Project Progress Report

Kaixin Wang / Xingchen Wu – CS410 Final Project

**Work in progress**

Those tasks divided below are our team bullet point and the primary focus at this moment is the second point: Sentiment and trending analysis on the provided data set. We are working with relevant Python libraries such as numpy and metapy to assist with the word-by-word processing and POS tagging of the sentences. Since this portion of the task involves the core concept of what we learned from this course, this is determined to have the biggest bottle neck and will take longer time relative to other tasks.

* Web Crawling to collect large sets of review data for specific product (10 hours)
* Sentiment and trending analysis on provided data set (20 hours)
* UI enhancement and project integration (10 hours)

To generate the training data for the project, we have self-generated some hypothetical comments/reviews of certain products on excel worksheet and feed them into the program as we are developing it for the purpose of development, testing, and debugging. As we progress through the sentiment analysis portion of the task, we will move on to use/develop certain tools that can help us crawl and collect reviews/comments automatically through a given URL.

**What has been done**

The initial UI of the program is designed and developed. The UI is at the initial stage with some user friend interface to guide the user to search and extract reviews. However, the actual integration will be done after the sentiment analysis is complete.

**Challenges**

One of the challenges we face during the development of sentiments analysis is the program processing speed. With manually created training data, the analysis speed varies between half a second to 2 seconds. After the program fetches real work data from various web site sources, one of the challenges is to balance the data amount fetched and processing time. We decide to use the brute force way to get the programing working first, and then depending on the processing time we will target and tune it afterwards.

The other challenge is to develop crawlers to mine targeted sections of the various URLs. Different websites have different structures, we are still deciding whether to work on a generic crawler that can adjust and adapt to extract useful information on various websites or just to design one that target certain ones or based on certain keywords. The decision will be made after the program sentiment analysis is complete to see how much effort and how much time we have left.