**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.

The question “Should Recreational Marijuana Be Legal?” is one of the most popular issues, mostly because of the current pollical discussions.

We collected the “pro marijuana” comments in the file “con\_ marijuana\_raw.txt” and the “against marijuana” in “pro\_marijuana\_raw.txt”. The 2 files are pure copy/paste/save from the website.

Goal of this assignment is to compare the 2 texts.

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

* Remove from both the files the stopwords, using the attached *stopwords\_en.txt* file
* Remove all the words and characters that are not relevant (and explain why/what/how you did it)
* Remove all the words that are obviously frequently used (like " marijuana")
* Remove punctuation (an easy way is to create a list of punctuation and perform a loop on the text eliminating characters that are in this list. See here for additional options: <https://stackoverflow.com/questions/15547409/how-to-get-rid-of-punctuation-using-nltk-tokenizer>)
* Remove end-of-line (“/n”) and blank lines
* Using the sample in sentiwordnet\_sample.py, calculate the sentiment for the 2 texts. If the script is not working on your computer, you can use a free online service, such as <https://www.danielsoper.com/sentimentanalysis/default.aspx>
* Extract bigrams. The easiest way (not necessarily the most accurate) is to use nltk. Attach the bigrams to the list of words for the 2 texts
* Using/modifying the attached cloud.py script, create word clouds for the 2 texts. If the script is not working on your computer, you can use <http://www.wordle.net>

Once you performed all the above steps, **write a brief report with your conclusions**. The report 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, …). **Report has to be human readable**, with a narrative supported by visualizations and results from the scripts. It should be **not shorter than 2 pages**, including visualizations and results from the scripts, as part of the narrative.

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