CS410 Progress Report

Dancing Text: Yuxiang Liu (yuxiang@illinois.edu, leader), Hongfei Ma (hongfei7@illinois.edu)

1. Progress

For the task of identifying in-demand skills, we have written an auto-crawler to crawl tweets related to a given subject "Computer Science" from a social media, Twitter, and saved attributes of tweets to a file. Due to the rules of Twitter, there is a rate limit with the crawling so we cannot get as many data as we want. However, we can obtain the most-recent tweets related to the CS subject, which helps to discover emerging keywords or topics. We have also processed the crawled tweets, and analyzed the hashtags, which represent the themes of corresponding tweets, to preliminarily extract top-rank topics of these tweets.

For the task of displaying relative documents of each topic, we have finished crawling PDF slides of several courses in UIUC. However, due to the difference of course websites, crawlers are running independently in different python scripts. We have also investigated some common ranking algorithms, including BM25, Jaccard index, cosine similarity and some provided algorithms in MeTA Toolkit.

2. Challenges

Since there are many tweets without hashtags, we cannot extract topics from such tweets, but it helps to utilize hashtags in extracting topics. Hence, instead of designing a topic discovery algorithm for general texts, we need to propose an algorithm which can extract topics from both tweets with and without hashtags, and combine these topics to get the final top-rank topics among these tweets.

Another challenge is how to demonstrate that the extracted topics are truly popular in these tweets, or how to evaluate the performance of our algorithm. Even if we have crawled tweets from Twitter, we do not know what topics are popular in these tweets.

Evaluation of different ranking algorithms seems time-consuming because it is hard to get all-round training dataset. In addition, the form of displaying our ranking results is still undecided.

3. Remaining work

To identify in-demand skills, the remaining work is to propose a topic discovery algorithm specifically for tweets, and design an evaluation method to demonstrate the performance of this algorithm.

To display relative documents of each topic, the remaining work is to design a workflow that can automatically crawl documents from various course websites, and evaluate the existing ranking algorithms' performance in our dataset.