CS6370-PROJECT
TEAM-21

S ANIRUDDH
Mechanical Engineering
Indian Institute of Technology Madras
Chennai,INDIA
me18b185@smail.iitm.ac.in

MOHSIN SACKEER
Chemical Engineering
Indian Institute of Technology Madras
Chennai,INDIA
ch19b065@smail.iitm.ac.in

T SIVAKUMAR
Biological Engineering
Indian Institute of Technology Madras
Chennai,INDIA
be18b014@smail.iitm.ac.in

***Abstract*—We are building up an information retrieval system using Natural Language Processing techniques in this project. In this report, we will discussing about the models we tried to address the limitation of the baseline model used in the assignments.**

I. Introduction

This project is an attempt to improve the previous version of the model used for the information retrieval task. We tried three different models to address various limitations of the vector space model. We improved pre-procesing in the baseline model and tried to compare its performance with the base line model. Then we tried Latent Semantic Analysis(LSA) technique to address the dimensionality limitation of the baseline model. These improvements have been made to improve the efficiency of the Information Retrieval system. We also tried clustering approach to improve the efficacy of the Information Retrieval system by decreasing the search time.

II. Models trained

A. Baseline Model-M1

A vector space model that uses TF-IDF score of the words in the documents to calculate the relevance of a document with the given query. The baseline model is the same vector space model used in the assignment.

Pre-processing techniques:

- Tokenization:We used Treebank tokenizer from the NLTK package, to tokenize documents into sentences and further into words.
- We removed the stopwords from the list before performing further processing as these stopwords doesn't affect the retrieval due to their high occurence frequency in the documents.
- InflectionReduction: We used lemmatization to reduce the words to its root form.

Ranking and Evaluation of the model:

- Ranking: After getting the query vector, we calculate the similarity of the query with the document using cosine similarity. We ranked the results based on the similarity score in the decreasing order with the document with highest similarity score being ranked one.
- Evaluation: We evaluated the IR system using metrics like Precision,Recall,MAP,n-DCG and F-Score.

B. Improved Vector Space Model-M2

This is an improved version of the base line model where we incorporated few other processing steps to clean the raw data.

Additional Processing step:

- Removing the punctuation marks from the sentence can improve the retrieval.
- We are converting all the upper case letters in the words to lower case, so as to avoid difference between identical words due to case differences.
- We have removed all the white spaces and numbers from the text.
- Then, we followed the same processing steps as in baseline model.

Ranking and Evaluation:
Ranking and evaluation are same as the baseline models.

Limitations of the Vector Space Model:

- In vector space models, all the words are assumed to be orthogonal. Similarity between words is not captured.
- Long documents have very poor representation and hence lesser cosine similarity.
- Synonymy and polysemy issues are not addressed in the naive vector space model.

C. Latent Semantic Analysis(LSA)

We are trying a new approach to solve the dimensionality problem and synonymy by using Latent Semantic Analysis technique(LSA).
We going to follow same procedures as in the baseline(M1) to arrive at the TF-IDF term document matrix.Since, we have the base line model as well as the improvised vector space model(M2), we will be applying LSA technique on both of these models.

Ranking and Evaluation:
Ranking and evaluation are same as the baseline models.

D. Clustering

We have tried to improve efficiency of the information retrieval system through the previous models. But the search time also plays an important role in determining the efficacy of the system. let us now change gears to clustering.
Clustering is an unsupervised technique in which the set of similar data points is grouped together to form a cluster.Clustering could also be viewed as a Data Compression technique in which the data points of a cluster can be treated as a group. Clustering is also called Data Segmentation because it partitions the data such that a group of similar data points forms a cluster.
In information retrieval task, data points refers to documents. This method will be highly beneficial once the number of documents is very large in the corpus. Clustering can reduce the search time to retrieve relevant documents.
We have used K-Means clustering method to reduce the search time.

Implementation:

- We cluster the documents by applying the algorithm on the term document matrix.
- We find the cluster center nearest to the query vector and select that cluster.
- Once we have selected the nearest cluster, we could compute cosine similarity of the query with all the documents within the cluster and rank them on this basis.

We will do hypothesis testing for checking whether clustering improves search time or not

III. Model's result and evaluation

Here, in this section, we will list out the results all of the models and their evaluation score, when trained on the cranfield dataset for an information retrieval task with the foresaid evaulation metrics.

A. Baseline Model-M1

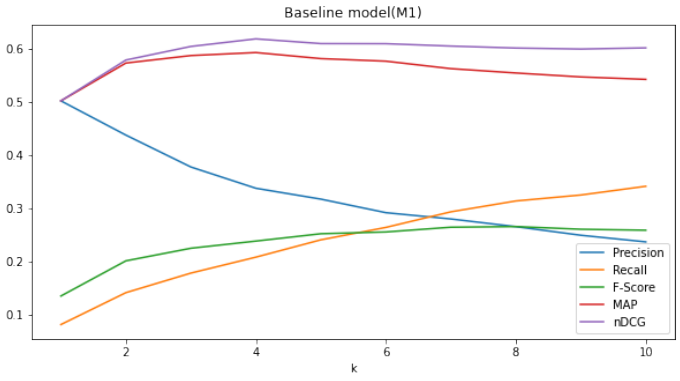


Fig. 1. Baseline model-M1

Figure 1 refers to the performance of the baseline model measured using the all the metrics.

We can see that as k increases, precision decreases while recall increases as expected. MAP@k is maximum at k=4.

B. Improved Vector space model-M2

Figure 2 refer to the performance characteristics of the vector space model 2 with respect to the evaluation metrics.

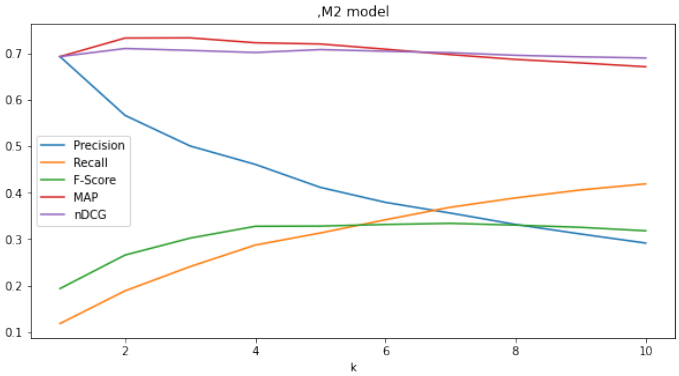
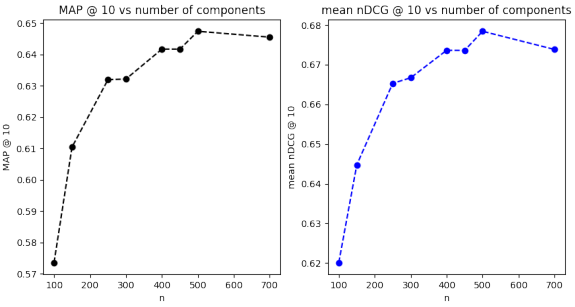


Fig. 2. Improved Vector Space Model-M2

C. Latent Semantic Analysis-LSA

As mentioned earlier, we will apply LSA on M1 and M2. To arrive at the number of latent dimensions, we randomly checked the evaluation score of the model for some arbitrary values of number of latent variables.
The resultant graph is shown below:



From the graph we can say that 500 latent dimension will be good enough.
So the number of latent dimensions = 500

- Fig 3 refers to the plot of LSA on M1
- Fig 4 refers to the plot of LSA on M2.

IV. Comparison of Model's performance and observation:

Here,we are going to compare the models performance using the evaluation metric scores rank wise and list out some of the key observations obtained from comparing models pairwise.

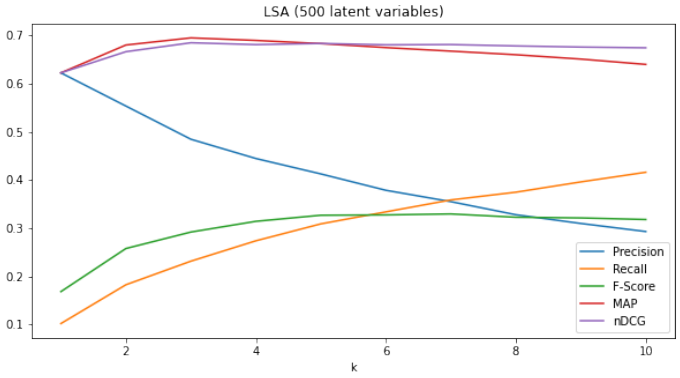


Fig. 3. LSA ON M1

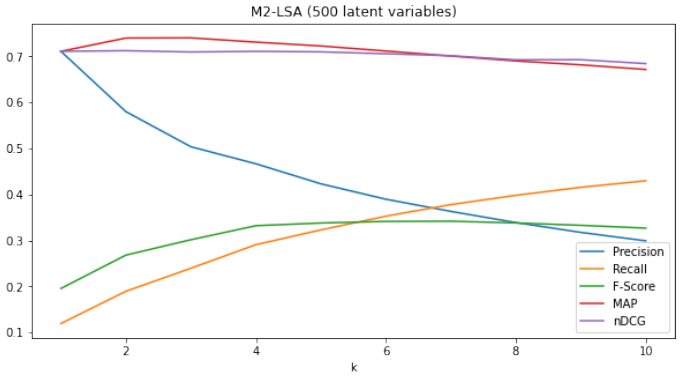


Fig. 4. LSA ON M2

Baseline Model

Rank(k)	Precision	Recall	F-Score	MAP@k	nDCG
1	0.50222	0.08135	0.13505	0.50222	0.50222
2	0.43777	0.14135	0.20120	0.57333	0.57899
3	0.377777	0.17816	0.22488	0.58740	0.60458
4	0.33777	0.20815	0.23836	0.593086	0.61881
5	0.31733	0.24074	0.25220	0.581691	0.60996
6	0.29185	0.26408	0.25552	0.57684	0.60977
7	0.28000	0.29340	0.26447	0.56280	0.60527
8	0.26555	0.31386	0.26569	0.55465	0.60157
9	0.24938	0.32506	0.26065	0.54722	0.59970
10	0.23688	0.34150	0.25884	0.54262	0.60186

Fig. 5. M1 Performance

M2:

Rank(k)	Precision	Recall	F-Score	MAP@k	nDCG
1	0.69333	0.11831	0.19345	0.69333	0.69333
2	0.56666	0.18859	0.26577	0.73333	0.71079
3	0.50074	0.24088	0.30223	0.73370	0.70666
4	0.46111	0.28726	0.32769	0.72320	0.70218
5	0.41155	0.31326	0.32816	0.72051	0.70849
6	0.37925	0.34171	0.33159	0.70933	0.70487
7	0.35619	0.36880	0.33406	0.69758	0.70173
8	0.33166	0.38892	0.33017	0.68729	0.69615
9	0.31111	0.40603	0.32549	0.67977	0.69282
10	0.29155	0.41912	0.31797	0.67148	0.69042

Fig. 6. M2 performance

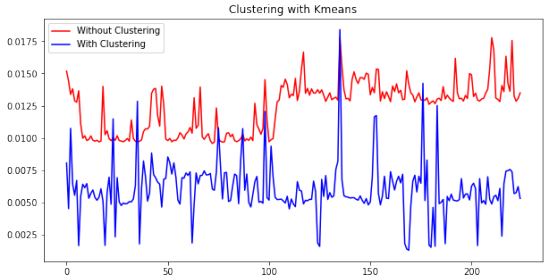
A. *Some Key Observations:*

- We can observe from clearly for the plots as well as from the table that LSA on M2 model's performance is very similar to M2 model which is better than LSA on M1 model and M1's performance when we consider MAP@K metric graph in the plots as well as values from the table.
- Similarly, Precision graph is highest on LSA on M2 model and similar to M2 model plots and lower for M1 model.
- Recall also follows the same trend as precision.
- We can speculate that LSA on M2 model is performing similar to M2 model. LSA on M1 model is performing better than M1(vector space model) but less than the previous two models.
- We can confirm or verify these speculations based on the results of hypothesis testing that is to discussed in the next section.

B. *Clustering model:*

Here, since we are trying to improve the efficacy. So, we cannot measure the performance of the model based on the evaluation metrics but can measure the performance based on time taken for retrieval.

We have clustering model with the naive vector space model without clustering.
To arrive at the number of clusters, we considered the trade off between the silhoutte distance and the number of clusters and we found a sweet spot at number of clusters=6.
Average time taken for per query without clustering =0.012363311979505751 seconds
Average time taken with clustering = 0.00603457980685763 seconds We can see that the time taken for retrieval gets reduced by 50 percentage when compared to the vector space model without clustering.



We can clearly see that the time taken for retrieval in the

without clustering case's graph is higher than the clustering model.

V. Hypothesis Testing

We are going to use two sample t-test to check our hypotheses.
We are going to check five cases corresponding to pairs of models involved in improving efficiency. Clustering is dealt at last.

A. Baseline Model(M1) vs M2 model

Null hypothesis: M1 model performs comparably similar to M2 model in terms of the evaluation metrics Precision, Recall, F-Score and n-DCG.
Alternate Hypothesis: M1 model performs comparably similar to M2 model in terms of the evaluation metrics Precision, Recall, F-Score and n-DCG.
The result of the hypothesis test is shown below:

```
Precision: t = -3.16902  p = 0.0016347
Recall: t = -3.07628  p = 0.00222461
Fscore: t = -3.46671  p = 0.000577746
nDCG: t = -3.28127  p = 0.00111684
```

Conclusion: Since all the p-values are less than 0.05, we reject the null hypothesis.
We can say that M1 and M2 doesn't perform similarly. We can infer from graph the M2 performs better than M1.

B. Baseline Model(M1) vs LSA on M1 model

Null hypothesis: M1 model performs comparably similar to the LSA on M1 in terms of the evaluation metrics Precision, Recall, F-Score and n-DCG.
Alternate Hypothesis: M1 model doesn't performs comparably similar to M2 model in terms of the evaluation metrics Precision, Recall, F-Score and n-DCG.
The result of the hypothesis test is shown below:

```
Precision: t = -3.17567  p = 0.00159919
Recall: t = -2.9365  p = 0.00349079
Fscore: t = -3.38497  p = 0.000774818
nDCG: t = -2.63744  p = 0.00864995
```

Conclusion: Since all the p-values are less than 0.05, we reject the null hypothesis.
We can say that performance of these model are not similar or they are not comparable. From the plots we can say that LSA on M1 model is performing better when compared to M1 model.

C. LSA on M1 model vs M2 model

Null hypothesis: LSA on M1 model performs comparably similar to the M2 in terms of the evaluation metrics Precision, Recall, F-Score and n-DCG.
Alternate Hypothesis: LSA on M1 model doesn't performs comparably similar to the M2 in terms of the evaluation metrics Precision, Recall, F-Score and n-DCG.
The result of the hypothesis test is shown below:

```
Precision: t = 0.0735069  p = 0.941436
Recall: t = -0.124429  p = 0.901032
Fscore: t = -0.00771477  p = 0.993848
nDCG: t = -0.649749  p = 0.516188
```

Conclusion: Since all the p-values are greater than 0.05, we cannot reject the null hypothesis.
We can say that performance of these model are similar.

D. M2 model vs LSA on M2 model

Null hypothesis: M2 model performs comparably similar to the LSA on M2 model in terms of the evaluation metrics Precision, Recall, F-Score and n-DCG.
Alternate Hypothesis: M2 model doesn't performs comparably similar to LSA on M2 model in terms of the evaluation metrics Precision, Recall, F-Score and n-DCG.
The result of the hypothesis test is shown below:

```
Precision: t = 0.493611  p = 0.621823
Recall: t = 0.453204  p = 0.650621
Fscore: t = 0.575398  p = 0.565312
nDCG: t = -0.208915  p = 0.83461
```

LSA ON VM1

Rank(k)	Precision	Recall	F-Score	MAP@k	nDCG
1	0.62666	0.10263	0.16942	0.62666	0.62666
2	0.56666	0.18544	0.26263	0.68444	0.66881
3	0.48148	0.22996	0.28991	0.69814	0.68256
4	0.44222	0.27060	0.31106	0.69246	0.68151
5	0.40977	0.30826	0.32525	0.68454	0.68501
6	0.3729	0.33296	0.32578	0.67963	0.68611
7	0.35111	0.35479	0.32565	0.67101	0.68357
8	0.33111	0.37693	0.32620	0.65829	0.67952
9	0.31209	0.40000	0.32371	0.64882	0.67657
10	0.29377	0.41447	0.317953	0.64188	0.67602

Fig. 7. LSA ON Base Line model performance

LSA ON VM2

Rank(k)	Precision	Recall	F-Score	MAP@k	nDCG
1	0.71111	0.11932	0.19555	0.71111	0.71111
2	0.57777	0.18937	0.26791	0.74222	0.71533
3	0.50962	0.24040	0.30359	0.74222	0.71255
4	0.46333	0.28804	0.32878	0.73296	0.71261
5	0.42311	0.322080	0.33742	0.72483	0.71240
6	0.39037	0.35423	0.34250	0..71582	0.71156
7	0.36317	0.37939	0.34247	0.70582	0.70744
8	0.34111	0.40106	0.34047	0.69221	0.69702
9	0.31654	0.41352	0.33156	0.68522	0.69331
10	0.29955	0.43059	0.32731	0.67285	0.68900

Fig. 8. LSA ON M2 performance

Conclusion: Since all the p-values are greater than 0.05, we cannot reject the null hypothesis.
We can say that performance of these models are similar.

E. LSA on M1 model vs LSA on M2 model

Null hypothesis: LSA on M1 model performs comparably similar to the LSA on M2 model in terms of the evaluation metrics Precision, Recall, F-Score and n-DCG.
Alternate Hypothesis: LSA on M1 model doesn't performs comparably similar to LSA on M2 model in terms of the evaluation metrics Precision, Recall, F-Score and n-DCG.
The result of the hypothesis test is shown below:

```
Precision: t = -0.411169  p = 0.681145
Recall: t = -0.574549  p = 0.565885
Fscore: t = -0.571451  p = 0.56798
nDCG: t = -0.447299  p = 0.654876
```

Conclusion: Since all the p-values are greater than 0.05, we cannot reject the null hypothesis.
We can say that performance of these model are similar.

F. M2 without clustering model vs M2 with clustering model

Null hypothesis: Mean Retrieval time with and without clustering are same.
Alternate Hypothesis: Mean Retrieval time with and without clustering are not same.
The result of the hypothesis test is shown below:

```
t = 32.9048  p = 2.78211e-121
```

Conclusion: Since all the p-values are less than 0.05, we reject the null hypothesis.
We can say that mean retrieval time is not same for both the models. In fact, from the graph we can say that clustering the mean retrieval time by 50 percent on the minimum side when compared with the model without clustering.

Observation after hypothesis test: Whatever we speculated regarding the performance of the model by seeing the evaluation plots are correct except that we speculated LSA on M1 model is not comparable to M2 and LSA on M1 model. But from hypothesis tests, it is evident that all three models perform similar when the efficiency of the system is concerned. Also clustering improves the search time of the retrieval as the same is confirmed with the hypothesis test conducted.

FINAL REMARKS ON THE PROJECT AND SUMMARY

Initially through the assignments we built vector space model for an information retrieval task. We understood the intricacies of how a search engine works. We built all the functions from scratch which helped us in getting a grip over the course. Later, we identified the limitations of these vector space models and also tried to improve the same in this project work.

In the project, we tried to address some potential limitations which could be improved included in the pre processing steps of the assignments, We addressed such limitations and found that it improved the efficiency of the IR system.
Then, we also tried to address limitations of the vector space model like the poor representation of long documents , orthogonality of words, synonmy and polysemy relations among words, which it could address through Latent Semantic Analysis(LSA).
It improved the efficiency of the system.
We then moved gears to apply LSA on the M2 model which is the pre processing improved vector space model.
There was an improvement in terms of efficiency when compared to the model which involved LSA applied on M1. But this model was similar in terms of efficiency when compared to M2 model without LSA.
In this project, we also tried to improve efficacy of the IR system by reducing the search time of the retrieval process. From the results and plots it was evident that search time got reduced by using the clustering technique.

REFERENCES

[1] https://en.wikipedia.org/wiki/Latentsemanticanalysis.
[2] https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
[3] https://www.scribbr.com/statistics/t-test/
[4] course lectures
[5] Research papers shared in the course and various other resources.