**COMP 6791 Project 3 Report**

**Name:**

**Mo Chen**

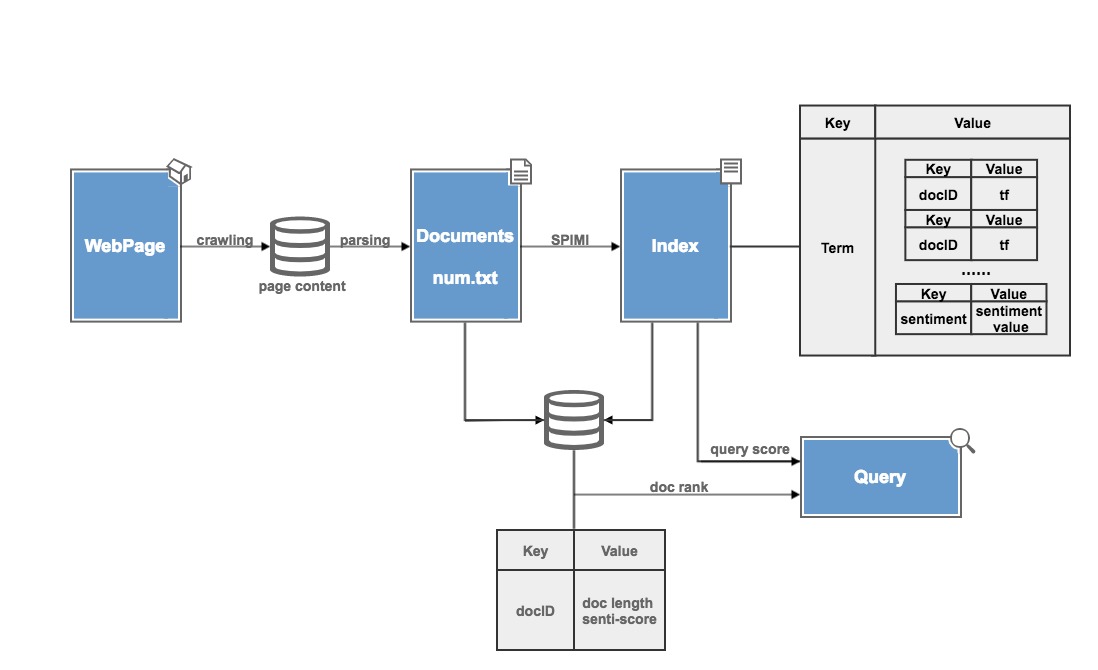
**Tianen Chen**

**Qi Xia**

**Introduction**

For this project, to crawl the web and construct an index for the web documents, we implement the program with Python, using the request and beautiful soup libraries to scrape and analyze the web content.

**Implementation**

****

1. Crawling and Parsing

Through using the request library, we can trace the web links and scrape the content of the website. Then, using the beautiful soup to parse the html documents, we take out the body of the content, and extract some useful information, apply normalization to these information, we can get the documents which meet the format requirements.

By the way, we only keep the English words while delete all other languages like Korean and special symbol like emoji.

1. Constructing Index

We use the SPIMI algorithm to construct the inverted index like what we did in project1 & project 2. The difference is that, in the last two projects, the value of the index dictionary is the docID-tf pairs (also saved as a dictionary, let’s say inner dictionary), but now, we add a “sentiment”-value pair into the inner dictionary, so, we can keep the sentiment value for every term, then calculate the document sentiment value and the query sentiment value. The sentiment value for each term comes from the sentiment dictionary aFinn, if term in aFinn has a score, then give it this score, or make the score as 0, which means neutral sentiment.

There is two additional procession we have done for this project:

The first one is abandoning the stemming, for stemming reduces the inflected or derived words to their root form, it will change some words’ sentiment value, so we haven’t done stemming this time.

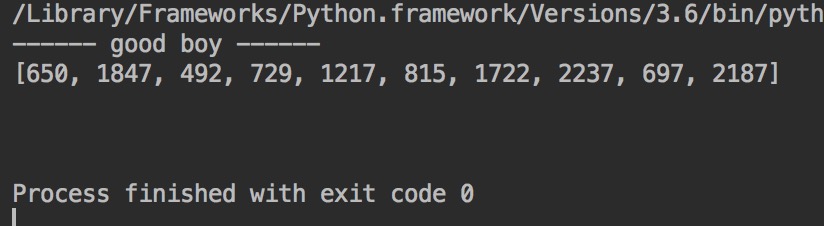
The second one is deleting all the terms those are not English, without this step, we will have many terms of other language even emoji in our index. This will produce some interferences to our result, and this is also a point which should be considered more and tried to fix out in the future.

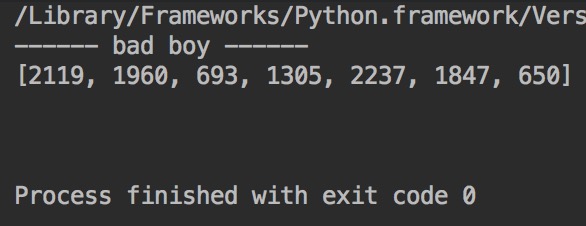
1. Score and ranking

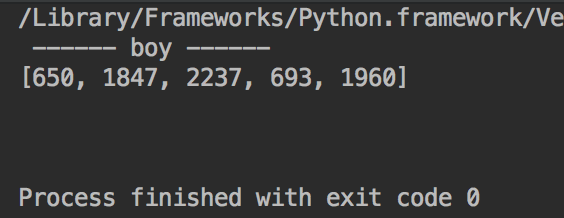
To calculate the sentiment score for a document, we also have used the formula of RSVd which we done in the project 2. The reason is we cannot only get the average sentiment score of all the terms in the document, but also should consider the relevance of the document to the query, to say, either the document is only relevant to the query or only has the same sentiment tendency is not entirely correct, thus, we do the multiply, the formula is:

For ranking, if the query is positive or neutral, rank the documents according to the decrease score, or, for the negative rank the documents according to the increase score. Finally, we give the top-10 rank result.

**Result Analysis**







|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Doc-ID | 650 | 1847 | 2237 | 693 | 1960 | 1305 | 2119 |
| # of “good” | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| # of “bad” | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| # of “boy” | 6 | 6 | 1 | 1 | 1 | 0 | 0 |

To get the result, we have done three queries for those have different sentiment tendency. Through compare the results, we can see those documents which rank higher in the result of query “good boy” have a lower rank in the result of query “bad boy”. The reason is they have a higher positive sentiment value, it’s obviously they will get a lower negative sentiment value.

We also have tried a neutral query “boy”, some of the documents appear in both “good boy” result and “bad boy” result have occurred here. Through checking the documents, they all have a high relevance to the word “boy”.

A high-light point here is in the result of query “bad boy”, for document #2119 has a higher rank than document #1305 though the latter has more “bad” than the former, because, after reviewing the documents, we found that document #1305 has a longer length than #2119, in our formula for calculating doc-score, the document length difference will bring some affects.

**Q&A**

1. What was the hardest step?

To calculate the sentiment score of document, we have thought about many approaches, such as, give the RSVd and sentiment score weights, then we can try different weights to give a better result. The problem is, sometimes the relevance score is too high, maybe it will reverse the sentiment tendency of the document. For example:

Query: I want to buy a car.

Document: I don’t want to buy a car.

There is only “don’t” is the difference, but it’s the key point of the sentiment guide. We cannot set the weight for this exception, but we also cannot ignore these special cases.

So, at the end, we do the multiply, to make the relevance and sentiment score in a harmonious.

1. How big is the index?

There are 21,877 unique terms in the index. Each term has its own posting list with term frequency and sentiment value. The format is like:

Term: {document\_id1: tf1, document\_id2: tf2, …, ‘sentiment’: sentiment\_value}

1. How did you define what constitutes a document in your index?

For each document, we have a document length and a sentiment score, document length is used to calculate both the relevance and sentiment value.

1. What observations did you make during your experiments?

We have found many useless web page contents, although we can filter most of them through the program, it still need some artificial recognize to delete those non-relevance content, such as, “page not found”.

1. What did you learn from your experience?

For this project, we exactly feel the complex of the natural language in real life, there is very little possibility that the task which you dealt with is only one language, clear formed, and always useful, a lot of effort should be made to analyze the documents. Besides, the standard how to construct the index and give the suitable rank score to the documents is also very important. Human language is complex, even the occurrence of a “very” in two documents different will bring some subtle gaps. We have considered many examples to evaluate our sentiment formula, and also have tried kinds of query on Google to see how results different, it’s really a difficult task to understand the real meaning of the words in our real life.