COMP 479/6791 **Lab Assignment 2** Fall 2023

Objectives: Downloading a text collection from an outside source, preprocessing the collection.

Getting started:

1. Download the Reuters collection Reuters-21578 (http://www.daviddlewis.com/resources/testcollections/reuters21578/). 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,...)?

Advanced:

1. What is a spliterator?
2. Using the Reuters collection Reuters-21578, use Java lists to associate each of the places in all-places-strings.lc.txt with all the NEWIDs of articles in which that place occurs. You may use the metadata or use the main text
3. Experiment with streaming the data
4. "Clean" the text using NLTK in a nondestructive fashion (i.e. without loosing the metadata)
5. Tokenize reut2-021.sgm using NLTK. What do you observe?