Durham University

ASSIGNMENT

Data Science – Data Collection and Data cleaning

The file keywords.xlsx includes a column of keywords (from the 2nd row to 11th row).

	A	B	C
1	Keywords		
2	targeted threat		
3	Advanced Persistent Threat		
4	phishing		
5	DoS attack		
6	malware		
7	computer virus		
8	spyware		
9	malicious bot		
10	ransomware		
11	encryption		

You are asked to investigate distances between those 10 keywords by collecting the data from BBC news, processing and analyzing it. You need to also submit a simple report to explain your algorithm (problem 3) and the visualization of the result (problem 4). The report should be as visual as possible because the visualization can enhance the reliability of your report.

Problem 1. (20%)

Create a Python program that downloads the webpage content (from BBC news) of the top 100 articles relevant to the given keywords from the following url:

https://www.bbc.co.uk/news

(The articles must actually contain the exact meaning of the keyword instead of the text string of the keyword.)

Problem 2. (20%)

Use BeautifulSoup, a library that facilitates scrapping information from web pages, to collect and process the articles contents. Save each article content to any reasonable data structure which can be conveniently use by your algorithm for Problem 3.

Reference:

https://www.crummy.com/software/BeautifulSoup/

Problem 3. (40%)

Create a Python program to calculate the semantic distances between each two keywords which belong to the list of keywords saved in keywords.xlsx. Develop your own algorithm to calculate the semantic distances. You are suggested to use the data you downloaded from BBC news but not limited to those data. You can also use other data collected from other source.

Save the results of this algorithm in another xlsx called distance.xlsx, whose format is as followed.

	A	B	C	D	E	F	G	H	I	J	K
1	Keywords	targeted threat	Advanced Persistent Threat	phishing	DoS attack	malware	computer virus	spyware	malicious bot	ransomware	encryption
2	targeted threat										
3	Advanced Persistent Threat										
4	phishing										
5	DoS attack										
6	malware										
7	computer virus										
8	spyware										
9	malicious bot										
10	ransomware										
11	encryption										

Problem 4. (20%)

Use seaborn (http://web.stanford.edu/~mwaskom/software/seaborn/index.html) to support the data processing and analysis for your algorithm addressing problem 3.

Also, visualize the distance between any two keywords.

(Please put the output of those visual graphs in the report with explanation. Matplotlib is not allowed to be used for problem 4)