# CMPT 733 Big Data Programming – Final Project

## Call for Project Ideas

**Title of the Project: Topic modeling and visualization of news comments**

**Description:**

The goal of this project is to extract topics from news articles and their comments, and to perform visualizations of the results.

Specific project ideas:

1. Explore and visualize which are the top topics covered by the articles, and which are the top topics by the comments, and what is the overlap.
2. Identifying what **entities** (including persons, issues, policies, places, and organizations) from the article get mentioned a lot in the comments. This will involve Named Entity Recognition and coreference resolution.
3. Clustering comment authors based on their commenting style or topics of interest. For instance, clustering comment authors based on article topics or most prominent entities in the article they comment on.

Other ideas:

- threads (which topics have the most embedded threads)
- article authors and comments (which authors get the most comments)
- sentiment, toxicity
- multiple visualizations

**Datasets**

SFU Opinion and Comments Corpus

The SFU Opinion and Comments Corpus (SOCC) is a corpus for the analysis of online news comments. Our corpus contains comments and the articles from which the comments originated. The articles are all opinion articles, not hard news articles. The corpus is larger than any other currently available comments corpora, and has been collected with attention to preserving reply structures and other metadata. In addition to the raw corpus, we also present annotations for four different phenomena: constructiveness, toxicity, negation and its scope, and appraisal.
The dataset contains isolated comments and comment threads posted in response to The Globe and Mail opinion articles. We have maintained two versions of the corpus. A cleaned version with minimal repetition of comments and an original version containing comment-thread information.

The clean version has ~663K comments, with ~273K top-level comments. This version has two CSVs: articles.csv and gnm_comments.csv.
Available from: https://github.com/sfu-discourse-lab/SOCC

**Contact person:**

Maite Taboada, mtaboada@sfu.ca or Varada Kolhatkar, vkolhatk@sfu.ca

**Contributor of the Project Idea:**

Maite Taboada
http://www.sfu.ca/~mtaboada/
http://www.sfu.ca/discourse-lab.html