Capstone Project 2 – Topic Modeling for Economic News Article

Milestone Report

1. Introduction

This project deals with the problem of building a machine learning model for news article classification – clustering the articles into a number of groups based on the underlying topic of each article. Therefore, the clients this project serves are the new agencies. The dataset used in this project is obtained from the *Data For Everyone Library*[1], which contains 8000 news articles related to economics ranging from 1951 to 2014. The sources of the articles are Wall Street Journal and Washington Post. Topic modeling is used to classify the economic articles into different categories and determine the membership of them by discovering the structure in the corpus.

2. Data Preprocessing

Preprocessing the text data is the essential step in building a good topic model. The following steps are performed:

- Remove numbers and special characters: all the kept words are transformed to lowercase
- Remove superfluous content: at the beginning of some articles, there is author information, e.g. "*Author: Author Name*", news source, e.g. "The Wall Street Journal Online", or location information of the news (which is often reiterated in the news content), e.g. New York or Washington. The content as such is removed.
- Lemmatize the tokens: the inflected forms of a word are grouped a single term. A lemmatizer is used in this case other than a stemmer since the words after lemmatization still remain readable.
- Remove the stop words and filter out the extreme words: the words in NLTK's dataset of English stop words and the words that appear in less 5 or more than 50% of the articles are filtered out.

There are 10558 unique tokens in the final dictionary created.

[1] https://www.crowdflower.com/data-for-everyone/