Text Analysis, Social Networks and Crowdsourcing

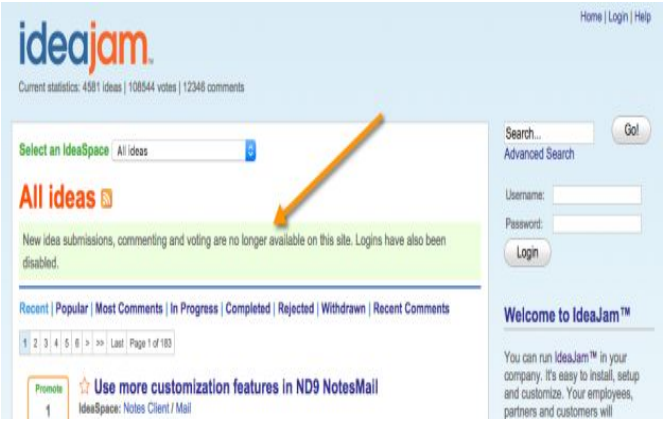
Students:
Dan Peng, MS CS
Gitanjali Kanakaraj, MS CS
Hanieh Arabzadehghahyazi, MS CS
Naiya Shah, MS CS
Nana Andriana, MS CS

Participating Faculty:
Daniel Edmund O'Leary
Keith Burghardt

Introduction

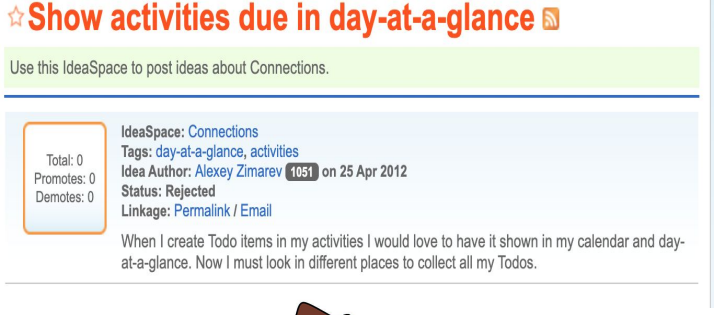
The main aim of project was to explore the dynamics within a crowdsourcing platform - how the activities on any post(like upvote,downvote etc) related to the language people used to support or oppose an idea and find out if we could trust the ‘Wisdom of the Crowds’.

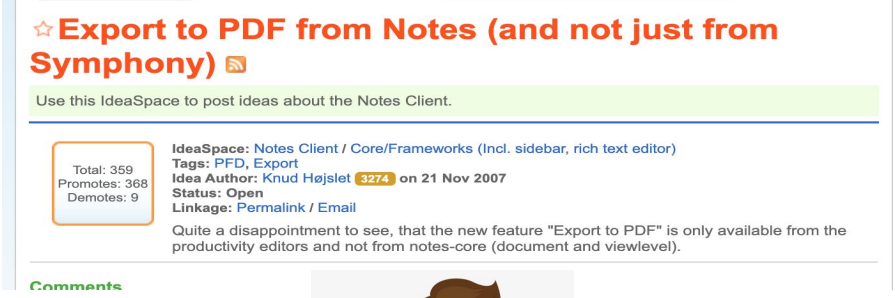
We used IdeaJam which was a crowdsourcing platform created by IBM to get ideas from their participants, who were working with many of their products, like LotusNotes. People would opine on the various ideas put forward by others in form of comments as well as upvotes/downvotes.



Motivation

- Ever wondered why a certain post gets more upvotes than other posts ?
- Ever wondered is it just about the content or are there hidden dynamics we are missing?
- Does the sentiment of comments, network of author, or the influence of commentator in any way affect the activities happening on post ?



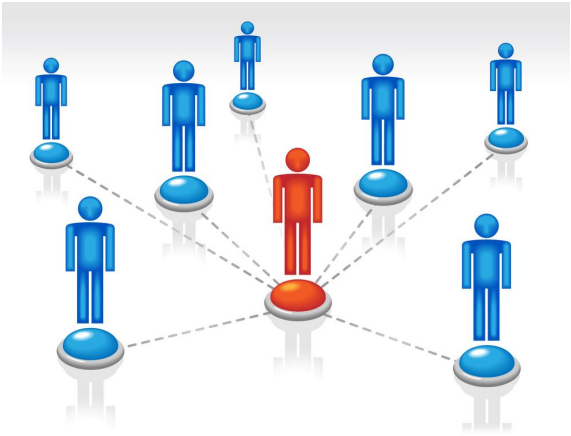


Motivation

Our main motive was to answer these questions; to understand the dynamics within a crowdsourcing platform - how people used language to either support or oppose ideas generated.

We also wanted to find out if there was any connection between the people making these opinions. For example, if there was a leader - follower relationship within groups of people, or if people formed opinions on their own.

Influence Of Language



Problems

During this semester we were able to tackle these questions

1. How to scrape a large amount of data from a website
2. Data Cleaning/Preprocessing - How to structure the data so that it can be easily used for analytical purposes and help us answer the questions more efficiently.
3. Were there any relations between the sentiment of the comments and the number of votes for an idea?

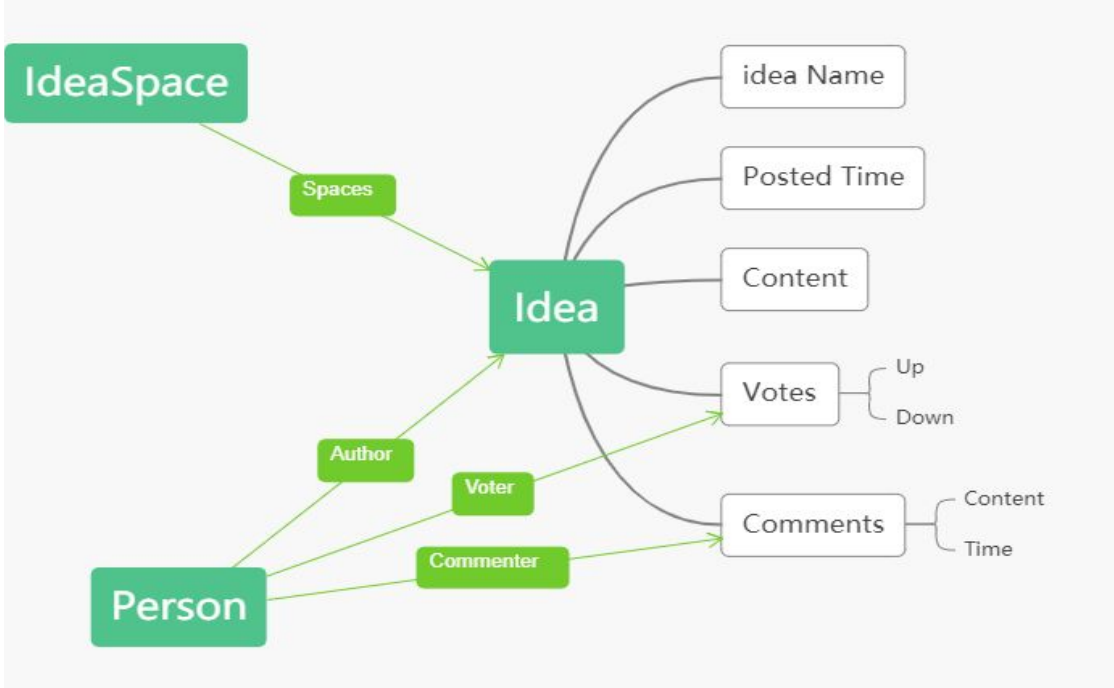
Data and resources available

Main Data Site ： IDEAJAM (http://ideajam.net/)

The data features are as follows : .

• Author • Idea name • Idea content • Tags of the idea • Comments	• Time of the idea posted • The number of upvotes and downvotes • People who voted • Commentators

Example of Data Scraped



For sentiment analysis, we primarily used the Comments and the Votes features.

Tools

Methods used to solve the problem

- Web-Scraping : Python and Selenium Web crawler
- Sentiment Analysis : NLTK, VADER
- Data Visualization : Pandas
- DataBase : MongoDB

Tools

MongoDB

- Used to store the data retrieved from Scrapy and we can output data in JSON format

Sentiment Analysis: VADER

- VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool.
- Used to determine if the comments are positive or negative and identifying commenters sentiment or emotion.

Conclusion

During the course of this project we were able to successfully

1. Scrape data from the website
2. Clean the data
3. Perform sentiment analysis on the data
4. Compare it to the number of votes it got in total to see of there was a correlation

Results

Many comments that were downvoted had neutral comments like “test comment”, or positive comments like “I support this idea! Let me know how I can help.”

The amount of actual negative comments in the downvoted ideas were actually very few.

Overall accuracy = 59.33% (Correctly matched/Total number of comments)

Accuracy of positive = 65.24% (Number of upvoted texts that matched the positive sentiments/Total Number of Comments that were downvoted)

Accuracy of negative = 2.69% (Number of downvoted texts that matched the negative sentiments/Total Number of Comments that were downvoted)

RESULTS

Total Number of Ideas having positive sentiment comments = 5905

Total Number of Ideas having More Upvotes than Downvotes = 5205

Total Number of Ideas having negative sentiment comments = 1574

Total Number of Ideas having more Upvotes than Downvotes = 1361

Future work

- Network Analysis
- Generalise the platforms. See if we find the same conclusions across variety of other crowdsourcing platform like Stack Overflow etc.
- Find if there was any impact of the actual content of the idea on sentiment as well.

Thank You :)