Topic Modeling for Economic News Articles

Data Wrangling

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 Wranling

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

1) Remove numbers and special characters; all the remaining words are transformed to lowercase.
2) 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 (e.g. New York or Washington, which is often reiterated in the news content). The content as such is removed.
3) Lemmatize the tokens to group the inflected forms of a word into a single term. A lemmatizer is used in this case other than a stemmer since the words after lemmatization still remain readable.
4) In order to apply unigram in topic modeling, only the nouns are kept as they are more useful in revealing the subject of the articles compared with the other forms of words, for instance, verbs and adjectives.
5) The set of English stop words from *nltk.corpus* are filtered out from the list of tokens. After step (4), most of the stops were already removed, and this step ensures that the tokens contain no stop words.
6) To further refine the quality of the remaining tokens, the 50 most frequent words in the collection of these 8000 articles are reviewed:

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

No.	Word	Occurrence	No.	Word	Occurrence	No.	Word	Occurrence
1	year	10189	2	market	8498	3	rate	8151
4	stock	7669	5	price	5592	6	economy	4097
7	interest	3985	8	bank	3978	9	month	3794
10	company	3762	11	week	3526	12	percent	3488
13	inflation	3256	14	investor	3048	15	time	2988
16	point	2952	17	government	2831	18	dollar	2818
19	president	2806	20	tax	2723	21	bond	2690
22	yesterday	2621	23	growth	2581	24	day	2567
25	reserve	2463	26	share	2454	27	fund	2439
28	state	2399	29	index	2344	30	business	2326
31	york	2296	32	quarter	2194	33	money	2148
34	increase	2106	35	job	1984	36	consumer	1929
37	fed	1904	38	report	1894	39	deficit	1869
40	policy	1855	41	sale	1845	42	economist	1805
43	trading	1797	44	analyst	1768	45	budget	1751
46	term	1750	47	today	1713	48	people	1699
49	security	1646	50	dow	1639			

Table 1: Word Count for the Top 50 Words

The following list of words from the 50 most popular words are excluded as they are considered superfluous:

['day', 'today', 'yesterday', 'week', 'month', 'quarter', 'time', 'percent', 'rate', 'point', 'economy', 'economist', 'growth', 'increase', 'york', 'report', 'analyst', 'term', 'people']

7) Eventually, the rare words that appear in less than 5 articles and the frequent words that appear in more than 50% of the articles are filtered out. On top of step (6), this step ensures the common words based document frequency are removed.

As the result of the above steps, there are 5293 unique tokens in the final dictionary created.