Information Retrieval (TDT4117)

Assignment 3

Submitted by

Name : Md Anwarul Hasan
Student ID : 583233

1. Data loading and preprocessing

1.0) Answer:

Simply initiated random function and random seed as well.

1.1) Answer:

Like the instruction provided did import codecs and also loaded the file.

1.2) Answer:

To read each line we divided the each line by split function of python where we found carriage return or newline character. Here we used a custom function "split_into_paragraphs"

1.3) Answer:

With the help of custom function "remove_gutenberg_word" I searched whether the word "Gutenberg" is present and if present we rremoved the word.

1.4) Answer:

Like the instruction provided here I divided the paragraphs into words. I took the help of "enumerate" function to find out the word counts and splits them by searching " "

1.5) Answer:

Then we eliminated the punctuations by a custom function called "eliminate_punctuations" where I used punctuation function of string library and ultimately eliminated the functions.

1.6) Answer:

Like the instruction implemented the PorterStemmer to stem each words. Here we also used a custom function "temmed_lists". This function simply used the code portion as instructed in the question.

2. Dictionary building

2.1) Answer:

Like the instruction provided imported the library genism and the dictionary was build with the code portion provided.

Also to filter the stop-words we downloaded the stop-word lists and like reading a file read the file and searched in the dictionary to find the locations of the stop-words and then finally filter-out the stop-words. This portion was also used the code snippet provided.

2.2) Answer:

The stemmer words was turned into bag-of-words using the dictionary and also by a custom function “bag_of_words”.

3. Retrieval Models

3.1) Answer:

Like the instruction provided bag-of-words were used to build TF-IDF model. The function “gensim.models.TfidfModel” was used to do this.

3.2) Answer:

Like the instruction I mapped the bag-of-words into TF-IDF weights by simply assigning the value of tfidfModel using the bag-of-words to a value called tfidfCorpus

3.3) Answer:

Using the result from 3.2 and put it in gensim.similarities.MatrixSimilarity function we obtain similarity matrix.

3.4) Answer:

Following the instruction also applied the point 3.1, 3.2 and 3.3 for LSI models and set the limit of the topics to 100 as provided.

3.5) Answer:

Then we printed three LSI topics. Which are given below.

[(0, '0.138*"price" + 0.134*"labour" + 0.121*"produc" + 0.120*"countri" + 0.119*"tax" + 0.119*"hi" + 0.115*"employ" + 0.114*"trade" + 0.114*"upon" + 0.111*"capit"'), (1, '-0.294*"silver" + -0.274*"gold" + -0.228*"coin" + -0.208*"price" + -0.177*"bank" + -0.165*"money" + 0.164*"capit" + -0.148*"0" + 0.143*"stock" + 0.137*"employ"'), (2, '0.284*"price" + 0.244*"labour" + -0.231*"0" + 0.198*"rent" + -0.181*"trade" + -0.163*"coloni" + -0.158*"duti" + 0.149*"quantiti" + 0.146*"wage" + 0.145*"land"')]

4. Querying

4.1) Answer:

The given query is "What is the function of money?"

For this part I remove punctuations, tokenize, stemmed and convert to bag-of-words like I did in part one

4.2) Answer:

Also converted bag-of-words into TF-IDF representation. This part is also similar to the task of previous parts.

4.3) Answer:

Finally we worked for to find out 3 most relevant relevant paragraphs for the query "What is the function of money?" according to TF-IDF model. This part used the code help provided by the instruction and simply printed the number of paragraph and upto five lines of code. The provided code portion was very helpful in this regard to work with.

Three most relevant topics according to TF-IDF are,

“[paragraph: 676]

The general stock of any country or society is the same with that of

all its inhabitants or members; and, therefore, naturally divides itself

into the same three portions, each of which has a distinct function or

office.

[paragraph: 677]

The first is that portion which is reserved for immediate consumption,

and of which the characteristic is, that it affords no revenue or

profit. It consists in the stock of food, clothes, household furniture,

etc. which have been purchased by their proper consumers, but which are

not yet entirely consumed. The whole stock of mere dwelling-houses,

...

[paragraph: 987]

That wealth consists in money, or in gold and silver, is a popular

notion which naturally arises from the double function of money, as the

instrument of commerce, and as the measure of value. In consequence of its being the instrument of commerce, when we have money we can more readily obtain whatever else we have occasion for, than by means of any ..."

4.4) Answer:

Like 4.3, I also used the code portion provided in 4.4 and figured out the 3 most significant weights which have largest absolute values. Here I also get the concept of doing that from the code portion that was provided.

Three most significant topics according to LSI are,

"[paragraph: 987]

That wealth consists in money, or in gold and silver, is a popular notion which naturally arises from the double function of money, as the instrument of commerce, and as the measure of value. In consequence of its being the instrument of commerce, when we have money we can more readily obtain whatever else we have occasion for, than by means of any ...

[paragraph: 861]

In some countries the interest of money has been prohibited by law. But as something can everywhere be made by the use of money, something ought everywhere to be paid for the use of it. This regulation, instead of preventing, has been found from experience to increase the evil of usury. The debtor being obliged to pay, not only for the use of ...

[paragraph: 684]

First, of the money, by means of which all the other three are circulated and distributed to their proper consumers."