CMPT 318 - Special Topics in Computing Science

Spring 2019

Instructor: Dr George Weir (gweir@sfu.ca)

Outline

Text analytics as a sub-domain of data mining, the role of corpus linguistics, types, tokens, n-grams and parts-of-speech, readability, sentiment analysis and text classification. The rise of Big Data and the recognized potential of data mining and machine learning, have increased research attention on collections of text as a readily available data type. In this course, students will be introduced to concepts and techniques for textual analysis, both as a basis for simple text mining and as input to text classification problems. Topics will include the roles and application of corpora, Unix-based techniques for quantitative analyses of textual data, part-of-speech tagging, document forensics, readability and sentiment analysis. * Electronic readings will be assigned.

Prerequisites

CMPT 225. Additional prerequisites to be determined by the instructor subject to approval by the undergraduate program chair.

Syllabus

- The nature of textual analysis
- Corpus linguistics
- Words, types and tokens
- Working with n-grams
- Features, meta-data and quantification
- Stemming and lemmatization
- Investigating a kidnapping
- Varieties of words
- Readability
- Sentiment analysis
- Text classification

Grading

Midterm exam 30%; Term paper 20%; Final exam 50%.