Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Text, Web and Social Media Analytics Lab

Prof. Dr. Diana Hristova

Exercise 2. Preprocessing

In this exercise, we will be using the 20-Newsgroups dataset. This version of the dataset contains about 11k newsgroups posts from 20 different topics. We will do following steps:

A) Import and examine data
B) Remove initial text metadata
C) Remove numbers, punctuation, tabs and convert to lower case with gensim
D) Remove stop words and short words
E) Stemming and Lematization

0. Open a Colab notebook

1. Import the following packages:
   a. pandas
   b. re
   c. from gensim.parsing.preprocessing import STOPWORDS, strip_tags, strip_numeric, strip_punctuation, strip_multiple_whitespaces, remove_stopwords, strip_short, stem_text
   d. pickle
   e. en_core_web_sm
   f. nltk
   g. Run nltk.download('stopwords')
   h. from nltk.corpus import stopwords after g.

Part A):

2. Read with panda's ***read_json()*** function the following dataset: *https://raw.githubusercontent.com/selva86/datasets/master/newsgroups.json* either from the url or from a file on Google Drive. Call it ***df***.

3. Print the head of ***df***. What kind of data does it contain? How many entries does it have? Which categories can be found in the column ***target_names***? What is their distribution (e.g. with ***value_counts()***)? Print the first ***content*** value. Does it match the target name? Which business question can this dataset address?

Part B):

4. Remove the lines beginning with any of the following: 'From:', 'Article-I.D.:', 'Organization:', 'Lines:', 'NNTP-Posting-Host:', 'Distribution:', 'Reply-To:', 'X-Newsreader:', 'Expires:', multiples (also one) of '-' preceded by space using the

package ***re***. Remove additionally any of the words ‘Subject:’, ‘Summary:’ or ‘Keywords:’. Both removals should be case insensitive. Call the new object ***data*** which is your corpus and display the first entry. Why are we doing this?

Part C):

5. Apply ***strip_numeric, strip_punctuation*** and ***strip_multiple_whitespaces*** to ***data*** and override it. What are those functions doing and why?

6. Transform all letters to lower case ones and override ***data*** with the result (Hint: Use ***string.lower()***).

Part D):

7. Print both the stopwords in *gensim* and those in *nltk*. Do you see a difference? Hint: it may make sense to sort the objects beforehand.

8. Remove the stopwords using gensim’s ***remove_stopwords*** and override ***data***.

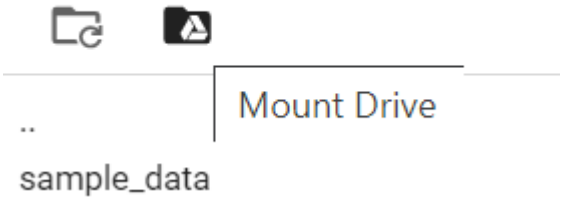
9. Apply ***strip_short*** to ***data***. What is this function doing and why?

Part E):

10. Apply ***stem_text*** to ***data***. What is this function doing and why?

11. Apply lemmatization with Spacy by:
    a. Initializing spacy’s 'en' model with en_core_web_sm.load()
    b. Applying the model to the documents in ***data***

12. If not done so, mount your GoogleDrive by



and running

```
from google.colab import drive
drive.mount('/content/drive')
```

13. Store both the stemmed and lemmatized data (your corpus) in Google Drive (you may wish to make an extra folder) using *pickle.dump*. Why does it make sense storing this data?