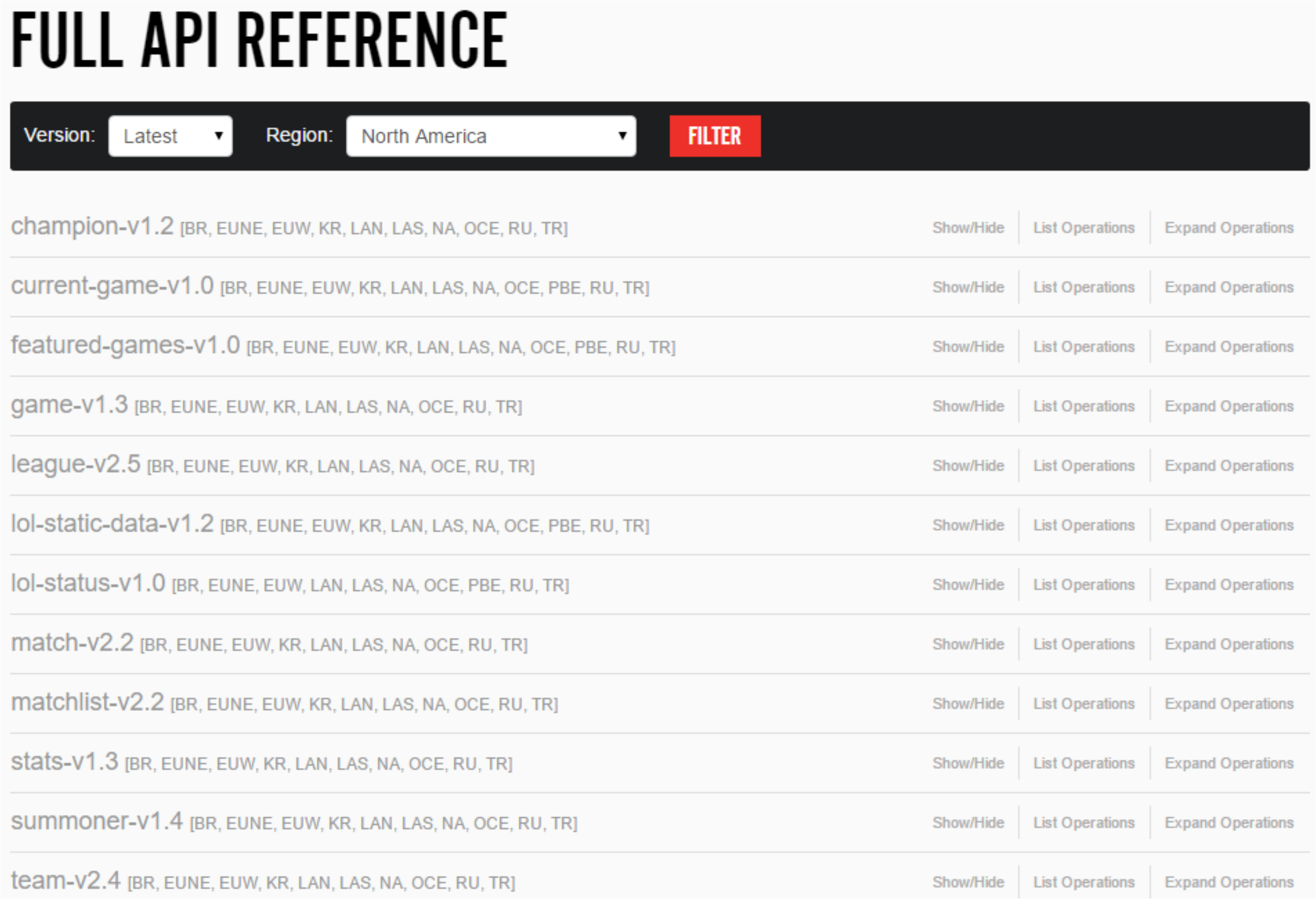
League of Legends is one of the most popular multiplayer online battle arena game. The decisive factors of a game’s result include: players’ performance, objective control, team strategy and team composition.

This project aims to provide a solution to building a team in League of Legends Fantasy, and similarly, provide professional teams a data based analytical method for better team management.

This project focuses on analyzing pro-players’ performance from past games, and learning each player’s playstyle. For example, aggressive players perform better in fast push strategy, while passive players fit better in a late game team composition.

The analysis helps fantasy users and team managers choose pro-players that fit best into their teams. It also helps with the decision of starting lineup based on players’ recent performance, opponent team, and game strategy.

Dataset Roit API will be used for collecting dataset. A Python script will keep tracking game statistics from chosen pro-players periodically. Data will also backup in MySQL database.
https://developer.riotgames.com/api/methods



Algorithms K-means Clustering, Recommendation, Logistic regression

Tools Python, Hadoop, Mahout, MySQL

Current Progress, Schedule and Expected Contributions

Current Progress:

Writing Python script to record statistics of most recent games from chosen pro-players. Most interested in champion selection, game time, KDA, objective control, gold earned, kill participation and so on.

Schedule:

Week_1 Finishing python script, import dataset.

Week_2 Data analyzing using different Algorithms. Compare Fantasy scores before/after applying the method.

Week_3 Apply prediction algorithm to increase win rate.

Week_4 Final Presentation

Expected Contributions:

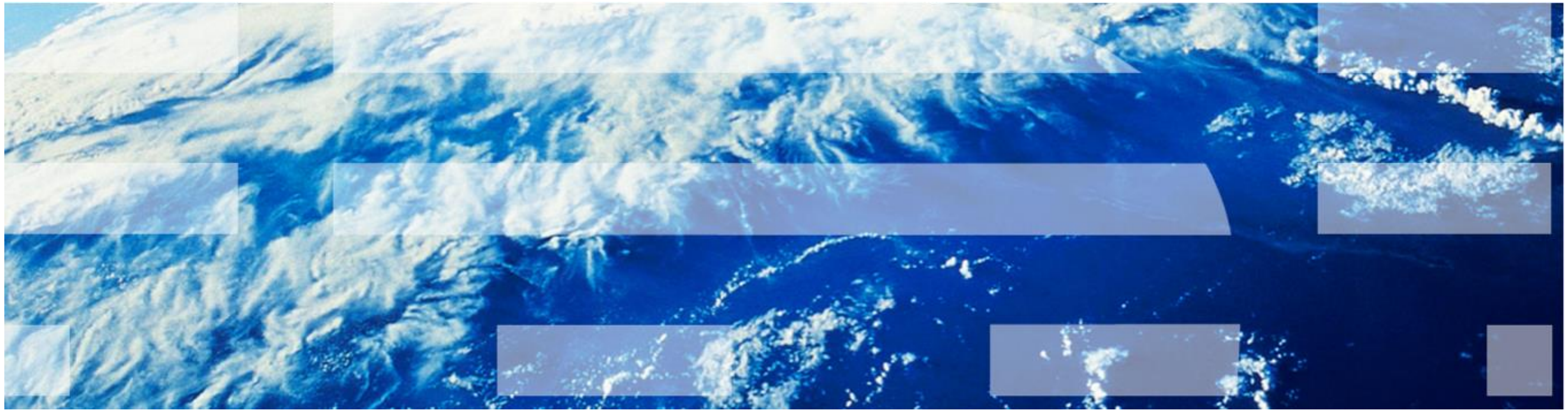
The goal is to achieve 5%-10% increased win rate in League of Legends Fantasy, and also provide a potential solution for better game strategy making in professional matches.

E6893 Big Data Analytics Project Proposal:

Big Data on RSS Feeds

Team Member:

Jing Chen (CVN)



November 19th, 2015

Motivation

RSS (Rich Site Summary) is utilized to publish frequently updated works, such as news/sports/journals. It allows you to stay informed by retrieving the latest content from the sites that you are interested in. That says, if we could apply Big Data Analytics strategies to RSS feed, it will help process RSS content faster and organize the RSS information better.

Dataset, Algorithms and Tools

✦Dataset: will be chosen from various RSS feeds.

✦Language: Python, Java, Hadoop

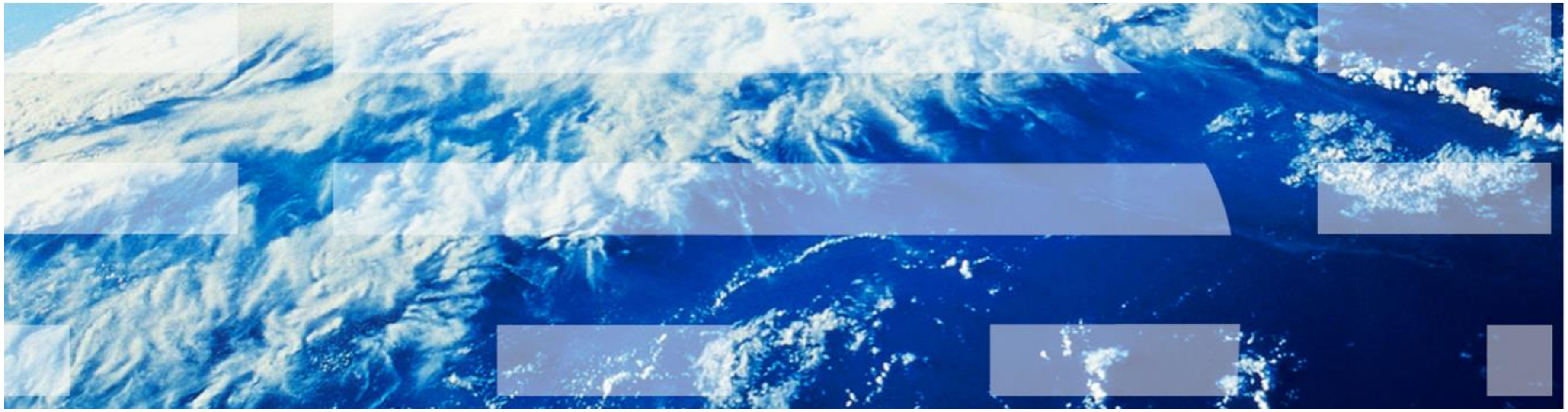
Current Progress, Schedule and Expected Contributions

Just brainstorming ideas and possible algorithms to add to the project, I'm the only person of the team so I will contribute to the whole project.

E6893 Big Data Analytics Project Proposal:

Predicting the United States Presidential election results based on Twitter sentiment

Kirill Alshewski



November 19th, 2015

Motivation

- At least 270 of Electoral College (EC) votes are required to win the election
  - *Each state is assigned a certain number of EC votes*
  - *Candidate with popular vote in a state (can be <50%) receives ALL state's EC votes*
  - *Nation's popular vote has no impact on results: In 2000, Al Gore won the popular vote by more than a half a million votes, but George W. Bush became President*
- Knowing state's public sentiment toward a candidate may help in running a successful campaign
  - *Public sentiment toward a candidate in each state allows campaign manager to maximize the effectiveness of the campaign*
  - *Campaign manager may pinpoint location where additional effort and funding is required to win the state's EC votes*
  - *Resources and funding may be re-allocated from "hopeless" to a "battlefield" states*
- Monetizing the predictions: IEM-Iowa Electronic Markets
  - *Online futures market where contract payoffs are based on real-world events*
  - *2016 US Presidential Election Markets http://tippie.uiowa.edu/iem/markets/pres16.html*
- Applicable to other areas where campaign management is important
  - *Marketing –> analyze target audience*
  - *Retail –> analyze market for current or future product*

Dataset, Algorithms and Tools

- Datasets:
  - *Twitter data via Twitter API*
  - *Training/Test dataset: Twitter data manually classified as positive/negative* *http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip*
- Algorithms**:
  - *Naïve Bayes/Complementary Naïve Bayes*
  - *SVM*
  - *Random Forests*
  - *Gradient-boosted trees*
  - *Monte Carlo for simulation of election result*
- Tools:
  - *Back end: Ubuntu Server 14.04 LTS on Amazon EC2*
  - *Language: Storage: MongoDB*
  - *Analytics**: Hadoop, Mahout, Spark, Python*
  - *GUI/Visualization: Javascript*

*** Currently I'm evaluating various classifiers and tools to produce the most accurate results. In addition, I'm looking at performance of "hybrid" algorithms i.e. Naïve Bayes for vectorization and SVM for classification*

Current Progress, Schedule and Expected Contributions

- Current Progress:
  - *Set up Amazon EC2*
  - *Set up Twitter API*
  - *Developed Monte Carlo model to simulate election result (using Python)*
  - *Started evaluation of various classifiers and analytic tools*
  - *Started working on architecture design*

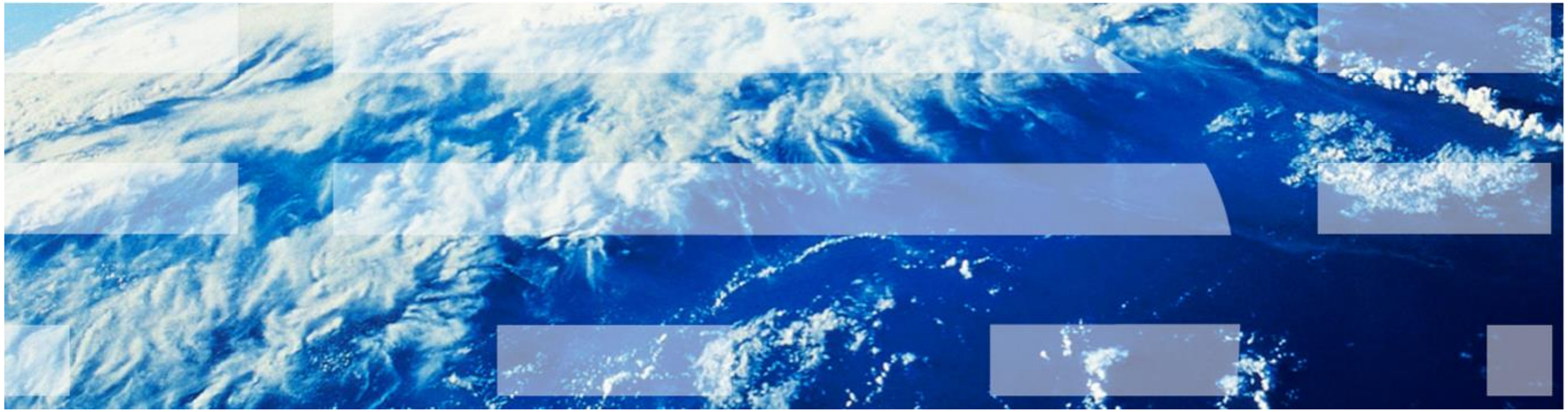
- Schedule:
  - *w/e November 27: Complete evaluation of classifiers and architecture design; start writing code; start collecting data via Twitter API*
  - *w/e December 4: Set up all required tools; implement classification algorithm(s); perform test run of entire application*
  - *w/e December 11: Work on fine-tuning the application*
  - *w/e December 17: Prepare project slides and report*

- Expected Contributions:
  - *All tasks are performed by Kirill Alshewski*

E6893 Big Data Analytics Project Proposal:

Peer and Trend Analysis of US Institutional Investors

Xin Luan Tan (xt2167)



November 19th, 2015

Motivation

In finance, a lot of times the narrative is from the investor's perspective, which is where to invest money or what stocks to buy. Therefore there has been quite a lot of work done to find stock peers to predict or benchmark performance, and to discover potential investment targets.

On the other hand, for a company who is looking for investors to inject capital, they also need a way to find and evaluate potential investors. Most of the peer analysis for an investor is done at a fund level, comparing portfolios of stocks. But not too in depth analysis is done at an institutional level. A company might want to target new investors who are similar to their existing investors. Also, once potential investors are found, the company needs a way to evaluate the list.

The purpose of this project is to find a way to determine peers of an institutional investor, discover trends, and to discover a way to evaluate potential investors.

Dataset, Algorithms and Tools

- Dataset consist of quarterly SEC Form 13F filings, which is required of institutional investment managers with over $100 million in qualifying assets
- I plan on using recommendation techniques in Mahout as a way of finding peers for institutional investors. Each investor has their investment strategies such as allocation across sectors, geography, or exposure to different asset classes. These can be seen as a rating. For example, a technology company would give an investor with 40% allocation in technology stocks and 20% allocation in energy stocks a higher rating as a potential investor than an energy company. The goal is to try to use a combination of different “preferences” to “recommend” similar peers for an investor
- Clustering would also be an interesting way to partition investors depending on the features used, or to find investors that demonstrate a certain feature profile. The features and parameters will need to be determined by playing around with the data.
- Lastly, I want to explore whether classification techniques can help with the evaluation of a list of investors. The plan is to use a company’s investors and their peers (from recommendation above) to train a classifier to determine whether an investor would invest in the company.
- If time allows, will explore some visualization or UI to display the results and functionality better

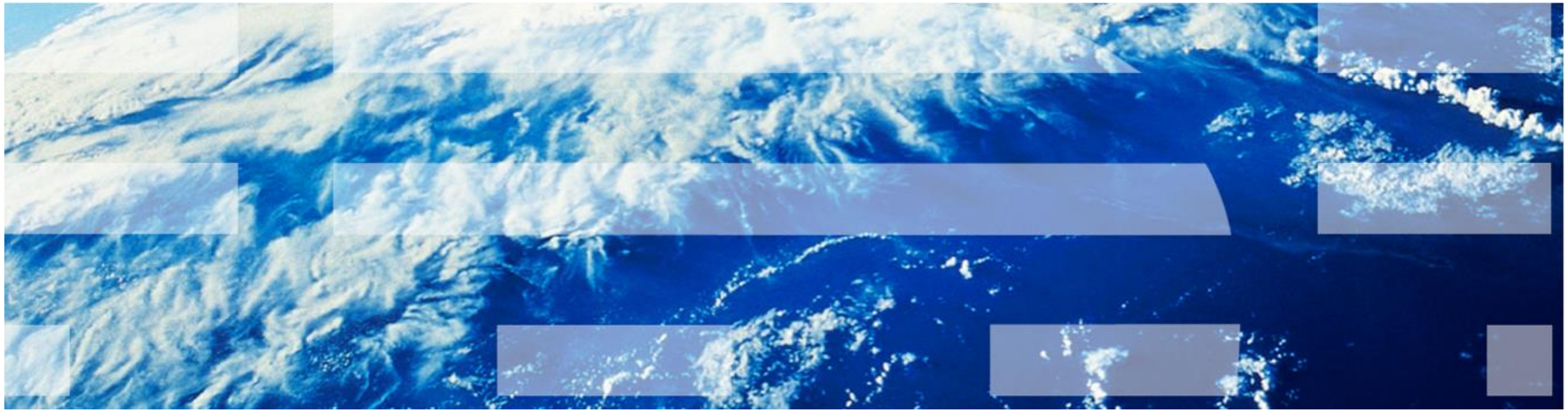
Current Progress, Schedule and Expected Contributions

- Currently in the data collection and formatting phase
- Will allocate a week each for each of the three items in the previous slide
- If successful, will discover a way to find high level peers of an investor based on a combination of features. Currently this is mostly done by matching discrete feature with no ordering in similarity.
- Provide insight into group trends for institutional investors over time
- If successful, provide a novel way to evaluate a list of potential investors

E6893 Big Data Analytics Project Proposal:

Achieving Greater Efficiency Using Machine Classification of Support Tickets

Sam Gabor (sg662) - CVN



November 19th, 2015

Motivation

A typical service organization can receive many free-form support requests emailed to a designated mailbox. Free-form requests received in this manner need to be reviewed, categorized and prioritized by support staff. The process can be very time consuming for a service organization which routinely receives hundreds or thousands of emailed requests per day.

A possible solution to this challenge is to employ machine learning algorithms to automatically classify and prioritize incoming requests. The ultimate goal is to build a streaming interface to examine incoming emails and automatically classify and prioritize in near-time.

Dataset, Algorithms and Tools

The dataset of this application will be many labeled examples of emailed support requests with associated classification and priority extracted from a production system. The data will be partitioned between training and test data.

NOTE: ***Since actual live data may be used to develop and experiment with the system, any data sets submitted with this project will be programmatically anonymized to protect the confidentiality of any participating clients.***

The current goal is to implement this project using Scala and Spark's MLib. However, given MLib's lack of maturity when compared to Mahout, the project may be implemented using the latter's environment. The final decision will be made in early December.

Current Progress, Schedule and Expected Contributions

Current progress has included:

- Implementation of a data preparation program to extract emails from a Microsoft Exchange mailbox.
- Experimentation with Scala as an implementation language
- Experimentation with Spark's MLib support
- Experimentation with third party support for text processing, such as Stanford's NLP libraries

Current schedule calls for:

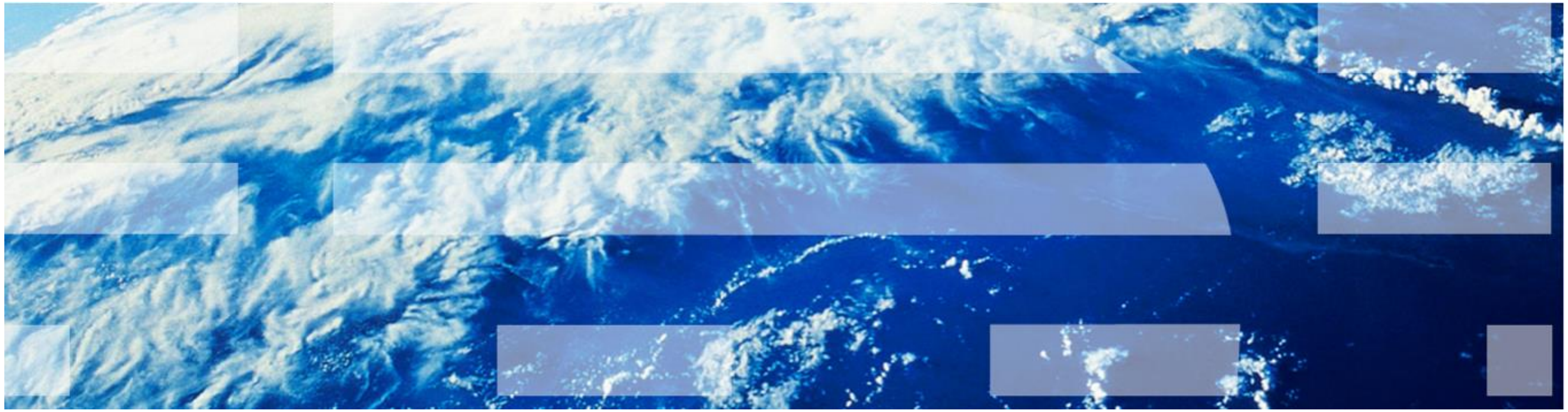
- Finalizing the choice between Spark and Mahout (12/1)
- Implementation (12/2-12/15)
- Final presentation preparation (12/16)

The project only has one participant responsible for all tasks.

E6893 Big Data Analytics Project Proposal:

Identifying Correlated Stock Pairs

Chris Rohlfs



November 19th, 2015

Motivation

The “pair trade” is a common trading strategy in equity markets.

- Pick two similar stocks (e.g., Target and Walmart)

- Predict “mean reversion”: if the one price deviates from the other, it's probably due to temporary mis-pricing –buy one and sell the other short to bet that this mis-pricing will self-correct.

- Traders often pick stock pairs based upon what “seem similar” -- same industry or the stocks have high correlations historically

The innovation of this study is to find a systematic way to select mean reverting pairs.

- Identify from historical data which pairs would have led to the most profitable pairs trading strategy on future dates → predict future performance of that trading strategy.

Dataset, Algorithms and Tools

CRSP data on daily stock prices.

Daily data on 124,750 stock pairs – each possible pair of stocks that are constituents of the S&P 500 Index.

For each one, x-variables are the daily closing prices of the two stocks over the past 90 days.

The variable to predict (y) is the amount of profit that a simple pairs trading strategy on that pair would have generated on the next 90 days.

Algorithms:

Test multiple classification methods including Support Vector Machines and Logistic Regression

- What method identifies the most profitable pairs? Would simpler approaches (picking same industry or correlated pairs) be as effective?

- Kernel modification of predictors to allow for potential nonlinearities and interaction effects.

Tools:

Programming will use a mix of C++, Python, and R.

Current Progress, Schedule and Expected Contributions

Progress so far:

- Have pairs trading strategy coded up
- Have dataset of S&P 500 constituents identified, downloaded and cleaned
- Some coding of predicting algorithms complete

Schedule:

- Continue to code, write, and perform data analysis over the next month to produce estimates of the effectiveness of different classification strategies.

Expected contributions:

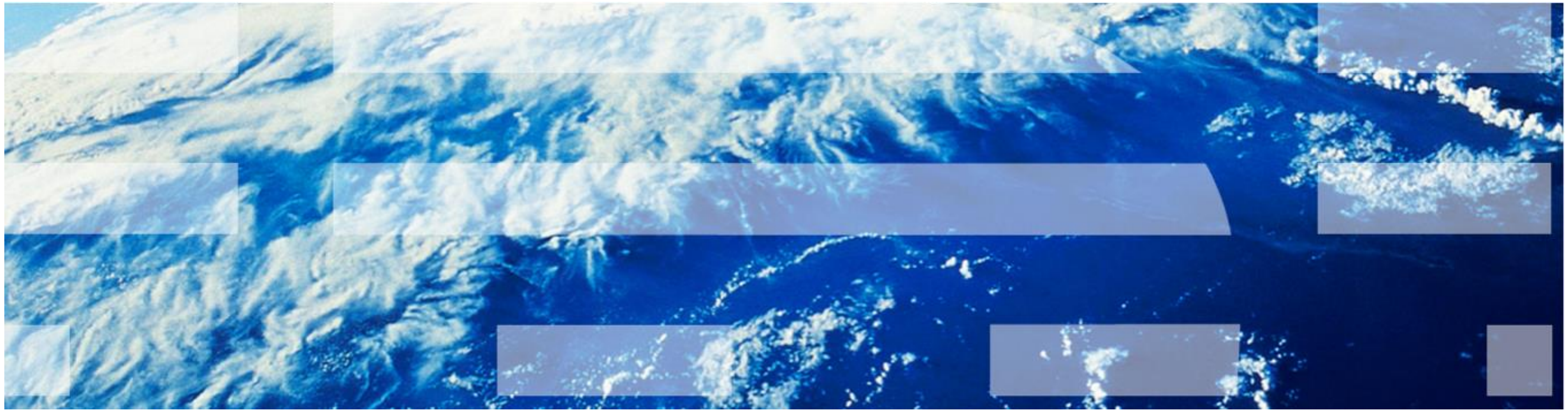
- Hope to answer the questions:
  - Which are the best pairs?
  - Are those pairs stable over time?
  - Can historical price movements accurately forecast the profitability of a pairs trading strategy?
  - Is there a simple pattern (e.g., correlated stocks, those from certain industries or geographic regions) that can predict which pairs are best for mean reversion trading?

E6893 Big Data Analytics Project Proposal:

Live Portfolio

Paresh Thatte – pat70

Manjiri Phadke – mp3212



November 19th, 2015

Motivation

Describe the motivation of your project

Rebalancing of a large set of portfolios is a periodic activity that large teams and in-house products are constructed around. Running these in batches is typically how this is done, and in some cases unavoidable.

Doing these in real-time and using the same operations to run scenarios as for batch processing allows for more nimble strategies and more confidence in performing actions.

Any change in position requires the portfolio to be recalculated. As the size of the portfolio and the number of portfolios affected grows, the number of operations that need to be performed keeps growing.

Using open source technologies that scale out and support automatic repartitioning allow implementations to focus on the task at hand – run the formulas.

Dataset, Algorithms and Tools

- ✦Spark (Streaming)
- ✦Messaging (Kafka)
- ✦RESTful backend (Vertx - Spring/Scala,Java/JavaScript)
- ✦MySQL in memory

- ✦Sample portfolios (mix of asset classes)

- ✦Streaming market data
- ✦Streaming trade confirmations
- ✦Research mode

- ✦Recalculate position based on trade
- ✦Recalculate value based on market data
- ✦Allow scenarios to be run in research mode

Current Progress, Schedule and Expected Contributions