10/16/24, 1:32 PM Project 1

WSM Project 1: Ranking by Vector Space Models

1. [40 points] Vector Space Model with Different Weighting Schemes & Similarity Metrics

The example codes given in Week 3 demonstrate how an IR system works via Vector Space Model. Below are some steps in the codes:

- 1. Stemming & Removing Stop Words (English Stop Words); & Indexing
- 2. Transfer Queries into a Vector
- 3. Transfer Documents into Vectors
- 4. Calculate the Similarity between the Query Vector and the Document Vectors
- Rank the Documents according to the Similarity scores

Now you are asked to develop a retrieval program that is able to retrieve the relevant news to the given query from a set of 7,875 English News collected from reuters.com according to different weighting schemes and similarity metrics. In the given dataset, each file is named by its News ID and contains the corresponding news title and content, as shown in below:

cat EnglishNews/News5005.txt

Foreign pair bring energy and fresh ideas to the table in Japan's tradit sake brewing industry as sales staff, they not only discovered the intri in one of the country's most traditional industries.

There are the four combinations you're asked to implement. For each combination, please retrieve the top 10 results and scores.

- [10/40 points] TF Weighting (Raw TF in course PPT) + Cosine Similarity
- [10/40 points] TF-IDF Weighting (Raw TF in course PPT) + Cosine Similarity
- [10/40 points] TF Weighting (Raw TF in course PPT) + Euclidean Distance
 [10/40 points] TF-IDF Weighting (Raw TF in course PPT) + Euclidean Distance

Here is an example result for the query "Typhoon Taiwan war":

TF-IDF Cosine	
NewsID	Score
News12780.txt	0.5671282
News10184.txt	0.3621771
News12428.txt	0.3444484
News13724.txt	0.2968883
News10152.txt	0.2700445
News10355.txt	0.1942214
News12944.txt	0.1547813
News10460.txt	0.1401013
News6715.txt	0.1318642
News6825.txt	0.1242267

TF-IDF Euclidean	
NewsID	Score
News561.txt	7.8418809
News13136.txt	13.7768485
News5680.txt	14.8391712
News12524.txt	15.1864187
News9700.txt	15.3223668
News13100.txt	15.5027014
News5668.txt	15.6400241
News13924.txt	15.6856955
News6486.txt	15.7987287
News11212.txt	15.9287906

2. [10 points] Relevance Feedback

Relevance Feedback is an IR technique for improving retrieved results. The simplest approach is Pseudo Feedback, the idea of which is to feed the results retrieved by the given query, and then to use the content of the fed results as supplement queries to re-score the documents.

In this work, you're asked to use the Nouns and the Verbs within the first document of the above **Method 1** (e.g. TF-IDF Weighting + Cosine Similarity) for Pseudo Feedback. The new query term weighting scheme is [1 * original query + 0.5 * feedback query]. Please try to use the new query to re-rank the documents.

For instance, suppose the index vector is ["network", "computer", "share", "ask", "soccer", "song"], the query is "network", and the content of the feedback document is:

Jimmy shares songs via the computer network.

Then we will get a new query vector like this:

```
1 * [1, 0, 0, 0, 0, 0] + 0.5 * [1, 1, 1, 0, 0, 1] = [1.5, 0.5, 0.5, 0, 0, 0.5]
```

In this work, you may need to use the Python NLTK package. For more details, please refer to this link.

3. [20 points] Vector Space Model with Different Scheme & Similarity Metrics in Chinese and English

In this part, you are asked to retrieve the relevant news to the query from a set of 2,589 News collected from according to different weighting schemes (TF and TF-IDF) and cosine similarity metric.

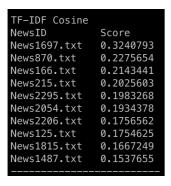
There are the two combinations you're asked to implement. For each combination, please retrieve the top 10 results and scores

• [10/20 points] TF Weighting (Raw TF in course PPT) + Cosine Similarity

10/16/24, 1:32 PM Project 1

∘ [10/20 points] TF-IDF Weighting (Raw TF in course PPT) + Cosine Similarity

Here is the example result of the guery "資安 遊戲":



Hint: You may use Jieba or CKIP to split the Chinese word segments.

4. [30 points] Evaluation IR system

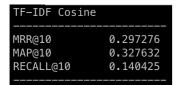
In this part, we will focus on another <u>smaller dataset</u>, which have 1460 documents, 76 queries and their labelled relevant documents.

You need to implement the following metrics on this dataset:

- [10 points] MRR@10
- [10 points] MAP@10
- [10 points] Recall@10

by using vector space model and trying some NLP technique e.g. stemming, remove stop word ...

Here is the example result:



Submission Details

- Due: 13:10, Wednesday, 23 October 2024
- · What to turn in:

Electrical submission: compress all the necessary fiels and data into a zip file, and submit it via the WM5 website. Please DO comment and format your codes to avoid any penalty imposed by the grader.

• Late policy:

In general, late homework may receive fewer points than incomplete homework. The penalty for late homework is 20 points per day.