**Exercise 9: Web page mining**

**Assignment Specification**

**Description**: This program will extract data from a web page and perform some analysis.

**Input**: No user provided input. Data will be collected from any news website.

**Output**:

Print the headlines

Generate a wordcloud for the words/bigrams in the headlines

Write an interpretation of the results

See details in the Procedure.

**Procedure**:

1. Import the needed libraries: urllib3 and BeautifulSoup
2. Define the target URL and open it, using urllib3
3. Load the page into your “soup”
4. Create an empty list to host the list of words from the headlines
5. Loop into the “soup”, looking for the section with headlines
6. Transform the story heading/headlines into a string first and a list then. You may want to use the following:

* story\_title = story\_heading.text.replace("\n", " ").strip()
* new\_story\_title = story\_title.encode('utf-8')

1. Print the headlines within the loop
2. Remove from the list you created all the non-semantically relevant words (the “stopwords”). Please use the attached file “stopwords\_en.txt” for the list of stopwords. Feel free to update the list, adding words that may be too frequent and – in your opinion – not too relevant (explaining the reason why you want to remove them). Filter out non-alphabetical elements
3. Extract bigram, generating a separate list. Consider bigrams 2 words appearing together more than 2 times in the whole text. Bigrams will be like “word1\_word2”, meaning you will create a new string composed by the 2 words, separated by an underscore (“\_”)
4. Merge the list of single words with the list of bigrams
5. Create a wordcloud with the resulting list. If wordcloud is not available on your computer, either use an online option (see previous assignments) or calculate the sentiment as in previous assignments
6. Write your interpretation in a separate word/pdf document (1-2 pages, including graphs)
7. Submit the py and the word/pdf files.