Comp 479/6791 # Lab Challenge 1 **Fall 2019**

You have to perform the following experiments during lab time. You may confer with fellow students in your class, you may search on the internet (including stackexchange), but you are not allowed to ask for outside help. You have to submit your code and a report. The report has to answer the questions, give the experiment results you obtained, clearly specify how to run the code and document the code (in addition to the inline documentation!).

---

1. Download the Reuters collection Reuters-21578. How many documents does reut2-020.sgm contain? How many does reut2-021.sgm contain? How can you find this information?

2. Break reut2-021.sgm up into articles, associating the NEWID with the article text.

3. Investigate, how to extract the text from the article document.

4. Answer the following question: For each of the orgs in all-orgs-strings.lc.txt, determine how often they occur in the corpus. How did you find your answer? Try another way. Do you get the same answer? What is the difference (time, effort,...)?