Text Mining Using R

Syllabus

Phoebe W. Ishak

Course's objective

Much of the data proliferating today is unstructured and text heavy. Yet, text data became increasingly useful to analyze public opinion and shed lights on economic, political and social preferences. That became much easier with the digitalizing of newspapers and the extensive use of social networks by both citizens and politicians. Think of Twitter becoming an important hub to communicate government policies, while allowing general public to express their reactions to these policies. But one question remains, that is how to analyze such data, especially if these data come in millions of unstructured and untidy text forms?

The objective of this course is to give students the computational tools to retrieve, manage and reshape text data. We will learn how to obtain data from either text files or from the web, and how to manipulate it, so that it can be used in practical applications. This covers text cleaning, search of patterns, text analysis, and data visualization. In the next step, students will learn how to perform sentiment analysis through understanding of the emotional intent of words to infer whether the text is positive or negative. Finally, we will introduce topic modeling as a method for unsupervised classification of text data, similar to clustering on numeric data.

Requirements

All codes and exercises will be written and executed using R. Some basic knowledge of R is recommended, but not required.

Outline

The course is organized in 4 three-hours sessions over 2 weeks period. We shall cover the below topics:

Lecture 1: Introduction to R and text data
Lecture 2: Text and sentiment analysis, and data visualisation
Lecture 3: Topic modelling and other data classification
Lecture 4: Web scrapping

Course materials

Lectures' notes and other materials are available on AMeTICE.

Grading

Students will be graded on the basis of continuous assessment that consists of handing in at least 4 take-home assignments covering the different topics discussed.

References

Paradis, Emmanuel. "R for Beginners." (2002).

Wickham , Hadley and Grolemund , Garrett (2016). “R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.

Silge, Julia, and David Robinson (2016). “tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” JOSS 1 (3).

Sanchez, Gaston. “Handling and processing strings in R”. Berkeley: Trowchez Editions (2013).