Assignment 2

Information Retrieval

- Mayank Chauhan, MT18008

Methodology

Preprocessing

- First, Extract title, file name, story title from index.html using BeautifulSoup.
- Apply tokenization - remove punctuation, accent marks and other non alphanumeric characters like (, >, $, \, \\n, /t, etc.
- Apply Normalization, Stemming and Lemmatization.
- Now, we have the final list of tokens for our index.

Building Inverted Index

- Create two separate Inverted Index, one using all the content of files and other one for only using title text.
- The tf-idf for each (term, document) pair is calculated using-

$$\text{tf-idf}_{(t,doc)} = 0.7*(\text{tf-idf}_{(t,\ Title)}) + 0.3*(\text{tf-idf}_{(t,Body)})$$

Analysis

The result for Case 1 and Case 2 are different because in the first case we finding the relevant document based on the each term w.r.t to a particular document considering only one at a time. Whereas, in the second case the we convert the query into a vector and we are considering the all the query terms for computing the relevance w.r.t to each document.

Query : 4 moons

Results :

Case1. Nigel.4
4moons.txt
timem.hac
gulliver.txt
hound-b.txt

Case2. 4moons.txt
nigel.4
quarter.c11
empsjowk.txt
imagin.hum