OS4118 **Text Mining Example** Fall AY 2020

**Introduction:** In this exercise we will construct a “corpus” (a set of documents) and use text mining tools to try to analyze it.

**The Data:** I have extracted the abstracts from all the NPS OR department Master’s theses from 2014 to 2018 (except for a few items that I couldn’t process for whatever reason). I believe there are 181 of these.

[Note: This is a web scraping exercise, in which I had to identify the names of the documents, which are PDFs; download them; convert them to text; and then figure out where the abstract started and ended. This is the sort of thing we analysts sometimes need to be able to do!]

Those abstracts are available on Sakai under Resources | Data | OR.Abstracts.zip. Each abstract is roughly 150-200 words long.

**Software:** Install the tm, cleanNLP and topicmodels packages first.

**Tasks:** Converting the abstracts into a corpus – specifically, a VCorpus like the crude data we looked at in lecture – in straightforward. But I have something smarter (?) in mind. We will first extract just the nouns, proper nouns, and verbs, and use lemmatization to extract just the stems of those words. Then we will build a corpus out of those words.

With the corpus in hand, we will then create the document-term matrix (DTM). When I did this, that matrix was 181×2781, but yours might differ.

Then we can perform LDA on the DTM and try to see if the abstracts get assorted into identifiable topics. Do they? Remember that the @gamma item in the LDA output shows the strength of each topic within each document (so for four topics, this will be a 181 × 4 matrix), while the @beta shows the strength of each word within each topic (so this would be 4 × 2781).