**EX07: Comparing text**

**Assignment Specification**

The website ProCon.org is collecting some of the most controversial issues in our society. They collect and publish both Pro & Cons quotes and readers’ comments.

Peak one of the following controversial issues, select/copy/paste the Pro & Cons in 2 separate txt files for analysis.

Controversial issues:

* Prescription Drug costs
* Private Prison
* Student Loan Debt
* Pit Bull Bans

Copy the pro and con from the website and paste them in 2 separate txt files.

**Goal of this assignment is to compare the 2 positions (Pro/Con) via the related texts.**

In order to perform the comparison, you will write a Python script to extract key information and then add a layer of interpretation.

You need to create the txt files first, by separately selecting the pro and con, copy, paste them in a text editor and save the 2 files as txt. A preliminary cleaning (before saving the files) is recommended. Please refer to what presented in class for examples.

Once you have the files, the script will do the following:

* Clean the texts by removing
  + the stopwords, using the attached *stopwords\_en.txt* file
  + words shorter than 3 characters
  + all the words that are obviously frequently used (like the name of the issue)
  + the punctuation
  + end-of-line (“/n”) and blank lines
* Using the library "vader", calculate the sentiment for the 2 texts
* Extract bigrams
* Create word clouds for the 2 texts.

Once you performed all the above steps, write a report with your conclusions. Reports...

* will be based on the results the above analyses (e.g.: an interpretation of the words that are most popular/frequent in the 2 texts from the word clouds, sentiment values, the word clouds, …)
* they should be not shorter than 3 pages, including visualizations and results from the scripts, as part of the narrative
* they must be human readable, with a narrative supported by visualizations and results from the scripts.
* They should not contain an explanation of the code, but an explanation of the findings. Check the discussion questions on the ProCon website for the different topic as a base for possible questions your report should answer.

Reports not containing a plain English interpretation of the results will lead to a low grade for this assignment.

Submissions that are essentially identical (either the report or the script) will get the overall grade divided by the number of identical instances (e.g.: assignment valued 100%, 5 identical submissions getting 20% each).

Please note: for this assignment, we will evaluate the similarity based on the most similar of the 2 (script or report).

Submit the report and the python script (2 separate files).