

CMSC 476/676 Information Retrieval

Homework 2 REPORT

JA52979

Steps to run the program:

In the IDE terminal, You can run the program described below:
python3 calcweights.py "input_direct_path" "output_direct_path".

Here, 'calcweights.py' is the main python program. You must have all libraries installed and your current working directory containing python files in order to run the method described above. The input directory is the output files(tokenized files) of phase 1.

Implementation:

Improvement of preprocessing:

For this phase of the project, to improve the preprocessing, I made a list to store all the stopwords given. As mentioned, making use of the list, I removed stopwords, then removed the words that occur only once in the entire corpus, and also the words of length 1. Thus, the resulting number of tokens for all the documents have been slightly decreased, making my term weights more effective.

Term Weighting:

For the TF-IDF, I calculated the product of term frequency(tf) and inverse document frequency(idf), later the TF-IDF score for each term is normalized by dividing each TF-IDF score by the square root of the sum of the squares of all TF-IDF scores in the document.

$$\text{TF-IDF} = (\text{TF} * \text{IDF})$$

For calculating the term frequency, i.e., the number of documents that a particular token is appearing in. I calculated using the below formula,

$$\text{TF} = (\text{Freq. of token within a doc}) / (\text{Total number of tokens in the same document})$$

Similarly, for calculating the inverse document frequency, using the below formula, calculations are made.

$$\text{IDF} = \log(\text{Total number of documents} / \text{Number of documents containing the term})$$

I have implemented the above logic using the 'calculate_tf_idf' function. By maintaining a two dimensional dictionary, where I stored all the tokens and each token maintains its own dictionary of all the documents it appears in and its frequency in each document it appears in. Later, corresponding document frequency of all the tokens is stored in a dictionary. The above mentioned function takes 'document_tokens' dictionary containing mappings of filenames to lists of tokens from each document, 'df' dictionary containing mappings of each term to its document frequency and total number of documents as inputs. Thus using all these inputs and

formulas each individual term weights are obtained, which is a dictionary mapping filenames to another dictionary of terms and their corresponding TF-IDF scores.

BM25:

For the BM25, I calculated the product of inverse document frequency(idf) and term frequency(tf).

$$\text{BM25} = (\text{IDF} * \text{TF})$$

Using the above inverse document frequency formula,

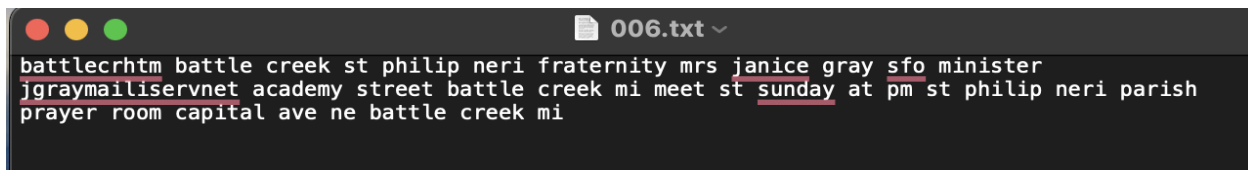
$$\text{IDF} = \log(\text{Total number of documents} / \text{Number of documents containing the term})$$

For term frequency, I used below formula,

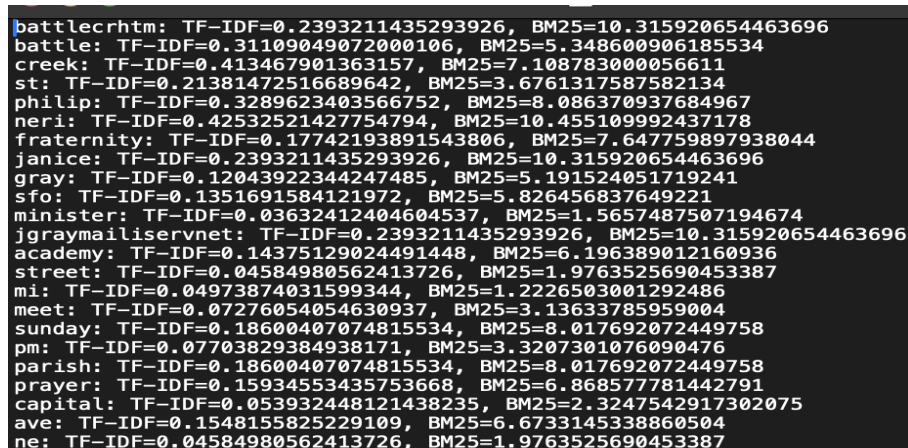
$$\text{TF} = ((\text{tf} * (\text{k1} + 1)) / (\text{tf} + \text{k1} * (1 - \text{b} + (\text{b} * (|\text{D}| / \text{avgl}))))$$

Using the same dictionaries mentioned in TF-IDF calculation, using the IDF and term frequency which is already calculated. Additionally, we need to calculate the average length of the documents, which we determine by multiplying the length of all documents in the corpus by the total number of documents in the corpus. To normalize the impact of a document's size on a term weight, we use this particular formula (length of document/average length of documents). In BM25, there are two tuning factors: k and b. term frequency scaling is controlled by the tuning factor k, where a larger k value indicates a higher priority for term frequency. To avoid that, I've used the recommended value of 1.2 for the k. The tuning factor b is used for managing the document length scaling where shorter documents are assigned greater weight when the value of b is set higher. In this case, the suggested value of 0.75 is used for b.

The following are the outputs I have got for the document 6 & 100..



```
006.txt
battlecrtm battle creek st philip neri fraternity mrs janice gray sfo minister
jgraymailiservnet academy street battle creek mi meet st sunday at pm st philip neri parish
prayer room capital ave ne battle creek mi
```



```
battlecrtm: TF-IDF=0.2393211435293926, BM25=10.315920654463696
battle: TF-IDF=0.31109049072000106, BM25=5.348600906185534
creek: TF-IDF=0.413467901363157, BM25=7.108783000056611
st: TF-IDF=0.21381472516689642, BM25=3.6761317587582134
philip: TF-IDF=0.3289623403566752, BM25=8.086370937684967
neri: TF-IDF=0.42532521427754794, BM25=10.455109992437178
fraternity: TF-IDF=0.17742193891543806, BM25=7.647759897938044
janice: TF-IDF=0.2393211435293926, BM25=10.315920654463696
gray: TF-IDF=0.12043922344247485, BM25=5.191524051719241
sfo: TF-IDF=0.1351691584121972, BM25=5.826456837649221
minister: TF-IDF=0.03632412404604537, BM25=1.5657487507194674
jgraymailiservnet: TF-IDF=0.2393211435293926, BM25=10.315920654463696
academy: TF-IDF=0.14375129024491448, BM25=6.196389012160936
street: TF-IDF=0.04584980562413726, BM25=1.9763525690453387
mi: TF-IDF=0.04973874031599344, BM25=1.2226503001292486
meet: TF-IDF=0.07276054054630937, BM25=3.13633785959004
sunday: TF-IDF=0.18600407074815534, BM25=8.017692072449758
pm: TF-IDF=0.07703829384938171, BM25=3.3207301076090476
parish: TF-IDF=0.18600407074815534, BM25=8.017692072449758
prayer: TF-IDF=0.15934553435753668, BM25=6.868577781442791
capital: TF-IDF=0.053932448121438235, BM25=2.3247542917302075
ave: TF-IDF=0.1548155825229109, BM25=6.6733145338860504
ne: TF-IDF=0.04584980562413726, BM25=1.9763525690453387
```

Above figure depicts the file-006 input and output files.

```
100.txt
sub themes sub themes processes at the organism leveladaptive evolution energeticslife history
adaptations ecological geneticsembiochemical adaptations trophic interactionsphysiological
adaptations nutrient dynamicsmorphological adaptations large scale patterns and processeslong
term change faunal and floral historyclimate change speciation and microevolutionssubdecadal
climate variability colonisation and recruitmentincreasing uvb community developmentlongterm
ecological research human impact environmental monitoringthere will be a session for papers that
environmental managementdo not fit into the above categories management of living resources
pollution ecotoxicology and introduced biota

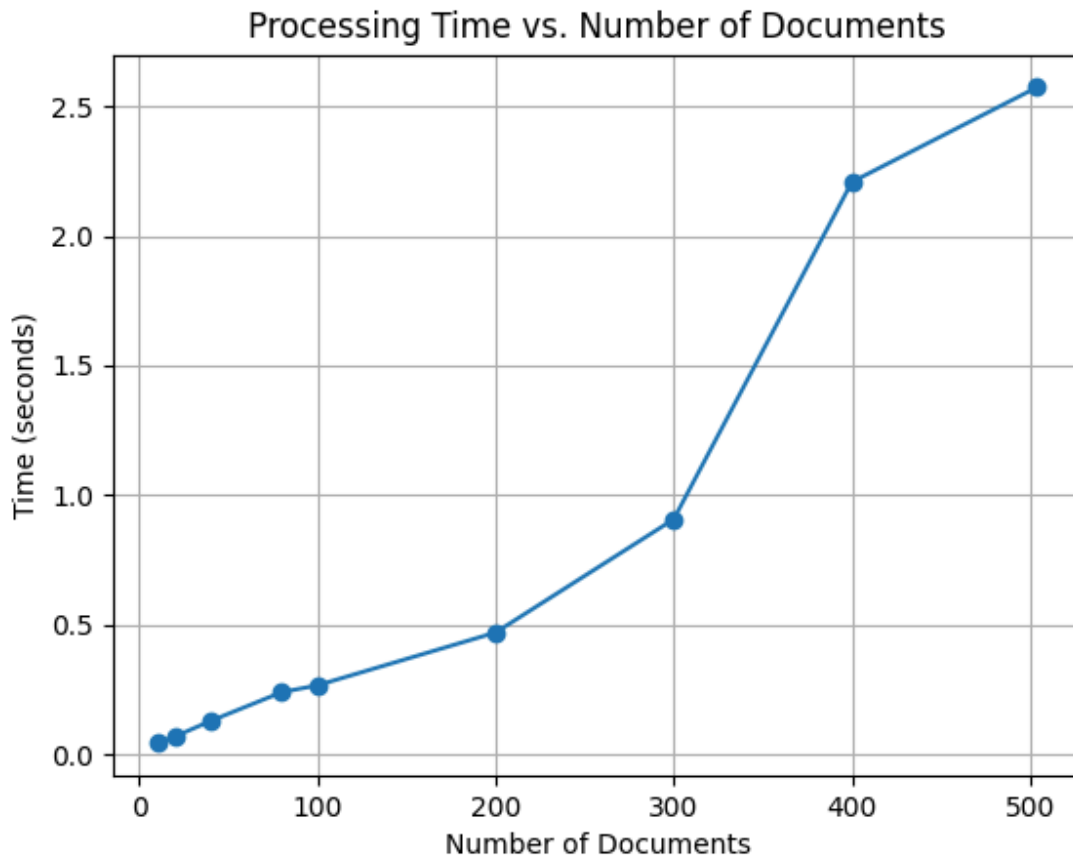
sub: TF-IDF=0.27388597057932973, BM25=9.601282002174168
themes: TF-IDF=0.09389737379552061, BM25=3.291644194726021
processes: TF-IDF=0.06516828500621759, BM25=3.9790612068701625
organism: TF-IDF=0.14777943975310845, BM25=9.023153453224955
leveladaptive: TF-IDF=0.16630448101658118, BM25=10.154259988253775
evolution: TF-IDF=0.05824995281888136, BM25=3.556639975127693
energeticslife: TF-IDF=0.16630448101658118, BM25=10.154259988253775
history: TF-IDF=0.0023343012675196752, BM25=0.14252834208924184
adaptations: TF-IDF=0.47367177610476846, BM25=8.967264072790245
ecological: TF-IDF=0.16273572299849337, BM25=5.7048251322667625
geneticsembiochemical: TF-IDF=0.16630448101658118, BM25=10.154259988253775
trophic: TF-IDF=0.16630448101658118, BM25=10.154259988253775
interactionsphysiological: TF-IDF=0.16630448101658118, BM25=10.154259988253775
nutrient: TF-IDF=0.12925439848963574, BM25=7.892046918196136
dynamicsmorphological: TF-IDF=0.16630448101658118, BM25=10.154259988253775
scale: TF-IDF=0.0728568718062467, BM25=4.448512834586483
patterns: TF-IDF=0.05091535775245487, BM25=3.108802461924231
processeslong: TF-IDF=0.16630448101658118, BM25=10.154259988253775
term: TF-IDF=0.014587643122726592, BM25=0.8906959089650698
change: TF-IDF=0.06442546907773518, BM25=2.25848405242985
faunal: TF-IDF=0.16630448101658118, BM25=10.154259988253775
floral: TF-IDF=0.16630448101658118, BM25=10.154259988253775
historyclimate: TF-IDF=0.16630448101658118, BM25=10.154259988253775
speciation: TF-IDF=0.16630448101658118, BM25=10.154259988253775
microevolutionssubdecadal: TF-IDF=0.16630448101658118, BM25=10.154259988253775
climate: TF-IDF=0.04516629421240369, BM25=2.7577747234181955
variability: TF-IDF=0.0890564482992758, BM25=5.437630568542316
colonisation: TF-IDF=0.16630448101658118, BM25=10.154259988253775
recruitmentincreasing: TF-IDF=0.16630448101658118, BM25=10.154259988253775
uvb: TF-IDF=0.16630448101658118, BM25=10.154259988253775
community: TF-IDF=0.027222275509221003, BM25=1.662144407070686
developmentlongterm: TF-IDF=0.16630448101658118, BM25=10.154259988253775
research: TF-IDF=0.01290286080655299, BM25=0.7878260550834212
human: TF-IDF=0.014771327592598927, BM25=0.9019113605962396
impact: TF-IDF=0.031861071765674535, BM25=1.945381172145411
environmental: TF-IDF=0.09149463354538746, BM25=3.207414299084758
monitoringthere: TF-IDF=0.16630448101658118, BM25=10.154259988253775
session: TF-IDF=0.07367927469921759, BM25=4.498727313109678
papers: TF-IDF=0.034043997610711065, BM25=2.0786667963816545
managementdo: TF-IDF=0.16630448101658118, BM25=10.154259988253775
fit: TF-IDF=0.08136786149924669, BM25=4.9681789408259975
categories: TF-IDF=0.0836933262696031, BM25=5.110167741898982
management: TF-IDF=0.01580479848548735, BM25=0.9650132810768761
living: TF-IDF=0.042178558754605924, BM25=2.575348835497725
resources: TF-IDF=0.0352061800672272, BM25=2.1496276192357042
pollution: TF-IDF=0.054331830542773976, BM25=3.3174062995576628
ecotoxicology: TF-IDF=0.16630448101658118, BM25=10.154259988253775
introduced: TF-IDF=0.08250530912274966, BM25=5.037629498149939
biota: TF-IDF=0.16630448101658118, BM25=10.154259988253775
```

Above figure depicts the file-100 input and output files.

Finally, after calculating the corresponding term weights, I maintained two dictionaries for storing the tf_idf and bm25 results. Using those dictionaries, respective term weights are written to the respective individual file. Also, during processing the documents, considering the mentioned document count set, respective timestamps are calculated and stored in a list. Using that list and the document count set, a graph is plotted.

Below figure shows the timings of term weight calculation on a varying number of documents.

Document Count	Timing (seconds)
10	0.05
20	0.07
40	0.12
80	0.25
100	0.26
200	0.49
300	0.91
400	2.20
504	2.60
Total Processing Time: 6.97 seconds	



The graph above depicts the plot between the time and number of documents processed.

Analysis:

From the graph, we can say that the processing of documents for TF-IDF and BM25 scoring is linearly scalable with the number of documents and the number of tokens within those documents. The execution times for varying numbers of documents have increased significantly, due to a significant rise in program complexity. Also, to run the program effectively, I have used a lot of temporary lists and dictionaries in addition to performing repeated reads and writes.

Thus, I can say that the overall time complexity can be approximated as $O(d \times t)$, where d is the number of documents and t is the average number of tokens per document. Since it involves processing of documents including tf-idf and bm25 scores calculation.

References:

- [1] <https://en.wikipedia.org/wiki/Tf%E2%80%93idf>
- [2] https://en.wikipedia.org/wiki/Okapi_BM25
- [3] <https://www.elastic.co/blog/practical-bm25-part-3-considerations-for-picking-b-and-k1-in-elasticsearch>