A report of project,part1

Member: Mingxin Lu Steven Hu 2/22/18

1. description of approach to construct feature vectors

First, we use python as programming language to process all data.

We import a module *sgmllib* to parse those files with sgm format. It could handle each article separately, construct a dictionary to record different tags and their contents. There are few important tags: *topic, data.* (other meaningful tags: place, orgs, companies, date etc. We probably will use them in the future.) They will becoming the most target of our first project, especially *data.*

Then we should remove all preps and words in common usage. We ().

Then we should remove all words containing digit.

We also need to clean punctuation for each word.

To calculate word’s frequency, we should transform all words to lower case. We construct dictionary for each article and combine them to an array. We do the same thing to the topics. Therefore, we get our feature vector. The index represent articles’ order.

1. Difficulty