CS 386: PROJECT 3 REPORT **Ishan Singhal (160010016)**
October 27, 2018

1 Introduction and Overview

Online consumer reviews (OCR) have helped consumers to increase their knowledge about different products/services and choose the ones that best suit their needs. Recommender systems are software agents that predict the interests or preferences of individual users for products and make recommendations accordingly. In this project, I have aimed at developing a recommender system for online reviews that is able to predict the perception of each user regarding the helpfulness of online reviews. The dataset that I have worked on is a Women Clothing E-Commerce dataset revolving around the reviews written by customers. Its nine supportive features offer a great environment to parse out the text through its multiple dimensions. This dataset includes 23486 rows and 10 feature variables.

2 Methods

2.1 Text Cleaning

The first part was to clean the reviews i.e. removing the stopwords and then standardizing the word tokens by using stemming and lemmatization. The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

2.2 Information Retrieval (IR)

After the cleaning process, IR was done using TF-IDF and LSA methods.

2.2.1 TF-IDF

In this method, we constructed a DxV document-term TF-IDF matrix based on our corpus and a 1xV query term vector based on the user input (D:no of documents, V:no of terms in vocabulary). Then based on the cosine approach between the query and document vectors, we selected the best response.

2.2.2 LSA

In this method, we applied SVD on the term-frequency matrix and then reduced the matrices from n to K dimensions. Then based on the cosine approach between the query and document vectors, we selected the best response.

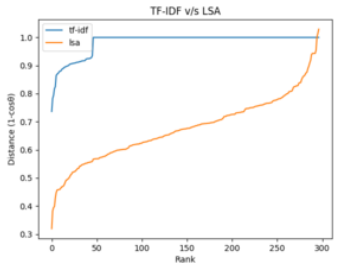
2.3 Recommender System

2.3.1 Statement

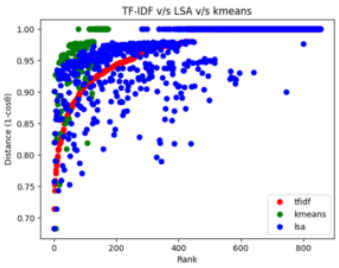
I wish to recommend the best products available to a user based on his search requirement. Since our dataset is too large, I have selected a particular Department Name (i.e. Clothing Type).

2.3.2 Solution

Initially, I have filtered the dataset by selecting a department and only those with Recommended IND = 1, since we want to recommend the best products. After processing the corpus, I've performed LSA using SVD on the term-frequency matrix. Having a vector representation of a document gives us a way to compare documents for their similarity by calculating the distance between the vectors. The



((a)) TF-IDF v/s LSA



((b)) TF-IDF v/s LSA v/s Kmeans

Figure 1: Comparison between different approaches

SVD step does more than just reduce the computational load, we trade a large number of features for a smaller set of better features.

Now K-means was used to group the similar products together and then find the cluster centre matching the users requirements the most. The products embedded in that cluster contains the required recommendable products. Once I have the cluster, using Euclidean distance I find the products which are most similar to the users requirements.

3 Analyses of Results

3.1 Compare the performance of Information Retrieval (IR) using both TF-IDF and LSA methods

We can see from the plot shown in the **Figure 1(a)** that for TF-IDF after a certain rank, the distance values goes to 1 and remains same i.e. after a particular document, all the other documents do not have the query keywords. Whereas for LSA, although the distribution seems much better than TF-IDF but since LSA reduces the dimensions(or vocabulary) from n to K words, it ends up giving a response which is not good enough. It clubs certain words to form phrases and then some of them match while some don't with our query keywords. Its accuracy is directly proportional to the number of dimensions. But TF-IDF does no such reduction and computes based on all the keywords, hence it gives more accurate result. This can be seen from the responses displayed on terminal.The best response for each of the methods is corresponding to the value of 1-Cosθ at rank=0 is minimum indicating maximum similarity between the query and the document.

3.2 Recommender System

To analyze my algorithm, I wanted to compare my approach with the LSA as well as the TF-IDF approach. So to achieve that, I have plotted all the results in the same space as shown in **Figure 1(b)**. As you can see at rank=0, lsa and my approach are overlapping and in further ranks, my approach and tf-idf are overlapping. So my result consists of both lsa as well as tf-idf responses. The best response for each of the methods is corresponding to the value of 1-Cosθ at rank=0 is minimum indicating maximum similarity between the query and the document.

References

[1] https://www.researchgate.net/publication/220980957_A_new_collaborative_filtering_algorithm_using_K-means_clustering_and_neighbors