**{Instructions and Guidance in this template**

You will find guidance and information in this template.

Instructions are highlighted in turquoise and inside curly brackets, {like this}. All such information is designed to aid you in complying with the University’s [guidance](http://go.soton.ac.uk/thesiscompletion) on producing a research thesis. This should be deleted when finalising your thesis.

Text placeholders, where you need to enter data, are highlighted in yellow and surrounded by square brackets. Remove both the yellow highlighting and the square brackets when finalising the document.

The template meets all of these requirements; however you are heavily advised to use the support materials to find out more about using Word to ensure your document continues to comply as you add your own content.

There are 3 different demonstration seminars entitled “[Word: creating a thesis - Save time and tears](https://www.southampton.ac.uk/doctoral-college/professional-development-programme/about-the-programme/knowledge-and-techniques-for-research/using-it-software.page)”. Web based support can be found at <http://go.soton.ac.uk/thesispc> and <http://go.soton.ac.uk/thesismac>

**Copyright Notice**

The copyright notice below should be left untouched. It should be included in your e-thesis but is not required in your print document.}

University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]

[**University of Southampton**](https://www.southampton.ac.uk/lynda)

Faculty of [Your Faculty]{ in Title Case}

[School]{in Title Case}

[Thesis Title and any Subtitle]{in Title Case and bold}

DOI [enter DOI] hyperlink, optional}

[Volume X of Y] required if more than 1 volume}

by

**[Your Full Name]**{e.g. Sarah Alison Watson in Title Case and bold  
Qualifications and distinctions can be added if wished}

ORCID ID [enter ORDID]{optional}

Thesis for the degree of [name of degree]{in Title Case and underlined}

[Month and Year of submission]

**University of Southampton**

**Abstract**

Faculty of [Your\_Faculty]{in Title Case}

[School]{in Title Case}

Thesis for the degree of [name of degree]{in Title Case and underlined}

[Thesis Title]{in Title Case}

by

[Your Full Name]{e.g. Arthur Francis Jones}

Table of Contents

[Table of Contents i](#_Toc16455136)

[Table of Tables iii](#_Toc16455137)

[Table of Figures v](#_Toc16455138)

[List of Accompanying Materials vii](#_Toc16455139)

[Research Thesis: Declaration of Authorship ix](#_Toc16455140)

[Acknowledgements xi](#_Toc16455141)

[Definitions and Abbreviations xiii](#_Toc16455142)

[Chapter 1 Introduction 1](#_Toc16455143)

[1.1 Inspiration 1](#_Toc16455144)

[1.2 Background Research 1](#_Toc16455145)

[1.3 Thesis Idea 2](#_Toc16455146)

[1.4 Scope 3](#_Toc16455147)

[Chapter 2 Topic Modelling 5](#_Toc16455148)

[2.1 Latent Dirichlet Allocation 5](#_Toc16455149)

[2.2 Coherence 6](#_Toc16455150)

[2.3 Hyperparameters 6](#_Toc16455151)

[Chapter 3 Text Classification 7](#_Toc16455152)

[3.1 Recurrent Neural Networks 7](#_Toc16455153)

[3.2 fastText 8](#_Toc16455154)

[3.3 Random Forests 9](#_Toc16455155)

[3.4 Multinomial Naïve Bayes 9](#_Toc16455156)

[3.5 SGD Classifier 9](#_Toc16455157)

[3.6 Logistic Regression 9](#_Toc16455158)

[Chapter 4 Methodology 11](#_Toc16455159)

[4.1 Data Source 11](#_Toc16455160)

[4.2 Data Pre-processing 11](#_Toc16455161)

[4.2.1 Feature Engineering 11](#_Toc16455162)

[4.2.2 Data Cleaning 11](#_Toc16455163)

[4.2.3 Stopwords Removal 11](#_Toc16455164)

[4.2.4 Word Tokenization 12](#_Toc16455165)

[4.2.5 Part of Speech Tagging 12](#_Toc16455166)

[4.2.6 Stemming and Lemmatization 13](#_Toc16455167)

[4.2.7 Bigram and Trigram 13](#_Toc16455168)

[4.3 Build Model 14](#_Toc16455169)

[4.3.1 Topic Modelling using LDA Mallet 14](#_Toc16455170)

[4.3.2 Text Classification 15](#_Toc16455171)

[Chapter 5 Results and Future Scope 17](#_Toc16455172)

[5.1 Results 17](#_Toc16455173)

[5.1.1 Topic modelling Results 17](#_Toc16455174)

[5.1.2 Text Classification Results 19](#_Toc16455175)

[5.2 Visualization 19](#_Toc16455176)

[5.3 Future Scope 19](#_Toc16455177)

[Appendix A [Your first appendix] 23](#_Toc16455178)

[Appendix B [Your second appendix] 39](#_Toc16455179)

[Glossary of Terms 41](#_Toc16455180)

[List of References 43](#_Toc16455181)

[Bibliography 45](#_Toc16455182)

Table of Tables

[**Table 1‑1** **Training Time for a single epoch on sentiment analysis dataset** 5](#_Toc16786890)

[**Table 3‑1** **fastText parameters** 13](#_Toc16786891)

[**Table 4‑1** **Coherence values for different number of topics** 19](#_Toc16786892)

[**Table 5‑1** **Topics for books published in 2010** 23](#_Toc16786893)

[**Table 5‑2** **Different Topics published between 1981 to 2010** 27](#_Toc16786894)

[**Table 5‑3** **Evaluation of Recurrent Neural Network with Glove** 27](#_Toc16786895)

[**Table 5‑4** **Evaluation of fastText for various hyperparameters** 29](#_Toc16786896)

[**Table 5‑5** **Evaluation of different classification algorithms** 29](#_Toc16786897)

[**Table 5‑6** **Count of books for each topic** 30](#_Toc16786898)

Table of Figures

[**Figure 2‑1** **Graphical Representation of LDA** 6](#_Toc16676114)

[**Figure 3‑1** **Recurrent Neural Network Architecture** 9](#_Toc16676115)

[**Figure 4‑1**  **Frequency Distribution of Most Common Words** 14](#_Toc16676116)

[**Figure 4‑2**  **Frequency Distribution of Most Common Words after Stopwords removal** 14](#_Toc16676117)

[**Figure 4‑3**  **Part of Speech Tagging** 15](#_Toc16676118)

[**Figure 4‑4**  **Selectivity as a function of the length of the reference list: unigrams, bigrams, and trigrams [8]** 15](#_Toc16676119)

[**Figure 4‑5** **Architecture of Text Classification Model** 19](#_Toc16676120)

[**Figure 5‑1** **Number of Topics from 1981 to 2010** 26](#_Toc16676121)

[**Figure 5‑2** **Number of Books per Topic (1981 - 2010)** 27](#_Toc16676122)

[**Figure 5‑3** **Topic vs Year Visualization** 27](#_Toc16676123)

List of Accompanying Materials

{This section can be deleted if there is no accompanying material

Include the data DOI, if available}

Research Thesis: Declaration of Authorship

|  |  |
| --- | --- |
| Print name: | Janani Sundaresan |
| Title of thesis: | Unsupervised Text Classification with Topic Modelling |

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. {Delete as appropriate}None of this work has been published before submission {or} Parts of this work have been published as:- [please list references below]:

[Add references here] {otherwise delete}

|  |  |  |  |
| --- | --- | --- | --- |
| Signature: |  | Date: |  |

**{Important note:**

The completed signed and dated copy of this form should be included in your print thesis.

A completed and dated but unsigned copy should be included in your e-thesis}

Acknowledgements

Definitions and Abbreviations

|  |  |  |
| --- | --- | --- |
| **#** | **Abbreviation** | **Expansion** |
| 1 | LDA | Latent Dirichlet Allocation |
| 2 | RNN | Recurrent Neural Network |
| 3 | GloVe | Global Vectors for Word Representation |
| 4 | POS | Part of Speech |
| 5 | CNN | Convolutional Neural Networks |
| 6 | LSTM | Long Short Term Memory |
| 7 | GRU | Gates Recurrent Units |
| 8 | NLP | Natural Language Processing |
| 9 | SGD | Stochastic Gradient Descent |
| 10 | SVD | Singular Value Decomposition |

# Introduction

## Inspiration

Everyday gigabytes of text data are generated in various forms, including chat messages, error logs, customer reviews, articles. These text data should be classified into a proper structure so that information can be retrieved. Majority of that data generated are unlabeled. Currently, there are many supervised, and semi-supervised techniques for classifying data which already have labels have associated with them. However, technology has not yet caught up with organizing the vast amount of unstructured, unlabeled data without the humongous effort of manual intervention.

The thesis is aimed at reducing the amount of manual intervention required in classifying unlabeled text data using a technique called Topic Modelling, which is an unsupervised machine learning method. Topic Modelling is used to identify a thematic structure in a sizeable unstructured corpus quickly. Then, build a Text Classification model, a supervised machine learning technique that would automatically classify new data based on the knowledge obtained from training data.

## Thesis Idea

Books form an important part of our lives. Before the internet age, they were the main source of information. Book reading continue to be the favorite hobby for many even at this internet age. Publishing industry is a very big market which deals with billions of dollars per year. Books are sold in physical and digital formats like e-books, audio books, online subscriptions. Publishers Association, represents the book and journal publishers in UK. They conducted a poll with 2000 UK residents and found that 72% of the population purchased at least one book either in physical or digital format in 2017. Under fiction, 43% preferred ‘Crime’, 38% preferred ‘Thriller’ and ‘25% preferred ‘best sellers’. Under non-fiction, 32% preferred autobirogaphy,27 % History and 25% preferred books related to Food and Cookery. As per the publishing Association yearbook, total income from publishing was 6 billion Great Britain Pounds and books worth 2.9 billion Great Britain Pounds were sold in the year 2018 alone. The physical book market value at end purchase prize in United Kingdom alone is estimated to be 3,306 million pounds. Even though this value is 4% down when compared to 2017 it is 14% up when compared to 2014. [21] United States is a much bigger market with net revenue of 26.23 billion United States Dollars and global e-book publishing revenue alone is around 12 billion United States Dollars. This is a globally thriving industry and there is possibility for companies to gain more money by tapping into its potential. But there is little to no public information about how the publishing industry decides on the topic of the books to be published the following year as per market conditions in order to keep the business thriving even at this electronic age. This information can be obtained by analyzing the trend in publishing industry over the decades and also analyzing the trend in the best-selling books. This can be done by analyzing the topics associated with the books published year by year.

There are currently over 21 million books in ISBN Database alone. Currently, only keywords are associated with each book, and they are not relevant to the topic of the book.  Since there are millions of books, it is not possible to read every book to identify the range of topics. Manually classifying them would be a hugely time-consuming task. The solution is to build a model that would be trained one a set of labelled data automatically classify the unlabeled books, in order to analyze the trend. In order to train the model, we need to have a large set of labeled documents. i.e., an extensive training set where a particular topic is assigned to each book based on its context. So far, unsupervised machine learning technique along with text classification has not been used to classify books.

The innovative idea proposed in the thesis is to analyze how the topics of the books published over the decades have changed and also classify the books into different genres based on the topics generated from the synopsis. Topic Modelling has been used in many applications like news/movie recommendations, analyzing tweets over the decades.  Many methods currently exist for text classification in case of supervised methods. However, in the case of unsupervised classification, it is a rigorous task to classify each observation with a label manually.

Objectives of Thesis includes the following:

* To extract topics from the books published between 1981 to 2010 using topic modelling, an unsupervised machine learning technique.
* To classify books based on the synopsis into different genre using different text classification algorithms and evaluate their performance based on selected metrics
* Analyze and look for pattern in topics of books over three decades, in the publishing industry.
* Look for a correlation between the topic of best -selling books and the topic of the books published the following year.

## Background Research

Unsupervised Machine Learning is a technique in which the machine has to teach itself to organize data into different clusters or groups based on similar features. Since the datasets are unlabeled, there is no clear-cut way to validate the accuracy of the prediction.

Topic modelling is an unsupervised classification technique which helps to group unstructured data. Numerous topic modeling technique exists like SVD, NMF. SVD is very useful in reducing the number of variables that needs to be considered. But it requires a lot of Random Access Memory and is not helpful in distinguishing documents since they are linear combinations of input terms. NMF are useful to identify the incoherent topics. In [20], three topic modelling techniques, namely, LDA, LSA and NMF are being compared. In [16], a topic modelling based on LDA for classifying oncology-based text is proposed. Here the authors discuss the difficulty involved in accurately interpreting the topic based on the keyword associated especially if the person is not familiar with the domain of the corpus.

Text Classification has emerged to be an essential task in the field of Natural Language Processing. Initially, Linear classifiers were used to classify text, but with an increase in complexity, the volume and lack of structure, more powerful classification techniques were required even though linear classifiers are faster to train. Many machine learning models currently exist including KNN, Decision Trees Classifier, Support Vector machines that achieve an excellent performance. Stochastic Gradient Decent Classifier and Random Forest Classifier are also used to classify text. SGD Classifier is known to perform better with sparse and high dimensional data. However, there is always a tradeoff between accuracy and time taken to train models when large volumes of data are involved. This is where Neural network comes into the picture. Recently, deep learning neural network models have shown a significant increase in accuracy by producing state-of-art results. Convolution Neural Network are known for their pattern matching ability and are widely used in image processing applications. Convolution Neural Network differs from Vanilla Neural Network from the fact that it consists of a convolution layer which applies filters made of kernels to extract features, a pooling layer which reduced dimensionality, a linear rectifying layer and a fully connected layer which aggregates information for the final classification. In [17], a convolution neural network is used to perform text classification with Term Frequency – inverse Document Frequency algorithm in order to calculate the weight associated with every word in a document in a corpus and a weight is assigned to vector based on the TF-IDF value. This text vector is passed as input to CNN which in turn extract features. This CNN model was able to classify text with an accuracy of 96.97%. However, even CNN has issues. “The CNN requires multilayers of CNN to capture long term dependencies because of the locality of the convolutional and pooling layers; as the length of the input grows, it becomes crucial to capture long-term dependencies.” [11]. Recurrent Neural Network overcomes these drawbacks with the ability to capture long-term dependencies when it comes to a single layer. Recently, fastText, a linear text classifier library introduced by Facebook, has been achieving state-of art results in classification. It is especially useful when large corpus of text data needs to be classified. The time taken by fastText in training the model is very less when compared with others deep learning techniques. “It can train more than one billion words in less than ten minutes using a standard multicore CPU, and classify half a million sentences among 312K classes in less than a minute.” [13]. In Table 3-1 from [13], we can see that fastText takes maximum of ten seconds to train data ranging between 125,000 to 4,000,000 observations, whereas other techniques take between minimum of 24 minutes to maximum of 5 days. In “Identification of Security Related Bug Reports via Text Mining Using Supervised and unsupervised Classification,” the author proposes an Unsupervised Classification using anomaly detection without the need for labeled datasets. The author concludes by saying that unsupervised classification results were not as good as supervised classification models.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Zhang and LeCun (2015)** | | **Conneau et al. (2016)** | | | **fastText** |
|  | Small char -CNN | Big char -CNN | Depth = 9 | Depth = 17 | Depth = 29 | h=10, bigram |
| AG | 1h | 3h | 24m | 37m | 51m | 1s |
| Sogou | - | - | 25m | 41m | 56m | 7s |
| DBpedia | 2h | 5h | 27m | 44m | 1h | 2s |
| Yelp P. | - | - | 28m | 43m | 1h09 | 3s |
| Yelp F. | - | - | 29m | 45m | 1h12 | 4s |
| Yah. A. | 8h | 1d | 1h | 1h33 | 2h | 5s |
| Amz. F. | 2d | 5d | 2h45m | 4h20 | 7h | 9s |
| Amz. P. | 2d | 5d | 2h45m | 4h25 | 7h | 10s |

**Table 1‑1 Training Time for a single epoch on sentiment analysis dataset**

## Scope

The scope of this project is to make use of the existing machine learning algorithms with different hyperparameters optimization. This thesis does not include invention of a new machine learning algorithm from scratch. The data used in thesis are not modified. The assumption made is there is only one topic associated with every document. Multiple topic per document is beyond the scope of this thesis.

# Topic Modelling

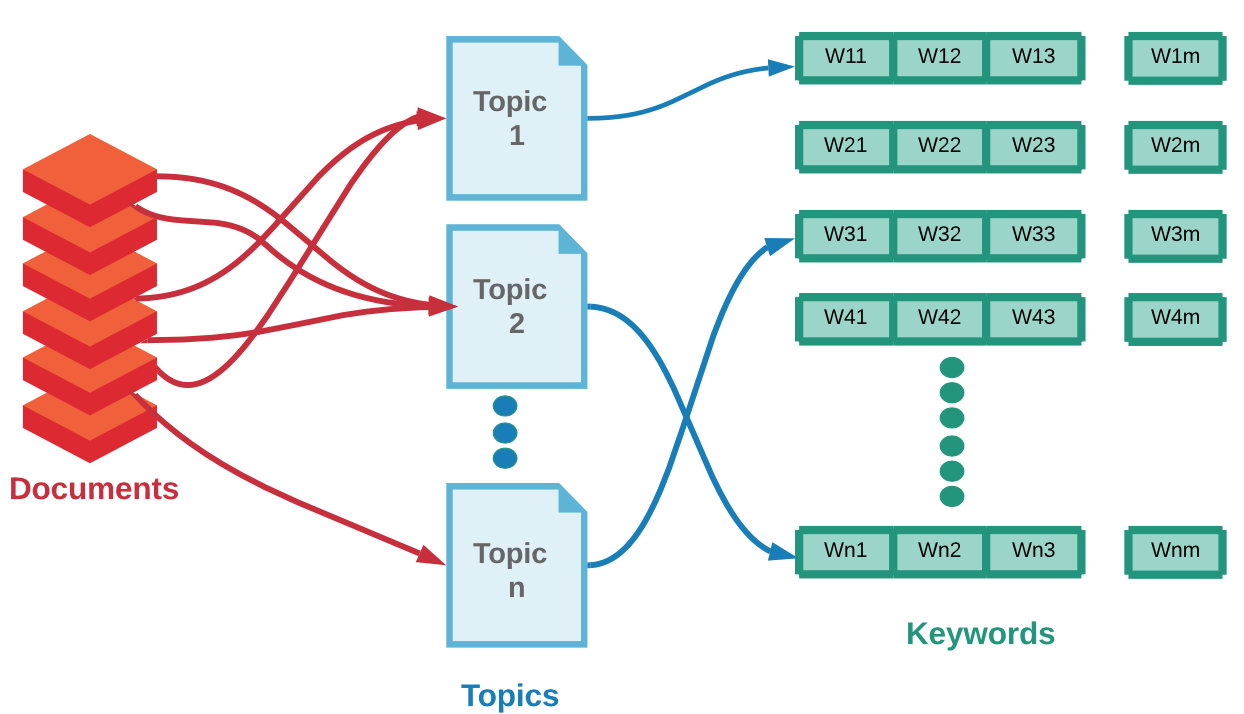
Topic Modelling is an unsupervised technique that can be used to extract useful sematic information. There are many topic modelling methods available like Latent Dirichlet Analysis, Latent Sematic Analysis, NMF etc. For our analysis, LDA with Mallet is considered. It can be carried out using different methods namely bag of words, frequency matrix, inverse document frequency, here LDA methodology is considered. The most predominant topic modeling technique is the Latent Dirichlet Allocation method. It is a hierarchical Bayesian model that represents a group of documents and the documents with words related to specific topics.

## Latent Dirichlet Allocation

Latent Dirichlet Allocation, abbreviated as LDA proposed by Blei et al [4] is a generative statistical model that identifies the predominant topics from a corpus of document [4]. There are numerous applications of LDA including sentiment analysis, classification of newspaper articles. LDA is based on two assumptions.

**Assumption 1:** The semantic information can be inferred from words and documents that revolve around similar topics will have words conveying similar semantic information. This helps to identify the documents that discuss about the same topic by identifying the group of words in a corpus of documents that frequently occur together within a document.

**Assumption 2:** Each document is a probabilistic distribution of related topics and each topic is probability distribution over a collection of keywords.



**Figure 2‑1 Graphical Representation of LDA**

In the figure 2.1.1, we can see that multiple documents can be related with a single topic and each topic has its own set of keywords. A keyword in one topic can also be a keywork in another topic. There are ‘n’ number of topics and ‘m’ number of keywords associated with each topic. There are totally ‘m x n’ keywords.

LDA assumes that entire corpus has a finite number of topics T, and each topic T is associated with a multinomial distribution over the words in the corpus called as vocabulary V.

## Coherence

Coherence is used as metric to assess the quality of the topics generated by the model. It is a quantitative measure of similarity of the list of keywords associated with each topic.  In particular, given a set V of topic words in a corpus (we will use the top 10 most frequent words in each topic), coherence is computed as a sum of similarity scores over all pairs of words in V. [6] Coherence can be measured using different similarity scores – u\_mass, c\_v, c\_uci, c\_npmi.

Intrinsic Measure – Uses the sum of the pairwise score of words used to describe a topic.

Extrinsic Measure -

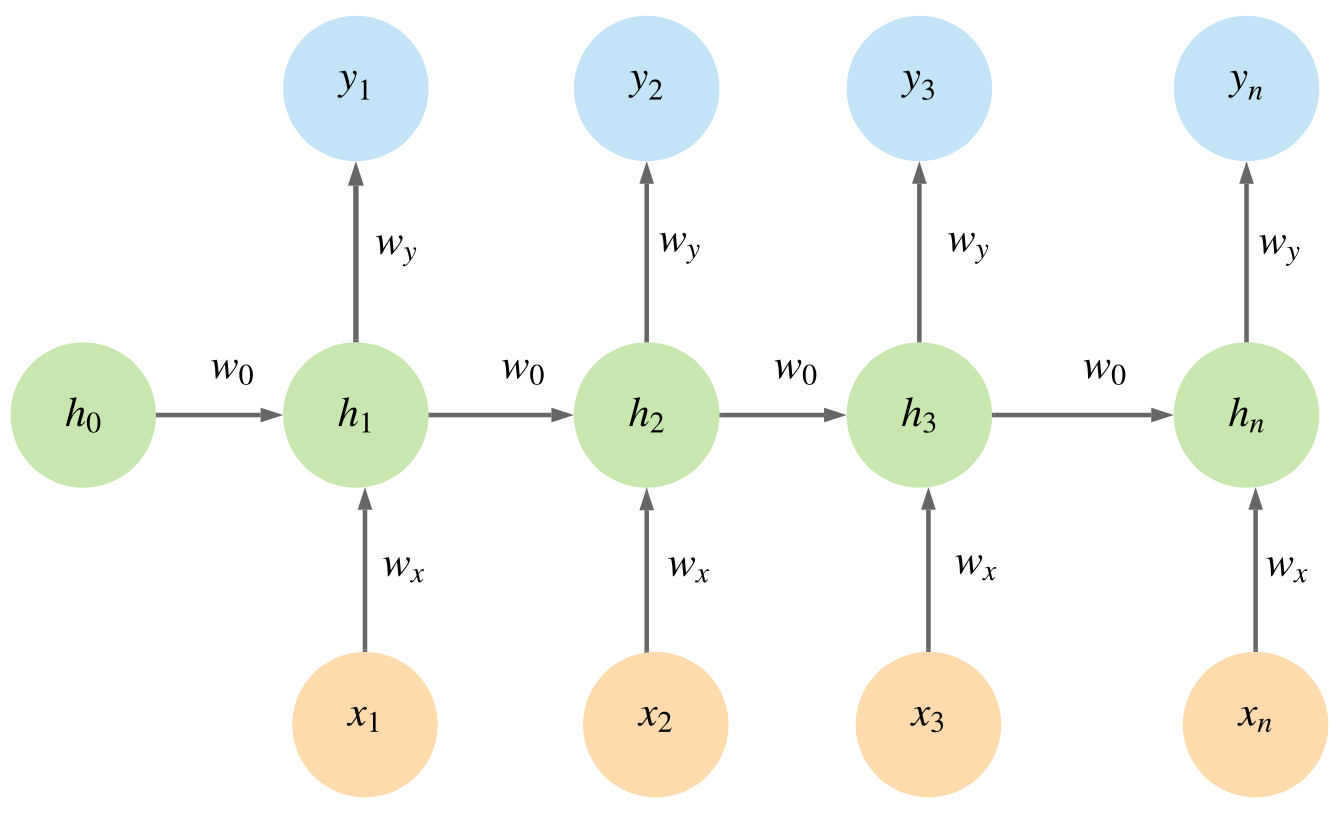
## Hyperparameters

The hyperparameters considered for topic modelling are number of topics, window size, intervals, iterations and number of threads.

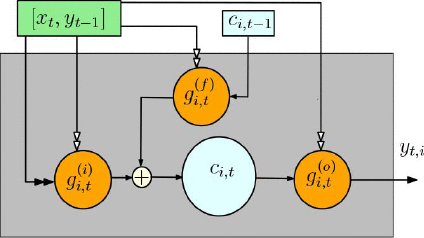
# Text Classification

## Recurrent Neural Networks

Recurrent Neural networks is widely used for NLP tasks due to its’ sequential structure. Vanilla Neural Network makes predictions based on the information gained from the training phase, whereas Recurrent Neural Network makes predictions using the training information along with the previous input with the help of a hidden state. The hidden state plays a key role in helping the neural network understand the context based on the previous inputs and outputs. The input consists of data at time t and also the output from the previous time step t-1. In the figure 3-1, we can see the hidden state receiving input from current state as well as previous states. The hidden state along with serves as input for the first hidden state in order to predict the output . along with serves as input for the next hidden state and this helps to remember the context. The RNN stores the previous information in memory which leads to the vanishing gradient descent problem. This problem is mitigated by Long Short Term Memory also called as LSTM.



**Figure 3‑1 Recurrent Neural Network Architecture**



“LSTM incorporates “gates”, which are neurons that use a sigmoid activation, and are multiplied with the output of other neurons. Using gates, the LSTM can adaptively ignore certain inputs. LSTM also maintains a set of values that are protected by the gates and that do not get passed through an activation function.” [12] In the below Figure 3-1 from [12], is the input data at time t, is the output at time t, and are the input and output gates which are you to regulate the follow of information.

Recurrent Neural Networks or RNN makes prediction by taking to account the past words in a sentence. They are very useful in predicting the next word. But in case of text classification, the prediction will be more accurate when RNN considers a context window of words around the target word. This is accomplished by Bi-directional LSTM.

## fastText

fastText is a sub-word embedding technique based on principle of morphology. The unique feature of fastText is that each word is represented as a bag of characters n-grams. word2vec only computes weight for whole word in a corpus. Whereas fastText creates vector for all n-gram combination including the word itself. These n-grams along with the whole word are used in the training and the error which is calculated at the end of training, is used to update the vectors. These n-gram word-vectors are more accurate than word2vec and are useful in generating word embedding for rare words. One outstanding feature of fastText is that it can construct a vector for the word using its n-gram characters even if the word does not appear in the corpus. On the other hand, word2vec and GloVe are not equipped to handle a new word that is not in vocabulary appears. However, this take a toll on the computation and also time consuming since the training is done at n-gram character level, more time is used to generate word embeddings than that of word2vec. “FastText utilizes the skip-gram objective with negative sampling” [15]

The labels should in the format “\_\_label\_\_<text>”. fastText is known to improve the representation of uncommon words.

fastText representors each word as a bag of characters n-grams in addition to the word itself. For example, the word ‘encyclopedia’ with n=3, will be presented by fastText as < ‘enc’, ‘ncy’,’cyc’,’ycl’,’clo’,’lop’,’ope’,’ped’,’edi’,’dia’,

fastText has various parameters

|  |  |
| --- | --- |
| Parameters | Description |
| -input | Link to the training folder or file |
| -output | Link to output folder or file |
| -wordNgrams | Maximum character length of a word to be considered |
| -label | Prefix to class type |
| -lr | Learning Rate |
| -dim | Dimensions of the word embedding |
| -epoch | Number of times learning algorithm runs through training data |
| -loss | Loss function , options – hs, ns, softmax |
| -pretrainedVectors | Pre-trained word embeddings |
| -bucket | Number of buckets to limit memory usage |

**Table 3‑1 fastText parameters**

## Random Forests

Random Forest is an ensemble of weaker models which together form a stronger model. Random Forest model is chosen as one the classification model because it is not sensitive to outliers and noise.

## GloVe

GloVe uses the co-occurrence probabilities of words. “GloVe is the best model as compared to other models as it is scalable to large corpus and works well even with small corpus. It improves the quality of learned representations by normalizing counts and log smoothing on them. [14]

# Methodology

## Data Source

Thesis is performed on two datasets. The first datasets that is used for analysis in this thesis were collected in “late 2017 from [goodreads.com](https://www.google.com/url?q=https%3A%2F%2Fwww.goodreads.com%2F&sa=D&sntz=1&usg=AFQjCNGVK7spOvUFEgelySj4wq_9zIyOcg), where we only scraped users' public shelves, i.e. everyone can see it on web without login. User IDs and review IDs are anonymized.” [1]. The data can be used for academic purposes. The second dataset is the list of New York Times Bestseller List, which is the list of best-selling books in United States. Every year since 9131, the list is being released.

## Data Pre-processing

### Feature Engineering

The Goodreads dataset consists of 23,60,655 rows of observations. For each book the dataset provides information about ISBN, Author, Description, Publishing Year, Publisher, Title etc. The detailed list of attributes in the data are shown in the Appendix B. The dataset consists of information about books published over two centuries. The New York Times Bestseller List dataset consists of 10195 rows of observations. For each bestseller-book, the dataset provides information about ISBN number, title, author, publisher, link to the amazon.com where the book is sold. The detailed list of attributes in the data are shown in the Appendix B. The dataset consists of information about best-selling books published between 2018 to 2018. Both of the datasets are initially pre-processed using techniques mentioned below and all irrelevant rows are removed before the model is built.

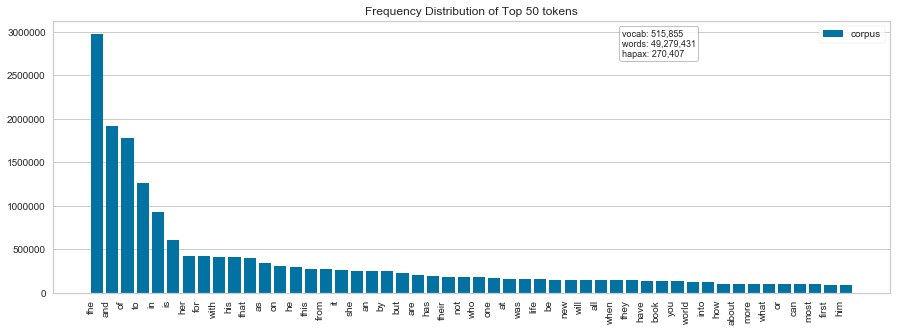
### Data Cleaning

The synopsis of the books in English language alone is considered for analysis. Language detect package is used to detect the synopsis in ‘English’ language. The number of words in the description of each book were calculated. The observations where the minimum length of the words is 30 and maximum length of the words is 250 were alone considered for analysis. The descriptions were cleaned by removing punctuation, making the words lower case, removing special characters, numbers, remove words which fall under pre-defined pattern. The words which had less than or equal to three characters were also removed.

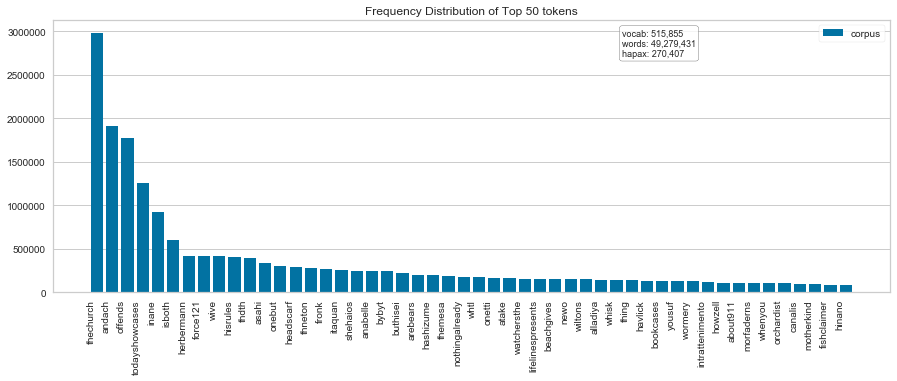
### Stopwords Removal

Stopwords refers to those words that occur most frequently in all the documents and can be a hindrance to identify the topics. They do not carry much lexical information. Examples of stopwords are ‘to’, ’and’, ’the’, etc. and removal of these words do not have profound impact on the prediction. ‘stopwords’ package from NLTK is used for Stopwords removal. Even after applying the technique, the synopsis still consisted of some frequently occurring unimportant words. So, a customized stop list of frequently occurring unimportant words were created and extended to the list of existing Stopwords. These customized list of Stopwords can be found in Appendix B Table < >. In the Figure 4-1 and Figure 4-2, we can see the frequency distribution of the words on corpus before and after removal of stopwords.

We see the word ‘the’ occurring more than 3\* 10 ^ 6 times, thereby emphasizing the importance of removal of stopwords.



**Figure 4‑1 Frequency Distribution of Most Common Words**



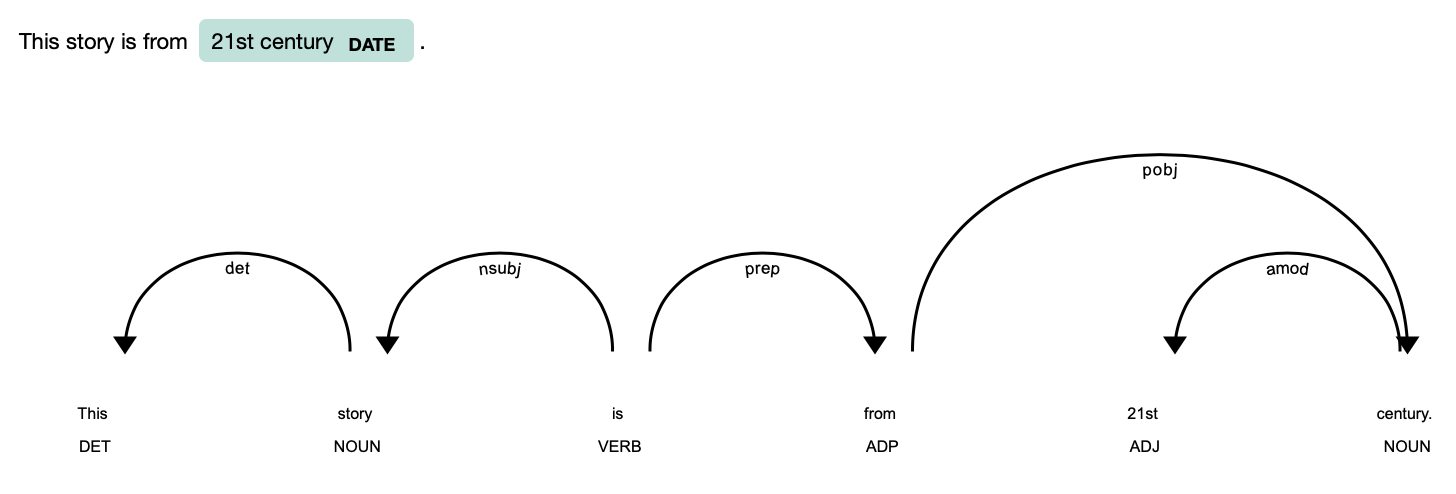
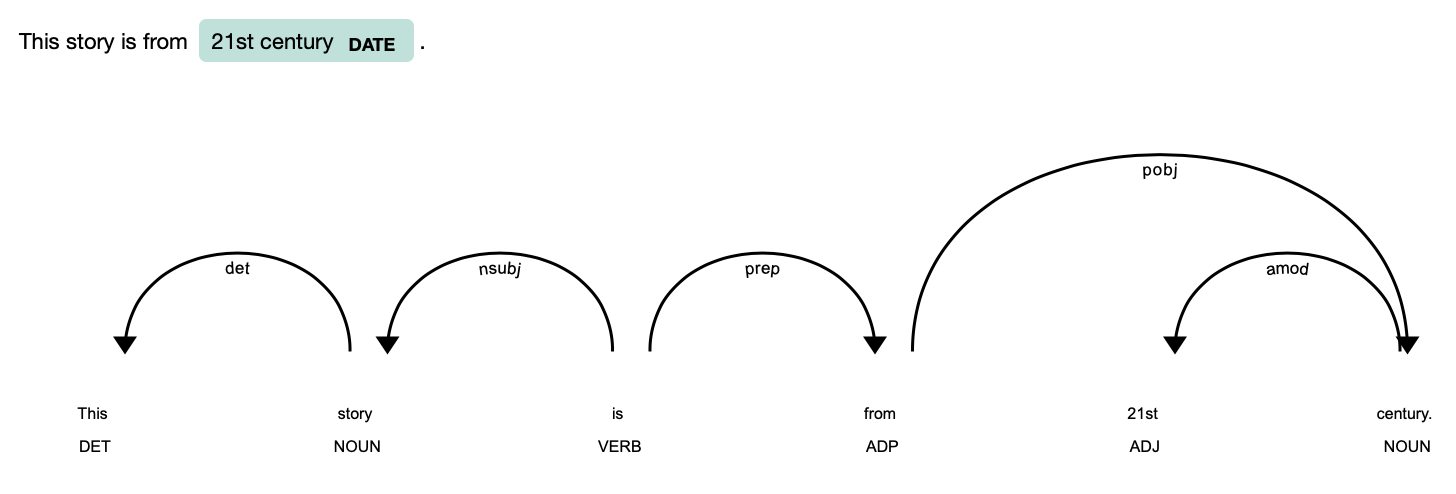
**Figure 4‑2 Frequency Distribution of Most Common Words after Stopwords removal**

### Word Tokenization

Words are the fundamental unit of a language. In order to build a better model, the text should be reduced to chunk of words. Work Tokenization refers to the process of extracting all the words from all sentences in a paragraph and converting them into a list. The words were tokenized using ‘word\_tokenize’ from nltk package.

### Part of Speech Tagging

Part of Speech tagging is a crucial step in any Natural language processing task. It is the process of assigning a part of speech to each word based on it’s context and position in a sentence. Discarded POS tags are listed in Table < >. In the below figure 4-3, we can see POS tagging performed for a sample sentence “This story is from 21st century.” The ‘21st century’ is marked as ‘DATE’ POS tag.



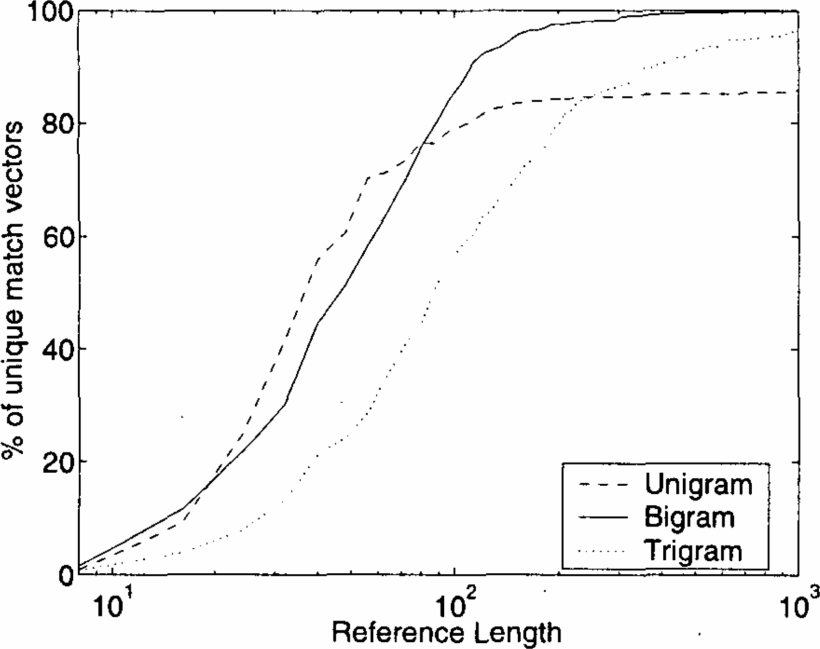
**Figure 4‑3 Part of Speech Tagging**

### Stemming and Lemmatization

Lemmatization is a normalization process. There are various algorithms to perform lemmatization. In [5], the performance and impact of using rule-based and dictionary-based algorithms are being compared. After analysis, a rule-based algorithm called ‘Snowball Stemmer’ is used to generate stemming for each word. Lemma for each word is determined using spacy.

### Bigram and Trigram

N-gram modelling is very useful in understanding the context of the sentence. This is done by considering a context window of varying sizes before and after a word. This is used to understand if the target word is a verb or noun or what context it is used. Bigram model is used to compute the probability of a words given the conditional probability of all the previous words. This assumption is called ‘Markov’ Assumption. Markov models are the class of probabilistic models that assume that we can predict the probability of some future unit without looking too far in the past [9].



**Figure 4‑4 Selectivity as a function of the length of the reference list: unigrams, bigrams, and trigrams [8]**

Based on Figure 4-4, we can see that unigram flattens out after a certain reference length. Hence, it is evident that bigram and trigram perform better than unigram model. Finally, a dictionary and corpus are created from the trigram list.

### Evaluation Metrics

The following metrics are used to evaluate the performance of each model.

1. Precision
2. Recall
3. Accuracy
4. F1-score
5. Balanced Accuracy

Accuracy is used to give a quick picture of how the model is performing. The drawback of using accuracy is that is does not give clear picture when the dataset has imbalanced classes.

Precision is a measure of how the model accurately classified as positives among the total number of outcomes predicted as positive.

Recall is a measure used when the number of predictions as True are actually False is high.

F1- score is used as measure to strike a balance between precision and recall.

The drawbacks of accuracy is overcome by using balanced accuracy as one of the metric for evaluating a model.

## Build Model

### Topic Modelling using LDA Mallet

The synopsis of the books for each year were pre-processed and passed through genism model. The assumption made for this thesis is there is only one topic related to every book.

The evaluation consisted of testing the LDA model for coherence using a wide range of hyper parameters with different set of values. Sliding window sizes varied between 50 to 350,

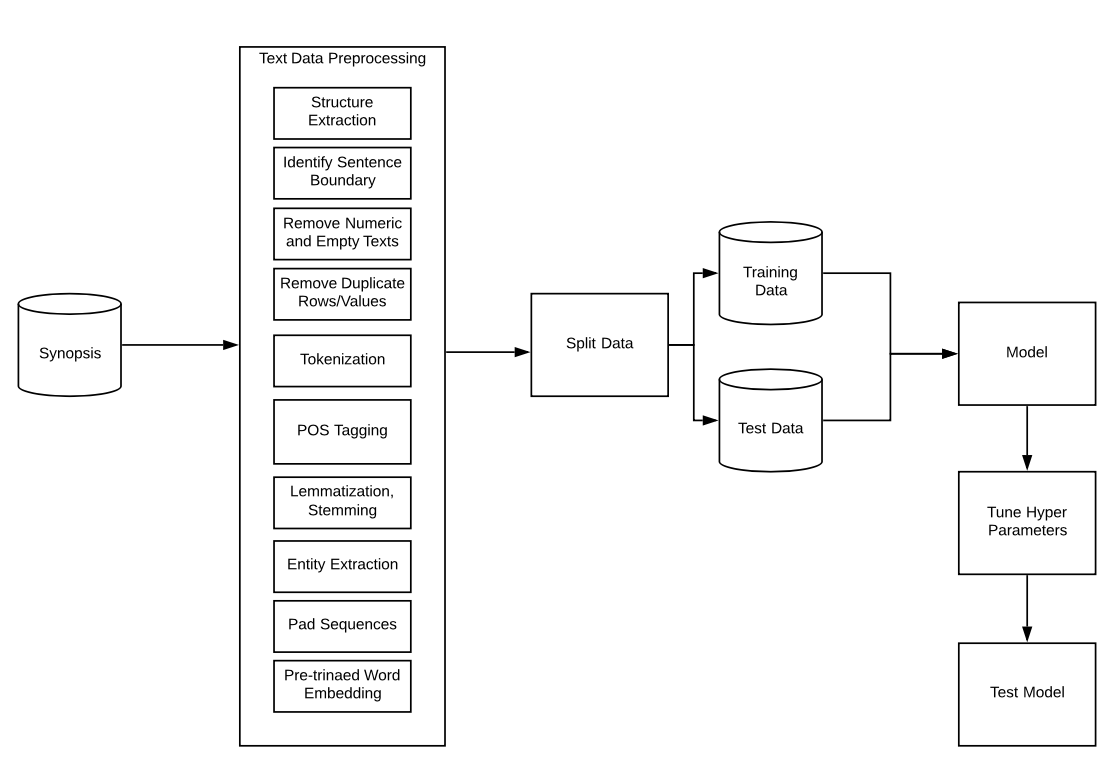
**Table 4‑1 Coherence values for different number of topics**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Year** | **Number of Topic** | | | | | | |
|  | **5** | **10** | **15** | **20** | **25** | **30** | **35** |
| 1981 | 0.481 | 0.451 | 0.447 | 0.408 | 0.412 | 0.393 | 0.398 |
| 1982 | 0.489 | 0.482 | 0.456 | 0.439 | 0.413 | 0.435 | 0.417 |
| 1983 | 0.491 | 0.502 | 0.481 | 0.455 | 0.410 | 0.418 | 0.429 |
| 1984 | 0.487 | 0.510 | 0.486 | 0.446 | 0.449 | 0.435 | 0.425 |
| 1985 | 0.501 | 0.497 | 0.497 | 0.485 | 0.483 | 0.445 | 0.431 |
| 1986 | 0.497 | 0.512 | 0.511 | 0.492 | 0.457 | 0.453 | 0.457 |
| 1987 | 0.514 | 0.527 | 0.534 | 0.510 | 0.501 | 0.475 | 0.466 |
| 1988 | 0.521 | 0.536 | 0.551 | 0.547 | 0.520 | 0.504 | 0.509 |
| 1989 | 0.521 | 0.540 | 0539 | 0.552 | 0.519 | 0.504 | 0.500 |
| 1990 | 0.547 | 0.545 | 0.544 | 0.526 | 0.547 | 0.535 | 0.517 |
| 1991 | 0.523 | 0.578 | 0.551 | 0.563 | 0.549 | 0.545 | 0.531 |
| 1992 | 0.536 | 0.543 | 0.559 | 0.572 | 0.558 | 0.557 | 0.542 |
| 1993 | 0.520 | 0.550 | 0.562 | 0.577 | 0.575 | 0.553 | 0.554 |
| 1994 | 0.541 | 0.582 | 0.576 | 0.595 | 0.583 | 0.569 | 0.556 |
| 1995 | 0.554 | 0.573 | 0.580 | 0.583 | 0.577 | 0.572 | 0.579 |
| 1996 | 0.521 | 0.572 | 0.576 | 0.583 | 0.576 | 0.580 | 0.577 |
| 1997 | 0.526 | 0.568 | 0.578 | 0.584 | 0.578 | 0.587 | 0.589 |
| 1998 | 0.525 | 0.556 | 0.580 | 0.577 | 0.580 | 0.579 | 0.586 |
| 1999 | 0.509 | 0.561 | 0.575 | 0.586 | 0.582 | 0.585 | 0.586 |
| 2000 | 0.544 | 0.569 | 0.574 | 0.580 | 0.593 | 0.604 | 0.600 |
| 2001 | 0.513 | 0.568 | 0.574 | 0.582 | 0.587 | 0.597 | 0.586 |
| 2002 | 0.530 | 0.565 | 0.575 | 0.573 | 0.573 | 0.588 | 0.577 |
| 2003 | 0.514 | 0.568 | 0.571 | 0.575 | 0.588 | 0.587 | 0.587 |
| 2004 | 0.538 | 0.557 | 0.576 | 0.592 | 0.580 | 0.584 | 0.588 |
| 2005 | 0.530 | 0.554 | 0.574 | 0.582 | 0.581 | 0.582 | 0.591 |
| 2006 | 0.534 | 0.571 | 0.565 | 0.574 | 0.593 | 0.581 | 0.592 |
| 2007 | 0.533 | 0.568 | 0.566 | 0.586 | 0.582 | 0.586 | 0.601 |
| 2008 | 0.526 | 0.553 | 0.562 | 0.575 | 0.587 | 0.573 | 0.574 |
| 2009 | 0.519 | 0.560 | 0.559 | 0.575 | 0.575 | 0.58 | 0.583 |
| 2010 | 0.499 | 0.552 | 0.569 | 0.567 | 0.564 | 0.580 | 0.583 |

### Text Classification

Text Classification is performed using different machine learning techniques and their performances are compared to find the best model. First, a classification model based on Recurrent Neural Network with bidirectional LSTM was built. RNN first tokenized all the words in the synopsis of the books into vector of words, together forming a 2-Dimesnional matrix. Bidirectional LSTM was used to capture the long-term sequence dependencies. Pre-trained Word embeddings are used. The dimensions of the GloVe word embedding varied between 840 Billion words to 6 Billion words with dimensions varying between 50 to 300. RNN model constructed for text classification has 150 hidden layer and 52 Dense Layer in output because of the 52 different types of labels. The dataset was split into training and test dataset. In order to build a better model between 50 to 60 books from each topic were randomly selected to train the model for classification. The sample data had 3828 observations. The sample data was split into training and test data set. The training data set had 3094 observations and test data set had 774 observations. The data was shuffled before splitting and they data split was stratified to ensure that the training data set had all the 52 labels and no new labels were introduced in test data set to avoid selection bias.

The RNN model performance was evaluated on different word embeddings of Glove and their accuracies were compared.



**Figure 4‑5 Architecture of Text Classification Model**

Next fastText embedding was used to classify the books into different genres. fastText has different set of parameters that can be configured to build an efficient model.

# Results and Future Scope

## Results

### Topic modelling Results

The results from LDA Mallet Model were evaluated using Coherence explained in Section 2.2. The keywords generated for each topic were manually analyzed. Each year contained books belonging to topics ranging anywhere between 5 and 35. A label was assigned as a topic manually based on the top 10 keywords associated. In the below Table 5-1, we can see the topics generated by LDA Mallet and the top 10 keywords associated with each topic for the books published in the year 2010. Totally 35 different topics were generated. Looking at the keywords, a topic label was manually assigned.

The first topic ‘Food’ contains the keywords like ‘receipe’, ‘cook’ , ‘food’ which are words often used in topics related to ‘Food’. The topic ‘Family’ has all keywords pertaining to the member in a family. The same process was repeated to find the predominant topics for the books published in a year between 1981 to 2010 and manually a topic label was assigned based on the keywords.

**Table 5‑1 Topics for books published in 2010**

| Topic # | **Topic** | **Top 10 keywords** |
| --- | --- | --- |
| 0 | Food | make, food, recipe, easy, good, grow, enjoy, garden, simple, cook |
| 1 | Dream | life, dream, make, find, face, hope, lose, choice, fear, back |
| 2 | Family | father, mother, daughter, wife, family, brother, sister, husband, woman, love |
| 3 | Battle | force, mission, fight, soldier, battle, group, military, enemy, attack, agent |
| 4 | Memory | secret, past, truth, discover, hide, dark, mysterious, memory, begin, reveal |
| 5 | Sports | play, game, world, team, time, star, great, year, high, rule |
| 6 | Movies | artist, work, music, film, movie, famous, scene, include, image, show |
| 7 | Poetry | poem, word, voice, write, poetry, collection, form, poet, sense, work |
| 8 | King | power, battle, force, great, fight, king, land, powerful, rule, warrior |
| 9 | Thriller | mystery, series, call, find, begin, adventure, miss, turn, late, bring |
| 10 | Magic Fiction | world, power, magic, ancient, ghost, evil, magical, spirit, dark, strange |
| 11 | School | book, student, reader, text, include, provide, chapter, language, teacher, study |
| 12 | Religion | experience, faith, life, personal, people, spiritual, journey, church, challenge, reader |
| 13 | Murder Mystery | murder, case, crime, killer, dead, kill, death, victim, police, body |
| 14 | Space Fiction | world, human, future, planet, power, race, space, humanity, scientist, earth |
| 15 | Discover | find, begin, lead, answer, question, discover, death, search, young, life |
| 16 | Vampire Fiction | vampire, save, blood, demon, dead, protect, human, dark, kill, soul |
| 17 | Relationship | love, heart, fall, meet, romance, find, relationship, true, expect, plan |
| 18 | Time | time, make, learn, part, give, good, start, show, thing, follow |
| 19 | Home | home, find, town, house, leave, summer, back, return, place, year |
| 20 | Comics | book, animal, reader, perfect, page, picture, funny, illustration, baby, hilarious |
| 21 | Information | include, guide, book, design, information, create, feature, project, cover, easy |
| 22 | Classic | story, author, tale, character, writer, fiction, short, reader, write, classic |
| 23 | Nature | city, travel, adventure, place, land, journey, ship, island, water, mountain |
| 24 | Night | thing, back, head, night, time, wrong, call, hand, walk, stand |
| 25 | Finance | business, work, make, money, company, career, sell, successful, market, offer |
| 26 | Health | mind, people, body, relationship, doctor, experience, care, deal, problem, health |
| 27 | Friendship | friend, good, girl, school, high, thing, turn, start, friendship, year |
| 28 | Learn | build, learn, develop, base, work, create, system, practical, show, guide |
| 29 | Social Culture | social, study, explore, examine, issue, science, contemporary, culture, development, modern |
| 30 | Love | woman, desire, passion, heart, night, sexy, lover, sexual, beautiful, beauty |
| 31 | Political History | history, American, country, political, state, nation, account, century, early, culture |
| 32 | Parenting | family, child, young, parent, year, grow, live, raise, home, bear |
| 33 | Life | life, live, change, world, year, time, people, bring, move, story |
| 34 | Series | book, work, include, write, edition, series, author, volume, publish, original |

| **Year** | **Topic List** | **Count** |
| --- | --- | --- |
| 1981 | Murder Mystery, Series, Space Fiction, Love, Classic | 5 |
| 1982 | Series, Classic, Love, Thriller, Time | 5 |
| 1983 | Marriage, Alien, Political History, Classic, Friendship,  Murder Mystery, Discover, Dream, Series, Space Fiction | 10 |
| 1984 | Thriller, Friendship, Classic, Series, Space Fiction | 10 |
| 1985 | Space Fiction, School, Love, Dark Mystery, Classic, Life,  Murder Mystery, Family, Battle, Political History | 10 |
| 1986 | Information, Battle, Series, Classic, Love, Murder Mystery, Thriller, History, Family, Life | 10 |
| 1987 | Literature, Series, Experience, Murder Mystery, Photography, Battle, Marriage, Travel, Youth, Alien, Parenting, Political History, Classic, Discover, Friendship | 15 |
| 1988 | Relationship, Social Culture, Classic, Series, Parenting, Information, Travel, Love, History, Battle, School, Dream, Murder Mystery,  Discover, Thriller | 15 |
| 1989 | Travel, Classic, Dark Mystery, Information, Battle, Love, History,  Murder Mystery, Friendship, Political History | 10 |
| 1990 | Murder Mystery, Battle, Information, Family, Literature, Life,  Political History, Classic, Series, City | 10 |
| 1991 | Political History, Friendship, City, Discover, Classic, Marriage,  Murder Mystery, Information, Experience, Time | 10 |
| 1992 | Youth, Discover, Information, Battle, History, Series, Movies,  Murder Mystery, Experience, Classic, Family, Life, Marriage, Poetry, School | 15 |
| 1993 | History, Experience, Murder Mystery, Marriage, Discover, Movies,  Classic, Information, Time, Science, Social Culture, Learn, Battle,  Parenting, Poetry | 15 |
| 1994 | Marriage, Family, Murder Mystery, Classic, Social Culture, Series,  Religion, Life, Sports, Time, Experience, Dark Mystery, Friendship,  Photography, History, Battle, Love, Poetry, Youth, Information | 20 |
| 1995 | Photography, Marriage, Poetry, Classic, Murder Mystery, Series,  Science, Political History, Religion, Parenting, Dark Mystery, Battle, Friendship, Government, Information, Youth, School, Time, Night,  Life | 20 |
| 1996 | Information, Murder Mystery, Religion, Classic, Science, Discover, Series, Poetry, Magic Fiction, Battle, Political History, History, School, Memory, Love, Parenting, Travel, Family, Comics, Home | 20 |
| 1997 | Movies, History, Youth, School, Murder Mystery, Poetry, Family, Travel, Dark Mystery, Classic, Learn, Marriage, Series, Thriller, Philosophy, Government, Health, Fiction, Religion, Finance, Nature, Explore, Experience, Battle, Food, Political History, Sports, Life, Photography, Information, Time, Home, Magic Fiction, Space Fiction | 35 |
| 1998 | Love, Classic, Home, Friendship, Murder Mystery, Literature, School, City, Space Fiction, Information, Battle, Discover, Religion, Marriage, Life, Dark Mystery, Travel, Photography, Family, Series,  Health, Magic Fiction, Social Culture, Food, History, Movies, Comics, Science, Political History | 30 |
| 1999 | Photography, Family, Dark Mystery, Sports, Magic Fiction, Food, Nature, Battle, Fiction, Government, Murder Mystery, Night, Biography, Learn, Literature, Home, City, Comics, Information, Religion, Life, Classic, Dream, Social Culture, Memory, Science,  Marriage, Series, Thriller, School, Health, History, Space Fiction, Time | 35 |
| 2000 | Time, City, Political History, Murder Mystery, Experience, Battle, Science, Classic, Information, Friendship, Movies, Food, Life, Parenting, Love | 15 |
| 2001 | Family, Murder Mystery, Battle, Memory, Travel, Relationship, Social Culture, Information, Parenting, Friendship, Food, City, Science, Classic, Movies | 15 |
| 2002 | Series, Movies, City, Poetry, Nature, Parenting, Murder Mystery, Family, Religion, Photography, Political History, Battle, Friendship, School, Relationship, Finance, Space Fiction, History, Information, Science, Magic Fiction, Comics, Social Culture, Health, Discover, Thriller, Fiction, Love, Dream, Sports | 30 |
| 2003 | Murder Mystery, Memory, Marriage, Health, Information, Learn,  Social Culture, Parenting, Food, Series, Religion, History, Battle,  Magic Fiction, Time, Science, Life, Poetry, Photography, Nature, School, Comics, Discover, City | 25 |
| 2004 | Murder Mystery, Religion, Dark Mystery, Nature, Marriage, Food  Political History, Poetry, Family, Battle, Space Fiction, Night, Learn, Information, Comics, Time, Science, Sports, Classic, Finance, School, History, Movies, Love, Experience | 25 |
| 2005 | Marriage, History, Travel, Food, Religion, Fashion, Classic, Family, Relationship, Experience, Health, Youth, Life, Social Culture, School, Murder Mystery, Finance, Night, Time, Poetry, Photography, Home Information, Learn, Dark Mystery, Battle, Movies, Comics, City,  Space Fiction, Series, Discover, Political History, Friendship, Dream | 35 |
| 2006 | History, Food, Poetry, Dream, Dark Mystery, Parenting, Movies, Friendship, Explore, Battle, Home, Experience, Classic, Discover, Photography, Family, Murder Mystery, Series, Religion, Love, Political History, Government, Space Fiction, Information, Finance, Comics, Nature, Sports | 30 |
| 2007 | Magic Fiction, Space Fiction, Explore, Murder Mystery, Family, Love, Classic, Friendship, Battle, Movies, Discover, City, Health, Photography, School, Political History, Food, Experience, Parenting, History, Dark Mystery, Series, Home, Finance, Marriage, Poetry,  Life, Time, Religion, Learn | 30 |
| 2008 | Comics, Vampire Fiction, Dream, Murder Mystery, Love, Marriage, Political History, Night, Religion, Health, Relationship, Food, Discover, Movies, Friendship, Nature, Series, Space Fiction, Poetry, Photography, Classic, School, Finance, Youth, History, Battle, King  Explore, Experience, Time, Social Culture, Information, Memory, Family, Sports | 35 |
| 2009 | Home, Vampire Fiction, Murder Mystery, Social Culture, Parenting, Learn, Nature, King, Friendship, Classic, Time, History, Thriller, Science, Space Fiction, Love, Life, Religion, Movies, Poetry, Family, Food, Health, Memory, School, Discover, Series, Finance, Night,  Political History, Comics, Relationship, Photography, City | 35 |
| 2010 | Vampire Fiction, Sports, Nature, Political History, Food, Relationship, Love, Battle, Religion, Night, Dream, Social Culture, Finance, Life, Family, Memory, Murder Mystery, Magic Fiction, Friendship, Home, Space Fiction, Thriller, Poetry, Series, Learn, Parenting, School,  Information, Comics, King, Movies, Discover, Classic, Health, Time | 35 |

**Table 5‑2 Different Topics published between 1981 to 2010**

### Text Classification Results

|  |  |  |
| --- | --- | --- |
| Algorithm | Dimensions | Accuracy |
| RNN with GloVe 840B | 300 | 0.661 |
| RNN with GloVe 42B | 300 | 0.651 |
| RNN with GloVe 6B | 50 | 0.574 |
| RNN with GloVe 6B | 100 | 0.601 |
| RNN with GloVe 6B | 200 | 0.660 |
| RNN with GloVe 6B | 300 | 0.668 |

**Table 5‑3 Evaluation of Recurrent Neural Network with Glove**

Comparing Table 5-3 and Table 5-4, we can see that fastText performs better than LSTM with GloVe word embeddings and also the time take by fastText is in minutes whereas RNN takes time in hours to classify text.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Epochs | Window size | Learning Rate | wordNgrams | Loss | Precision |
| 50 | 2 | 0.01 | 1 | hs |  |
| 100 | 2 | 0.01 | 1 | hs |  |
| 50 | 2 | 0.001 | 1 | hs |  |
| 100 | 2 | 0.001 | 1 | hs |  |
| 50 | 2 | 0.01 | 1 | Softmax |  |
| 100 | 2 | 0.01 | 1 | Softmax |  |
| 50 | 2 | 0.001 | 1 | Softmax |  |
| 100 | 2 | 0.001 | 1 | Softmax |  |
| 50 | 5 | 0.01 | 1 | hs |  |
| 100 | 5 | 0.01 | 1 | hs |  |
| 50 | 5 | 0.001 | 1 | hs |  |
| 100 | 5 | 0.001 | 1 | hs |  |
| 50 | 5 | 0.01 | 1 | Softmax |  |
| 100 | 5 | 0.01 | 1 | Softmax |  |
| 50 | 5 | 0.001 | 1 | Softmax |  |
| 100 | 5 | 0.001 | 1 | Softmax |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Epochs | Window size | Learning Rate | wordNgrams | Loss | Precision |
| 50 | 2 | 0.01 | 2 | hs | 0.666 |
| 100 | 2 | 0.01 | 2 | hs | 0.673 |
| 50 | 2 | 0.1 | 2 | hs | 0.669 |
| 100 | 2 | 0.1 | 2 | hs | 0.672 |
| 50 | 2 | 0.01 | 2 | Softmax | 0.751 |
| 100 | 2 | 0.01 | 2 | Softmax | 0.786 |
| 50 | 2 | 0.1 | 2 | Softmax | 0.793 |
| 100 | 2 | 0.1 | 2 | Softmax | 0.793 |
| 50 | 5 | 0.01 | 2 | hs |  |
| 100 | 5 | 0.01 | 2 | hs |  |
| 50 | 5 | 0.1 | 2 | hs |  |
| 100 | 5 | 0.1 | 2 | hs |  |
| 50 | 5 | 0.01 | 2 | Softmax | 0.751 |
| 100 | 5 | 0.01 | 2 | Softmax | 0.786 |
| 50 | 5 | 0.1 | 2 | Softmax |  |
| 100 | 5 | 0.1 | 2 | Softmax |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Epochs | Precision | Recall | Learning Rate |
| 50 | 0.787 | 0.787 | 0.1 |
| 100 | 0.791 | 0.791 | 0.1 |
| 500 | 0.795 | 0.795 | 0.1 |
| 50 | 0.788 | 0.788 | 0.01 |
| 100 | 0.791 | 0.791 | 0.01 |
| 500 | 0.788 | 0.788 | 0.01 |
| 50 | 0.614 | 0.614 | 0.001 |
| 100 | 0.684 | 0.684 | 0.001 |
| 500 | 0.787 | 0.787 | 0.001 |

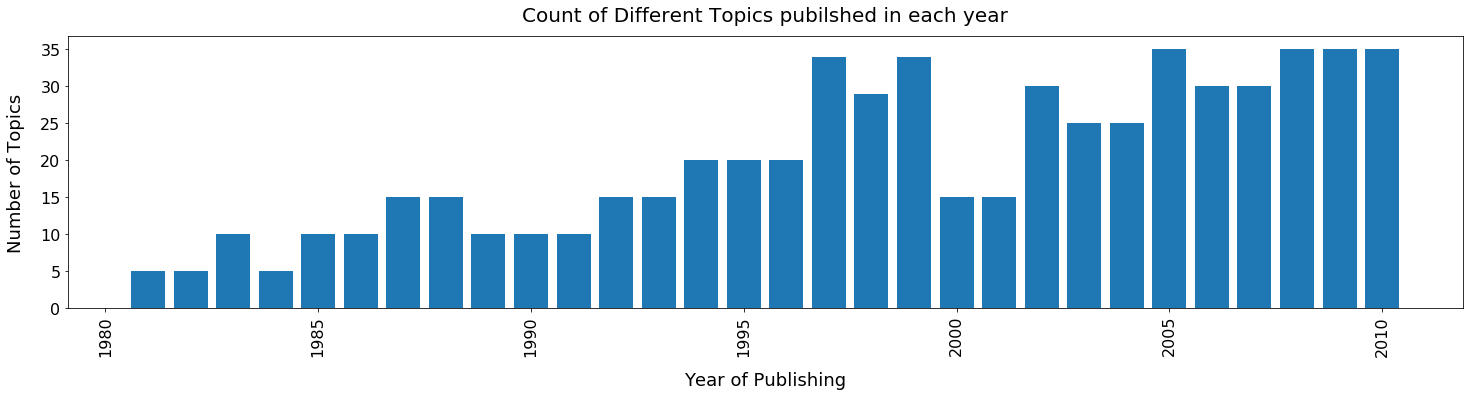
**Table 5‑4 Evaluation of fastText for various hyperparameters**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Precision | Recall | Accuracy | Balanced Accuracy | F1-score |
| Random Forest | 0.593 | 0.659 | 0.593 | 0.603 | 0.594 |
| SGD Classifier | 0.746 | 0.810 | 0.746 | 0.759 | 0.750 |
| Logistic Regression | 0.785 | 0.814 | 0.785 | 0.790 | 0.776 |
| Multinomial NB | 0.531 | 0.803 | 0.531 | 0.536 | 0.763 |

**Table 5‑5 Evaluation of different classification algorithms**

### Topic Modelling on Best-Selling books

## Visualization

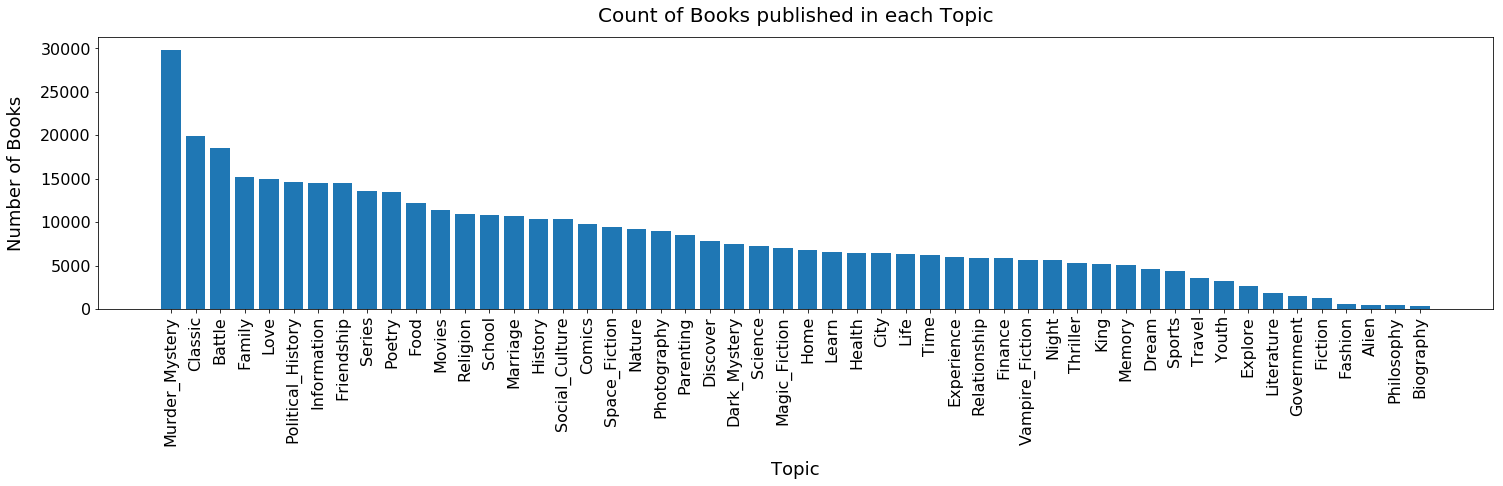


**Figure 5‑1 Number of Topics from 1981 to 2010**

In Figure 5-1, we can see the bar chart of number of topics published every year between 1981 to 2010. Number of topics published range between a minimum of 5 to a maximum of 35. The books published between 1981 to 1990, belonged to an average of 10 topics per year. Between 1991 to 2000, the average number of topics published doubled to 20. Between 2001 to 2010, the average number of topics published tripled to 30 topics per year. We can see a drastic drop in the number of topics from 35 in 1999 to 15 in 2000 and 2001 and doubles to 30 topics within tine year in 2002.

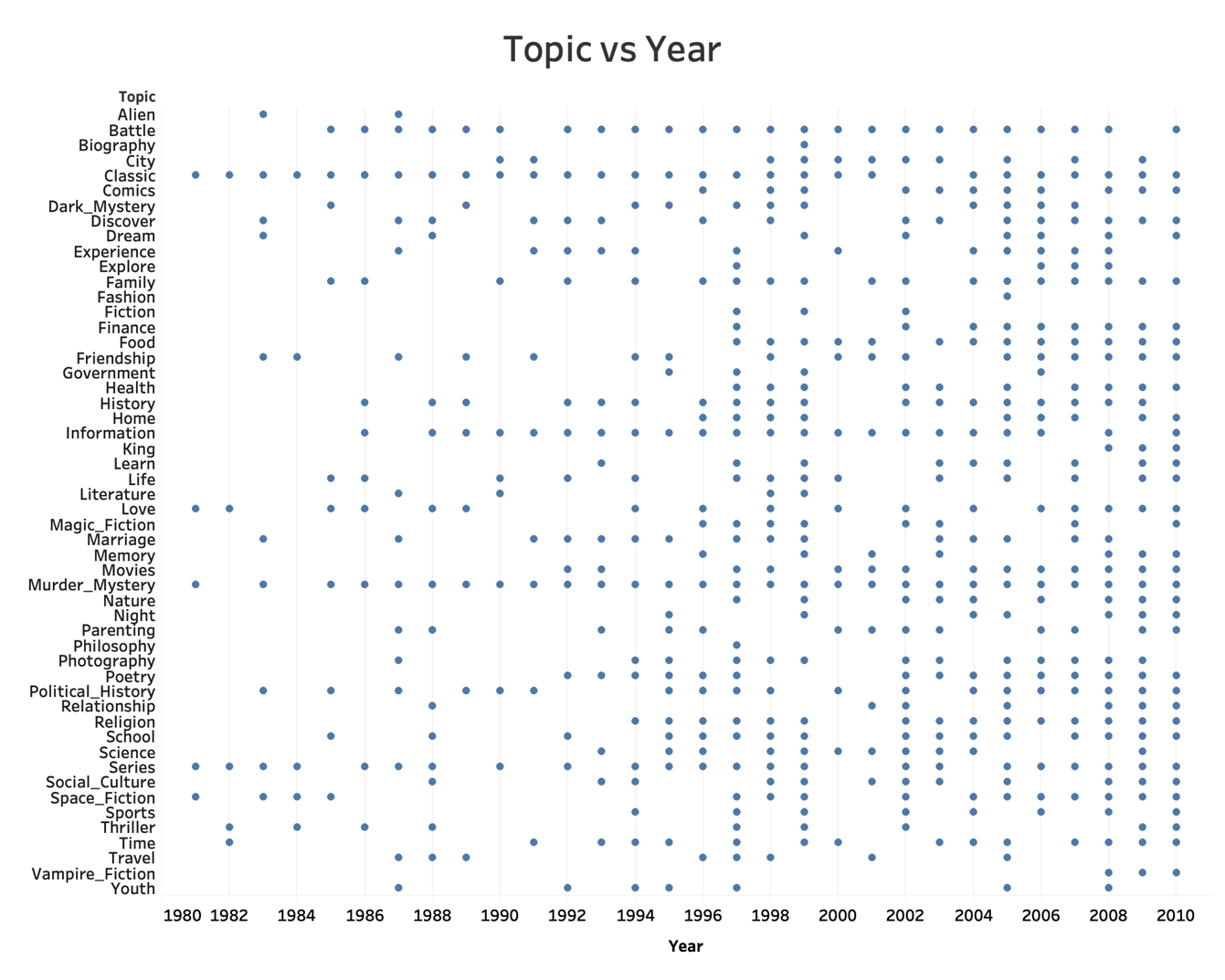
| Topic | **Count** | Topic | **Count** | Topic | **Count** | Topic | **Count** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Murder Mystery** | 29811 | **School** | 10810 | **Home** | 6808 | **Memory** | 5080 |
| **Classic** | 19935 | **Marriage** | 10716 | **Learn** | 6540 | **Dream** | 4628 |
| **Battle** | 18499 | **History** | 10397 | **Health** | 6449 | **Sports** | 4420 |
| **Family** | 15188 | **Social Culture** | 10330 | **City** | 6408 | **Travel** | 3569 |
| **Love** | 14987 | **Comics** | 9748 | **Life** | 6276 | **Youth** | 3220 |
| **Political History** | 14614 | **Space Fiction** | 9397 | **Time** | 6183 | **Explore** | 2665 |
| **Information** | 14508 | **Nature** | 9245 | **Experience** | 5926 | **Literature** | 1820 |
| **Friendship** | 14467 | **Photography** | 9006 | **Relationship** | 5888 | **Government** | 1453 |
| **Series** | 13578 | **Parenting** | 8507 | **Finance** | 5840 | **Fiction** | 1211 |
| **Poetry** | 13512 | **Discover** | 7791 | **Vampire Fiction** | 5674 | **Fashion** | 543 |
| **Food** | 12221 | **Dark Mystery** | 7442 | **Night** | 5625 | **Alien** | 449 |
| **Movies** | 11376 | **Science** | 7191 | **Thriller** | 5242 | **Philosophy** | 409 |
| **Religion** | 10882 | **Magic Fiction** | 6981 | **King** | 5224 | **Biography** | 385 |

**Table 5‑6 Count of books for each topic**



**Figure 5‑2 Number of Books per Topic (1981 - 2010)**

In Table 5-6, we can see out of total 429074 books, 29811 books belonged to Murder Mystery topics. 50% of the books published between 1981 to 2010 belonged to ‘Murder mystery’ , ‘Classic’ ,’Battle’ , Family’, Love’, ‘Political History’, ‘Information’, ‘Friendship’, ‘Series’, ‘ Poetry’, ‘ Food’, ‘Movies’, ‘Religion’ and ‘School’ Topics.



**Figure 5‑3 Topic vs Year Visualization**

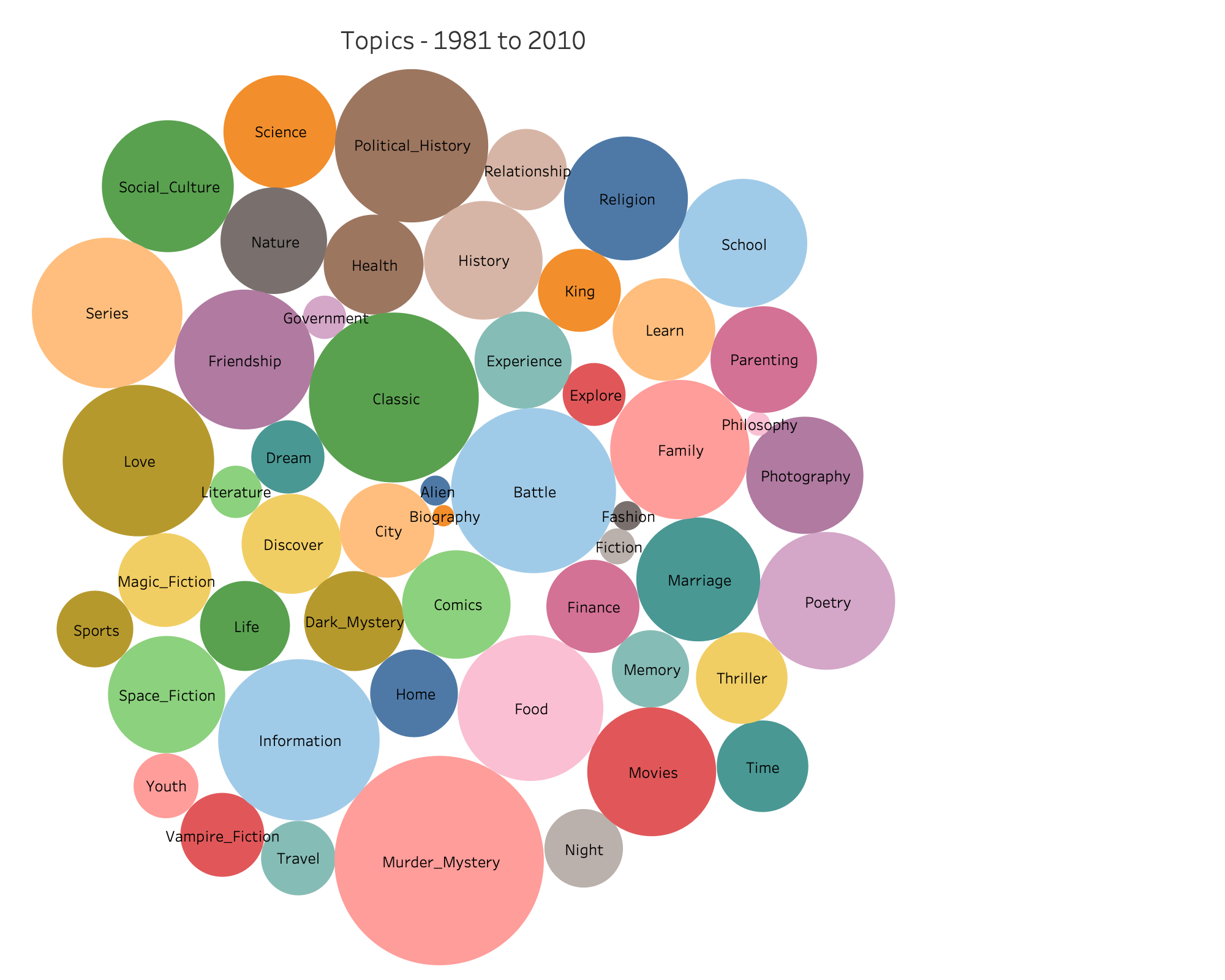
The Figure 5-3 shows the plot of topic for each year. Books related to ‘Alien’ topic first appeared in 1983 and then after a four-year gap in 1987, after which looks like alien topic detrended. Books related to topics ‘Biography’, ‘Philosophy’, ‘Fashion’ were published only once in the thirty decades. Books related to ‘Vampire Fiction’ became a trend in late 2000s and has been steadily published ever since 2008. This is when the movies related to Vampire Fiction ‘Twilight’ was released indicating the sudden trend. Books related to. ‘Finance’ also has been continuously published since 2004 which is also coincidently along the same timeline as the huge global recession period. Books on ‘Murder Mystery’ and ‘Classic’ were published every year except for two years in the thirty-year period, indicating that these genres are the most successful.

## Future Scope

Text Classification has scope in many areas including information retrieval, analyzing the hidden messages, search and indexing, document classification,

## Conclusion

I sincerely hope that my findings on the trend of the topics published over the decades found provide an insight into the success of the publishing industry and also instigate them to dive further deep, find and compare the trends in journals, iteratures as well.



|  |  |
| --- | --- |
| **#** | **Topic - Alien** |
| 1 | world, human, future, city, time, seek, survive, alien, small, mysterious |
| 2 | power, planet, fight, space, battle, enemy, ship, survive, alien, hand |
| **#** | **Topic - Battle** |
| 1 | force, mission, fight, soldier, battle, group, military, enemy, attack, agent |
| 2 | world, force, powerful, land, fight, battle, face, journey, save, hope |
| 3 | force, land, evil, power, battle, magic, save, rule, destroy, lead |
| 4 | power, battle, fight, force, evil, magic, enemy, save, soldier, fantasy |
| 5 | adventure, force, world, land, lose, battle, evil, game, fight, back |
| 6 | battle, fight, force, british, mission, attack, enemy, military, face, soldier |
| 7 | power, force, powerful, battle, face, fight, magic, evil, save, hero |
| 8 | battle, fight, enemy, force, mission, soldier, military, hero, attack, survive |
| 9 | battle, fight, force, world, human, planet, soldier, mission, enemy, military |
| 10 | force, face, fight, battle, destroy, escape, enemy, save, mission, planet |
| 11 | find, time, long, battle, return, fight, face, hold, meet, leave |
| 12 | force, fight, battle, mission, soldier, rescue, attack, ship, enemy, survive |
| 13 | battle, fight, force, soldier, enemy, military, army, ship, join, lead |
| 14 | battle, fight, side, force, lead, group, military, soldier, survive, british |
| 15 | force, fight, battle, enemy, military, face, soldier, join, army, british |
| **#** | **Topic - Biography** |
| 1 | work, artist, music, famous, career, biography, portrait, image, interview, great |
| **#** | **Topic - City** |
| 1 | city, street, hand, turn, ride, train, fire, place, light, horse |
| 2 | city, year, place, call, live, black, bear, begin, street, white |
| 3 | year, place, city, people, town, find, home, land, journey, long |
| 4 | find, begin, live, back, call, bring, strange, land, force, turn |
| 5 | place, city, home, people, country, town, travel, live, year, house |
| 6 | face, struggle, lead, begin, power, year, force, follow, powerful, bring |
| 7 | life, woman, love, live, year, power, struggle, bring, bear, city |
| 8 | year, follow, call, lead, line, stand, turn, open, long, cross |
| 9 | make, find, thing, meet, happen, leave, night, feel, fall, hard |
| **#** | **Topic - Classic** |
| 1 | story, tale, author, character, adventure, reader, series, classic, collection, short |
| 2 | story, good, death, tale, classic, american, bring, writer, character, write |
| 3 | book, reader, series, story, tale, author, read, adventure, animal, character |
| 4 | story, world, time, tale, place, collection, adventure, short, series, good |
| 5 | book, author, edition, series, reader, classic, popular, original, include, read |
| 6 | story, make, book, good, publish, write, classic, american, character, reader |
| 7 | people, power, bring, great, life, classic, american, world, century, powerful |
| 8 | book, write, story, author, work, publish, edition, classic, writer, character |
| 9 | story, world, time, great, tale, character, play, classic, young, adventure |
| 10 | work, write, writer, collection, include, author, volume, publish, classic, edition |
| 11 | story, tale, author, character, collection, classic, work, short, fiction, death |
| 12 | work, edition, book, include, classic, volume, publish, original, text, author |
| 13 | story, book, write, author, reader, publish, classic, tale, read, character |
| 14 | story, author, tale, character, writer, fiction, short, reader, write, classic |
| 15 | story, life, tale, character, young, short, collection, brilliant, rich, reader |
| 16 | story, series, tale, adventure, character, classic, reader, author, feature, hero |
| 17 | story, tale, illustration, adventure, reader, classic, young, full\_color, child, illustrate |
| **#** | **Topic - Comics** |
| 1 | book, reader, illustration, artist, full\_color, create, animal, illustrate, page, picture |
| 2 | story, tale, character, animal, young, reader, fill, bring, world, humor |
| 3 | child, book, read, reader, animal, word, learn, page, young, perfect |
| 4 | book, animal, reader, perfect, page, picture, funny, illustration, baby, hilarious |
| 5 | good, thing, make, friend, start, funny, trouble, perfect, talk, feel |
| 6 | time, world, great, create, brilliant, rich, humor, bring, sense, master |
| 7 | reader, animal, child, picture, perfect, adventure, wonderful, illustration, funny, delight |
| 8 | make, thing, good, perfect, turn, hilarious, funny, back, full, ride |
| 9 | humor, funny, part, hilarious, character, time, full, drive, head, back |
| 10 | story, life, year, memoir, funny, extraordinary, journey, tale, portrait, move |
| 11 | life, world, make, time, real, turn, give, sense, voice, funny |
| **#** | **Topic – Dark Mystery** |
| 1 | force, discover, land, adventure, face, back, long, lose, dark, future |
| 2 | secret, find, past, discover, dark, truth, mysterious, search, hide, escape |
| 3 | dark, evil, save, vampire, night, power, force, ghost, blood, strange |
| 4 | evil, power, magic, save, powerful, vampire, protect, magical, ancient, land |
| 5 | secret, dark, mysterious, past, force, seek, hide, power, dangerous, face |
| 6 | secret, truth, dark, city, discover, mysterious, street, hide, past, place |
| 7 | secret, truth, discover, hide, mysterious, dark, close, seek, dead, mystery |
| **#** | **Topic – Discover** |
| 1 | life, year, death, bear, live, world, black, begin, struggle, place |
| 2 | find, discover, secret, search, truth, mystery, lead, miss, mysterious, hide |
| 3 | begin, secret, year, discover, leave, mysterious, face, future, strange, night |
| 4 | discover, secret, young, long, bring, house, begin, back, night, dark |
| 5 | year, time, call, discover, century, body, truth, find, world, death |
| 6 | secret, discover, begin, long, force, back, strange, evil, turn, journey |
| 7 | secret, heart, passion, discover, hide, desire, determine, trust, save, danger |
| 8 | secret, dangerous, truth, hide, find, turn, deadly, street, protect, stop |
| 9 | make, find, year, time, real, good, discover, learn, turn, hard |
| 10 | find, discover, begin, return, secret, search, truth, home, leave, past |
| 11 | year, power, world, past, journey, give, lead, discover, truth, long |
| 12 | find, begin, lead, answer, question, discover, death, search, young, life |
| 13 | find, discover, secret, town, home, house, night, leave, strange, turn |
| **#** | **Topic – Dream** |
| 1 | life, year, live, dream, bring, hope, escape, begin, bear, survive |
| 2 | life, live, change, dream, learn, woman, make, true, time, discover |
| 3 | year, make, child, give, time, life, dream, play, good, learn |
| 4 | secret, past, dark, fear, face, begin, back, find, heart, leave |
| 5 | find, dream, lose, heart, past, back, return, hope, discover, long |
| 6 | life, love, dream, heart, hope, true, bring, live, share, feel |
| 7 | life, dream, make, find, face, hope, lose, choice, fear, back |
| 8 | year, past, life, death, lose, find, dream, leave, long, return |
| **#** | **Topic – Experience** |
| 1 | experience, problem, offer, body, health, people, relationship, show, common, advice |
| 2 | book, offer, guide, experience, problem, show, personal, provide, research, understand |
| 3 | book, make, work, good, part, experience, give, time, real, great |
| 4 | life, story, people, year, author, tale, great, experience, move, place |
| 5 | book, work, business, change, success, show, experience, successful, create, company |
| 6 | life, time, good, give, make, experience, true, mind, author, reveal |
| 7 | life, people, experience, show, personal, world, change, reveal, insight, relationship |
| 8 | describe, author, make, experience, mind, fact, research, human, draw, result |
| 9 | life, american, travel, personal, experience, account, reveal, portrait, early, create |
| 10 | book, reader, experience, word, read, personal, spiritual, offer, author, insight |
| 11 | life, story, year, live, account, death, people, reveal, experience, personal |
| **#** | **Topic – Explore** |
| 1 | study, explore, important, work, book, modern, provide, idea, theory, present |
| 2 | woman, social, society, political, cultural, sexual, explore, movement, identity, study |
| 3 | human, question, world, mind, science, answer, nature, explore, work, knowledge |
| 4 | world, people, human, mind, explore, nature, sense, place, draw, power |
| **#** | **Topic – Family** |
| 1 | family, child, father, mother, daughter, parent, husband, wife, year, baby |
| 2 | family, house, move, home, visit, arrive, ghost, sister, friend, lktb |
| 3 | family, child, father, mother, young, parent, home, year, brother, sister |
| 4 | family, young, home, father, leave, girl, mother, grow, wife, house |
| 5 | love, woman, heart, young, daughter, husband, wife, brother, sister, beautiful |
| 6 | family, father, mother, home, year, find, brother, sister, daughter, young |
| 7 | live, child, year, family, friend, home, life, find, leave, young |
| 8 | life, family, time, live, friend, year, child, home, people, leave |
| 9 | family, child, father, mother, young, love, parent, daughter, brother, sister |
| 10 | family, father, child, mother, young, parent, brother, sister, daughter, home |
| 11 | family, child, young, father, mother, girl, parent, daughter, home, leave |
| 12 | family, year, mother, father, life, daughter, brother, sister, love, young |
| **#** | **Topic – Fashion** |
| 1 | make, hand, gift, room, perfect, piece, project, pattern, fashion, wear |
| **#** | **Topic – Fiction** |
| 1 | story, tale, character, time, reader, author, great, write, fiction, read |
| 2 | write, writer, story, collection, work, short, fiction, literary, american, writing |
| 3 | great, world, time, tale, power, bring, legend, young, epic, create |
| **#** | **Topic – Finance** |
| 1 | make, question, answer, high, choice, give, choose, money, plan, decision |
| 2 | make, business, show, success, high, career, money, company, give, real |
| 3 | business, success, company, make, successful, show, money, create, people, career |
| 4 | business, make, work, money, show, company, real, successful, success, create |
| 5 | business, work, people, show, offer, change, problem, success, company, make |
| 6 | business, show, success, successful, work, build, make, company, high, money |
| 7 | business, make, work, money, time, plan, company, start, good, career |
| 8 | business, work, make, money, company, career, sell, successful, market, offer |
| **#** | **Topic – Food** |
| 1 | make, perfect, animal, simple, food, picture, good, create, recipe, illustration |
| 2 | food, make, easy, recipe, home, simple, good, include, plant, fresh |
| 3 | make, food, recipe, good, easy, holiday, home, enjoy, perfect, time |
| 4 | make, food, recipe, good, easy, plant, garden, simple, create, home |
| 5 | make, food, recipe, color, create, simple, perfect, easy, tree, style |
| 6 | make, food, good, easy, recipe, simple, perfect, holiday, include, home |
| 7 | make, good, food, recipe, time, party, home, prepare, holiday, easy |
| 8 | make, play, game, food, simple, recipe, easy, garden, home, favorite |
| 9 | make, food, recipe, easy, good, grow, enjoy, garden, simple, cook |
| 10 | make, food, good, recipe, grow, garden, easy, special, prepare, plant |
| 11 | book, make, reader, animal, simple, illustration, picture, food, create, fill |
| 12 | make, food, animal, perfect, recipe, favorite, simple, good, home, enjoy |
| **#** | **Topic – Friendship** |
| 1 | life, great, time, people, live, story, good, city, home, friend |
| 2 | young, friend, girl, house, find, brother, sister, adventure, meet, grow |
| 3 | friend, good, thing, girl, turn, start, make, perfect, summer, plan |
| 4 | life, make, year, time, world, live, great, place, play, back |
| 5 | friend, good, girl, school, high, thing, turn, start, friendship, year |
| 6 | make, child, family, young, friend, year, home, girl, leave, thing |
| 7 | friend, good, thing, school, find, turn, girl, start, back, decide |
| 8 | find, friend, turn, house, year, town, life, leave, time, young |
| 9 | friend, good, find, make, time, thing, turn, back, start, school |
| 10 | friend, good, girl, school, find, start, turn, thing, meet, decide |
| 11 | friend, good, school, girl, thing, high, start, make, boyfriend, meet |
| 12 | good, find, make, friend, time, thing, turn, start, plan, back |
| 13 | family, year, child, life, young, live, find, friend, home, girl |
| 14 | good, thing, friend, time, back, happen, turn, start, night, trouble |
| 15 | friend, good, girl, school, thing, find, meet, decide, start, happen |
| **#** | **Topic – Government** |
| 1 | lead, black, white, give, attempt, order, free, country, group, half |
| 2 | political, american, society, social, black, politic, community, state, movement, public |
| 3 | people, country, world, group, community, state, event, nation, government, american |
| 4 | life, year, change, time, country, grow, lose, continue, past, generation |
| **#** | **Topic – Health** |
| 1 | body, show, book, change, problem, health, doctor, energy, care, mind |
| 2 | relationship, experience, body, people, offer, change, show, mind, problem, health |
| 3 | problem, care, show, body, relationship, doctor, health, brain, deal, common |
| 4 | mind, people, body, relationship, doctor, experience, care, deal, problem, health |
| 5 | offer, show, body, problem, common, mind, practical, guide, medical, care |
| 6 | body, problem, learn, health, people, mind, book, relationship, disease, physical |
| 7 | body, care, life, face, deal, problem, doctor, people, relationship, work |
| 8 | body, problem, health, care, show, energy, doctor, treatment, behavior, medical |
| 9 | problem, book, body, people, offer, guide, practical, health, approach, experience |
| **#** | **Topic – History** |
| 1 | great, story, history, american, extraordinary, remarkable, life, black, portrait, narrative |
| 2 | follow, author, word, great, history, present, thought, continue, generation, scene |
| 3 | work, history, american, writer, part, account, reveal, literary, society, event |
| 4 | work, history, write, author, american, volume, modern, early, account, century |
| 5 | history, american, account, event, early, reveal, fascinating, detail, present, culture |
| 6 | history, society, political, american, early, account, social, event, culture, show |
| 7 | american, great, country, bring, challenge, brilliant, black, lead, figure, history |
| 8 | history, modern, century, historical, ancient, early, tradition, present, religious, religion |
| 9 | history, great, modern, century, early, historical, american, culture, fascinating, ancient |
| **#** | **Topic – Home** |
| 1 | home, find, town, house, leave, summer, back, return, place, year |
| 2 | good, make, fill, home, perfect, follow, food, gift, special, long |
| 3 | back, house, head, turn, real, home, room, hand, watch, black |
| 4 | find, home, town, house, back, night, summer, plan, friend, room |
| 5 | town, home, house, find, village, grow, summer, local, place, horse |
| 6 | town, find, house, home, night, time, hear, local, leave, thing |
| 7 | home, find, city, house, town, place, small, begin, country, live |
| 8 | friend, good, town, house, girl, home, find, move, summer, turn |
| **#** | **Topic – Information** |
| 1 | book, include, make, text, read, give, word, show, cover, language |
| 2 | include, guide, book, design, information, create, feature, project, cover, easy |
| 3 | book, study, provide, present, offer, language, problem, show, work, student |
| 4 | book, study, present, show, provide, modern, problem, people, offer, question |
| 5 | book, guide, include, information, easy, offer, design, provide, practical, chapter |
| 6 | book, include, guide, provide, information, student, problem, offer, chapter, explain |
| 7 | book, show, offer, guide, learn, problem, understand, student, design, explain |
| 8 | book, guide, provide, student, work, information, study, field, design, practical |
| 9 | book, include, guide, provide, cover, series, information, read, feature, reader |
| 10 | include, guide, provide, book, information, cover, offer, easy, technique, advice |
| 11 | book, include, reader, text, guide, cover, provide, read, information, series |
| 12 | book, include, edition, original, volume, make, feature, page, present, complete |
| 13 | guide, include, book, information, provide, design, easy, technique, cover, practical |
| 14 | include, book, edition, text, provide, cover, chapter, guide, language, information |
| 15 | include, text, provide, edition, introduction, student, chapter, comprehensive, cover, information |
| **#** | **Topic – King** |
| 1 | power, battle, force, great, fight, king, land, powerful, rule, warrior |
| 2 | battle, fight, power, force, land, enemy, face, great, king, rule |
| 3 | power, magic, world, ancient, powerful, save, magical, battle, king, evil |
| **#** | **Topic – Learn** |
| 1 | make, learn, people, give, feel, real, time, life, person, matter |
| 2 | book, learn, guide, student, practical, design, create, teacher, skill, provide |
| 3 | guide, learn, skill, develop, practical, build, create, tool, provide, strategy |
| 4 | book, student, system, design, provide, develop, program, technique, basic, computer |
| 5 | build, learn, develop, base, work, create, system, practical, show, guide |
| 6 | learn, business, teacher, practice, skill, student, show, teach, practical, book |
| 7 | business, work, book, system, develop, process, build, computer, technology, create |
| 8 | make, find, time, life, thing, learn, people, feel, good, give |
| 9 | find, make, town, small, plan, friend, star, good, learn, head |
| 10 | book, reader, read, author, learn, write, make, experience, teach, page |
| **#** | **Topic – Life** |
| 1 | story, tale, character, life, rich, extraordinary, author, portrait, remarkable, brilliant |
| 2 | life, change, live, year, time, bring, long, forever, face, lose |
| 3 | time, life, story, live, good, people, long, bring, career, share |
| 4 | year, turn, life, make, find, play, good, move, give, high |
| 5 | life, live, change, world, year, time, people, bring, move, story |
| 6 | life, time, world, great, people, year, bring, good, make, change |
| 7 | make, good, time, people, year, life, child, turn, friend, learn |
| 8 | life, live, personal, experience, people, spiritual, world, faith, heart, true |
| 9 | life, make, learn, give, school, change, follow, time, hard, choose |
| 10 | story, life, move, year, memory, bring, childhood, live, struggle, remarkable |
| 11 | life, love, live, change, year, begin, dream, death, face, struggle |
| 12 | make, people, life, live, change, work, give, real, world, hand |
| **#** | **Topic – Literature** |
| 1 | work, write, author, writer, publish, reveal, literary, volume, literature, collection |
| 2 | work, writer, character, write, play, century, fiction, great, modern, literature |
| 3 | work, write, writer, poem, collection, writing, reader, publish, literary, literature |
| 4 | work, write, poem, writing, collection, poet, language, literature, writer, poetry |
| **#** | **Topic – Love** |
| 1 | woman, desire, passion, heart, sexy, marriage, marry, sexual, beauty, handsome |
| 2 | love, woman, life, make, dream, heart, beautiful, fall, give, rich |
| 3 | woman, desire, passion, sexual, lover, give, body, female, pleasure, break |
| 4 | love, woman, young, father, beautiful, mother, heart, daughter, passion, brother |
| 5 | life, love, woman, young, family, live, child, death, friend, leave |
| 6 | make, heart, love, long, night, passion, secret, body, meet, hand |
| 7 | love, passion, marriage, heart, husband, beautiful, woman, wife, fall, meet |
| 8 | love, heart, passion, desire, woman, dream, sexy, beauty, meet, beautiful |
| 9 | love, woman, young, father, mother, beautiful, heart, daughter, wife, husband |
| 10 | love, heart, dream, passion, desire, beautiful, meet, beauty, feel, handsome |
| 11 | woman, love, husband, heart, marriage, passion, beautiful, desire, fall, wife |
| 12 | woman, desire, passion, heart, night, sexy, lover, sexual, beautiful, beauty |
| 13 | woman, life, face, young, fear, draw, struggle, past, long, sexual |
| **#** | **Topic – Magic Fiction** |
| 1 | world, power, magic, ancient, ghost, evil, magical, spirit, dark, strange |
| 2 | world, power, human, future, powerful, magic, great, force, struggle, planet |
| 3 | power, magic, evil, world, dark, adventure, save, land, rule, powerful |
| 4 | world, power, magic, ancient, evil, great, land, fantasy, adventure, rule |
| 5 | power, magic, evil, land, powerful, great, bring, rule, epic, young |
| 6 | world, human, great, power, give, reality, vision, powerful, brilliant, earth |
| 7 | save, human, evil, destroy, power, force, powerful, magic, face, race |
| 8 | world, adventure, ancient, great, magic, power, magical, mysterious, discover, land |
| **#** | **Topic – Marriage** |
| 1 | woman, love, young, husband, wife, fall, marriage, beautiful, meet, heart |
| 2 | husband, wife, marriage, marry, find, plan, meet, young, wedding, leave |
| 3 | love, woman, beautiful, heart, dream, passion, meet, marriage, husband, beauty |
| 4 | love, heart, beautiful, passion, marriage, young, woman, desire, fall, beauty |
| 5 | love, woman, heart, beautiful, dream, passion, husband, beauty, marriage, desire |
| 6 | love, heart, husband, wife, marriage, beautiful, passion, fall, woman, romance |
| 7 | love, woman, heart, passion, marriage, husband, fall, desire, beautiful, wife |
| 8 | love, passion, marriage, husband, heart, fall, wife, beautiful, desire, marry |
| 9 | woman, husband, love, marriage, wife, beautiful, meet, marry, heart, fall |
| 10 | woman, love, life, wife, young, husband, marriage, fall, beautiful, daughter |
| 11 | love, woman, heart, beautiful, marriage, passion, husband, fall, desire, wife |
| **#** | **Topic – Memory** |
| 1 | secret, past, truth, discover, hide, dark, mysterious, memory, begin, reveal |
| 2 | story, life, character, tale, reader, extraordinary, memory, bring, portrait, voice |
| 3 | secret, find, discover, truth, past, begin, search, mysterious, hide, mystery |
| 4 | life, story, move, experience, memory, memoir, write, powerful, year, extraordinary |
| 5 | life, find, face, call, place, search, past, truth, live, memory |
| 6 | year, time, life, death, long, begin, face, memory, past, month |
| **#** | **Topic – Movies** |
| 1 | star, film, music, show, famous, career, scene, movie, time, record |
| 2 | play, film, career, part, star, music, game, famous, stage, team |
| 3 | play, game, star, team, film, music, great, time, famous, career |
| 4 | star, music, play, film, show, career, scene, time, movie, famous |
| 5 | play, character, great, film, follow, reveal, create, portrait, music, myth |
| 6 | play, great, year, time, world, make, career, game, star, film |
| 7 | artist, work, music, film, movie, famous, scene, include, image, show |
| 8 | film, career, include, play, music, famous, show, star, interview, record |
| 9 | film, music, star, character, popular, series, stage, late, movie, scene |
| 10 | play, film, star, music, scene, time, career, movie, stage, famous |
| 11 | play, film, show, star, career, business, game, music, make, scene |
| **#** | **Topic – Murder Mystery** |
| 1 | murder, case, mystery, find, death, crime, dead, killer, victim, police |
| 2 | death, murder, find, secret, mystery, kill, case, body, dead, dark |
| 3 | murder, town, turn, mystery, kill, dead, call, case, island, street |
| 4 | murder, case, crime, dead, death, killer, kill, victim, body, police |
| 5 | murder, case, mystery, dead, crime, find, death, killer, victim, kill |
| 6 | find, death, murder, body, mystery, kill, night, thing, leave, mysterious |
| 7 | find, death, discover, murder, secret, call, mystery, wife, kill, dead |
| 8 | find, death, murder, lead, begin, back, dead, secret, case, threaten |
| 9 | death, year, murder, country, struggle, body, case, mystery, mind, people |
| 10 | find, murder, death, dead, case, mystery, kill, detective, lead, crime |
| 11 | murder, case, mystery, dead, crime, killer, find, kill, death, victim |
| 12 | murder, case, crime, dead, death, kill, killer, find, mystery, victim |
| 13 | murder, find, mystery, dead, case, death, kill, crime, killer, body |
| 14 | murder, case, killer, crime, death, dead, body, victim, kill, detective |
| **#** | **Topic – Nature** |
| 1 | animal, adventure, water, bear, wild, tree, bring, mountain, creature, young |
| 2 | travel, animal, land, adventure, water, place, journey, island, trip, mountain |
| 3 | world, travel, journey, adventure, place, island, land, trip, water, discovery |
| 4 | city, place, travel, country, journey, small, town, people, land, adventure |
| 5 | travel, city, place, land, country, journey, ancient, adventure, island, water |
| 6 | city, travel, adventure, place, land, journey, ship, island, water, mountain |
| 7 | journey, travel, adventure, land, village, people, country, island, place, water |
| 8 | animal, land, place, bear, wild, small, mountain, make, ride, water |
| **#** | **Topic – Night** |
| 1 | find, year, call, night, discover, back, long, happen, begin, place |
| 2 | thing, turn, back, find, friend, start, night, good, trouble, head |
| 3 | time, back, money, night, leave, turn, good, long, late, sleep |
| 4 | town, house, night, home, find, place, room, back, local, walk |
| 5 | thing, back, head, night, time, wrong, call, hand, walk, stand |
| 6 | thing, find, turn, night, back, happen, head, wrong, trouble, start |
| **#** | **Topic - Parenting** |
| 1 | family, child, young, parent, year, grow, live, raise, home, bear |
| 2 | child, make, parent, baby, perfect, special, food, home, recipe, favorite |
| 3 | child, life, relationship, parent, grow, live, learn, experience, baby, young |
| 4 | family, child, home, young, mother, girl, friend, father, grow, year |
| 5 | family, year, life, child, home, live, young, father, mother, move |
| 6 | life, year, child, family, people, experience, grow, change, move, long |
| 7 | life, child, year, family, learn, make, live, young, home, mother |
| 8 | book, child, author, read, write, reader, time, world, give, parent |
| 9 | child, young, life, parent, grow, make, family, learn, time, live |
| 10 | family, child, father, young, mother, life, year, parent, home, daughter |
| 11 | family, child, year, father, home, mother, young, life, grow, parent |
| **#** | **Topic – Philosophy** |
| 1 | work, study, theory, important, offer, provide, development, understanding, text, philosophy |
| **#** | **Topic – Photography** |
| 1 | cover, work, base, illustrate, style, simple, detail, picture, photograph, artist |
| 2 | artist, create, include, design, book, image, work, feature, photograph, color |
| 3 | child, book, make, full\_color, animal, picture, gift, illustration, simple, color |
| 4 | create, artist, design, image, photograph, style, color, project, work, illustrate |
| 5 | artist, create, include, book, design, illustration, image, photograph, project, feature |
| 6 | create, artist, book, include, design, image, feature, work, photograph, illustration |
| 7 | book, illustration, reader, artist, page, picture, create, photograph, image, color |
| 8 | book, include, design, illustrate, color, create, guide, style, show, illustration |
| 9 | make, film, illustration, fill, picture, perfect, artist, page, color, create |
| 10 | work, create, artist, film, photograph, image, style, music, include, show |
| 11 | book, series, include, artist, illustrate, page, feature, create, photograph, popular |
| 12 | book, artist, include, work, create, image, photograph, feature, unique, color |
| **#** | **Topic – Poetry** |
| 1 | poem, word, voice, write, poetry, collection, form, poet, sense, work |
| 2 | work, write, poem, voice, poet, poetry, collection, form, spirit, song |
| 3 | work, poem, poet, volume, edition, publish, language, write, original, great |
| 4 | work, include, edition, text, volume, provide, introduction, language, poem, book |
| 5 | work, include, volume, edition, write, publish, introduction, writing, poem, original |
| 6 | life, year, place, bring, poem, give, travel, letter, poet, voice |
| 7 | work, write, writer, collection, author, short, writing, poem, publish, story |
| 8 | write, great, writer, poem, collection, american, fiction, language, poet, poetry |
| 9 | work, write, writer, collection, poem, writing, author, literary, publish, fiction |
| 10 | write, work, writer, author, collection, poem, writing, literary, poetry, publish |
| 11 | work, write, volume, publish, edition, poem, writer, collection, include, writing |
| 12 | write, work, writer, story, collection, author, poem, short, writing, character |
| 13 | work, write, writer, collection, publish, short, poem, include, author, writing |
| **#** | **Topic – Political History** |
| 1 | history, work, study, social, political, book, culture, relationship, theory, important |
| 2 | history, book, include, study, provide, political, important, account, social, historical |
| 3 | history, human, american, social, political, modern, explore, culture, society, historical |
| 4 | political, social, culture, cultural, issue, examine, study, society, movement, economic |
| 5 | history, american, century, modern, political, social, society, change, question, describe |
| 6 | history, social, political, american, society, account, change, role, state, community |
| 7 | history, culture, political, historical, society, modern, account, american, social, century |
| 8 | political, society, american, change, social, group, state, community, conflict, country |
| 9 | political, country, american, state, history, nation, government, politic, conflict, rise |
| 10 | american, political, history, social, society, state, culture, nation, government, politic |
| 11 | history, american, country, political, state, nation, account, century, early, culture |
| 12 | country, american, lead, nation, government, soldier, military, event, british, political |
| **#** | **Topic – Relationship** |
| 1 | show, offer, create, nature, provide, understand, mind, relationship, insight, result |
| 2 | experience, make, personal, people, learn, mind, relationship, give, offer, body |
| 3 | love, heart, fall, meet, romance, find, relationship, true, expect, plan |
| 4 | love, woman, life, young, heart, find, relationship, fall, meet, bring |
| 5 | love, heart, life, relationship, hope, true, fall, feel, lose, find |
| 6 | love, heart, dream, life, fall, find, true, meet, relationship, chance |
| 7 | woman, life, change, live, make, give, relationship, people, lead, free |
| **#** | **Topic – Religion** |
| 1 | question, answer, faith, spiritual, church, personal, offer, people, christian, wisdom |
| 2 | question, faith, religious, church, philosophy, religion, christian, book, answer, today |
| 3 | life, experience, live, spiritual, personal, people, share, give, heart, wisdom |
| 4 | life, experience, spiritual, faith, bring, personal, insight, vision, death, inspire |
| 5 | life, spiritual, personal, experience, faith, offer, wisdom, spirit, church, insight |
| 6 | history, culture, modern, historical, explore, tradition, century, early, religious, ancient |
| 7 | book, nature, experience, explore, present, offer, question, science, spiritual, show |
| 8 | life, relationship, experience, question, people, spiritual, mind, personal, faith, offer |
| 9 | life, experience, personal, faith, question, offer, people, church, spiritual, share |
| 10 | life, experience, personal, spiritual, insight, offer, teach, today, give, wisdom |
| 11 | life, woman, experience, reveal, personal, relationship, struggle, spiritual, work, explore |
| **#** | **Topic – School** |
| 1 | school, high, girl, game, play, year, student, team, class, college |
| 2 | friend, thing, girl, school, night, back, town, dream, happen, start |
| 3 | student, study, research, provide, approach, text, teacher, understand, theory, field |
| 4 | make, friend, good, give, find, learn, school, thing, plan, girl |
| 5 | student, guide, provide, include, information, chapter, practical, resource, essential, teacher |
| 6 | student, book, chapter, study, provide, teacher, approach, text, research, development |
| 7 | book, student, reader, text, include, provide, chapter, language, teacher, study |
| 8 | child, make, find, friend, thing, school, learn, begin, back, good |
| 9 | make, good, play, learn, give, show, school, animal, game, business |
| 10 | book, child, reader, read, learn, school, word, parent, make, author |
| 11 | good, make, friend, thing, time, start, head, school, happen, perfect |
| 12 | make, learn, give, school, teach, teacher, answer, good, show, student |
| **#** | **Topic – Science** |
| 1 | study, book, text, problem, issue, idea, work, approach, provide, theory |
| 2 | student, idea, work, science, research, show, study, provide, develop, explain |
| 3 | study, question, theory, work, idea, science, issue, nature, development, view |
| 4 | science, book, research, work, provide, student, base, system, development, language |
| 5 | study, human, issue, research, book, science, theory, idea, question, philosophy |
| 6 | idea, book, nature, research, theory, study, science, question, culture, show |
| 7 | study, book, science, theory, research, explore, field, idea, understand, work |
| 8 | world, human, city, people, mind, game, control, create, race, science |
| 9 | human, question, world, science, nature, explore, work, mind, answer, scientist |
| 10 | human, nature, science, theory, idea, modern, philosophy, explore, question, scientific |
| 11 | human, work, world, science, idea, nature, explore, theory, mind, philosophy |
| **#** | **Topic – Series** |
| 1 | work, book, include, history, edition, century, volume, present, study, write |
| 2 | book, work, history, include, author, write, reader, edition, publish, present |
| 3 | book, write, include, author, read, publish, edition, volume, reader, text |
| 4 | book, study, provide, work, offer, important, language, nature, present, question |
| 5 | book, work, series, include, study, part, make, present, provide, reader |
| 6 | book, include, reader, edition, read, series, word, full, page, cover |
| 7 | write, book, play, great, writer, time, publish, century, volume, early |
| 8 | include, provide, edition, text, student, chapter, study, book, cover, guide |
| 9 | work, text, include, provide, introduction, student, volume, present, essay, history |
| 10 | book, include, edition, series, reader, author, work, volume, introduction, text |
| 11 | story, author, writer, series, character, tale, write, collection, short, fiction |
| 12 | book, include, edition, reader, series, text, read, volume, introduction, feature |
| 13 | study, question, issue, examine, important, essay, critical, work, challenge, book |
| 14 | book, reader, series, author, read, original, word, page, write, feature |
| 15 | book, reader, edition, series, include, read, author, feature, classic, text |
| 16 | book, reader, series, author, read, write, classic, original, volume, popular |
| 17 | edition, include, series, volume, reader, feature, classic, original, text, present |
| **#** | **Topic – Social Culture** |
| 1 | social, political, society, culture, examine, issue, cultural, today, role, explore |
| 2 | political, american, social, state, country, public, community, nation, politic, society |
| 3 | history, culture, study, modern, cultural, explore, examine, historical, early, major |
| 4 | history, political, social, culture, explore, society, study, examine, issue, role |
| 5 | social, study, explore, examine, issue, science, contemporary, culture, development, modern |
| 6 | history, culture, modern, explore, cultural, examine, historical, social, contemporary, study |
| 7 | history, social, american, political, society, cultural, culture, examine, historical, study |
| 8 | history, social, political, society, culture, study, cultural, role, examine, historical |
| 9 | history, work, study, modern, social, historical, political, author, culture, film |
| 10 | social, culture, study, examine, cultural, explore, issue, political, development, economic |
| 11 | study, culture, issue, examine, explore, cultural, provide, social, book, contemporary |
| **#** | **Topic – Space Fiction** |
| 1 | world, time, ancient, people, light, earth, future, space, form, universe |
| 2 | world, time, city, power, future, begin, change, bring, save, earth |
| 3 | human, world, planet, rule, star, fantasy, part, prove, race, king |
| 4 | story, world, time, human, tale, adventure, place, land, planet, lead |
| 5 | power, woman, world, begin, human, call, people, force, adventure, long |
| 6 | world, human, future, planet, power, race, space, humanity, scientist, earth |
| 7 | world, human, future, earth, space, group, race, reality, year, light |
| 8 | world, future, time, human, people, race, power, planet, space, earth |
| 9 | world, human, time, future, race, light, science, space, year, earth |
| 10 | world, human, time, future, space, people, nature, reality, race, planet |
| 11 | world, human, time, future, year, begin, planet, place, space, discovery |
| **#** | **Topic – Sports** |
| 1 | star, game, team, hear, lktb, sound, hero, discover, roman, angel |
| 2 | play, game, team, time, ride, line, make, horse, rule, give |
| 3 | time, play, make, year, high, game, good, call, team, great |
| 4 | play, game, world, team, time, star, great, year, high, rule |
| 5 | world, play, great, game, year, high, team, career, real, bring |
| 6 | time, make, game, play, team, back, turn, call, break, head |
| 7 | play, world, game, time, team, high, give, make, call, action |
| 8 | year, game, great, world, high, team, full, lead, call, island |
| **#** | **Topic – Thriller** |
| 1 | find, back, discover, force, year, turn, human, murder, planet, people |
| 2 | woman, young, love, dream, beautiful, fall, past, return, escape, plan |
| 3 | love, find, young, family, child, friend, murder, heart, meet, turn |
| 4 | secret, past, death, dark, begin, truth, find, hide, mysterious, strange |
| 5 | save, mission, dangerous, force, protect, stop, danger, deadly, rescue, steal |
| 6 | death, begin, city, past, fear, force, dark, young, confront, evil |
| 7 | mystery, series, call, find, begin, adventure, miss, turn, late, bring |
| 8 | life, find, live, time, young, death, past, place, home, search |
| **#** | **Topic – Time** |
| 1 | human, power, time, great, century, change, city, struggle, lead, people |
| 2 | world, time, people, work, part, author, create, fascinating, place, full |
| 3 | make, time, give, part, call, place, hand, follow, long, work |
| 4 | make, time, give, people, learn, word, kind, hear, feel, call |
| 5 | people, life, live, time, face, black, find, year, white, free |
| 6 | world, human, time, future, great, begin, year, adventure, race, ancient |
| 7 | time, long, place, play, give, thing, make, true, back, bring |
| 8 | time, make, learn, part, give, good, start, show, thing, follow |
| 9 | year, begin, follow, word, people, time, story, death, bring, speak |
| 10 | life, time, live, change, world, people, year, make, place, long |
| 11 | year, time, long, find, back, begin, past, leave, place, lose |
| 12 | make, time, world, life, give, real, people, question, answer, fact |
| **#** | **Topic – Travel** |
| 1 | adventure, travel, journey, ship, water, companion, island, train, land, encounter |
| 2 | world, time, journey, light, control, encounter, ancient, choose, travel, century |
| 3 | world, great, tale, adventure, powerful, create, call, power, travel, human |
| 4 | city, adventure, journey, travel, place, find, world, search, island, encounter |
| 5 | world, adventure, find, land, journey, power, travel, magic, powerful, save |
| 6 | year, young, bring, return, begin, long, adventure, travel, back, time |
| 7 | find, dream, true, long, journey, life, travel, back, call, discover |
| **#** | **Topic – Vampire Fiction** |
| 1 | dark, vampire, power, magic, evil, demon, human, blood, soul, world |
| 2 | vampire, save, blood, demon, dead, protect, human, dark, kill, soul |
| 3 | secret, dark, mysterious, hide, vampire, strange, ghost, haunt, dead, deadly |
| **#** | **Topic - Youth** |
| 1 | year, make, people, find, live, place, part, past, young, answer |
| 2 | world, human, change, people, powerful, future, call, create, control, group |
| 3 | world, human, time, year, future, change, mind, power, return, begin |
| 4 | find, brother, leave, meet, young, return, lose, turn, plan, sister |
| 5 | year, time, good, back, find, make, turn, thing, call, high |

1. [Your first appendix]

**No table of figures entries found.**

1. [Your second appendix]

| Serial Number | **Columns** | **Data Type** | **Missing Values** |
| --- | --- | --- | --- |
| 1 | isbn | Integer | 983373 |
| 2 | text\_reviews\_count | Float | 524 |
| 3 | series | Float | 0 |
| 4 | country\_code | String | 490 |
| 5 | language\_code | String | 1060153 |
| 6 | popular\_shelves | String | 0 |
| 7 | asin | String | 1891138 |
| 8 | is\_ebook | String | 490 |
| 9 | average\_rating | Float | 524 |
| 10 | kindle\_asin | String | 1345725 |
| 11 | similar\_books | String | 0 |
| 12 | description | String | 412249 |
| 13 | format | String | 646754 |
| 14 | link | String | 524 |
| 15 | authors | String | 0 |
| 16 | publisher | String | 654372 |
| 17 | num\_pages | Float | 764133 |
| 18 | publication\_day | Integer | 1024429 |
| 19 | isbn13 | Integer | 780263 |
| 20 | publication\_month | Integer | 882945 |
| 21 | edition\_information | String | 2142646 |
| 22 | publication\_year | Integer | 599625 |
| 23 | url | String | 524 |
| 24 | image\_url | String | 490 |
| 25 | book\_id | Integer | 0 |
| 26 | ratings\_count | Integer | 524 |
| 27 | work\_id | Integer | 524 |
| 28 | title | String | 11 |
| 29 | title\_without\_series | String | 11 |

Glossary of Terms

List of References

**[1]** Mengting Wan, Julian McAuley, "[Item Recommendation on Monotonic Behavior Chains](https://www.google.com/url?q=https%3A%2F%2Fgithub.com%2FMengtingWan%2Fmengtingwan.github.io%2Fraw%2Fmaster%2Fpaper%2Frecsys18_mwan.pdf&sa=D&sntz=1&usg=AFQjCNGGcNRW1tSZKPWO0yZsr8mj7MkWuw)", in RecSys'18. [[bibtex](https://www.google.com/url?q=https%3A%2F%2Fdblp.uni-trier.de%2Frec%2Fbibtex%2Fconf%2Frecsys%2FWanM18&sa=D&sntz=1&usg=AFQjCNEy2HDVC1K59JJWybzGgq3MafQcWQ" \t "_blank)] <dataset>

**[2]** Mengting Wan, Rishabh Misra, Ndapa Nakashole, Julian McAuley, "[Fine-Grained Spoiler Detection from Large-Scale Review Corpora](https://www.google.com/url?q=https%3A%2F%2Farxiv.org%2Fpdf%2F1905.13416.pdf&sa=D&sntz=1&usg=AFQjCNFunKWkvxTa1tQ_xAMJdcgDfcWEig)", in ACL'19. <dataset>

**[3]**  M. A. Wajeed and T. Adilakshmi, "Semi-supervised text classification using enhanced KNN algorithm," 2011 World Congress on Information and Communication Technologies, Mumbai, 2011, pp. 138-142.

**[4]** David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993-1022,2003.

**[5]** P. Han, S. Shen, D. Wang and Y. Liu, "The influence of word normalization in English document clustering," 2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE), Zhangjiajie, 2012, pp. 116-120.

[6] R. Pandey and G. O. Mohler, "Evaluation of crime topic models: topic coherence vs spatial crime concentration," 2018 IEEE International Conference on Intelligence and Security Informatics (ISI), Miami, FL, 2018, pp. 76-78.

**[**7] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15). ACM, New York, NY, USA, 399-408. DOI: <https://doi.org/10.1145/2684822.2685324>

[8] A. El-Nasan, S. Veeramachaneni and G. Nagy, "Word discrimination based on bigram co-occurrences," Proceedings of Sixth International Conference on Document Analysis and Recognition, Seattle, WA, USA, 2001, pp. 149-153.

**[9]** CHAPTER DRAFT — | Stanford Lagunita. <https://lagunita.stanford.edu/c4x/Engineering/CS-224N/asset/slp4.pdf>

**[10]** B. V. Barde and A. M. Bainwad, "An overview of topic modeling methods and tools," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, 2017, pp. 745-750.

[**11]** A. Hassan and A. Mahmood, "Efficient Deep Learning Model for Text Classification Based on Recurrent and Convolutional Layers," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, 2017, pp. 1108-1113.

**[12]** A. Pulver and S. Lyu, "LSTM with working memory," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 845-851.  
doi: 10.1109/IJCNN.2017.7965940

[**13]** A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, "Bag of tricks for efficient text classification", Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2 Short Papers, pp. 427-431, April 2017.

**[14]** S. Bhoir, T. Ghorpade and V. Mane, "Comparative analysis of different word embedding models," 2017 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, 2017, pp. 1-4.

[15] El-Kishky, Ahmed et al. “Entropy-Based Subword Mining with an Application to Word Embeddings.” (2018).

[16] M. Allahyari and K. Kochut, "Automatic Topic Labeling Using Ontology-Based Topic Models," 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, 2015, pp. 259-264.

[17] L. Li, L. Xiao, N. Wang, G. Yang and J. Zhang, "Text classification method based on convolution neural network," 2017 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, 2017, pp. 1985-1989.

[18] A. B. Prasetijo, R. R. Isnanto, D. Eridani, Y. A. A. Soetrisno, M. Arfan and A. Sofwan, "Hoax detection system on Indonesian news sites based on text classification using SVM and SGD," 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), Semarang, 2017, pp. 45-49.

[19] S. Arora, R. Ge and A. Moitra, "Learning Topic Models -- Going beyond SVD," 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science, New Brunswick, NJ, 2012, pp. 1-10.

[20] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CoNLL '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 952-961.

[21] PA Publishing Yearbook 2018

Bibliography