Intermediate Report

Yongjian Mu / u1010337 Shan Wei / u0974032

Since our goal is to compare the similarity of two news websites(new york times and time), so we have extracted the titles and key words from datasets we collected from the main page of each website.

Firstly, we used the pure K-gram algorithm and let k = 2 words. In this case, we got the result that the Jaccard similarity for the two websites is about 0.64 daily and 0.68 for ten days. This result was too high since there were too many meaningless combination like “of the”, “is of”, etc.

Secondly, to reduce this meaningless combination, we used the pure K-gram algorithm and let k = 1 word. In this way, we got the result that the Jaccard similarity for the two websites is about 0.08 daily and 0.2 for ten days. This seemed much better than k = 2 words, but it still had some problems.

1. We still calculated the similarity including some meaningless key words like “is”, “of”, “the”, etc.
2. There were also some meaningless words (not key words) like “day”, “month” calculated.

In the first approaches, we use the k-gram of jaccard distance of 1 word among several same day separately and several days together. But the result is around 0.1, which is not good. I also tried k=2 characters k-gram on jaccard distance. Result is up to 0.68, but we can’t figure out any intuitive information from the intersection set. Thus I think k-gram for character does not work well. From another perspective, 1 word k-gram should be good, if I properly use well-processed data. Then I think there are some further optimization works we need to do before calculate the similarity.

Then I am trying to build up a stop-word list by applying the TF-IDF to text information in order to filter the some high frequency but meaningless works like “the”, “we”, ”year”, ”only”, “did”.

There are some features we could assume they are meaningless:

1. From a large dataset (such as two weeks or more data), those words that only appears once. It would probable be a really special word or it might be a not that important thing like ”date”, ”Monday”, “hair”.
2. From a comparably large data, those words show a really high frequency. It would more likely to be a “common”, necessary but do not have meaningful meaning like “a”, ”not”, “like”, “but”.
3. There are also some exception signal that I did not notice and remove from the dataset like” ”the ”……

After TF-IDF preprocess, the similarity between two website in the same day is going up form 0.1 to 0.16-0.18.

Finally, by using TF-IDF we removed the very high percentage key words like “of”, “the” and manually removed the obviously meaningless words like “day”, “month”, etc. Detailed introduction of TF-IDF can be found in Shan Wei’s report. Therefore, we got the result that the Jaccard similarity is 0.16~0.18 daily and 0.4 for ten days, which seems more reasonable.

In conclusion, to make this result more accuracy, we can improve the TF-IDF algorithm like not to distinguish the singular and plural of the words, and based on the more and more data we collected, remove more meaningless key words. To demonstrate these results, we can use Python or JavaScript to draw some charts or graphs.