Computitional social science - Assignment 3

JProf. Dr. Claudia Wagner
clwagner@uni-koblenz.de

Nada Beili
nbeili@uni-koblenz.de

May 27, 2020

Submission until: June 22, 2020 - 11:00 a.m.

Instructions:

1. Send your solution before the deadline to nbeili@uni-koblenz.de
2. Use as subject of the email "CSS Assignment 3 " + your full name and immatriculation number.
3. For the programming tasks, please do not add your code to the PDF. You need to submit only the .ipynb file. **The file name has to be the same as the email subject otherwise will not be accepted.**
4. Do not work in groups or copy from other students. The submissions of all students that are involved in a plagiarism case will not counted (independent of who copied from whom)
5. You are not allowed to use any library to do the most of the tasks
6. You are allowed to use math, NTLK and/or spacy

In this assignment, you are provided with a file 'Article.txt'. This file contains an article about covid-19 from CBC newspaper. The goal of this assignment is to summarize automatically the given text using TF-IDF algorithms.

1 Tasks

1. Convert the input text to a list of sentences. Then, compute the number of sentences in the given Text.
2. Calculate the frequency of words in each sentence: the output should be a dictionary where each key is a sentence and the value is also a dictionary of word frequency.
3. Calculate Term frequency for each word in a sentence:

TF(word)=(Number of times term “word” appears in a sentence) / (Total number of terms in the sentence)

4. Create a matrix termFrequency: The termFrequency matrix should be a dictionary where each key is a sentence and the value is also a dictionary of word frequency.

5. For each word compute how many sentences contain that word.

6. Calculate IDF for each word in a sentence.

IDF(word) = log_e(Total number of sentences/ number of sentences with term word in it)

7. Compute the TF-IDF for each word in each sentence.

8. Use the TF-IDF computed in (7) and give a weight for each sentence.

9. Threshold: compute the average sentence weight

10. Generate the summary : select a sentence for summarization if the weight of the sentence exceeds the threshold.