UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI
UNDERGRADUATE SCHOOL

Research and Development

BACHELOR THESIS

by

PHAN Manh Tung

USTHBI8-160

Information and Communication Technology

Title:

Topic Modelling and Trend Detection

Supervisors: Dr. DOAN Nhat Quang

Lab name: ICT Lab

Hanoi. July 2020

Declaration

I would like to declare this thesis totally belongs to my work under the guidance of Dr. Doan Nhat Quang. I certify that this work and the according results are honest and unprecedented previously published in the same or any similar form. Furthermore, any assessment, comment and statistics from other authors and organizations would be indicated and cited. If any fraud is found, I would take full responsibility for the content of my thesis.

Lời cam đoan

Tôi cam đoan luận án này hoàn toàn thuộc về tôi dưới sự hướng dẫn của tiến sĩ Đoàn Nhật Quang. Tôi xác nhận công việc này và những kết quả tương ứng hoàn toàn trung thực và chưa từng được xuất bản trước đây dưới bất kỳ hình thức nào tương tự. Ngoài ra, bất kỳ đánh giá, nhận xét, thông số nào được lấy từ tác giả hay tổ chức nào khác đều được chỉ định và trích dẫn. Nếu có bất kỳ sự gian lận nào bị phát giác, tôi xin chịu hoàn toàn trách nhiệm về nội dung luận án của mình.

(PHAN Manh Tung)
Hanoi, July 2020

Acknowledgement

I would love to express great thanks to my research supervisor Dr. Doan Nhat Quang for his meticulous guidance and support during the whole period of the internship. With his careful instruction, I not only successfully completed the research topic but also learnt a lot of new knowledge in the field of natural language processing. Furthermore, I also appreciate having been a USTH student who has all the advantages of being supported by all admirable professors, staffs. Last but not least, I want to show great gratitude to my family who backed me up during my university study period.

Lời cảm ơn

Tôi xin được gửi lời cảm ơn chân thành đến tiến sĩ Đoàn Nhật Quang đã tỉ mỉ hướng dẫn, chỉ bảo cũng như hỗ trợ tôi trong suốt kỳ thực tập. Với sự dẫn dắt vô cùng cẩn thận, tôi không chỉ thành công hoàn thiện dự án, mà còn được đào sâu vào lĩnh vực xử lý ngôn ngữ tự nhiên. Ngoài ra, tôi rất trân trọng khi được là sinh viên của trường Đại học Khoa học và Công nghệ Hà Nội, có được sự trợ giúp nhiệt tình từ những thầy cô giáo đáng kính cũng như các anh/chị nhân viên ở các phòng ban. Cuối cùng, xin được thể hiện sự trân trọng của mình tới gia đình tôi, nhưng người thân đã hỗ trợ tôi cả về vật chất lẫn tinh thần trong suốt quãng thời gian học đại học.

(PHAN Manh Tung)
Hanoi, July 2020

Contents

List of Acronyms

API Application Programming Interface. 4

LDA Latent Dirichlet Allocation. 2

LSTM Long short-term memory. 2

RNN Recurrent Neural Network. 7

TF-IDF Term Frequency – Inverse Document Frequency. 11

WCSS Within Cluster Sum of Errors. 14

List of Figures

Chapter 1

Introduction

1.1 Definitions

In the modern era, there has been a burst of new data and information, especially from the Internet. For instance, common social media platforms such as Facebook, Instagram, TikTok produce paramount amounts of data every single day through comments, status and posts. In consequence, the demand for taking advantage of the aforementioned massive unstructured data to find patterns, generate useful insights is rising in all areas. Text mining techniques, which use large collections of text from various formats, such as web pages, emails, social media posts, journal articles, in order to extract significant information and get useful insights, have become increasingly widespread in this modern era. Some practical applications of this technology include knowledge discovery, risk management, resume filtering in business, email spam filtering, fraud detection, social media analysis for daily purposes and many more.

A text mining process is to apply various techniques such as categorization, entity extraction, sentiment analysis and natural language processing to transform the text into useful data for further analysis. When dealing with a large amount of corpus, text mining could be implemented to turn unstructured data into more accessible and useful forms, so as to extract hidden trends, patterns or insight *(Team. 2016)*. One of the common text mining techniques called topic modelling, which clusters word groups and then formalize them into different topics, is the major step in this thesis. These are followed by trend detection, which is a process to determine how ubiquitous each topic is during a span of time. Lastly, time-series forecasting is a procedure of predicting future values based on the observation of past data points, which is resulted from the trend detection task.

Text mining is becoming a powerful tool for any organization because it provides the capability of digging deeper into unstructured and complex data to understand and identify relevant business insight. As a result, with the help of text mining, many businesses are able to fuel their own business processes or to form their own strategies for market competition *(Team. 2016)*. Furthermore, in this day and age, the amount of information is significantly growing and diversifying. Any organisation that could conquer and automate these resources would take great advantage to effectively compete in every field.

In this context, we are interested in politics due to the data available in the ICTLab. Thus, the collection of articles is chosen totally from the political area, though various international

sources. The main reason for choosing only one particular field is to make the analysis process become simpler and more effective. In the future, the study is expected to be into a broader number of fields and also with more complicated real-world databases. Besides, the duration of the study is a three-month timespan, within my internship period.

The internship objectives have two-folds:

- Finding major topics in the big collection of articles during a period of time.
- Predicting the changes in topic trending (increasing, decreasing or fluctuating) one to two week ahead.

In order to achieve the objectives, some common text mining techniques would be applied: for topic modelling part, K-means clustering is the main algorithm for dividing the data into several groups, and LDA is another method is implemented simultaneously to evaluate the results of K-means algorithm. Here, we have to take into account the fact that due to the data nature, normal clustering methods cannot be applied directly for text data. Afterwards, the trend detection problem would be solved using bar chart visualization and a deep learning technique for time-series prediction named LSTM to forecast the upcoming trend of each topic.

1.2 Objectives

The internship objectives include:

- Preprocess the huge corpus of text data into usable forms, eliminate unnecessary words and characters for further analysis.
- Identify how many topics there are in the big collection, group those with the same topics and label each article accordingly.
- Determine the trend of each topic over time. The ubiquity of each cluster is determined by the number of daily articles on a specific topic.
- Predict the future quantity of articles on each topic will be produced in the next couple of days. Verify and calculate the accuracy of the predictions.
- Visualise results from each process with appropriate visualization tools.

1.3 Thesis Structures

The structure of this thesis report is as follows:

- Chapter 1: Introduce definition, the importance of the research topic, scope, aims and major problems of this thesis.
- Chapter 2: Introduce all pre-built python packages and raw data and all main methods that are used in this project.
- Chapter 3: Propose a framework of the project and describe all the implementations in a logical sequence and obtained results.
- Chapter 4: Briefly summarize the experiment, results, discussion and future expectation.

Chapter 2

Material and Methods

2.1 Python Packages

2.1.1 pandas

pandas is one of the most common packages in python, which is used for data analysis and data manipulation. The key structures in pandas are Series and DataFrame which is in form of tabular data in rows of observations and columns of variables.

2.1.2 matplotlib

matplotlib is a visualization tool in python, which creates charts and graphs in order to visualize data.

2.1.3 numpy

numpy is a fundamental library in python, it is a scientific computing package. It provides multidimensional array, matrices and a handful of mathematical operations.

2.1.4 nltk

Natural Language Toolkit (nltk) is a leading package for natural language processing, working with human language. It provides a variety of techniques such as classification, tokenization, stemming, tagging, parsing, and semantic reasoning, and many more.

2.1.5 gensim

gensim is also a natural language processing package, a python toolkit developed by Radim Řehůřek, specifically for unsupervised topic modelling technique call Latent Dirichlet allocation (LDA).

2.1.6 spaCy

spaCy is an NLP package which resembles nltk, but in a higher industrial level, providing advanced techniques for processing human language.

2.1.7 collections

This package provides common data structure forms and useful calculation within them. In our project, we use Counter structure for word elimination.

2.1.8 sklearn

scikit-learn is one of the most common packages for machine learning in python, which provides various models for classification, regression (supervised learning); clustering, dimensionality reduction (unsupervised learning); and some model selection and data preprocessing techniques.

2.1.9 TensorFlow

Created by the Google Brain team, TensorFlow is an open-source library for numerical computation and large-scale machine learning and deep learning algorithms.

2.1.10 Keras

Keras is a high-level API for TensorFlow - one of the most universal packages for deep learning. It provides short, comprehensive functions to implement complicated artificial neural networks. Some highlighted modules within the package are neural layers, cost function, optimizers, initialisation schemes, activation functions, regularization schemes.

2.1.11 pickle

This library provides functions for saving tabular data, graph or any processing results into pickle files, which could be opened later in further processing steps.

2.2 Raw Data

Our raw data consists of over 28 thousands articles coming from a handful of ubiquitous international journal/newspapers, namely VOA Voice of America, The Guardian, Associated Press, etc, during the span of three months starting from the end of August to the beginning of December.

The statistics details below demonstrate the distribution of our raw data from various press agents:

	Number of Article
Agence France Presse (AFP)	1669
Associated Press	9617
CNN International	283
RFI	167
Radio Free Asia	379
Reuters	5879
The Diplomat Magazine	511
The Guardian	6380
VOA Voice of America	3212
globaltimes.cn	182

Figure 2.1 – Statistical table depicting the number of articles from each source.

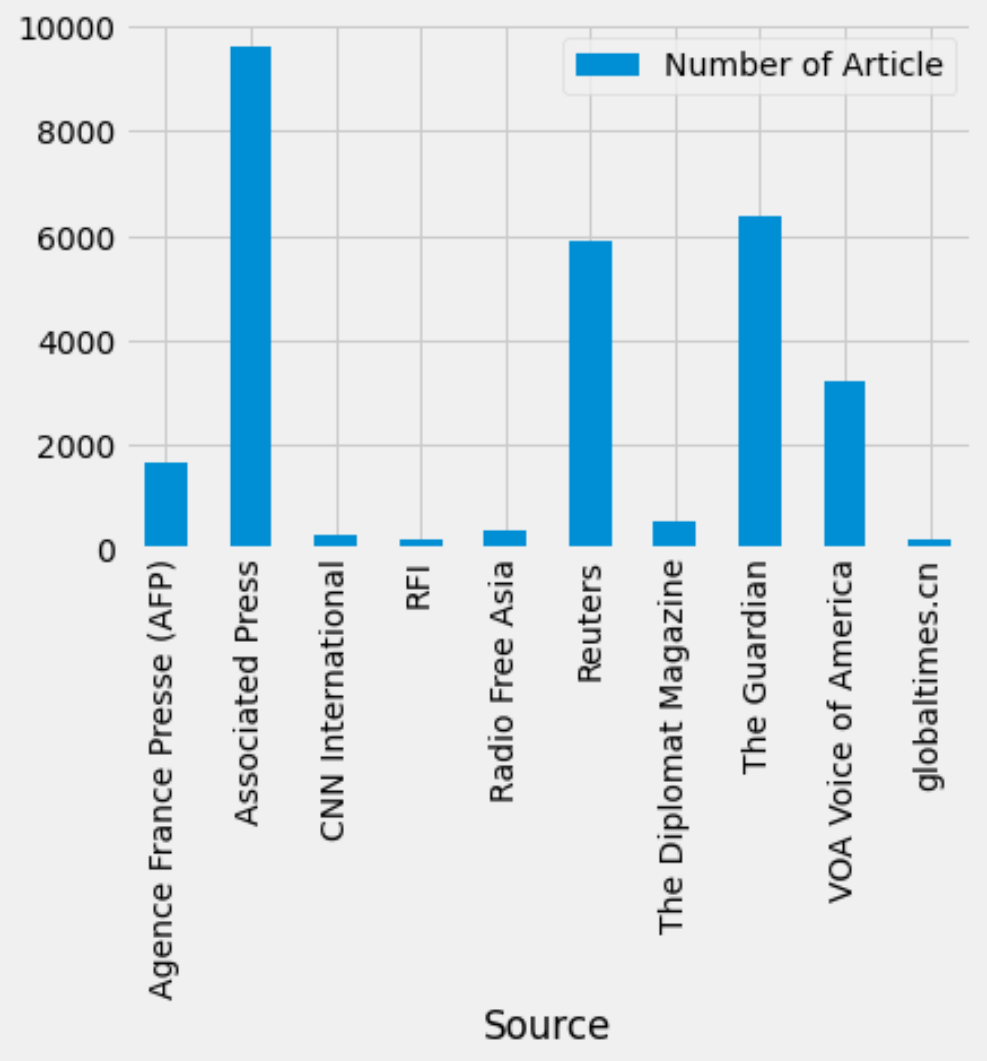


Figure 2.2 – Bar chart represents the distribution of raw data.

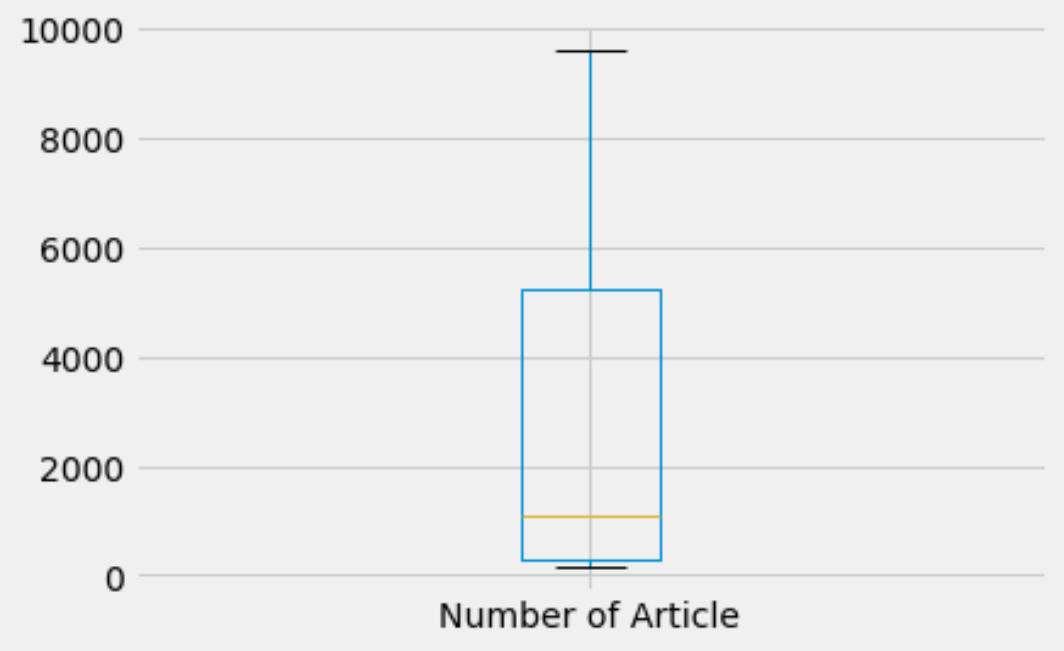


Figure 2.3 – Box plot represents the statistical distribution of raw data.

It is easily recognized that the Associated Press, Reuters and The Guardian accounted for the largest proportion of the data. In Figure 2.3, the median is just around 1000 articles, indicating the majority of data sources contain little amount of data.

2.3 Methods

Our major methods consist of two common clustering techniques and an artificial neural network technique for time-series prediction.

2.3.1 K-means clustering

K-means clustering is an unsupervised machine learning algorithm. The objective of K-means is to group similar data points together to recognize underlying patterns. To achieve the target, the K-means algorithm attempts to find out a fixed number (k) of clusters in the dataset (k refers to the number of centroids). Afterwards, K-means tries to allocate all the data points to their nearest centroids, creating k different clusters *(Dr. Michael. 2018)*.

2.3.2 LDA

Latent Dirichlet Allocation (LDA), implemented using python toolkit gensim, is a generative statistical model that could discover hidden topics in various documents using probability distributions *(Alice. 2018)*.

- Choose a fixed number of topics n.
- Randomly assign each word in each article to one out of n topics.
- Go through every word and its topic assignment in each document. Based on the frequency of the topic in the article and the frequency of the word in the topic overall, assign the word to a new topic.

- Go through multiple iterations of this process.
- After the whole process, we get 10 words representing each topic and the probability distribution of them.

The main reason for this method is to compare the clustering results with K-means, with different values of k to finalize the most appropriate number of topics within the article collection.

2.3.3 LSTM

The chosen time-series prediction method is LSTM, which stands for Long Short Term Memory, is an upgraded version of the Recurrent Neural Network Recurrent Neural Network (RNN) - a powerful deep learning technique in dealing with sequential data. The common problem in the traditional RNN is Vanishing Gradient Problem, in which the very beginning memory is lost when progressing along a fairly long sequence, is effectively solved with the LSTM gate ideas. Therefore, the deep learning technique LSTM has been one of the most effective algorithms to process sequential data recently.

Chapter 3

Proposed framework

3.1 Framework

Below is the proposed framework of the thesis. The large collection of raw data will go through 4 major proposed steps in order to achieve expected results.

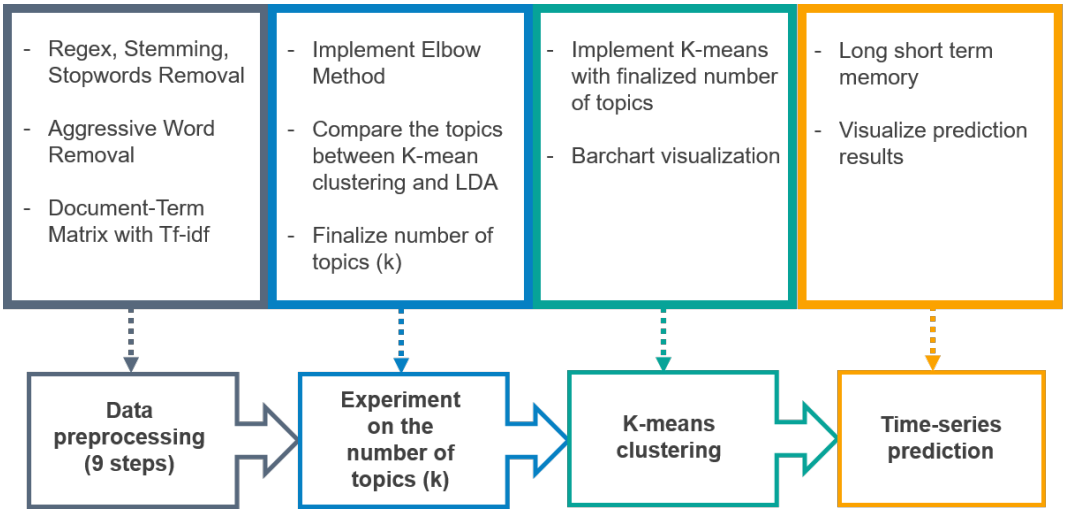


Figure 3.1 – Proposed framework of the text mining project

At first, the corpus coming from 28k articles is huge and almost words do not contain the important information to formulate major topics. Besides, due to the large scale of our dataset, it is nearly impossible for us to implement any models without reducing the magnitude of the corpus. Therefore, preprocessing data is an essential process to proactively reduce the corpus, get rid of all unnecessary information and turn the corpus into an appropriate form to fit into models.

Our experiment is mainly based on K-means clustering algorithm, with the aim of finding major topics among the articles. After the preprocessing steps, we implement the Elbow Method to find out which number is best for determining the number of clusters/topics in the collection. Then, the K-means algorithm is executed with different k to obtain different results. Afterwards, we use the LDA topic modelling method to evaluate the results from K-means and conclude the best k and visualize the final answer in a stacked bar chart.

An expansion of the project is acknowledged after the trending data is obtained. The clus-

tering process gives us typical time-series data (quantities of produced articles in each topic for 90 days). Thus, we have enough quantity of data points for future prediction. LSTM implementation is the final procedure in this project.

Chapter 4

Experiment and Discussion

4.1 Data preprocessing

4.1.1 From raw data to pandas dataframe

The first step is to take the 28k articles from 28k different files and transfer them into one pandas dataframe consisting of 5 columns for 5 attributes: Article ID, Source, Date, Title, Body. This process is rather slow because of the huge amount of data. After the process, the dataframe is saved into an excel file for conveniently calling it out later.

4.1.2 Regex

Regular expressions (or regex) is a widespread programming tool used for purposes such as feature extraction from text, string replacement or other string manipulations. A regular expression is a set of characters, or a pattern, to find substrings in a given string. For instance, extracting all hashtags from a tweet, eliminating all numbers from large unstructured text content *(Niwratti. 2019)*.

Our implementation includes making text lowercase, removing text in square brackets, removing any punctuation or any special symbols, removing all numbers and any words containing numbers, getting rid of blank lines. The more necessary regular expression filters are applied, the cleaner and better the corpus becomes.

4.1.3 Stopword Removal

The most commonly used words in the English language is called stop words, whereas these words do not contain any important meaning and ideas. For instance, ‘a’, ‘the’, ‘is’, ‘in’, ‘for’, ‘where’, ‘when’ are stop words *(SINGH. 2019)*. From this point, we easily realize that most of the words that are adverb, preposition, conjunction, determiner could be considered as stop words, because these contain less or no meaning in a specific context. Even verbs can be listed on the stop word collection.

Because our corpus, which contains over 28k articles, is extremely large, the more words we could eliminate, the faster and easier the implementation of later models would be. This motivation leads

to extremely aggressive word removal processes later. But at first, common stopword removal is implemented, with the help of spaCy package - there are 326 stop words in the collection.

4.1.4 Counter-less-than-k Word Removal

An aggressive move for more word elimination. With the help of library collection in python, any words that are counted less than number k are excluded from the corpus. This algorithm is also referred as K-core algorithm. As the result, we can filter out over 55k words from the corpus.

The risk of this idea is the possibility of important context-containing words to be filtered. Due to this issue, we make a careful choice that the number of k is equal to 5, which is a pretty small number in order not to damage the final result of the analysis, because those words only appear 5 times in the whole collection.

4.1.5 Stemming

Stemming is a text normalization technique that reduces any words to their root form, in other words, eliminates all prefix, suffix or infix of the words. For example, a list of words including "interesting", "interested", "uninterested", interestingly" would be reduced to the root form of "interest" through the stemming process *(Hafsa. 2018)*.

Our chosen method is Snowball Stemmer, which actual name is English Stemmer or Porter2 Stemmer. It is an improvement over the most common stemmer - Porter Stemmer, with more precision. The implementation of stemming is via the nltk package in python.

4.1.6 TF-IDF

TF-IDF stands for term frequency-inverse document frequency. TF-IDF weight is a statistical measure for determining how significant a word is to a document in a corpus. The significance rises proportionally to how many times a word appears in the document but is offset by the frequency of the word in the corpus. With the ability of importance evaluation, TF-IDF can be used for stopwords filtering in text summarization and classification *(Sailaja et al. 2015)*.

How to Compute:

The TF-IDF weight is composed by two terms: the first computes the normalized Term Frequency (TF): the number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF(t) = log e (Total number of documents / Number of documents with term t in it).

4.1.7 TF-IDF low-value Word Removal

Leveraging the calculation of the weight of importance, we take the TF-IDF weights for the next aggressive word removal step. First, we aggregate the sum of all TF-IDF weights of each word in all the articles. The range is from roughly 0.05 to 1200. Then, the filtering threshold is set to 8. There are 19k words in the corpus, which have the aggregated TF-IDF weights less than 8, being filtered out, leaving the remaining vocabulary of about 5k words.

4.1.8 Verbs, adverbs, conjunctions, prepositions, determiners elimination

In most of the contexts, verbs, adverbs, conjunctions, prepositions, determiners do not contain information about topics or main content of the texts. Therefore, these should be excluded from the word collection. We choose to take this step as the final one because executing this elimination process at the beginning is impossible. After this step, we complete our corpus with about 4500 vocabularies, ready for the model implementation.

4.1.9 Document Term Matrix

Document Term Matrix is simply an implementation of the Bag of Words concept, transforming the vocabulary into vectors containing the weight of each word, in form of a matrix *(Sailaja et al. 2015)*. This process is executed simultaneously with the TF-IDF process, turning words from corpus into a pandas DataFrame, with calculated TF-IDF weights. Both this step and the TF-IDF calculation are conveniently executed via the pre-built sklearn package.

A case in point of a piece of text after being processed through 9 different steps is shown below. The length of the text is 1657 and 716 words before and after the preprocessing steps, respectively. It is evident that all unnecessary and less meaning containing words are filtered out.

Before:
'President Donald Trump issued a federal permit for the expansion project in 2017, after being rejected by the Obama administration.

Together, the massive Keystone and the Keystone XL network would be about five times the length of the trans-Alaska oil pipeline.

The original Keystone is designed to carry crude oil across Saskatchewan and Manitoba, and through North Dakota, South Dakota, Nebraska, Kansas and Missouri on the way to refineries in Patoka, Illinois, and Cushing, Oklahoma. It has experienced problems with spills in the past, including one in 2011 of more than 14,000 gallons (53,000 liters) of oil in southeastern North Dakota, near the South Dakota border.

In 2017, the pipeline leaked an estimated 407,000 gallons (1.5 million liters) of oil onto farmland in northeastern South Dakota, in a rural area near the North Dakota border. The company had originally put the spill at about 210,000 gallons (795,000 liters).

Federal regulators said at the time the Keystone leak was the seventh-largest onshore oil or petroleum product spill since 2010.

North Dakota's biggest spill , and one of the largest onshore spills in U.S. history, came in 2013, when 840,000 gallons (3.1 million liters) spilled from a Tesoro pipeline in the northwestern part of the state. The company spent five years and nearly $100 million cleaning it up.

The Sierra Club pounced on the latest spill as an example of why the Keystone XL should not be built.

"We don't yet know the extent of the damage from this latest tar sands spill, but what we do know is that this is not the first time this pipeline has spilled toxic tar sands, and it won't be the last" '

After:
'presid donald trump issu permit expans project obama administr massiv network time length alaska oil pipelin origin design carri crude oil north dakota south dakota nebraska kansa missouri way illinoi oklahoma experienc problem spill past includ liter oil southeastern north dakota near south dakota border pipelin leak estim million liter oil northeastern south dakota rural area near north dakota border compani origin spill liter regul time leak seventh largest oil product spill north dakota s biggest spill largest spill u s histori million liter spill pipelin northwestern state compani spent year near million clean club latest spill exampl don t extent damag latest sand spill time pipelin spill toxic sand t'

4.2 Topic Modelling and Text Clustering

4.2.1 Elbow Method

Elbow Method is a common solution to determine the most appropriate number of clusters for the K-means algorithm, which simply tries out a different number of cluster, then quantifies the "badness" of each option using Within Cluster Sum of Square (WCSS), or basically, the sum of total variation squared *(Starmer. 2018)*. We then visualize the result via an Elbow plot.

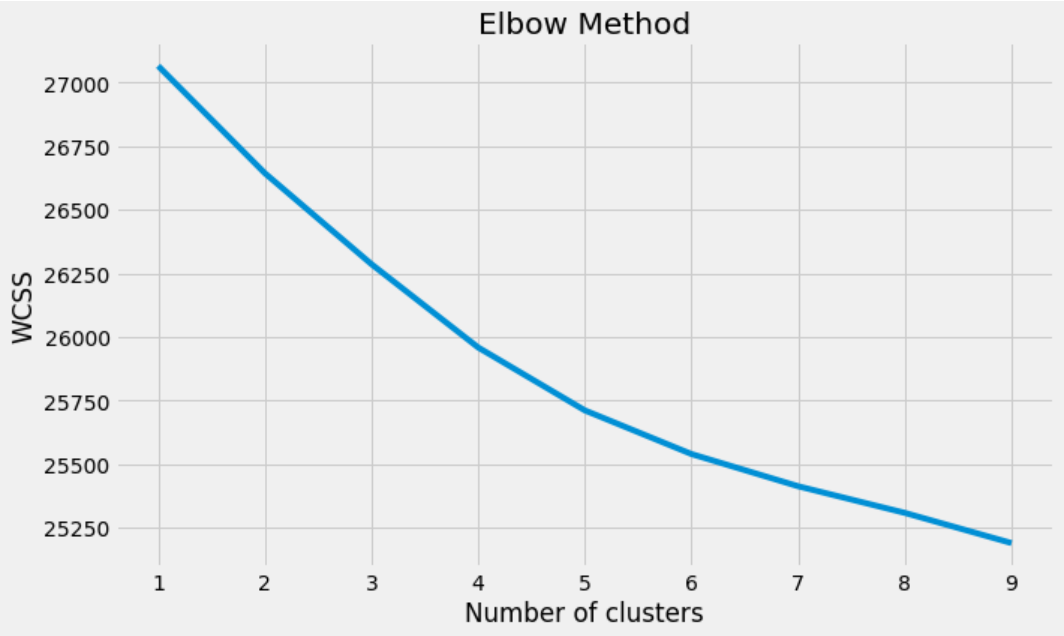


Figure 4.1 – Elbow plot for K-means algorithm

It could be seen that after the point of 5 clusters, WCSS begins to go down slowly, that turning point is the critical value that we attempt to find. However, the slope is just slightly less steep from the point of 5 and number 5 is too small compared to the number of articles we have (28k). Therefore, we could not immediately conclude that 5 is the best number of k. Instead, we conduct the K-means algorithm in 5, 6, 7, 8 and 9 to compare among them and also to the results from another model (LDA) to finalize the result accordingly.

4.2.2 K-means

To implement K-means, k is set to be equal to the critical value of the elbow method, but the critical value, as mentioned above, is not obvious. Thus, we choose k equal to 5, and the other three values 6, 7, 8 and 9 for comparison purposes. We train the K-means algorithm with the obtained data from the preprocessing steps. The algorithm results in an array of labels indicating which cluster each article belong to. We use the array to get 6, 7, 8 and 9 different clusters, then finding the top 10 most common words, representing the topic of each group.

4.2.3 Comparison with LDA

To then compare the LDA topics to the K-means topic, we use the Jaccard Distance (or Jaccard Index), which is a measure of the similarity between 2 sets, ranges from 0% to 100%. The higher the percentage is, the closer two sets of words are. The calculation of this measure is Jaccard Index = (the intersection of 2 sets) / (the union of 2 sets). We compare the result from K-means and LDA one by one. As the result, the comparing procedure generates 5, 6, 7, 8 and 9 values of Jaccard Index for each number of k topics, respectively. The results are shown in the table below:

LDA Topics	K-Means Topics	Jaccard Distance
['brexit', 'johnson', 'elect', 'eu', 'vote', 'deal', 'parliament', 'minist', 'conserv', 'corbyn']	['johnson', 'brexit', 'deal', 'eu', 'parliament', 'minist', 'elect', 'vote', 'govern', 'prime']	0.6666666666666666
['turkey', 'china', 'syria', 'korea', 'russia', 'turkish', 'kurdish', 'syrian', 'nato', 'russian']	['turkey', 'syria', 'turkish', 'syrian', 'kurdish', 'forc', 'state', 'trump', 'presid', 'border']	0.3333333333333333
, 'ukrain', 'impeach', 'presid', 'hous', 'investig', 'biden', 'inquiri', 'committe', 'ukrainian']	['trump', 'presid', 'hous', 'impeach', 'ukrain', 'investig', 'biden', 'republican', 'white', 'inquiri']	0.6666666666666666
['iran', 'govern', 'protest', 'kill', 'attack', 'israel', 'taliban', 'india', 'countri', 'state']	['year', 'state', 'govern', 'peopl', 'presid', 'new', 'trump', 'countri', 'elect', 'report']	0.17647058823529413
['polic', 'year', 'state', 'protest', 'hong', 'kong', 'citi', 'court', 'peopl', 'new']	['hong', 'kong', 'protest', 'polic', 'china', 'govern', 'citi', 'peopl', 'beij', 'chines']	0.42857142857142855

Figure 4.2 – Comparison between 5 pairs of LDA topics and K-means topic.

	Jaccard Index	Ratio of similar topics
5	[0.66666667 0.33333333 0.66666667 0.17647059 0.42857143]	0.8
6	[0.53846154 0.25 0. 0.05263158 0.33333333 0.33333333]	0.5
7	[0.17647059 0.66666667 0.66666667 0.33333333 0.53846154 0.05263158 0.33333333]	0.7142857142857143
8	[0.53846154 0.17647059 0.81818182 0.42857143 0.81818182 0.25 0.66666667 0.11111111]	0.5
9	[0. 0.25 0.66666667 0.17647059 0.42857143 0. 0. 0.33333333 0.66666667]	0.3333333333333333

Figure 4.3 – Jaccard Index values and similarity ratio for k=5,6,7,8,9.

Assuming that any Jaccard Index values are above 0.33 justifies topic similarity, in Figure 4.3, we obtained the best results of 0.8 and 0.71 topic similarity ratio of k=5 and k=7, respectively, in

which the topic similarity ratio = (number values greater than 0.33) / (number of topics). Thus, the number of k is best equal to 5 or 7.

4.2.4 Grouped Bar chart Visualization

After the process of comparing between 2 models, we could choose the number of k equal to 5 or 7. With k=5, we use the results from K-mean clustering for visualization. Our expected visual solution should be effective at describing the trend of each topic through time. As a result, the chosen one is a grouped bar chart, with the x-axis is the number of articles in each cluster, the y-axis represents the three months starting from the end of August. There are 5 differently coloured columns in each day, accounting for 5 different topics.

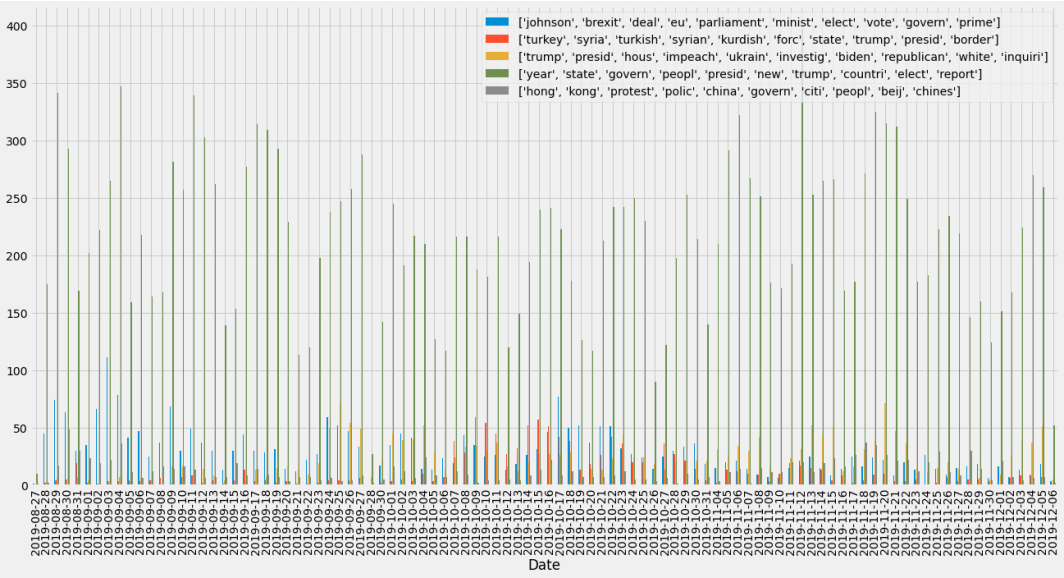


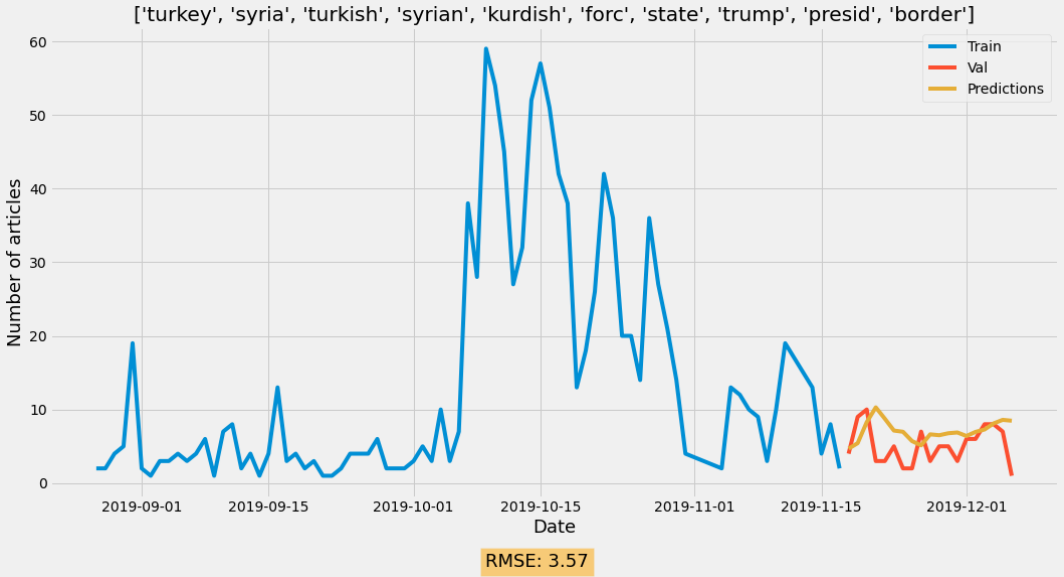
Figure 4.4 – Grouped Bar Chart with k=5.

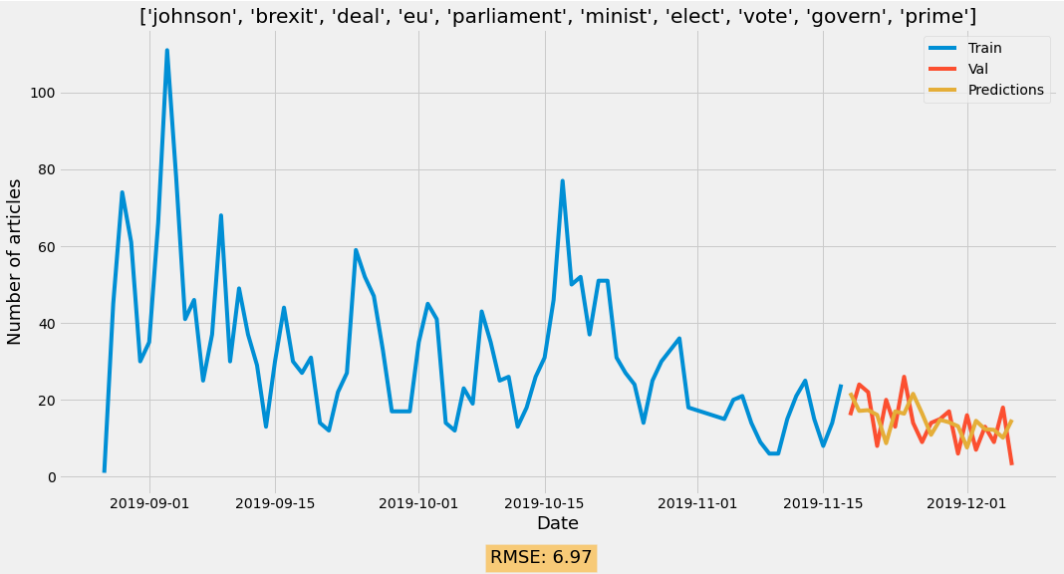
Generally speaking, it is obviously evident that all the figures experienced a fluctuation pattern over the time frame. Moreover, the most significant quantity belongs to the final cluster talking about the inner state of America.

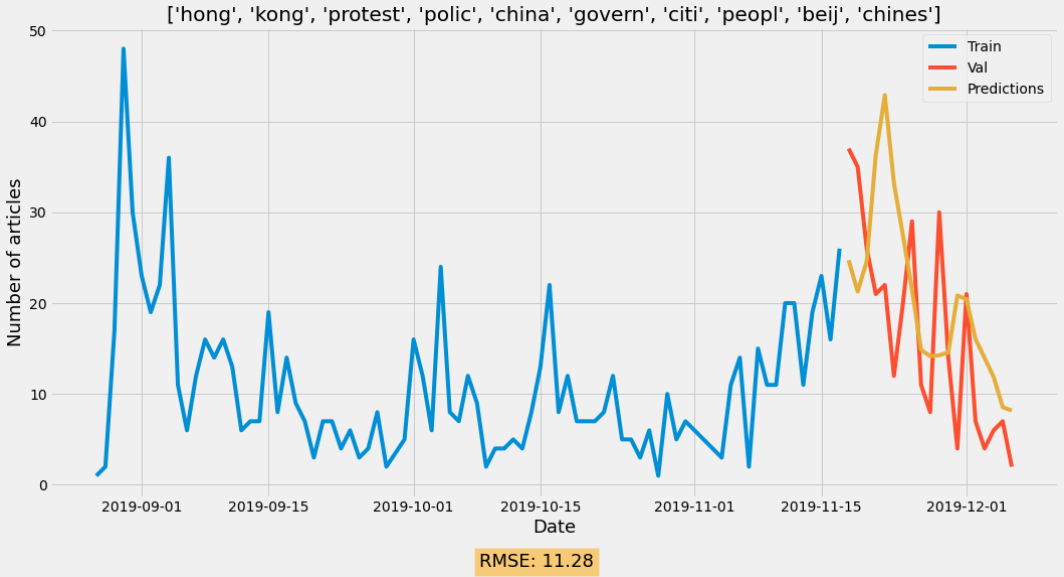
4.3 Time-series prediction with LSTM

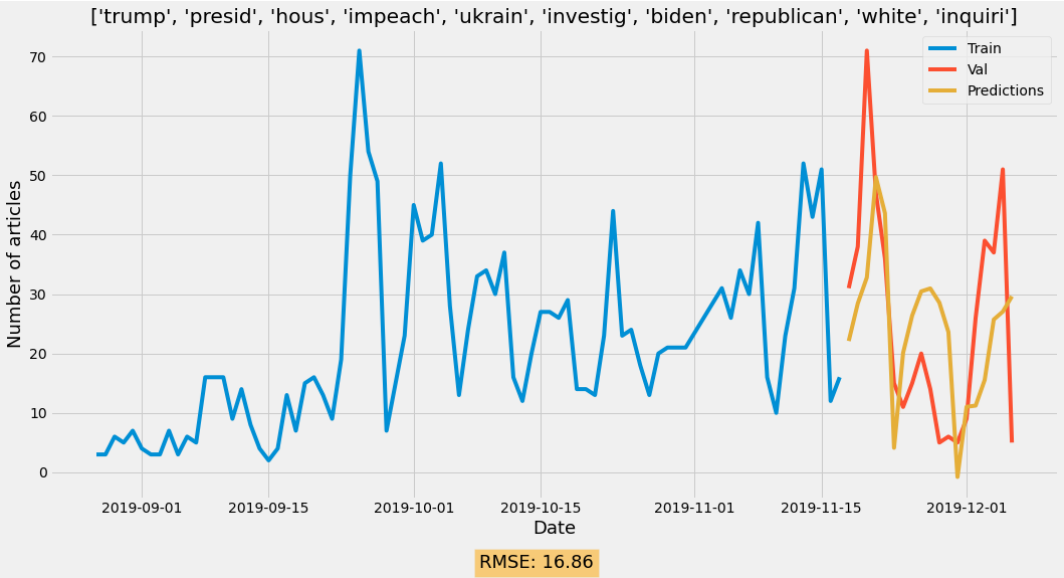
The LSTM implementation is executed following these steps in sequence: training/test set split with a ratio of 80:20; scale the data within range 0 and 1; prepare the right input format of the training data; build the stacked LSTM model with two 50-neuron LSTM layers and two additional dense layers; fit, train the model using the training data; finally, make predictions with the model and visualize the result simultaneously with the testing data. The implementation of LSTM is via Keras - a high-level API of TensorFlow *(randerson112358. 2019)*. We repeat the process on each identified topic, thus obtaining 5 different forecasting results.

We implement this model using the dataset from the result of the K-means model with k=5. After the process, we achieve five different predictions from each topic. The visualizations are shown below:


['turkey', 'syria', 'turkish', 'syrian', 'kurdish', 'forc', 'state', 'trump', 'presid', 'border']
Train
Val
Predictions
Number of articles
0
10
20
30
40
50
60
2019-09-01
2019-09-15
2019-10-01
2019-10-15
2019-11-01
2019-11-15
2019-12-01
Date
RMSE: 3.57


['johnson', 'brexit', 'deal', 'eu', 'parliament', 'minist', 'elect', 'vote', 'govern', 'prime']
Train
Val
Predictions
Number of articles
0
20
40
60
80
100
2019-09-01
2019-09-15
2019-10-01
2019-10-15
2019-11-01
2019-11-15
2019-12-01
Date
RMSE: 6.97


['hong', 'kong', 'protest', 'polic', 'china', 'govern', 'citi', 'peopl', 'beij', 'chines']
Train
Val
Predictions
Number of articles
0
10
20
30
40
50
2019-09-01
2019-09-15
2019-10-01
2019-10-15
2019-11-01
2019-11-15
2019-12-01
Date
RMSE: 11.28


['trump', 'presid', 'hous', 'impeach', 'ukrain', 'investig', 'biden', 'republican', 'white', 'inquiri']
Train
Val
Predictions
Number of articles
0
10
20
30
40
50
60
70
2019-09-01
2019-09-15
2019-10-01
2019-10-15
2019-11-01
2019-11-15
2019-12-01
Date
RMSE: 16.86

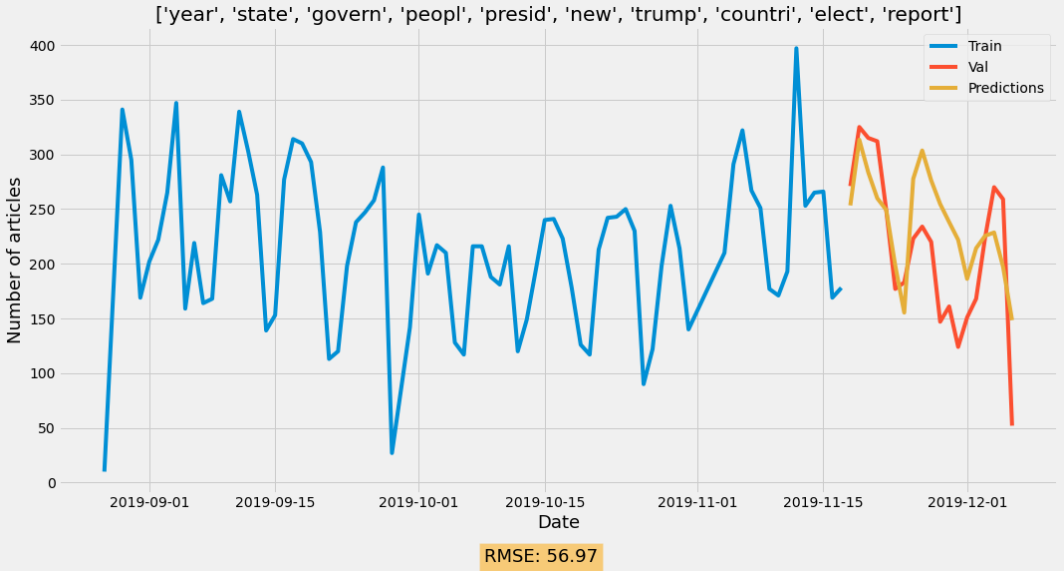


Figure 4.5 – LSTM time-series forecasting results with 5 topics.

The forecasting results of the LSTM are positively estimated due to the fairly low root mean squared errors. The difference in RMSE is due to more complex patterns in later topics. Furthermore, the most popular topic has an RMSE value of 56.97.

Chapter 5

Results and Future Works

5.1 Summary of Results

The experiment is carried out on a collection of 28k articles and it successfully completes the objectives of finding major topics and their trending through time. During the research, for the topic modelling target, we try out a different number of topics with two different methods (LDA and K-means) to then conclude the best values are 5 and 7. Afterwards, to depict the trend of the identified topics, the bar chart visualization is implemented to illustrate how popular each topic is over the time frame. Finally, we expand to experiment for future prediction using LSTM.

5.2 Future Works

Firstly, we state that there are more modern and advanced techniques for text mining than those that are used in this project. we try our best to both implement and interpret the methods so that traditional and convenient techniques are chosen. We hope that in the next attempts, with more time and dedication, should be conducted with state-of-the-art technologies. Secondly, for the scope of this thesis, it is promising to conduct this research in other fields such as economics, business or medical documents. Finally, our experiment is time-invariant, which means that we are not able to eliminate old data points or add new data points to the system without re-training all the process. The time-variant trait is extremely important with the real-world database and therefore, we hope to achieve that capability in the upcoming projects.

Bibliography

Alice, Zhao (2018). *Natural Language Processing in Python*. URL: https://www.pyohio.org/2018/schedule/presentation/38/.

Dr. Michael, J. Garbade (2018). *Understanding K-means Clustering in Machine Learning*. URL: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.

Hafsa, Jabeen (2018). *Stemming and Lemmatization in Python*. URL: https://www.datacamp.com/community/tutorials/stemming-lemmatization-python.

Niwratti (2019). *Regular Expressions — An excellent tool for text analysis or NLP*. URL: https://medium.com/analytics-vidhya/regular-expressions-an-excellent-tool-for-text-analysis-or-nlp-d1fa7d666cb9.

randerson112358 (2019). *Stock Price Prediction Using Python Machine Learning*. URL: https://medium.com/@randerson112358/stock-price-prediction-using-python-machine-learning-e82a039ac2bb.

Sailaja, D, M.V.Kishore, B.Jyothi, and N.R.G.K.Prasad (2015). "An Overview of Pre-Processing Text Clustering Methods." In: *International Journal of Computer Science and Information Technologies* 6. ISSN: 3119-3124.

SINGH, SHUBHAM (2019). *NLP Essentials: Removing Stopwords and Performing Text Normalization using NLTK and spaCy in Python*. URL: https://www.analyticsvidhya.com/blog/2019/08/how-to-remove-stopwords-text-normalization-nltk-spacy-gensim-python/.

Starmer, Josh (2018). *StatQuest: K-means clustering*. URL: https://www.youtube.com/watch?v=4b5d3muPQmA.

Team, Expert System (2016). *The value of text mining for unstructured information*. URL: https://expertsystem.com/value-text-mining-unstructured-information/#:~:text=Applied\%20to\%20a\%20corpus\%20or,in\%20large\%20amounts\%20of\%20information..