Task1

Nawar Saeed

Artificial Intelligence for the Web, VT21
DT506A

16 april 2021

For this task, oop has been used where a simple class for inverted indexes has been created. In the construction function, an empty dictionary has been initialized to be used for storing the inverted lists and the variable called total-filmes is used to count the number of movies in the file. It has also created a parser to read from a txt file. It starts with opening and reading the file, and this can be done with open(file-name).
The next step is to read the file line by line, because the file is constructed so that each movie has its own line. When done, tota-filmes by one is incremented to calculate the number of movies in the document. The next step is to split the name of the movies from its description, but since the task asks for different things, a variable called separator has been used and the purpose of it is to split after a special request. When this has been done, the next step is to include only words, and this can be done according to line 40.
The next step will be to convert all uppercase letters to lowercase letters to avoid having problems with e.g. a search query is written in capital letters, etc. The next step is the most important because here it is checked when an inverted list is to be created. If a word is seen for the first time, create an empty inverted list for it, otherwise assign a word a document id. And in this way the inverted index is created.
What is considered a token depends entirely on how it splits. I have used e.g. $[A-Za-z0-9]$ to split the words. The word $acconci's$ is considered as a one token just because I have specified how I want to split the words.

Total films in the file	Total different tokens in the file	Total tokens in the file
189897	214656	8837198

Results of Task1