COMP 479/6791 **Lab Assignment 1** Fall 2023

Objectives: Getting started with the NLTK environment. Exploring the Reuter's corpus. "Text cleaning"

Due: second week of September

Readings: Chapters 1,2,3 in NLTK book: *Natural Language Processing with Python — Analyzing Text with the Natural Language Toolkit*, by Steven Bird, Ewan Klein, and Edward Loper(https://www.nltk.org/book/)

Description:

1. install NLTK3, the Natural Language Toolkit
2. download the version of the Reuters corpus available in NLTK
3. In the Reuters corpus, what are the number of
   (a) documents? (b) words? (c) sentences?
4. For the text with fileID 'training/9920', determine the number of
   (d) words? (e) single word prepositions? (see
   https://dictionary.cambridge.org/grammar/british-grammar/prepositions)
5. In NLTK, create a table that lists fileIDs for each of the 90 categories. Retain a copy of this index.
6. Write a function word_freq() that takes a word and a fileID, and computes the frequency of the word in that file.
7. Download NLTK Reuters and

   inspect the files (maybe in Emacs or textedit, etc). Is a file equal to a newspaper article?

   inspect a newspaper article (maybe 9920?). Are all the characters part of the printed article?

   begin data preprocessing (or cleaning). What is the data that is not part of the printed article? Why is it there? Can you extract the original printed text? Can you do it without losing the extra information? What could that be used for? (This component will not be tested later but is essential now.)

Outcome: Compare results with colleagues during lab time.