**Comp 472** # Mini-Project 2 **Fall 2022**

---

**Due date:** April 19, 2022

**Goal:** NER + Sentiment Analysis

**Input:** SpaCy, sciKit Learn, AP text snippet, aFinn sentiment lexicon, AP text to analyze: APonTrump (Trump suit against Clinton part of longtime legal strategy. By JILL COLVIN, ERIC TUCKER and BERNARD CONDON, 30.3.2022)

**Output:** NER annotation, selected dependency graphs, sentiment labeling of named entities, analysis

**Description:** Working from Lab10, Question 4 and Lab 11, Question 3, run the text snippet through SpaCy for preprocessing and obtain the NER and the dependency graphs for each sentence. You now have certain token (sequences) labelled as a named entity and typed as a person or an institution. You also have the dependency graph, that connects each named entity with words in the graph (each token has exactly one *governor*. Some nodes have one or more dependents, some have none.

Perform sentiment analysis with aFinn (https://github.com/fnielsen/afinn) on all sentences. You now have a prediction whether each sentence is neutral, positive or negative. Through lookup in the aFinn lexicon, you also know the sentiment value of each word (if it is not present, mark the word as neutral, i.e. not carrying sentiment).

Put all tokens and all available features (NE?, NEtype, Governor, ListofDependants, SentimentValueofToken, SentimentValueofSentence) into a table T1 for classification. For a second experiment, put only named entities in table T2 (no other tokens) and reduce the set of features to (NEtype, Governor, ListofDependants, SentimentValueofToken, SentimentValueofSentence).

Perform a k-means clustering on both input tables. Experiment to find a good value for $k$.

Discuss the output. Note that A2 will ask you for more in-depth analysis, here, a brief discussion of the quality of the obtained clusters and a justification of your choice of $k$ will suffice.

**Deliverables:** Submit your well documented code and Project Report in Moodle.

For demonstrating your work in your submission, use the text snippet S1 (when asked): *U.S. intelligence agencies concluded in January 2017 that Russia mounted a far-ranging influence campaign aimed at helping Trump beat Clinton. And the bipartisan Senate Intelligence Committee, after three years of investigation, affirmed those conclusions, saying intelligence officials had specific information that Russia preferred Trump and that Russian President Vladimir Putin had "approved and directed aspects" of the Kremlin's influence campaign.*

1. (1 pt) well documented code (Attrib. 1, 5)
2. Report including
   - (a) (0.5 pts) SpaCy sentence and token splits for S1 (Attrib. 5)
   - (b) (0.5 pts) a graphical representation of the two dependency graphs for S1 (Attrib. 5)
   - (c) (5 pts) $T1_{S1}$ and $T2_{S1}$ for the snippet (Attrib. 1, 5)
   - (d) (1 pt) 3-means clusters for T1 (for the entire text) (Attrib. 1, 5)
   - (e) (1 pt) 2-means clusters for T2 (for the entire text) (Attrib 1, 5)
   - (f) (1 pt) 1 page discussion of of results, observations, issues (Attrib. 6, 7)