Intermediate Report

Yongjian Mu / u1010337

Continue the previous work, we grabbed the information (both original html files and filtered data) from New York Times and Time websites.

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.

Finally, we used TF-IDF to avoid this miss calculation. 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. This seemed 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.