Project Proposal: Latent Topic Modelling for Financial News

Siddharth Nand

Introduction

In this project, I propose a latent topic model to identify the underlying themes in a collection of financial news articles. The goal is to develop a model that given a news article/news heading, can identify the probability of the article belonging to each of the predefined categories.

Team: Individual

Project Theme: Topics models

Repo Link: https://github.com/sidnand/Topic-Modelling-Financial-News

Potential Approaches to the Problem

I want to model this problem using a variation of Latent Dirichlet Allocation (LDA), but for supervised learning. For each label y_i, the prior distibution of each would be uniform, so $\mathbb{P}(y_i) = \frac{1}{k}$, where k is the total number of labels. The posterior would be $\mathbb{P}(y_i | x_1, \dots, x_n)$, where $x_1, \dots, x_n$ are the n words in a document. Then for updating the posterior, for each word in the document, I'd update the probability of each label based on the observation. If the word belongs to a document with label y_i increase the probability of y_i linearly by the number of documents with label y_i that contain the word.

I would also need to do a lot of pre-processing of the data, such as removing unnessary words. I also need to structure the data such that each column is a label, a row is a word and the value at word i and label j is the proportion of number of times word i appears in documents with label j. If a word does not appear in a document with label j, then the value will not be 0 due to Cramwell's rule, but will be a small number.

Datasets

I have two proposed datasets. One is the Reuters-21578 dataset, which is a collection of 21,578 news articles, which labelled by predefined topics. The second is a collection of 211 financial news headings, each labeled by a number from 0 to 2, where 0 means bad news, 1 is neutral news and 2 is good.

Reuters-21578 Dataset:

```
## # A tibble: 6 x 2
##   text                                                  topics
##   <chr>                                                 <chr>
## 1 "Mounting trade friction between the\nU.S. And Ja..." ['trade']
## 2 "A survey of 19 provinces and seven cities\nshowe..." ['grain']
## 3 "The Ministry of International Trade and\nIndustr..." ['crude' 'nat-gas']
## 4 "Thailand's trade deficit widened to 4.5\nbillion..." ['trade' 'grain' 'rice'~
## 5 "Indonesia expects crude palm oil (CPO)\nprices t..." ['veg-oil' 'palm-oil']
## 6 "Tug crews in New South Wales (NSW),\nVictoria an..." ['ship']
```

News Headings:

```
## # A tibble: 6 x 2
##   text                                                                 sentiment
##   <chr>                                                                <chr>
## 1 The housing market tells the whole story                             1
## 2 U.S. equity futures were trading higher the morning after the S&P 5~ 2
## 3 LG Display posts strong Q1 profit on panel price boost               2
## 4 Brazil’s Covid inquiry puts Bolsonaro on the back foot               0
## 5 Indian doctors protest against guru who claims yoga can defeat Covid 2
## 6 Slower population growth, ageing need not be alarming: economists    1
```