CMPE493 - Information Retrieval - Assignment 3

Berk Atıl - 2016400102

Introduction

In this assignment, we are asked to build book recommendation system based on a content based filtering. We are required to use the descriptions and genres of books in "Goodreads". The program can take 2 different kinds of parameters. 1 of them is for building the system and the second one is for getting the recommendations. In the first case, a path to a file consisting of urls of books in the desired corpus is given. In the other one, a single url for which recommendations are desired is given. I differentiate these files by checking if the input contains **.txt** if so I assumed it is a file. In addition, when the file path is given, corpus and the models are created from scratch. That is, if two different files are given to the program sequentially, the first books and models for them are overriden by the second one. The details of the system are provided in the next sections.

Preprocessing

The descriptions are retrieved from the web sites. Then, new line characters(\n) are replaced with the space characters. Afterwards, all letters are made lower case and finally punctuations are removed. For the genres, I retrieved them from the website. There were some cases which show the sub-genre(i.e Business > Entrepreneurship). For these cases, I decided to take them separately. Afterwards, set of genres for each book is extracted and they are made lower case. Sometimes, some books cannot retrieved. For these cases, the program tries to retrieve it for 3 times and if it cannot be retrieved after 3 times, this book is skipped. After informative words are determined, 2415 unique words exist for the "books.txt" file.

Building the System

In order to look at the similarities between books, descriptions and genres are used. In order to represent descriptions, **log-frequency weighting** variant of **tf-idf** is used. Its formula is the following:

$$w_{t,d} = \begin{cases} 1 + \log_{10} tf_{t,d} & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

On the other hand, for genres, **binary-term document incidence matrix** in the lecture nodes is used. That is, if a genre exists for the document its value in the vector is 1 and 0 otherwise. In order to select the informative words **2 thresholds** are used. Min threshold is selected as 0.006 of the total number of books and max threshold is selected as 90% of the total number books in the corpus. This means that, the words that occur in less than 0.6% of the total number of books or greater than 90% of the total number of documents are discarded. Lastly, the effects of both genres and descriptions are assumed as the same so **0.5** is used as an α

Recommendations

18 recommendations are made for each book. This is done by looking the cosine similarities between the desired book and all books in the corpus and top 18 are returned. Afterwards, precision and average precision metrics are also printed assuming Goodreads recommendations as ground truth.