Term paper instructions for class: Introduction to Text Mining and Natural Language Processing

Goal of the term paper is to convince me that you have understood the material that we discussed in class and are able to integrate with a topic that you are interested in. You will form teams of up to 3 students to complete the task. You should write between 5 and 10 pages describing main idea of the project, the motivation for your question, the text data you used and results.

I am like Trump. I function through Figures. Good figures are key. Don't make figures that are not necessary and you cannot explain. This is a general point: being able to "explain" your own results is key – don't show things you don't reason about.

The presentation of the project takes place on the 12th of March. In your presentation you are supposed to discuss the main idea of the project, the motivation for your question, the text data you want to use (and results). I will ask questions and try to give feedback. Part of the presentation is that you write my feedback to me in an email.

Hand-in of the pdf containing the term paper, the codes used and the raw data you used is the 24th of March. Send one zip file with everything in. If the data is large do not send but you should send jupyter notbooks that allow us to follow the code output if possible.

Additional points:

I know you have little time. Hand-in is in a month from now but you have several term papers. So effectively you have like 5 days to work on this. But people have done AMAZING thing in this course so my bar is high (sorry).

Creativity wins. Human creativity will be rewarded – especially because it is hard with so little time.

Optimally you use one of the methods we discussed in class to generate the text (i.e. scraping, reading files, pre-processing, tf-idf, dictionaries, LDA, supervised learning). Do NOT use LLMs.

Optimally you do not only show something descriptive but derive some testable hypothesis that you test on the data you generate, or show the performance of a classifier you trained on the text data. I dislike clueless descriptive work even if technically done well - stay clear of ChatGPTable stuff.

Optimally you pay special attention to pre-processing steps and discuss these in your code and paper and do some analysis that shows what changes when you change pre-processing steps.