Data mining and visualization — social media orientated

Project Report: Proposal - January 31, 2020

Team members :

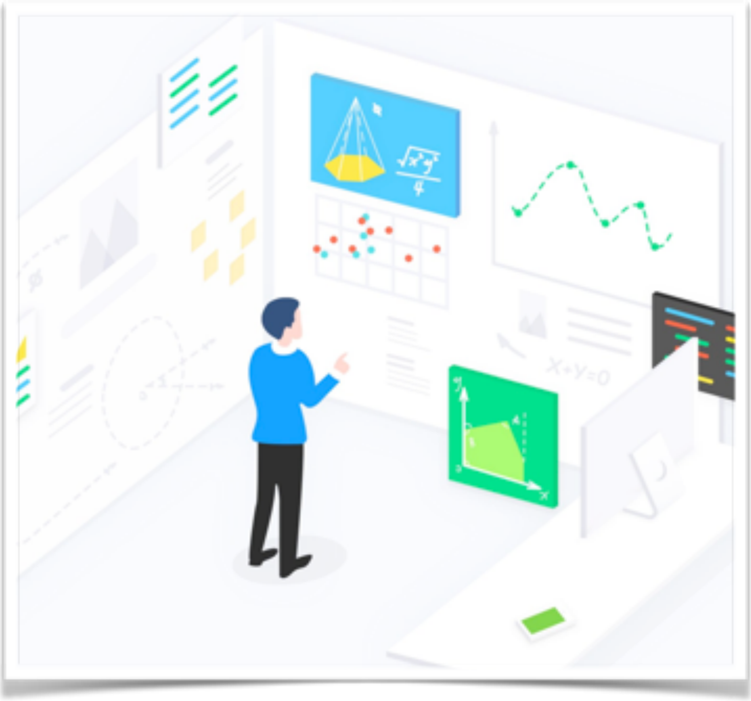
Jiqing Zhu,	105092951
Jiaxiu Li,	110008532
Weiwei Cao,	105171487
Hang Zhu,	105170587

1.Overview

Social media has become the major platform for people in modern society to express their opinions and comments. Such a large amount of information implies a very high commercial value. However, due to the dispersal and chaos of those information, manual processing is not possible. Thus, a product that can explore the text information from social media network, visualizing the analyzed data based on a variety of diagrams, is conducive for people to understand the trend of public opinion on the Internet.

Due to the volume of data produced on social media is extremely huge, we expect to implement the system on the basis of easily fetching popular topics and handily extracting subtle information to help people knowing what they are talking about on social media. We propose to achieve the project from the following aspects:

- Ability to process real-time data and update the model periodically.
- Topics are integrated with the past and current data in order to observe the trend of the topic.
- The system capability can process volume data under the stable situation to guarantee efficiency and sensitivity of emerging topics.
- Visualization of statistical results.



How Data Visualization Works

(https://twitter.com/ruanyf/status/1050263106989912065/photo/1)

2.Project Goals

Get insights from social media, visualize the audience's emotion. as follows:

1. Functional goals

- Easy to track, automated process of analyzing text data and classifying opinions;
- AI powered algorithms for improving system accuracy;
- Geospatial location data displayed on map;
- Data mining visualization presented by diverse diagrams in order to emphasize on tendency of entities differentiation.

2. Strategic goals

- Benchmark against competitors, finding keywords of your competitors' audiences to know their segments;
- Analyze customer's feedback, tracking your product keywords on websites to understand customer's feelings and what they talk about;
- Evaluate campaign impact, typing the keywords of the latest event to review how your target client's response to your promotion.

3. Technological goals

- HTML5 Presentation, capability of displaying content across all platforms, providing consistency and improved accessibility;
- Natural language processing (NLP), using NLP techniques such as such as part-of-speech tagging (POS tagging), stemming, lemmatization and semantic reasoning implemented by python to handle textual data;
- Machine Learning based algorithms for improving the accuracy, support vector machine (SVM) is used to filter the text, then Naive Bayes model and AdaBoost algorithm are applied to analyze the text.
- Several models based data mining for analyzing valuable units, including public opinion frequency, attitudes trend on timeline, regionalism sentiment analysis, etc.

3.Project Team

Name	Responsibility	Availability	Comment
Jiqing Zhu	Developer & Tester, Presenter	30hrs/week	
Jiaxiu Li	Developer & Tester, Designer	30hrs/week	
Weiwei Cao	Developer & Tester, System Analyst	30hrs/week	
Hang Zhu	Developer & Tester, Document Controller	30hrs/week	

4.Schedule and Milestones

Milestones	Description	Milestone Criteria	Planned Date
M0	Project Confirmation	Topic Clarification	2020-01-15
M1	Kick-off Meeting	Draft version of Proposal	2020-01-25
M2	Requirement Analysis and Data Model Design	Data Model design and table structure confirmation	2020-02-03 ~ 2020-02-09
M3	Development	Product with core function	2020-02-10 ~ 2020-02-28
M4	Testing and Bug fix	Test case completion with acceptable pass rate, No catastrophic bugs exist	2020-03-02 ~ 2020-03-08
M5	Documentation	Lists of delivery documents, e.g. product description, user guide, etc.	2020-03-09 ~ 2020-03-15
M6	Presentation Preparation	PowerPoint and presentation practices	2020-03-16 ~ 2020-03-27

5. Communication and Reporting

It is generally recognized that instant communication can be ameliorative and proactive for scrum programming during every iteration. In other words, the effective reporting between team members will improve and promote performance in several aspects such as conquer technical barriers or modify the specific requirements.

First of all, in the initial plan, a face-to-face regularly meeting will be held weekly to avoid ineffective methods as well as misconducting comprehension. Also, members can present whether they obey the schedule of the project timeline during this meeting; moreover, obstacles met through last week will be discussed to seek a better solution. After that, a meeting minutes and reports will be generated by participants.

In addition, it is essential for daily contacts among members. It means members will keep frequent associations whenever they want via different channels such as email, facebook or IM, which can provide a more reliable and conducive circumstance to keep high-quality efficiency. Exchanging ideas and inspiration will also be recorded as the table format in case of repetitive problems.

6. Delivery Plan

No.	Delivery Items	Receivers	Planned Delivery Date
1	Proposal	Pooya	2020-01-31
2	Data Model	Pooya, Team Members	2020-02-09
3	First progress report	Pooya	2020-02-14
4	Second progress reprot	Pooya	2020-03-13
5	Product	Pooya	2020-03-27
6	Documentation	Pooya	2020-03-30
7	Final report	Pooya	2020-04-03