Project Report -- part1

Member: Mingxin Lu Diwen 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.

Then we should remove all words containing digit.

We also need to clean punctuation for each word.

Finally, we need to get rid of all the common words like 'are', 'is', 'and' etc.

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 represents articles' order.

At the end, we collect all the data and output to the file called output.txt. We extract the doc_id, title, and body information as the output.

2. Difficulty

We used some fancy package to preprocess the data, but it would not work in stdlinux system. So we have to find some basic package to implement the data. It takes us time to look up many python package and materials about how to formatting the raw data.

We need to diminish the some data which are unnecessary common words in the future, it has difficulty to choose the important data since there are many features.

3.

We get rid of all the tags including<title><Date> etc.

We get rid of all the numbers in the article

We remove all the attributes in the tag

We get rid of some unnecessary tags like <People><Exchanges><Companies><unknown>

In the body part, we make the whole article to the lowercase.

We remove all the punctuation in the article.

We remove many common words including “are”, “is”, “and” etc.

In this project, we learn how to fluently use basic grammar of Python. We improve our understanding of basic database including map, array, list, iterator, and set. We understand basic function of using system package to read and write. And the most important thing is that we possess and review knowledge of pre-processing.

To improve our work, we would use more tag to analyze in the future to get better classification and algorithm. We would do more work on transforming our knowledge to practical usage.