**Axel Dzeukou(40089940)**

**PROJECT 4-REPORT**

**Notes:**

**Assumptions**

-For project 4 extracting the text from each webpage is is extracting all the text in between the html tags <p></p>

-the code only extracts links to web pages and if it is a web page that cant stand on its own the string 'https://www.concordia.ca' is appended to the link to make it a complete webpage that can be scrapped

-if you want to run the code for the project uncomment everything in between comments #for project and #\for project. Else if you want to run code for the demo uncomment everything in between comments #for demo and #\for demo.

-the demo file contains the link to the demo website that the scraping will start from which is ‘https://www.concordia.ca/campus-life.html

-the variable ‘webpage\_beginning’ is different depending on the webpage the scraping is about to start from. It is appended to links in attribute ‘href’ of ‘a’ tags in html to turn the link into an actual web page link

**Results**

After comparing both k\_means with different number of clusters when the number of clusters was set to 3 it had the most positive affine score and lower document spread among clusters and when it was set to 6 it had the most negative score. Sentiment analysis is useful to check if a string of text is overall positive or negative. Crawling was a bit difficult at first but now its fine

**Output:**

FOR PROJECT

For 3 clusters:

[0 0 0 1 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 2 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0

0 0 1 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

For 6 clusters:

[2 0 2 5 2 2 2 2 2 2 2 2 3 2 2 1 1 2 2 2 2 2 2 0 2 5 2 2 2 2 2 2 2 2 3 2 2

1 1 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 0 2 5 2 2 2 2 2 2 2 2 3 2 2 1 1 2 2 2 2 2 2

0 2 5 2 2 2 2 2 2 2 2 3 2 2 1 1 2 2 2 2 1 1 1 2 2 2 1 1 1 1 1 1 2 2 2 2 1

2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2 4 2 2 2 2 2 2 2]