National Library of New Zealand
Papers Past
Project Report

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Company

National Library of New Zealand - Emerson Vandy, Manager of *Papers Past*.

Students

Xiandong Cai - University of Canterbury - xca24@uclive.ac.nz
Yujie Cui - University of Canterbury - ycu23@uclive.ac.nz
Ben Faulks - University of Canterbury - bmf43@uclive.ac.nz

Abstract

Topic Modeling is an important tool in natural language processing. It is used to gain insight and makes inference on a corpus of documents. This report outlines the process in which we use topic modeling and some other tools and apply them to the Papers Past data set. Papers Past is a collection of New Zealand newspapers, journals and letters. This project focuses on solely the newspapers, here our aim is to gain a better understanding of the dataset so that we can make recommendations for applications of topic modeling for the National Library of New Zealand. As well as provide some future research leads for individuals willing to explore Papers Past. We began by loading and wrangling the dataset to get our 'clean' dataset while adding features like the title and regions columns. Then preprocess the data set, with tokenization, removal of stopwords and outputting bigrams. We chose to use the Latent Dirichlet Allocation for our algorithm, specifically the MALLET implementation. Due to resources and time limitations, we randomly sampled 20% of the dataset. The algorithm was able to find some interesting results. The significant topics given reflected our reference events. It gave us some topics we did not consider beforehand such illness epidemics. Expectedly, we found that topics vary over time and by period, some findings raised more questions than conclusions. Future work will have to be undertaken for definitive answers regarding this dataset. The subjectivity of text analysis makes difficult to make concrete conclusions, leaving room for further research.

Contents

Background

The New Zealand National Library has created a website called ‘*Papers Past’*. The aim of the website is to provide access to millions of digitised New Zealand newspapers. Although newspapers are the primary focus of the website, magazines and journals, diaries and letters, and parliamentary papers are also collected and accessible. The goal of *Papers Past* is continually updating the website to have new features and add more material.

As previously mentioned newspaper articles are the main focus of the website and contribute 99.4% to the collection. The collection includes digitised New Zealand and Pacific newspapers from the 19th and 20th century. Currently, users of the website can explore the collection by title or region or date. Users are also able to search and retrieve metadata through the DigitalNZ API.[1]

Papers Past has been focused on improving their website for their users’ benefit, focusing on clean user interface and making searching the website’s resources simpler to search and filter, as well as making it more stable and ready for the continuous growth of resources being added to the website.

To our understanding *Papers Past* and the National Library are working with local libraries to obtain sufficient licensing for the documents to be able to host them on the website. Some licensing agreements are easy to gain than others. The *Papers Past* dataset is in no way complete. “The newspapers on the Papers Past are only a proportion of all New Zealand newspapers”.[2]

[1] API information. Retrieved from: https://natlib.govt.nz/about-us/open-data/papers-past-metadata

[2] The quote is taken from the *Papers Past* ‘about’ page .Retrieved from https://paperspast.natlib.govt.nz/about

Introduction

The purpose of the New Zealand National Library is to “collect, connect and co-create knowledge”[3] all of the publications and resources created in New Zealand. The amount of digital content created is ever increasing, so the archiving workload is always increasing. Although the national library has ways of storing and providing open data in spreadsheets. There have also thought ahead improved their websites (including *Papers Past*) to handle the increase in resources.

New Zealand Library has made significant progress with the digitization of all the newspaper articles and with the current capabilities of the *Papers Past* website. Unfortunately, researchers lack the tools to handle large datasets and face the problem of every growing dataset. Basic tools limit queries to one or two words, which may still result in tens of thousands of matches. Examining such a large number of matches is a colossal task for researchers.

The proposed goal of this project is to use topic modeling techniques for the *Papers Past* Dataset. Topic Modeling is a well-known machine learning technique used in text mining. The dataset is a collection of 16 million texts. It includes newspapers articles and advertisements from New Zealand publications ranging from the years 1843 through to 1945. Topic modeling with ideally identify the broad-running topics and niche topics in New Zealand’s history.

The goal of this project will give better tools for researchers, historians, and the general public of New Zealand to easily analyze and query the dataset. Topic Modeling, and other tools will aid in granting insight into New Zealand’s publication history and allow the users to gain more understanding about the past. Ordinarily, a historian would have read through hundreds to thousands of newspapers and resources, to answer one question, and they aren’t guaranteed to find it.

So throughout this project, we wanted to use Topic Modeling to identify the interesting or informative topics over a period of time. This would save researchers or historians numerous hours of their time and reduce wasted and/or ineffective search time. So instead of the researchers searching for an interesting topic, Clusters of documents relating to a topic could be viewed, and documents with topics in common could be retrieved. We automatically cluster the papers into topics and we could then provide this to the researcher to greatly narrow their search to the most useful resources.

Another benefit of topic modeling it may reveal topics that weren’t considered as ‘important’ before. Picking up topics in the natural discourse that were previously missed, which could spark previously unconsidered areas of research. Topic Modeling can be done on decades or single years. This can show how discourse evolves, whether it be on a local or national level. This type of meta information about how a topic progresses over time would be beneficial to researchers. The traditional method of scrolling through the newspaper, may miss the greater picture.

Clearly, topic modeling has enormous advantages in the text mining field. New Zealand Library, researchers, historians, and the New Zealand public would benefit from Topic Modeling. There are lots of different approaches to Topic Modeling whether it be LDA or LSA methods to name two. Presenting our data in a meaningful way is absolutely imperative. Whether it be in a table, listing the topics by decade. We could also represent results visually such a word cloud or scatter plot for example.

[3] Quote is taken from the New Zealand Library Website, Retrieved from https://natlib.govt.nz/about-us

Literature & Existing Methods

Topic Modeling is a statistical modeling tool that can be used to abstract 'topics' or focus of a collection of documents (or a corpus). There are different approaches to Topic Modeling. It is important to discuss the different methods of Topic Modeling and cite some of the literature that discusses each approach. We can discuss the pros and cons of each method and why we chose one over another, with respect to the *Papers Past* dataset.

We will begin with Latent Semantic Analysis (LSA). LSA is a fundamental technique for topic modeling. The main idea of LSA is to take a matrix of documents and terms. It will then create a document-topic matrix and topic-terms matrix. LSA models will replace the raw counts of words the document-topic matrix into a term frequency-inverse document frequency (TF-IDF), .simply a weight balancing the word's frequency in the corpus against its frequency in the document. A term that has a large weight would appear frequently a given document and infrequently in the entire corpus, and therefore may tell us something distinctive about the document. So the rarity of a term across the corpus awards that term a higher weight.

This is important because normally machine learning tools aren't particularly good at capturing nuance and differentiating of a word used in context. For example, when speaking about a '*bank'* we could be talking about a financial institution where you keep your money. Or you could be we are talking about a '*bank*' next to a river. This is the concept of polysemy and it's an important distinction to make. So mapping all the tokens won't be helpful, the reason why LSA creates matrices and applies weights to terms, LSA will try to reveal topics that would have been missed.

LSA has many benefits, it is quick and simple to implement. It gets some good results too. However, some downsides. It is a linear model and won't perform well on a dataset with non-linear features. LSA also uses a singular value decomposition (SVD) which is quite computationally intensive and does not work well with adding in new data.
Mentioned previously, the National Library will need to constantly add in new resources to the dataset. Meaning LSA may not be the appropriate choice for this project.

Given we have covered so much of LSA, we can speak about pLSA or probabilistic Latent Semantic Analysis. To be expected, the two models act in a very similar way. pLSA uses a probabilistic method instead of using an SVD. pLSA can solve many of the problems that LSA has. But doesn't come without other issues. It can be difficult to assign probabilities to new documents. Also, the new parameters for pLSA grow linearly with the number of documents and can be prone to overfitting the data. pLSA does solve some problems of LSA but raises new ones of its own, while having similar results to an LSA model.

Latent Dirichlet Allocation is another probabilistic model and is one of the most popular in the text analysis field. The basic idea is that documents are represented as word probabilities. The words with the highest probabilities will give an overall idea of what the topic is. LDA is an unsupervised model and is very commonly used in topic modeling. It is a Bayesian version of pLSA.

Boyd-Graber (2009)[4] compared numerous models, including LDA. The paper finds that LDA generally performed better than LSA and pLSA and did a better job at predicting topics that matched the topics chosen by human users. Obviously having human-interpretable topics is a benefit and would be ideal for this project when it comes to analysis of the topics. However, some papers claim that is is difficult to know when LDA is performing optimally, and there is no objective metric to aid in hypermeter tuning. Like most topic modeling problems, training the model to get

[4] Reading Tea Leaves: How Humans Interpret Topic Models http://users.umiacs.umd.edu/~jbg/docs/nips2009-rtl.pdf

optimal results is a matter of trial and error. For this project, we chose to use the MALLET implementation of LDA.

The MALLET topic modeling package has an extremely efficient and scalable version of Gibbs sampling[5]. The package also contains methods for hyperparameter optimization (tuning every 10 iterations for example). The package is also good at inferring topics for new documents given the training model. The package also contains numerous ways to output the data, for example by topic keys or doc topics. The MALLET Implementation of LDA seems to be the ideal choice for the choice of the current topic modeling tools that exist. It meets most of the requirements that we need for this project and allows us to tune and adapt the project's design as we see the results. The only thing that may be an obstacle. Journals about MALLET state that the package is not designed for shorter documents. A significant proportion of *Papers Past* is made up of small documents[6]. It is unclear at this stage whether this will cause significant issues. We will discuss this idea further in the 'dataset summary' section and later on in our 'findings' section.

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

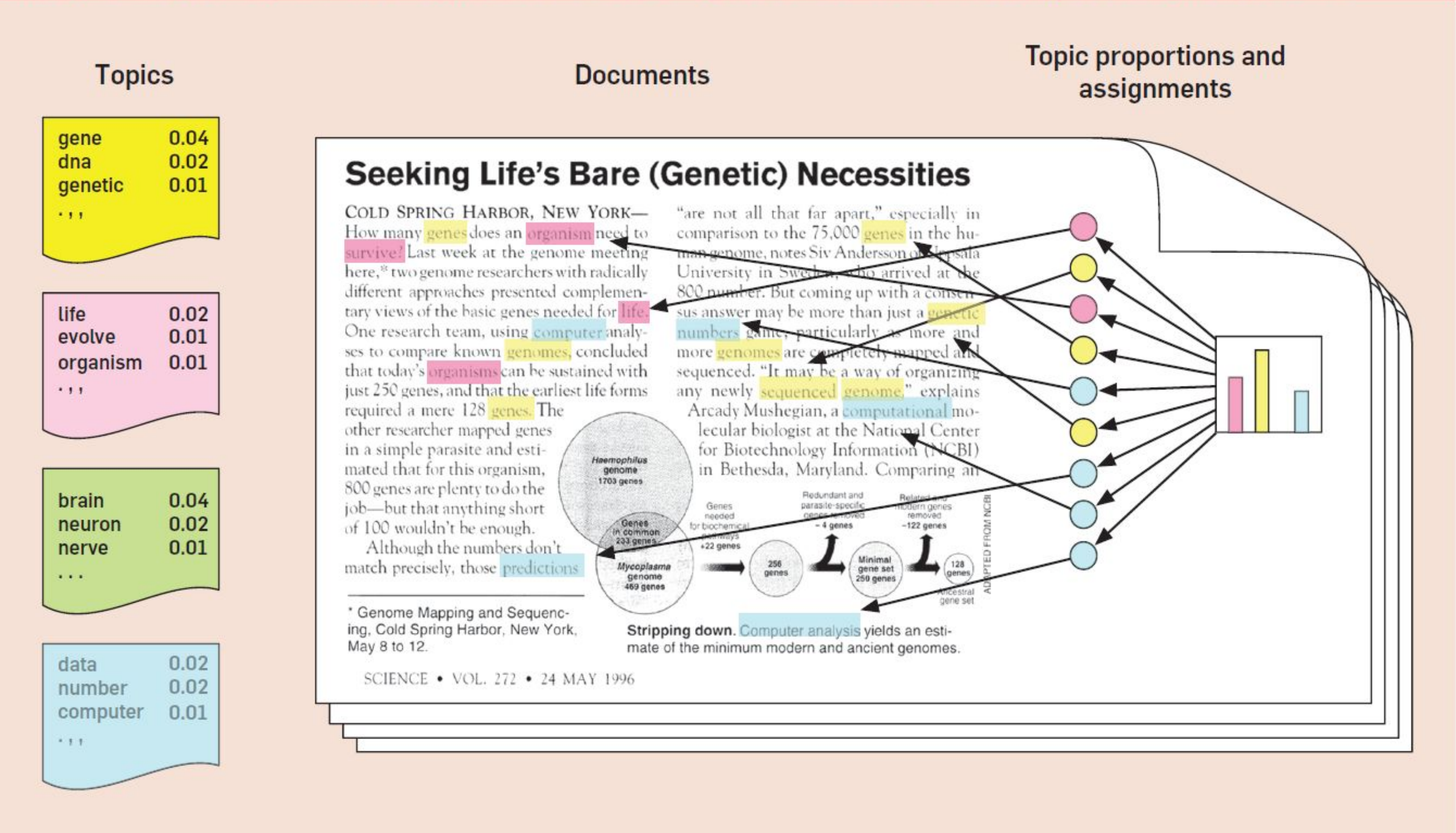


Fig 1: Introduction to LDA algorithm[7].

[5] Gibbs Sampling is discussed further in Bayesian Inference: Gibbs Sampling. http://www.mit.edu/~ilkery/papers/GibbsSampling.pdf

[6] More information found at Hong, L., & Davison, B. D. (2010) Empirical Study of Topic Modeling in Twitter

[7] Image retrieved from: https://www.google.co.nz/search?q=lda+algorithm&rlz=1C1GCEA_enNZ835NZ835&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjxguDE7qjgAhXZWisKHbQMC1kQ_AUIDigB&biw=1920&bih=938#imgrc=wZSOLc5fW1nbvM:

Research Questions

Our research questions will be divided into sections; implementation, interpretation, and application.

Implementation:

Part of the project to compare current topic modeling algorithms. We want to provide the National Library/*Papers Past* team with the best way to implement topic modeling dataset. In the literature section, we discussed some of the current models available to us. From our research, we saw that the MALLET LDA model was likely going to be the ideal solution for this project. We may find that this isn't the case in reality and some models may work better than others for this specific dataset.

Our questions regarding implementation:

- **How to identify "bad OCR" topics and eliminate the negative impact?** The OCR documents of historical newspapers are full of typos and missing letters, which would lead to meaningless topics after topic modeling.

- **How are we going to measure results and tune our parameters to obtain the best results?** Can we measure by topic comprehension or should we just use subjective measures?

- **How are we going to ensure that the code can handle the increasing size of the dataset?** We need to make it so that the same code is scalable.

Interpretation:

Another important aspect of this project is the interpretation of the results. In topic modeling, we get topics given to us based on what the model chooses. However, these topics are essentially meaningless unless we can try to match them to real historical events.

Our questions regarding interpretation:

- **What topics are dominant over the *Papers Past* corpus?** Are they what we would 'expect' and do they help us make sense of historical events?

- **How do the topics vary over the region?** Do different regions discuss the same topics more or less? For example, we could expect Otago publications to discuss the gold rush more than Auckland publications.

- **How do the topics vary over time?** How do topics fade from discourse, is there a sudden drop-off in discussion or does it slowly trend over decades.

- **How will we present our results?** We can show them in line graphs or scatter plots over time. We can see how many documents with a certain topic occur in a certain of time and how it trends over time. We could use compare the same topics across regions and decades. Finally, what would be most useful for likely users of this research?

Application:

Mentioned in the introduction our overall goal of this project. Is provide the National Library with advice on how they could implement Topic Modeling on the *Papers Past* website. Again, it is vital that it is still viable as the dataset increases in size.

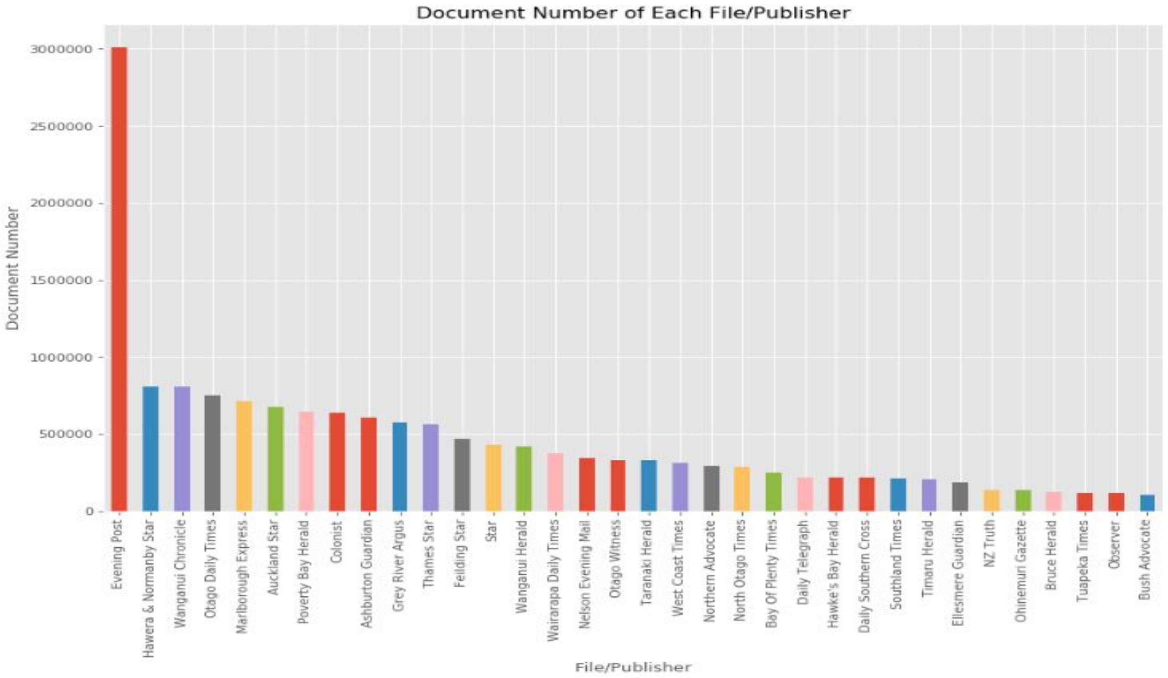
Our questions regarding application:

- **How we could efficiently organize the topics so you can make them easy to query?** What sort of interfaces could we the website use to effectively use topic modelling? Similar results recommendations or pronoun detection, for example.

- **What other future applications could there be for Topic Modeling?** Topic Modeling is the first step. There are some other future features that are outside the scope of this project but will use Topic Modeling. Such as ‘Spelling Correction’, ‘Pronoun detection’ and ‘Top topic contributors’. We will discuss features like this in the ‘future work’ section.

Dataset Summary

Our dataset is a collection New Zealand newspapers articles and advertisements. Gathered from 1839-1945. The total number of documents is 16,731,578 and equates to 33 gigabytes worth of data. There are 68 unique publishers in the dataset. There is a vast range of publication contribution. The least lines contributed by '*Albertland Gazette'* at 112 lines and the most lines were contributed by the '*Evening Post'* with 3,007,465 lines.



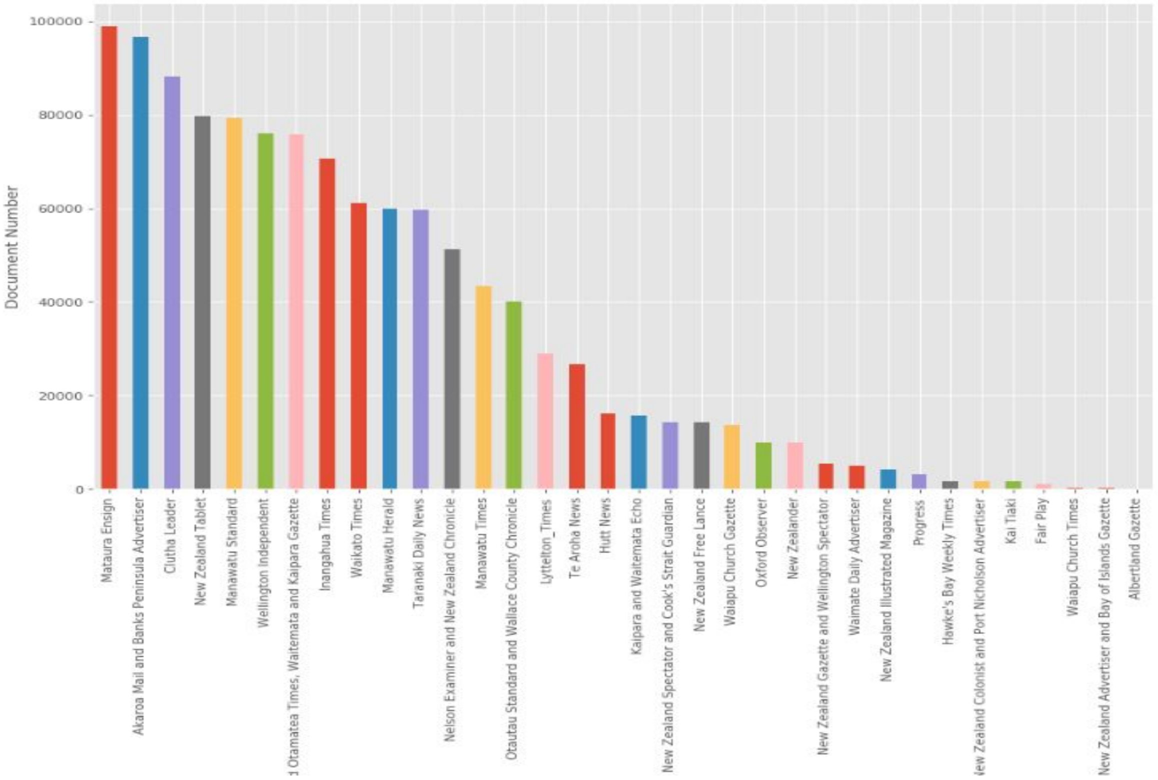


Fig 2: Document number of each file/publisher

In the previous section, it was mentioned that shorter texts do not contribute effectively to topic modeling. We need to find an appropriate number for the minimum number of words and remove the documents below the threshold from the dataset.

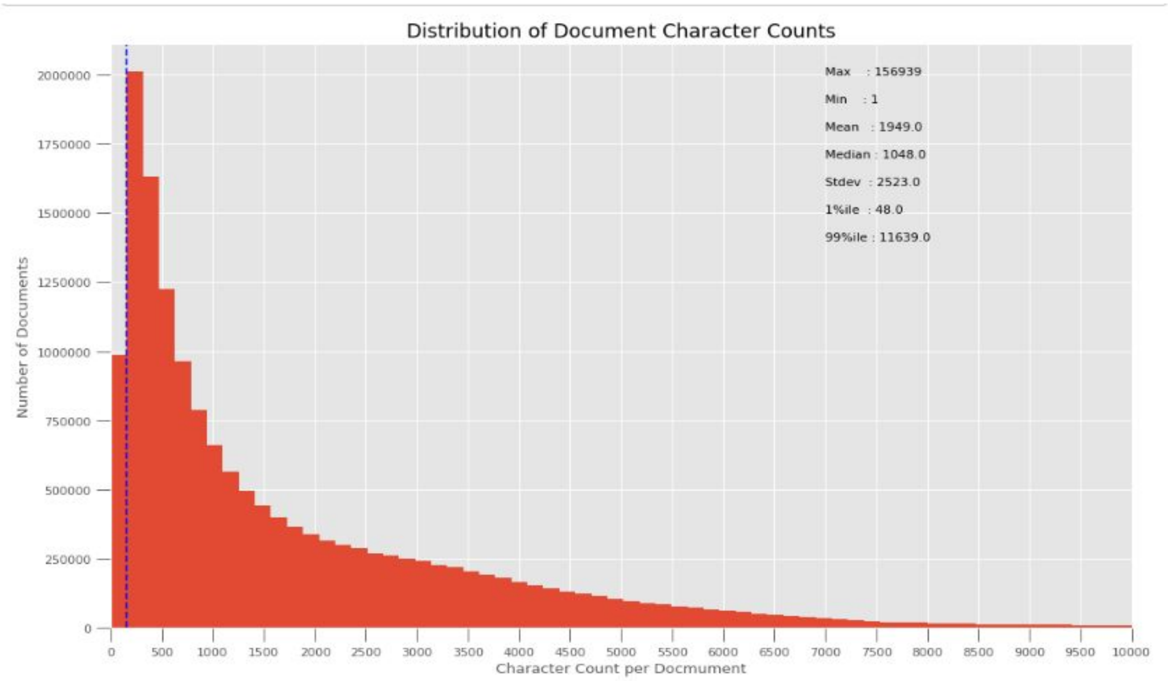


Fig 3: Distribution of Documents Character Counts.

The median character count per document is 1048 and the mean is 1949 characters. We can see that the character count is heavily skewed to the right. Most of the documents in the dataset are on the shorter end of the spectrum. It is vital that we do not remove a too high proportion of the dataset.

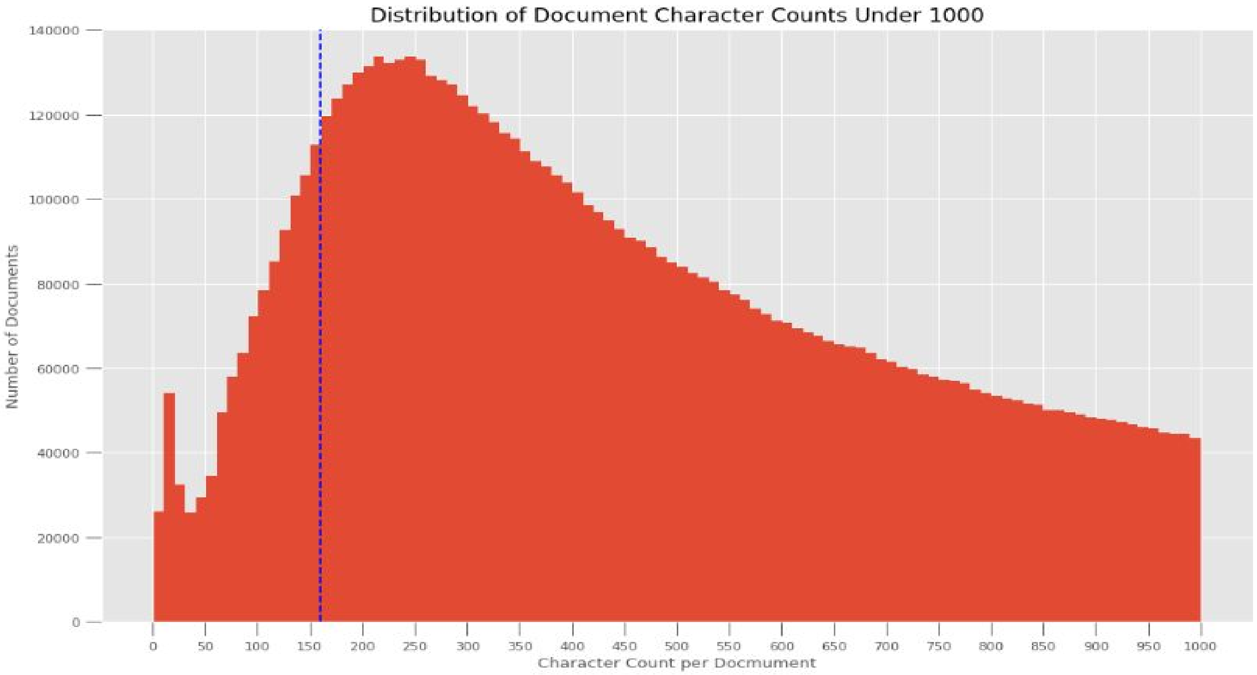


Fig. 4: Distribution of Documents Character Counts under 1000.

Figure 3 displays a magnified look at the distribution of word counts in documents. It displays the documents with a word count below 1000. From the graphs, we can see the number of documents rapidly decreases below 200 characters. So we will choose to set the lower character threshold to be 160. So any documents below 160 will not be in the subset, in order to select documents more suitable for the LDA algorithm to optimize the LDA algorithm.

We have a wide range of dates in the dataset and we have to sample multiple New Zealand historic events to evaluate that the topic modeling is working correctly. We can try and match the model outputs with where we might expect certain topics to be prominent/salient.

We did not have constant access to a historian who can explain these events. However, there are numerous events in New Zealand history that stand-out and we can use there to evaluate the relevance of our topic modeling results.

Some historical events we can use as reference points:

1. **New Zealand Immigration Boom (1843 - 1853):** After the signing on the treaty in Waitangi 1840. New Zealand saw a gigantic influx of non-Maori residents, going from 2,000 to 28,000 European immigrants.[8] There is also another population 'boom' known as the Vogel boom. The Vogel boom refers to the eighth prime minister of New Zealand, Julius Vogel. Vogel borrowed funds to support public projects and heavily supported immigration to help increases the economic growth. This took place in the 1870s. During these periods we could probably expect words like; "migrants", "shipping", "cargo", possibly the mention of European countries or city names.

2. **The New Zealand Wars (1845 - 1872):** Another consequence of the treaty signing was war. Disputes of land and sovereignty occurred between the European immigrants and the Maori. The wars were quite intermittent and were spaced out between the 1840s and 1860s -1870s. An interesting aspect of the New Zealand Wars is that they mostly took place in the north island, specifically Waikato and Taranaki. So we could expect mentions wars of these wars in more in the north island publications over the south island. For example, we could expect place names of battle sites and "muskets".

3. **Gold Rush - Otago and West Coast (1860s):** The gold rush caused a huge increase in interest for New Zealand attractive people from all over the world to chase the gold, including people from Europe and China. It was mostly in Otago but some gold mines were in the West Coast of the south island. This again led to a surge in population. Dunedin increased from 35,000[9] to 100,000[10] over one year in 1863. This is another example of a region-specific event. Although it attracted people worldwide, we could probably expect a higher count of Otago and west coast publication discussing the gold rush versus a northland publication.

4. **Women's Suffrage Movement (1893):** The suffrage movement was led by the activist Kate Shepard. It was a movement to give females the right to vote in parliamentary elections. The significance of this event was that New Zealand was

[8] Statistics retrieved from: https://www.radionz.co.nz/national/programmes/afternoons/audio/201836026/the-history-of-immigration-booms-in-nz

[9] Statistics retrieved from: https://www.radionz.co.nz/national/programmes/afternoons/audio/201836026/the-history-of-immigration-booms-in-nz

[10] Statistics retrieved from: https://www.radionz.co.nz/national/programmes/afternoons/audio/201836026/the-history-of-immigration-booms-in-nz

the first country in the world to give pass the law securing female's right to vote. This obviously led to global news and we can expect this topic to be highly discussed in this time period.

5. **The Boer War (1899 -1902):** The Boer war was fought between the British Empire and the Boer South African Republic. New Prime Minister, Richard Seddon, was quick to support the British Empire. Over the 2.5 years of war, New Zealand set 6,500[11] troops. 133[12]troops died due to poor conditions such as poor rations and disease, a large proportion of the 230[13] troops who died. It was a significant event, the New Zealand soldiers were considered strong soldiers, despite inadequate training and equipment. It sparked a wave of patriotism and showed New Zealand's commitment to the British. This led to New Zealand's involvement in the two world wars in the following decades.

6. **World War 1 (1914 - 1918):** This is one of the most important events of the nineteenth century. Not just in New Zealand but the whole world. Following the trend of the Boer war. New Zealand committed troops to support the British (and allies). The Gallipoli campaign when down as a tragic defeat for the ANZAC troops. New Zealand suffered 2,779[14] casualties in Gallipoli. One in four of the troops aged 20 to 45[15] that were deployed, were either killed or injured. There were devastating losses and we could expect World War 1 to be a higher discussed topic in the publications.

7. **The Great Depression (the 1930s):** The great depression was the greatest crash of financial systems in world history. The depression led to run on banks, mass unemployment. There were government relief schemes put in place after there were riots in the major cities. This led to labour restructures and the creation of unions. Another important aspect of this period is that New Zealand was heavily depended on for its agriculture exports for Britain. So we could expect the mention of; 'exports', 'wages', 'unemployment', 'unions' during this time period.

8. **World War 2 (1939 - 1945):** The biggest war in world history. New Zealand again followed Britain and Australia to war. New Zealand had a contributed to the war effort in both European and the Pacific theatres. New Zealand sent about 140,000[16] people in numerous services to help in the war. It is also estimated that New Zealand suffered the most casualties (6,684)[17] per capita of the commonwealth countries. Versus 5123 and 3232[18] for Britain and Australia, respectively. As World War 1, we could probably expect this to be a higher discussed topic nationwide.

These are just some of the events we can look at to check the relevancy of our topic modelling results. The events mentioned are merely speculative and may not have the dominance in articles that we predict. We will discuss compare these events with the findings from our model in a later section.

[11] Statistics retrieved from: https://teara.govt.nz/en/south-african-war
[12] Statistics retrieved from: https://teara.govt.nz/en/south-african-war
[13] Statistics retrieved from: https://teara.govt.nz/en/south-african-war
[14] Statistics retrieved from: https://nzhistory.govt.nz/war/first-world-war-overview/defending-our-shores
[15] Statistics retrieved from: https://nzhistory.govt.nz/war/first-world-war-overview/defending-our-shores
[16] Statistics retrieved from: https://nzhistory.govt.nz/war/new-zealand-and-the-second-world-war-overview
[17] Statistics retrieved from: https://nzhistory.govt.nz/war/new-zealand-and-the-second-world-war-overview
[18] Statistics retrieved from: https://nzhistory.govt.nz/war/new-zealand-and-the-second-world-war-overview

Project design and methodology

Project structure

Based on the application scenario, we partition the system into three layers, from top to bottom they are service layer, inferring layer and training layer.

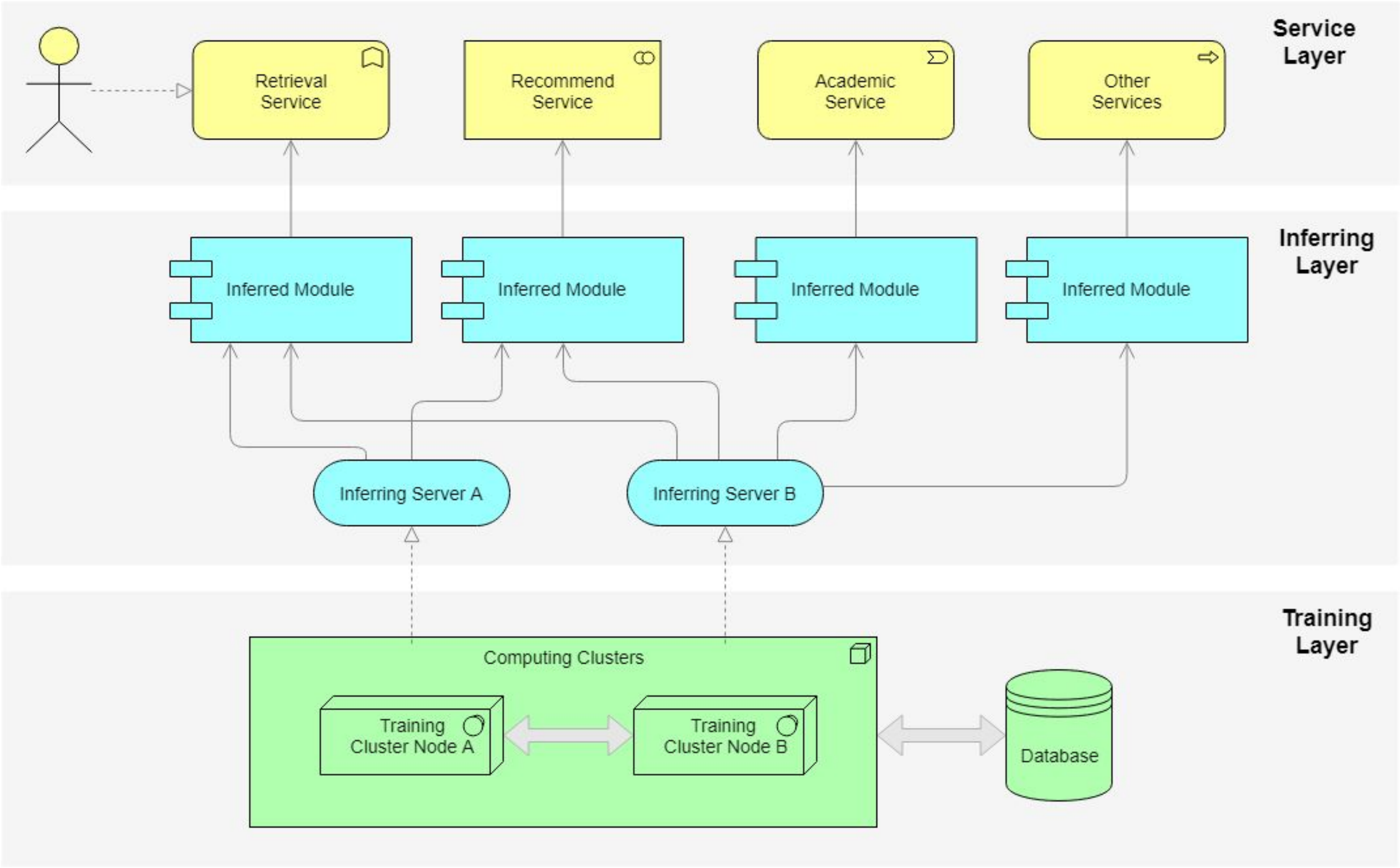


Fig. 5: Application Data Flow of the project

The Service Layer provides an interface to users and to Inferring Layer, it handles the requests from users and displays or returns the results from Inferring Layer. The Inferring Layer consists inferring servers which response the request from Service Layer, generate inferred models using topic models from Training Layer, and processes the inferred model to get the results to return. The Training Layer retrieves database in general, process the data from database and training corpora to generate trained models for Inferring Layer, this layer undertakes heavy computing tasks for training large scale corpora, which is not synchronous with other two layers.

A typical application scenario, e.g. retrieving documents by topics and region: the user input constraints and click searching on *Papers Past* webpage, the webpage sends a request to one of the inferring servers, the server starts a series of processing as below flowchart, and return the documents link to the webpage and the user.

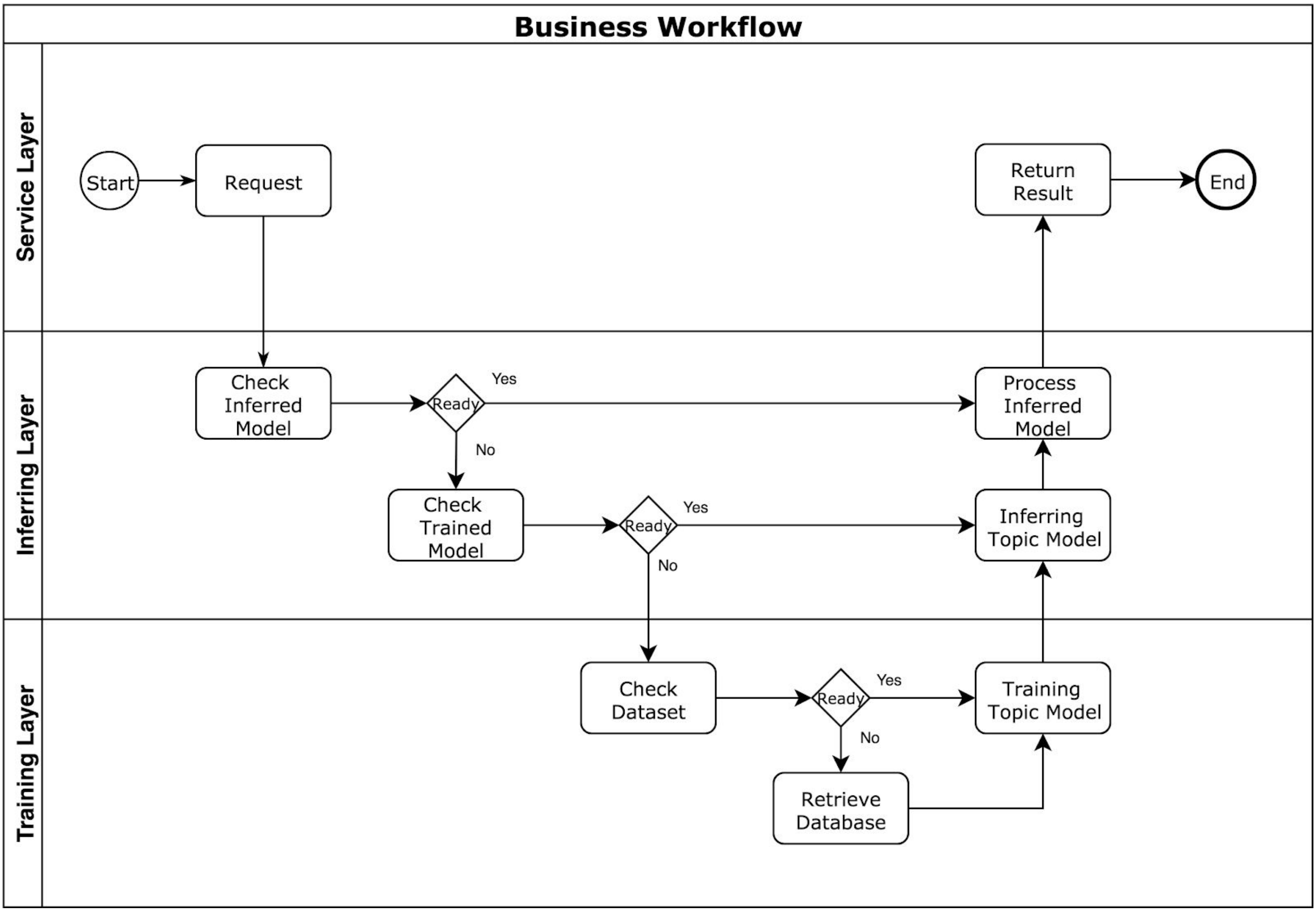


Fig. 6: Business flowchart

In practice, based on the data type of input and output, we transform above logic structure and workflow to four major task blocks, each block has specific tasks, input and output, as shown below:

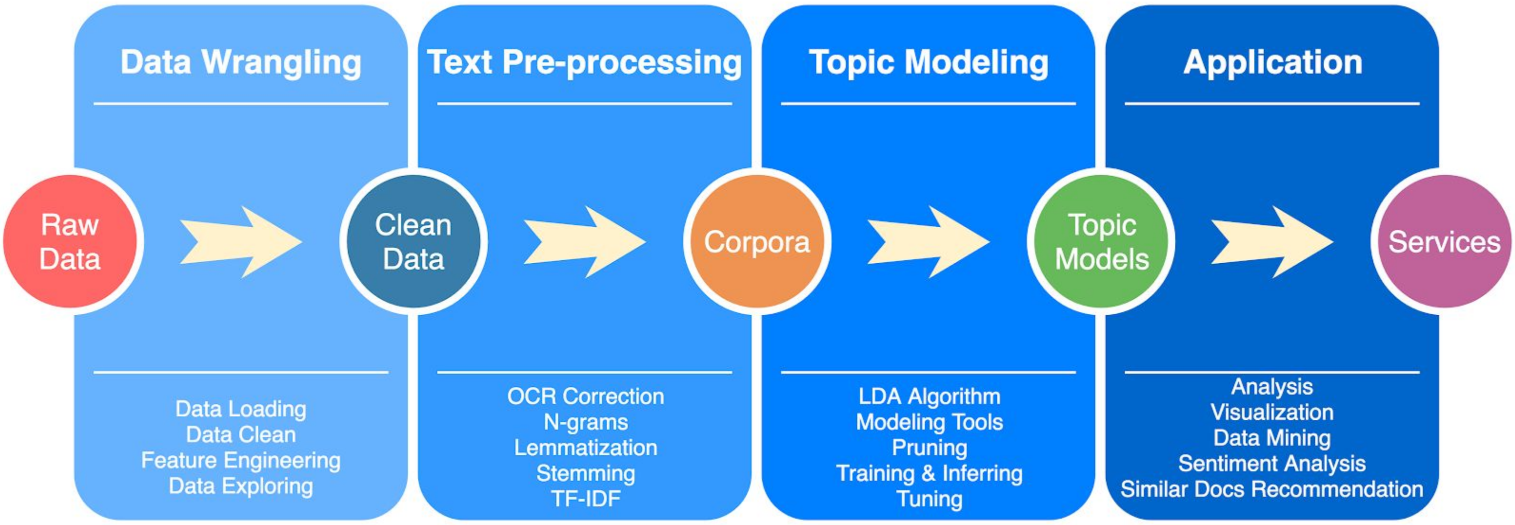


Fig. 7: Major Blocks

Implementation Process

After we were clear about the structure and workflow, defined tasks, input and output, we start implementing the project using python code. We split all tasks into seven parts based on detail function, the data model is shown in Figure 6.

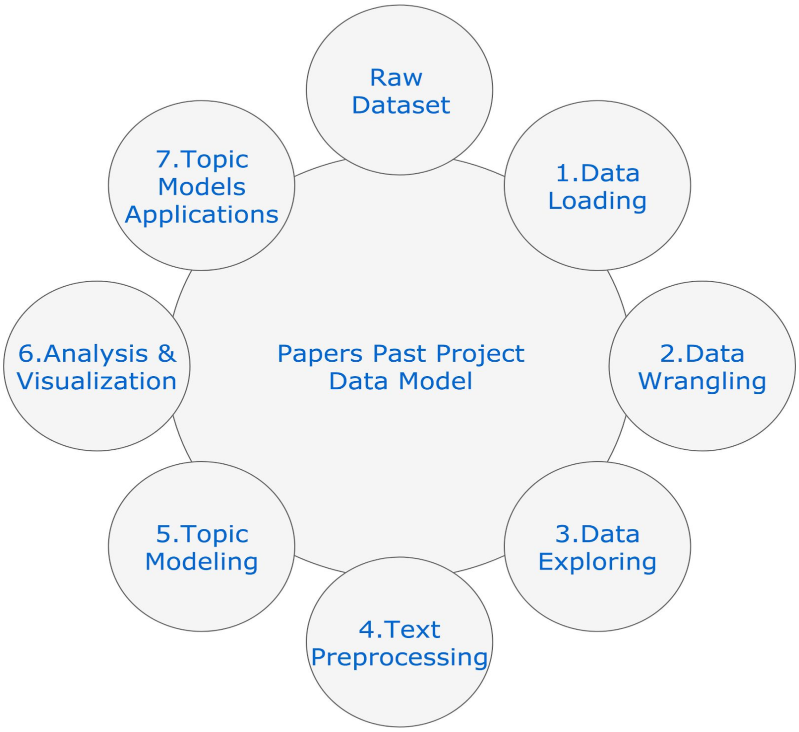


Fig. 8: Project Data Model

Part 1. Data Loading

In this part, after we did experiments to compare Pandas and PySpark, we found Pandas is good at rich methods on data manipulation and visualization, while PySpark is good at high performance on data manipulation and data access. Since the raw dataset size is 33GB, we select PySpark as the tool for large datasets/files manipulation and access, and select Pandas as the tool for subsets or relative small data frame manipulation, access and for visualization.

Part 2. Data Wrangling

We performed data clean and feature engineering to get clean dataset for further process. The workflow is shown in Figure 9.

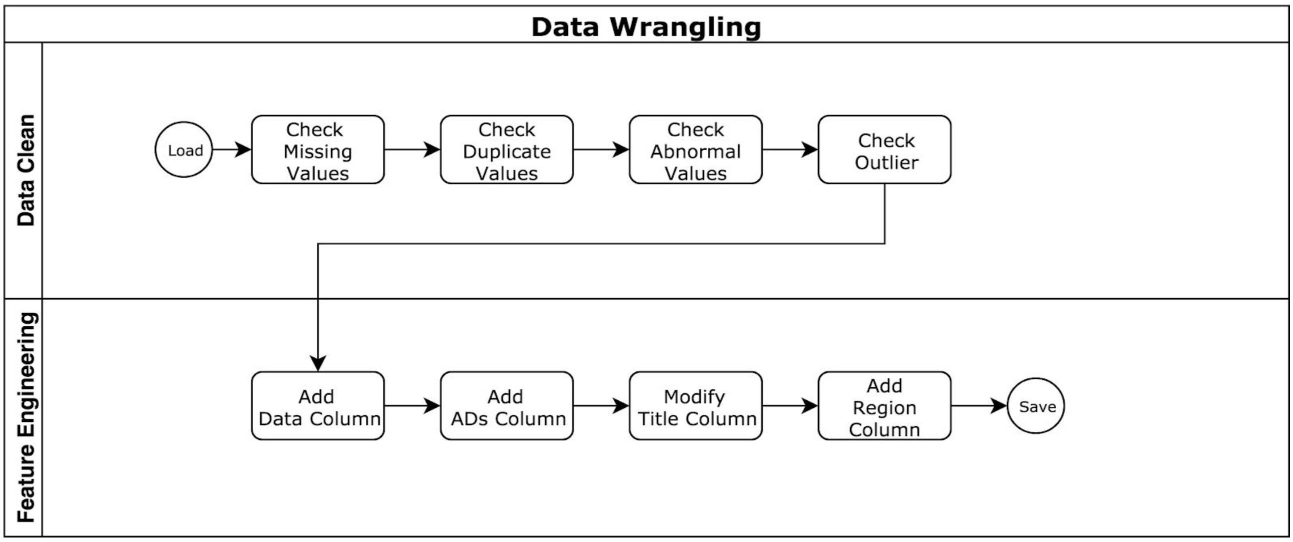


Fig. 9: Data Wrangling Flowchart

Part 3. Data Exploring

In this part, we analyzed and visualized the clean dataset after data wrangling, to deeply understand the distribution and features of the dataset. Based on the label, samples, papers and regions we did further analysis and visualized them.

Part 4. Text Preprocessing

After wangling and exploring the dataset, we need to preprocess the dataset for training topic modeling appropriately. In this part, we evaluated the quality of OCR, we experimented spelling correction and some other NLP processes. The workflow chart is shown in Figure 8.

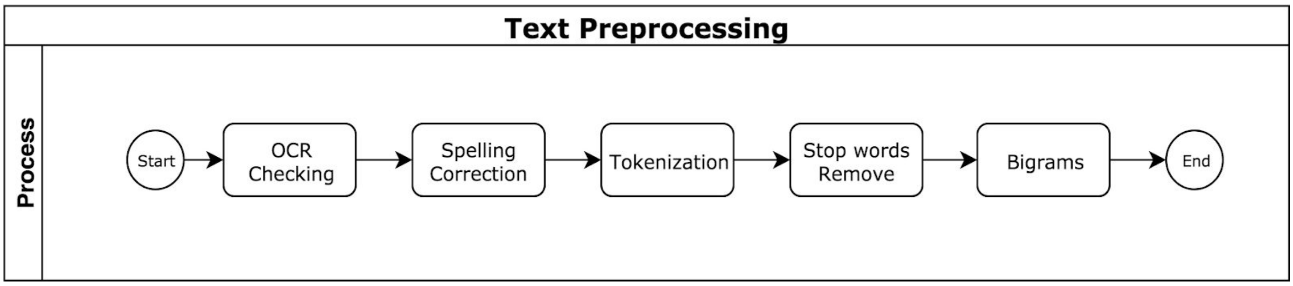


Fig. 10: Text Preprocessing Flowchart

In practice, we did not add OCR checking and spelling correction for time and resource limitations, we consider them as future work. And other NLP processes are performed by MALLET.

Part 5. Topic Modeling

Before training topic models, we need to sample the clean dataset to get a sample set, due to the limitation of computing resources. We attempted to train the clean dataset (14 GB) but the training process did not finish after 5 days waiting. Since the project only has 2 months time, we would not bear the period of around regression over 2 days. So we used random sampling to reduce the dataset size for training. After we get the sample set, we split it to train set and several subsets for a range of time, regions and label inferring and analysis.

The workflow of topic modeling is shown in Figure 11:

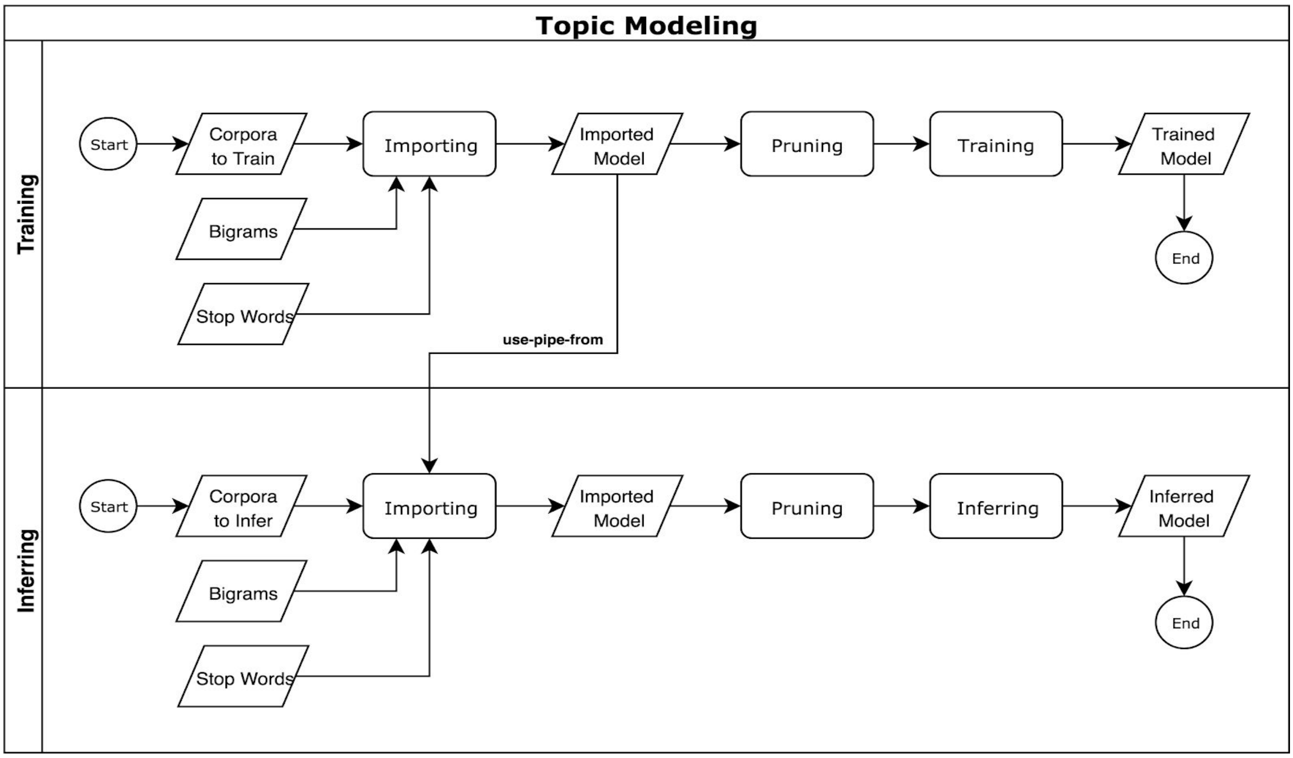


Fig. 11: Topic Modeling Flowchart

In this part, we trained the sample set to get a trained topic model and inferred the subsets to get several inferred topic models for further applications.

Part 6. Analysis and Visualization

After training and inferring process we get several topic models, the trained topic model contains a topic keyword list and a doc-topic matrix (row: document, column: topic weight), the other inferred topic models contains a doc-topic matrix (share the same topic keyword list with trained topic model). We manipulated and transformed above data of each topic model and got the data frame for analysis and visualization.

We used dominant topic count (sum of a dominant topic in a unit of time) and topic average weight (average weight of a topic in a unit of time) as base metrics for evaluating and analyzing topics. Combining with some statistical thoughts, we filtered, sorted and extracted the most distinct documents, the most popular topics over time, dominant topics distribution over time, annual average weight distribution, annual average weight of each topic, the most variant topics over time, the most significant topics over time, etc. to analyze and visualize, and interpreted finding from them.

Part 7. Applications

Utilizing topic models to find patterns and features are not enough for the project, we designed three applications for topic models.

The first one is the application of data mining in topic models. We applied linear regression in monthly average weight matrix, to evaluate the correlation between topics over time.

The sconed is the application of sentiment analysis in topic models. We employed a sentiment evaluation tool on annual average weight matrix, to evaluate the sentiment of publications over time.

The last one is the application of similarity calculation in topic models. We used Jensen-Shannon Divergence to calculate the similarity of an unseen article among the trained topic model and made a recommendation of the similar documents.

We have compiled all of our code in notebooks. As an overall project output, we will be providing the complete notebooks in a Github repository. The source for the notebooks will be provided in the appendix at the end of this report.

For more detail please check the repository of the project.

Findings

Overview

It is important to have an understanding of the distribution of documents before we examine the distribution topics. We can see in Figure 12 that most of the documents are published between 1880 through to 1920.

There are two important years we can see two noteworthy points in Figure 12.

- In 1907 there is a significant drop in the number of papers. Then 1908 has the second largest count overall. We believe this is due to the API collected the dataset from the *Papers past* website initially. This significant drop seems to be present in any percentage of the dataset (1% to 20%). This will obviously have a flow-on effect on every topic. We are attributing to an unseen dataset and will not reflect the discourse of 1907.

- The other point is 1921, again another decrease in count. With a significant decrease in count from 1920. To our understanding, this is due to some publications ceasing production during this time. But we could also attribute this to have an out of date dataset. *Papers past* may not have completed all of the digitisation processes for 1921 onward.

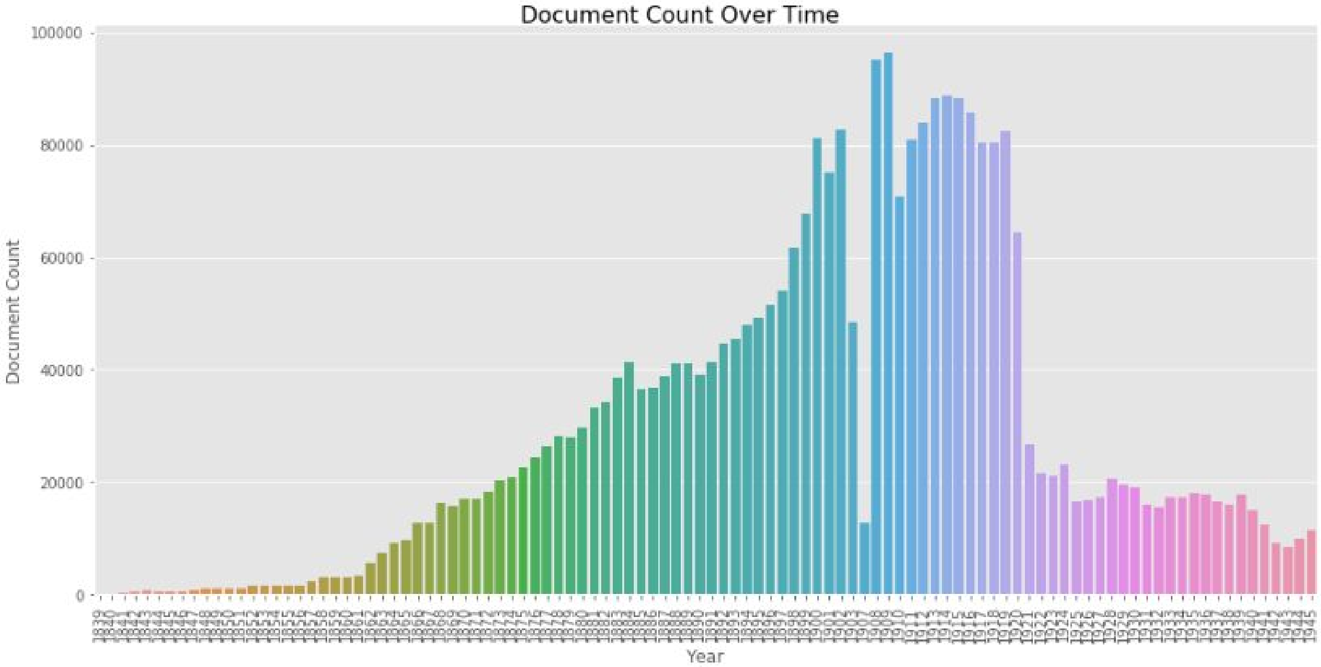


Fig. 12: Document Count over year

Now, we can look at the Dominant topic distribution. The dominant topic distribution tends to follow the document count over year. The density of documents with a topic weight over 0.5 experiences a huge increase after 1860. The same time that the document count rapidly tends upwards.

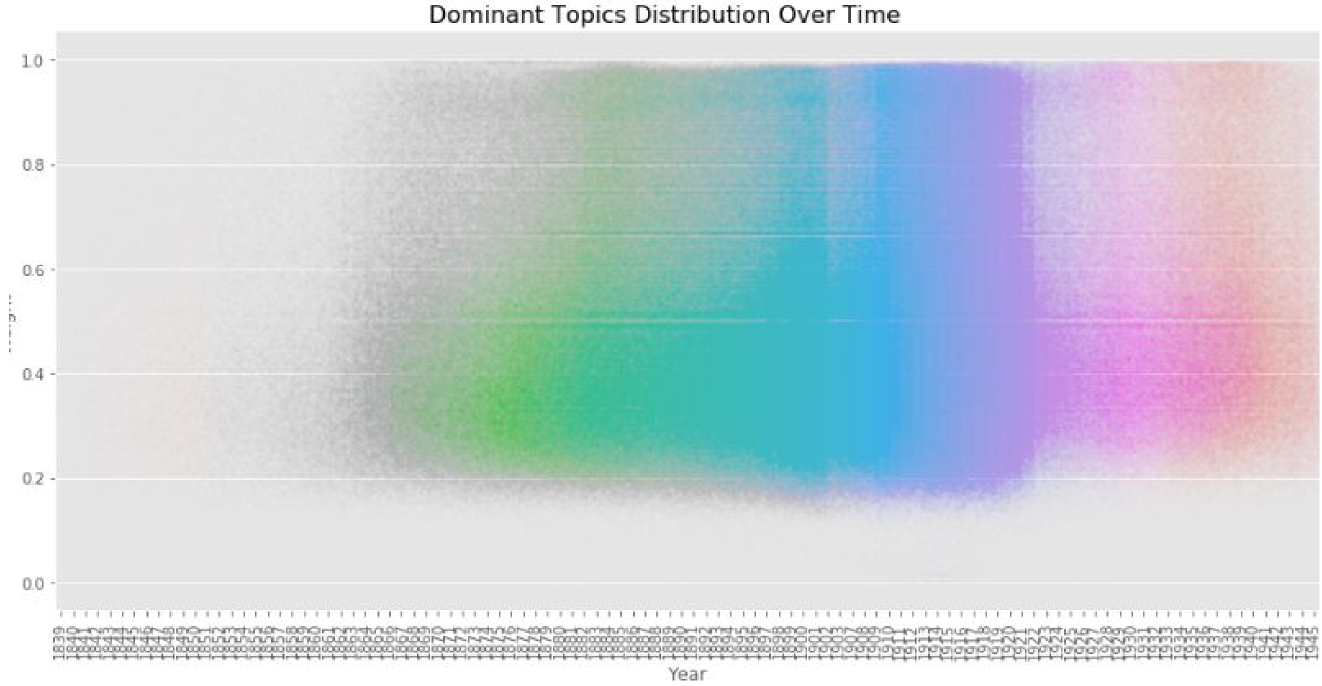
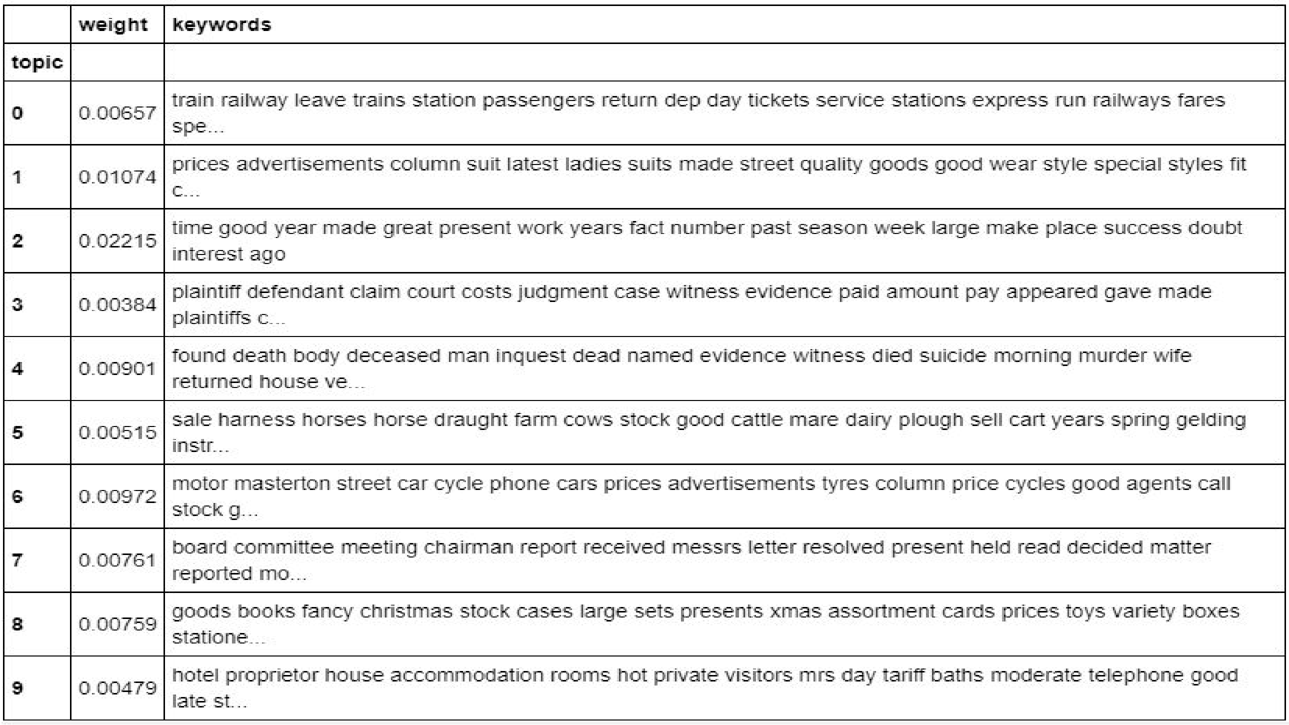


Fig. 13: Dominant Topics Distribution over years in general

One of the metrics for evaluating topics has dominant topics. In Figure 13, we have the distribution of dominant topics over the entire range of the dataset. The horizontal axis is years from 1839 - 1945 and weight on the vertical axis. The weight is the amount that the dominant topic attributes to the document and each point on the plot represents a single document.

We now want to take a look a the first ten topics (0 - 9). Before we check any of the specific topics we mentioned in the *dataset summary* section. We can plot them in a scatter plot.



	weight	keywords
topic		
0	0.00657	train railway leave trains station passengers return dep day tickets service stations express run railways fares spe...
1	0.01074	prices advertisements column suit latest ladies suits made street quality goods good wear style special styles fit c...
2	0.02215	time good year made great present work years fact number past season week large make place success doubt interest ago
3	0.00384	plaintiff defendant claim court costs judgment case witness evidence paid amount pay appeared gave made plaintiffs c...
4	0.00901	found death body deceased man inquest dead named evidence witness died suicide morning murder wife returned house ve...
5	0.00515	sale harness horses horse draught farm cows stock good cattle mare dairy plough sell cart years spring gelding instr...
6	0.00972	motor masterton street car cycle phone cars prices advertisements tyres column price cycles good agents call stock g...
7	0.00761	board committee meeting chairman report received messrs letter resolved present held read decided matter reported mo...
8	0.00759	goods books fancy christmas stock cases large sets presents xmas assortment cards prices toys variety boxes statione...
9	0.00479	hotel proprietor house accommodation rooms hot private visitors mrs day tariff baths moderate telephone good late st...

Fig. 14: First ten topics

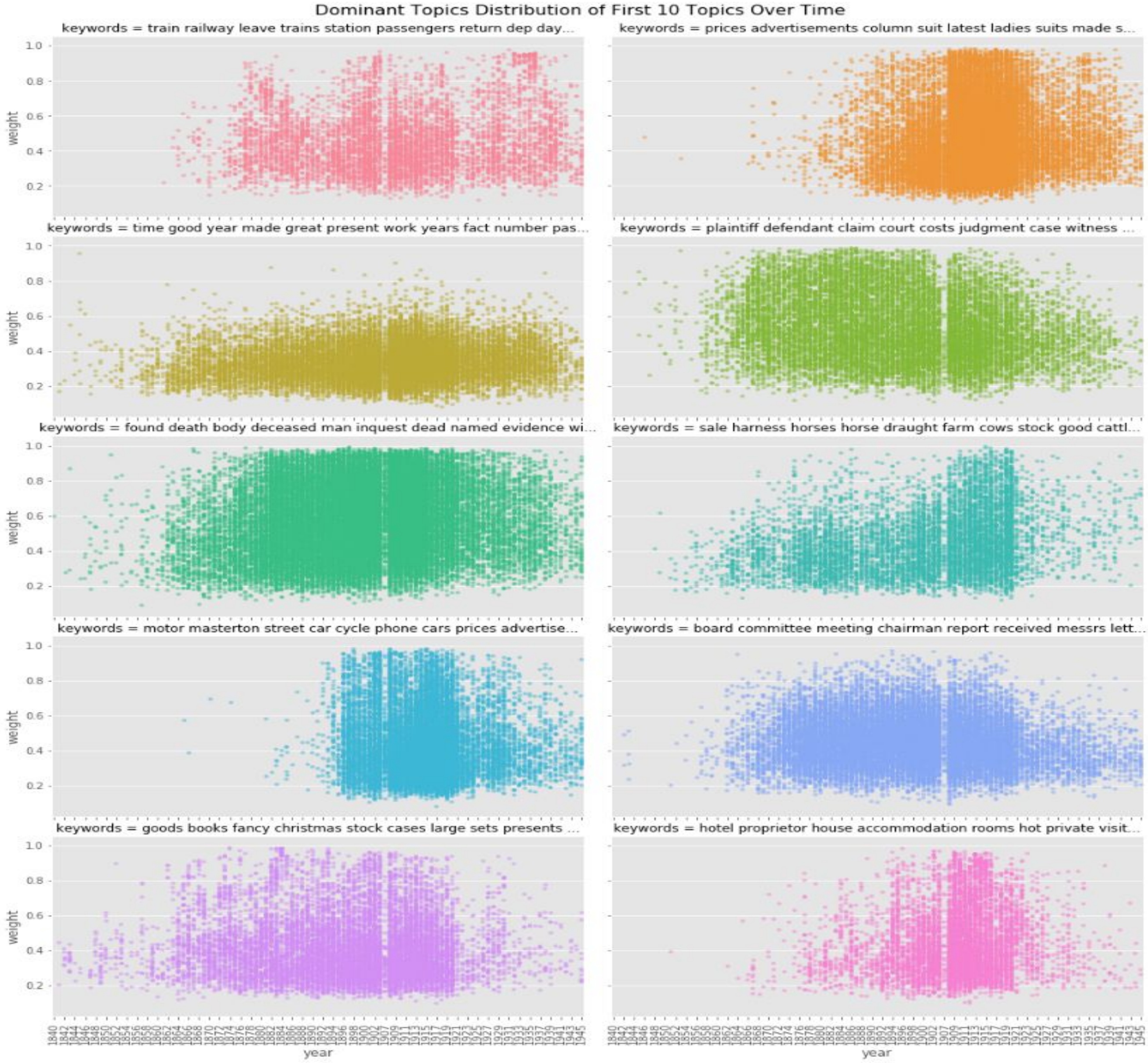
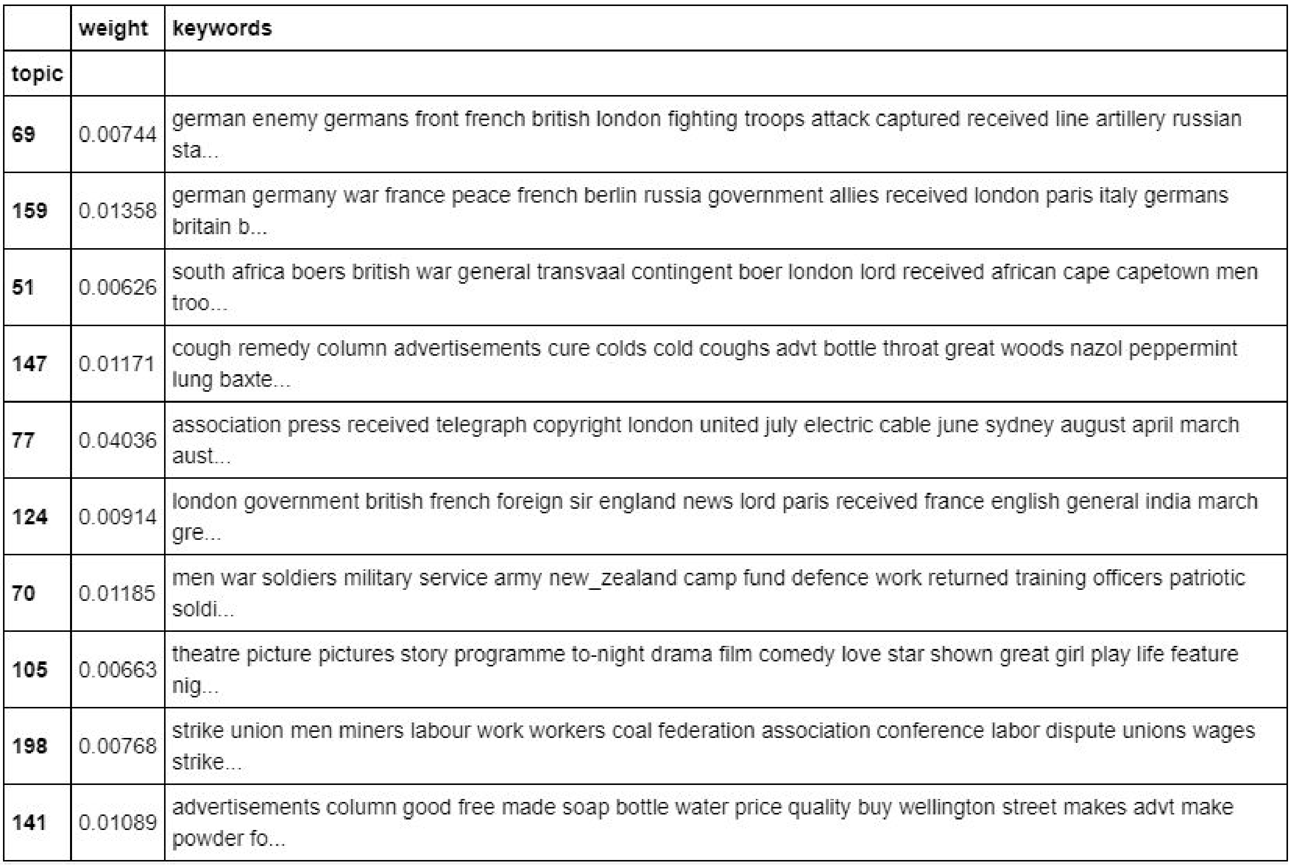


Fig. 15: Dominant Topics Distribution of First 10 topics over year.

The plots of the first ten topics are a great start. The scatter plots are a superb way to represent how a topic changes over time. It is easy to see the number of documents mentioning a particular topic (with the density of plot point) but also the weight proportion of that topic. Worth noting, the 1907 document count is apparent and can be seen clearly in the scatter plots. With the first ten topics is difficult to evaluate the accuracy of the topics, since there generic and are not tied to historical events.

We can look at the most variant topics. We do this by calculating the standard deviation of the average weight of each topic, we get the most variant topics. Some weights of the topic may vary following the overall trends, it's better to remove the effect of an overall trend for more accurate evaluation. Here we ignore the effect of overall trends.

The variance does not consider the document number. If a topic has a high variance but there are a few documents, that topic is not significant. To obtain the significant topics, we must figure out the document topic count and multiply that by the average weight of the topic. This considers both numbers of support documents and the average weight of the topic. This is the best way to find significant topics. In figure 16 is a table of the ten most *‘significant topics’.*



	weight	keywords
topic		
69	0.00744	german enemy germans front french british london fighting troops attack captured received line artillery russian sta...
159	0.01358	german germany war france peace french berlin russia government allies received london paris italy germans britain b...
51	0.00626	south africa boers british war general transvaal contingent boer london lord received african cape capetown men troo...
147	0.01171	cough remedy column advertisements cure colds cold coughs advt bottle throat great woods nazol peppermint lung baxte...
77	0.04036	association press received telegraph copyright london united july electric cable june sydney august april march aust...
124	0.00914	london government british french foreign sir england news lord paris received france english general india march gre...
70	0.01185	men war soldiers military service army new_zealand camp fund defence work returned training officers patriotic soldi...
105	0.00663	theatre picture pictures story programme to-night drama film comedy love star shown great girl play life feature nig...
198	0.00768	strike union men miners labour work workers coal federation association conference labor dispute unions wages strike...
141	0.01089	advertisements column good free made soap bottle water price quality buy wellington street makes advt make powder fo...

Fig. 16: Ten most significant 10 topics over time.

Now we have the ten most significant topics in the dataset. These topics are a little more of what we expected. Many of the topics are incredibly military or war focused. Topics 44, 100, 37, and 151 looks to be dominated by discussing either World War 1 or 2. Topic 73 is heavily focused on the Boer War in South Africa. Topic 123 looks focused on strikes and labour unions, possibly relating to the great depression and reforms. These were all reference events that were mentioned in the *dataset summary.*

However, there are some unexpected results. Topic 83 is about sickness and antidotes, this could be referencing the health epidemic between the 1890s and 1920s (known as the influenza era)[19]. While topic 184 is about advertisements/sales and topic 24 is about performing arts. This gives an excellent overview over the entire the dataset.

[19] Time Periods and information retrieved from: https://teara.govt.nz/en/epidemics/page-4

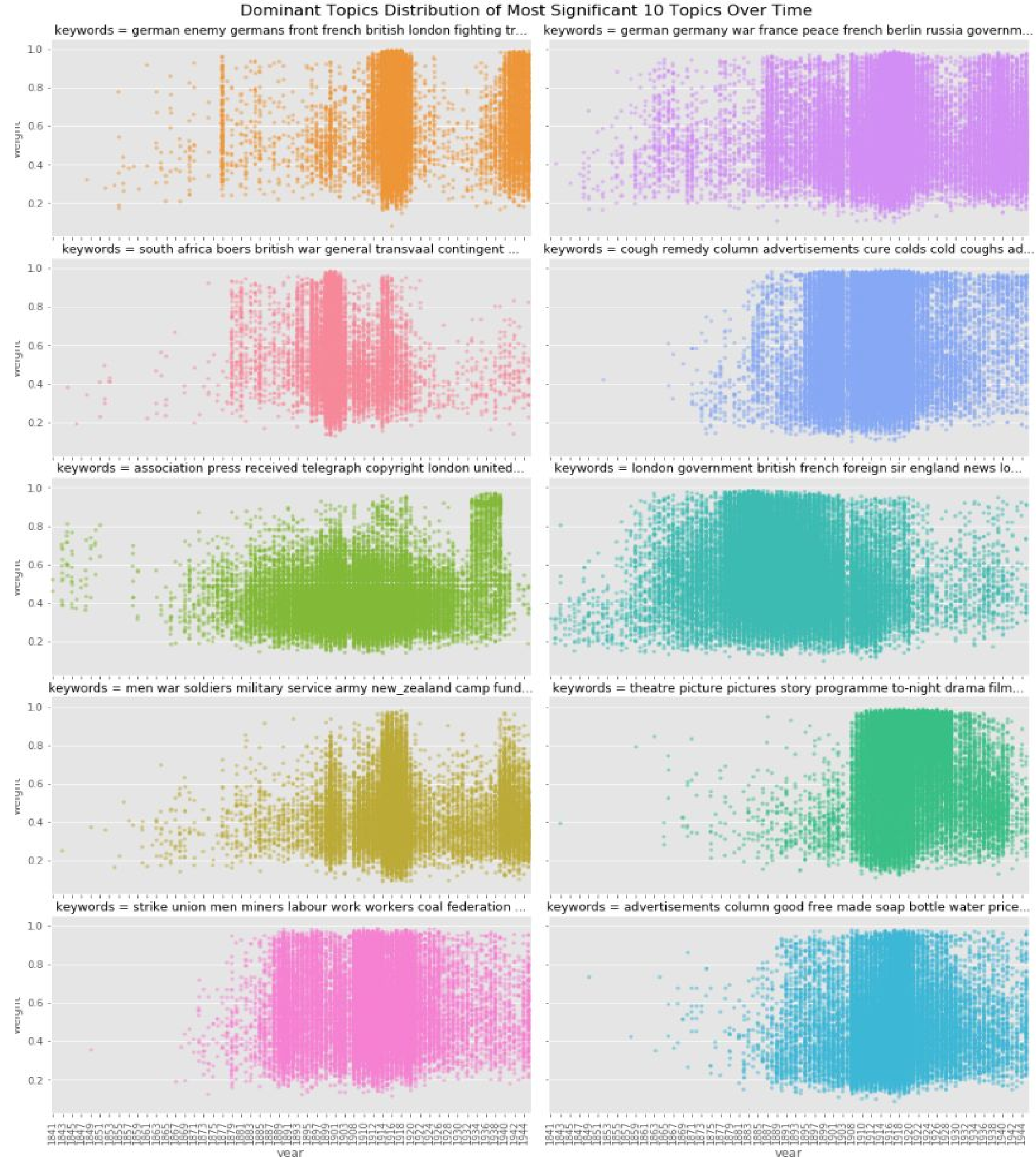


Fig. 17: Dominant topics distribution of most significant 10 topics over year scatterplots.

Five of the ten topics were about wars or military based on the topic keywords. We can check the accuracy by checking if the document distribution and weight match the known dates. The world war topics such as 44 (top-left plot) and 151(bottom-right plot) appear to line up quite nicely with our anticipated pattern. With an obvious clear result increase in the period between 1914 and 1918. Then again in the period 1939-1945. This appears to be the same as for topic 100 (second from the top-left plot). The keywords indicated that it is about the Boer war and the document count and weight are during the correct dates, 1899-1902. In topic 83, depicting an illness and symptoms matches the time period which the influenza era occurred. This is clearly shown in the second-right plot. The distribution of documents about the illness dissipates after 1920, around when the crises were ending. The significant topic results are promising and match up well with important events.

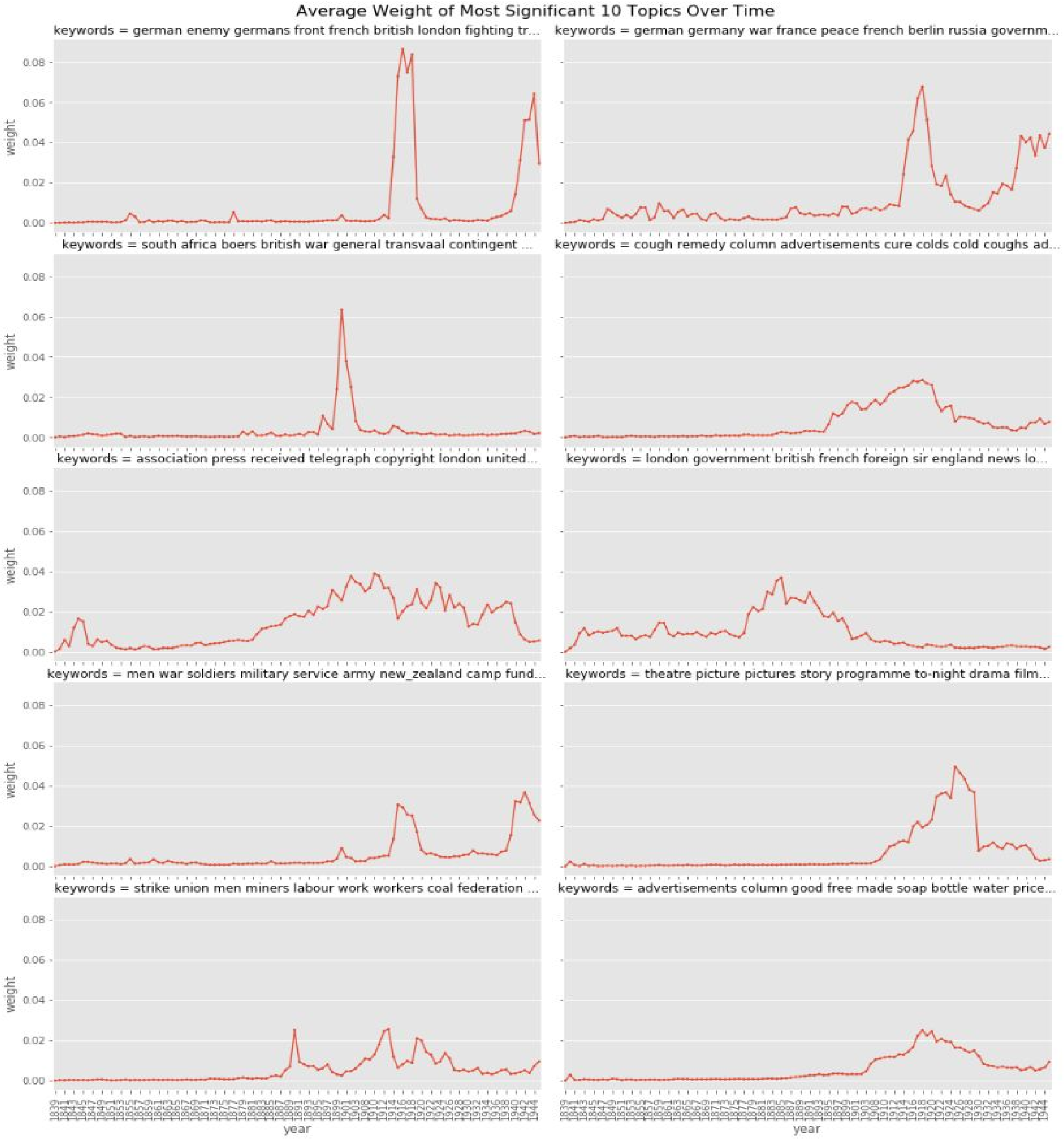


Fig. 18: Average weight of most significant 10 topics

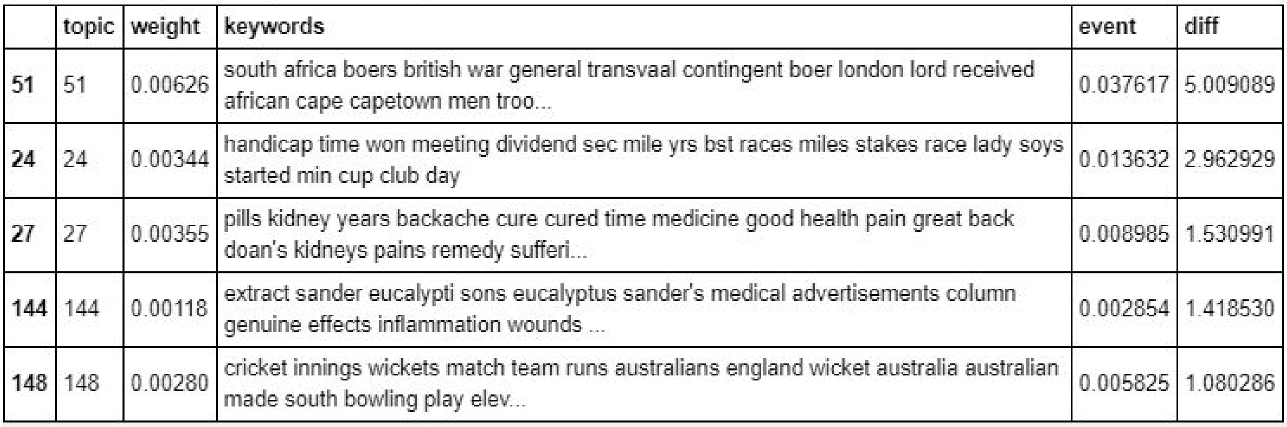
By examining and plotting the same significant topics we often get a clearer picture of the topics and how they trend. The dominant topic will tend to ignore the non-dominant topic. Average weight will reflect the topic variety and how it trends.

- This metric is the average weight of a topic through all documents in a range of time, it avoids the issue above, reflect the topic variety and trend over time. However, it might give a frequent/normal topic (which has a small weight in each document) a high weight because the topic appears in a large number of documents when summing all weights of each topic and dividing by total weight of all topics, this normal topic may get a high weight.

In the Specific Time Period

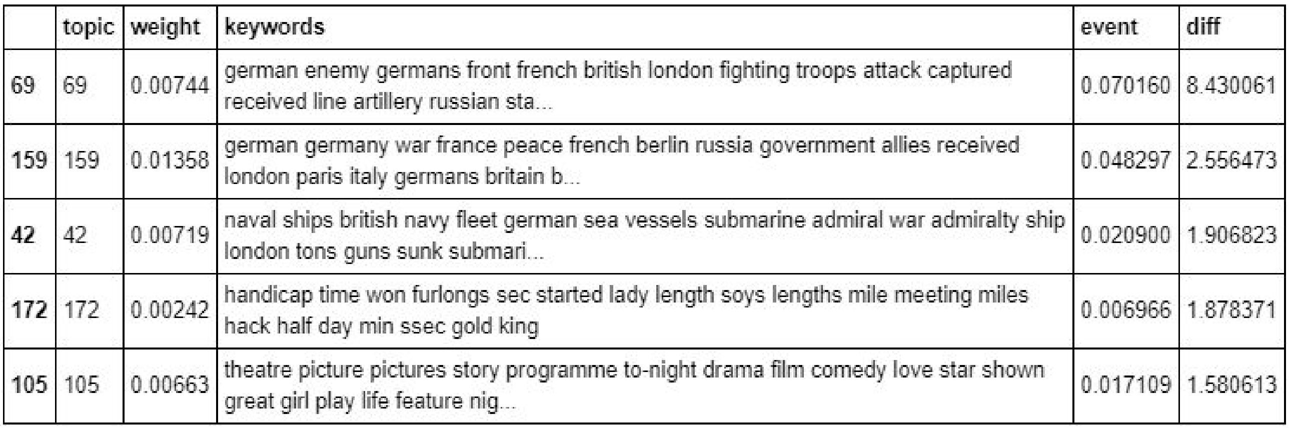
In dataset summary, we mentioned some events to use as reference points. We went through the time periods for which the events occurred and examined whether the dominant topics in the documents mentioned the reference events.

When we used the Boer War (1899 - 1902), we found that the war was mentioned as the ‘hottest’ topic and we had a similar result with World War 1 (1914-1918).



	topic	weight	keywords	event	diff
51	51	0.00626	south africa boers british war general transvaal contingent boer london lord received african cape capetown men troo...	0.037617	5.009089
24	24	0.00344	handicap time won meeting dividend sec mile yrs bst races miles stakes race lady soys started min cup club day	0.013632	2.962929
27	27	0.00355	pills kidney years backache cure cured time medicine good health pain great back doan's kidneys pains remedy sufferi...	0.008985	1.530991
144	144	0.00118	extract sander eucalypti sons eucalyptus sander's medical advertisements column genuine effects inflammation wounds ...	0.002854	1.418530
148	148	0.00280	cricket innings wickets match team runs australians england wicket australia australian made south bowling play elev...	0.005825	1.080286

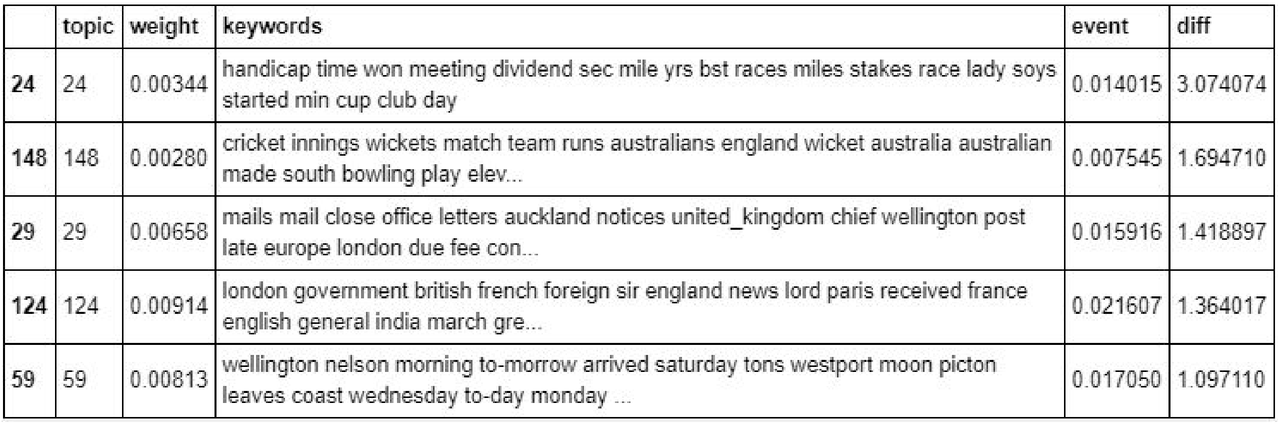
Fig. 19: Dominant topics for 1899-1902



	topic	weight	keywords	event	diff
69	69	0.00744	german enemy germans front french british london fighting troops attack captured received line artillery russian sta...	0.070160	8.430061
159	159	0.01358	german germany war france peace french berlin russia government allies received london paris italy germans britain b...	0.048297	2.556473
42	42	0.00719	naval ships british navy fleet german sea vessels submarine admiral war admiralty ship london tons guns sunk submari...	0.020900	1.906823
172	172	0.00242	handicap time won furlongs sec started lady length soys lengths mile meeting miles hack half day min ssec gold king	0.006966	1.878371
105	105	0.00663	theatre picture pictures story programme to-night drama film comedy love star shown great girl play life feature nig...	0.017109	1.580613

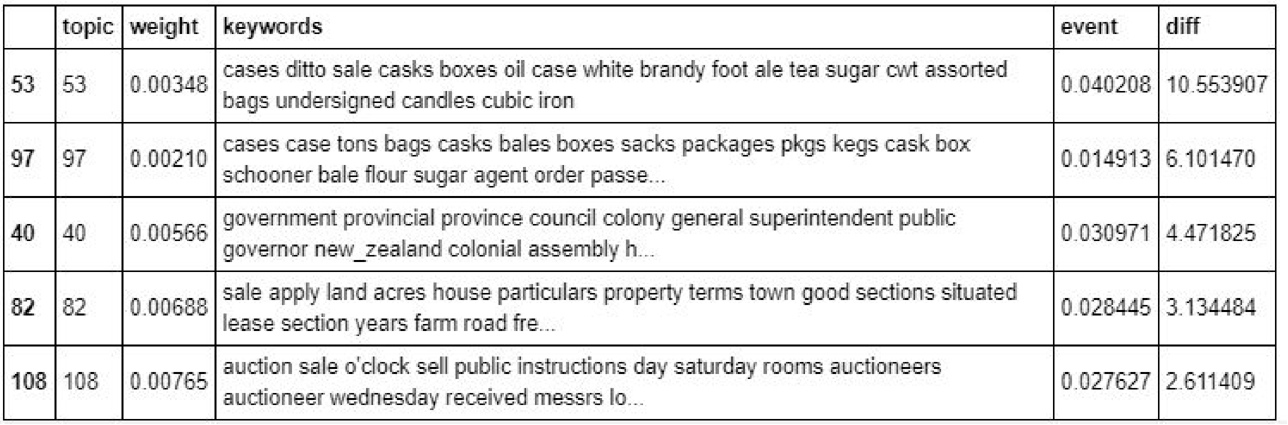
Fig. 20: Dominant topic for 1914-1918

Some other examples that were not so successful were, The Women’s suffrage movement (1893) and Otago & West Coast Goldrush (the 1860s).



	topic	weight	keywords	event	diff
24	24	0.00344	handicap time won meeting dividend sec mile yrs bst races miles stakes race lady soys started min cup club day	0.014015	3.074074
148	148	0.00280	cricket innings wickets match team runs australians england wicket australia australian made south bowling play elev...	0.007545	1.694710
29	29	0.00658	mails mail close office letters auckland notices united_kingdom chief wellington post late europe london due fee con...	0.015916	1.418897
124	124	0.00914	london government british french foreign sir england news lord paris received france english general india march gre...	0.021607	1.364017
59	59	0.00813	wellington nelson morning to-morrow arrived saturday tons westport moon picton leaves coast wednesday to-day monday ...	0.017050	1.097110

Fig. 21: Dominant topics for 1893.



	topic	weight	keywords	event	diff
53	53	0.00348	cases ditto sale casks boxes oil case white brandy foot ale tea sugar cwt assorted bags undersigned candles cubic iron	0.040208	10.553907
97	97	0.00210	cases case tons bags casks bales boxes sacks packages pkgs kegs cask box schooner bale flour sugar agent order passe...	0.014913	6.101470
40	40	0.00566	government provincial province council colony general superintendent public governor new_zealand colonial assembly h...	0.030971	4.471825
82	82	0.00688	sale apply land acres house particulars property terms town good sections situated lease section years farm road fre...	0.028445	3.134484
108	108	0.00765	auction sale o'clock sell public instructions day saturday rooms auctioneers auctioneer wednesday received messrs lo...	0.027627	2.611409

Fig. 22: Dominant topics for the 1860s.

There are no mentions of either event in the dominant topics for these periods. We think some possible explanations for the lack of mention are:

- When an event takes place over a long time, such as a decade, it may fall out of the news cycle and the initial excitement of a gold rush may dissolve.
- If an event does not span over a few years, it has a chance to be dominated by other topics, despite overall significance.
- Both the gold rush and suffrage movement examples are localized to a particular region. If we were to filter it by Otago and Canterbury, respectively, over the same periods we might expect different results.
- From the reference periods we used, the ones that mentioned the event in a topic were all of national importance. Examples of this are the wars that occur overseas and the great depression. This calls back to the localized argument. Events that affect the entire nation are more like to become dominant topics. Worth noting, the women's suffrage movement would be considered a national event but like we mentioned it was started in Canterbury.

The reasons for these results could be a mixture of these explanations or potential unknown ideas. Another point is that war and peace have been significant topic across any time period. Times of war are incredibly emotional, as families are separated and as news of casualties arises. People following the war closely may result in more publications writing about the events of the war more.

This may unearth a deeper question and topic for research as the people tend very differently to negative news to they do positive. However, this type of research is far beyond the scope of this project.

In a Specific Region

We performed a lot of analysis for the separate regions; Otago, Christchurch, Manawatu-Wanganui, and Wellington. In the interest of not making this section too long, we will only include and discuss some of the most important region findings. We employ readers to visit the Github repository for more information.

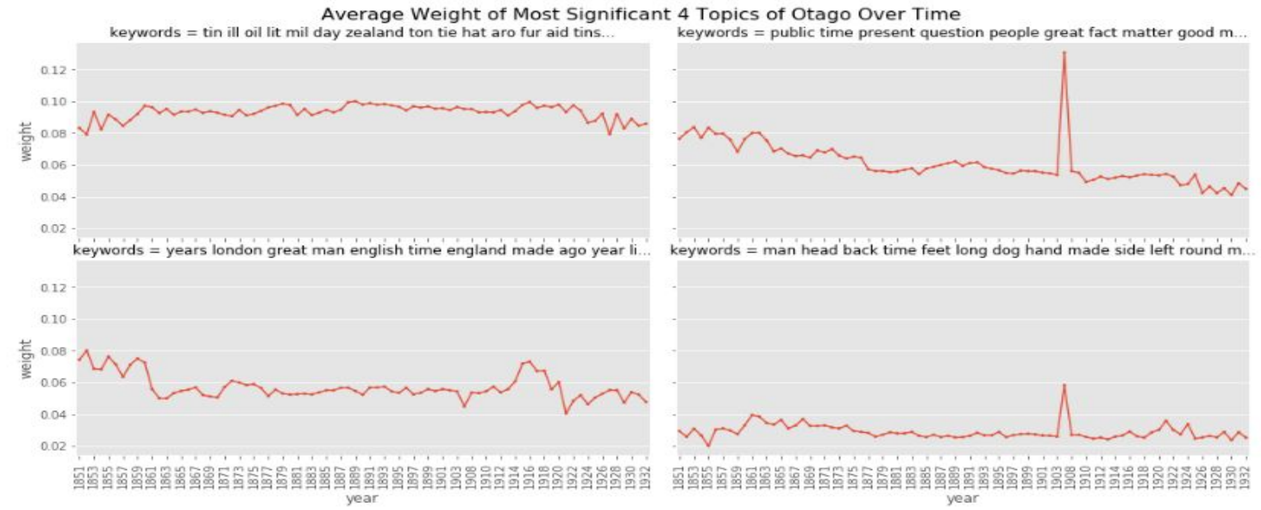


Fig. 23: Average weight of the most significant topics in Otago over time.

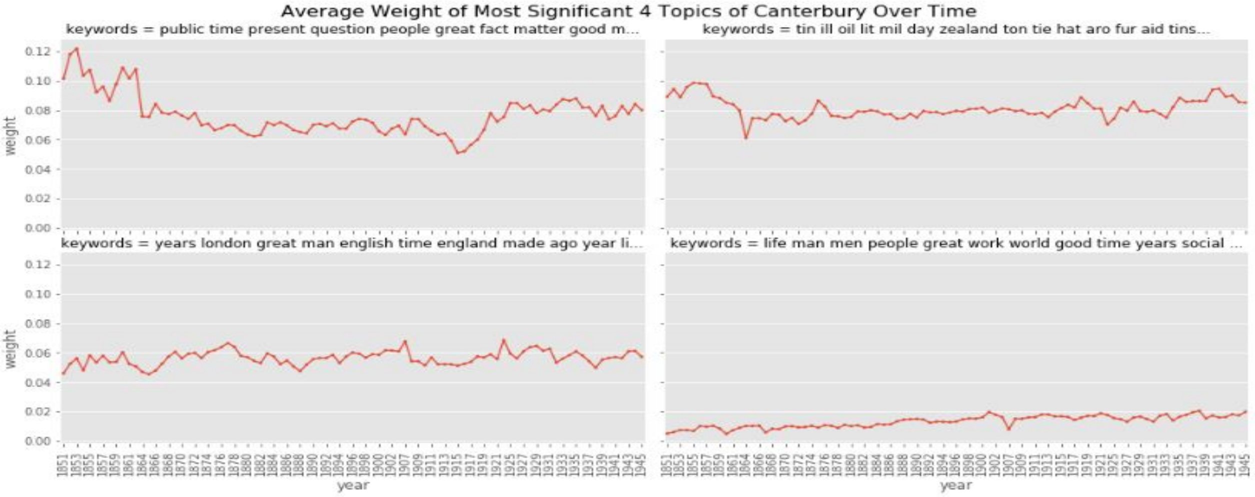


Fig. 24: Average weight of the most significant topics in Canterbury over time.

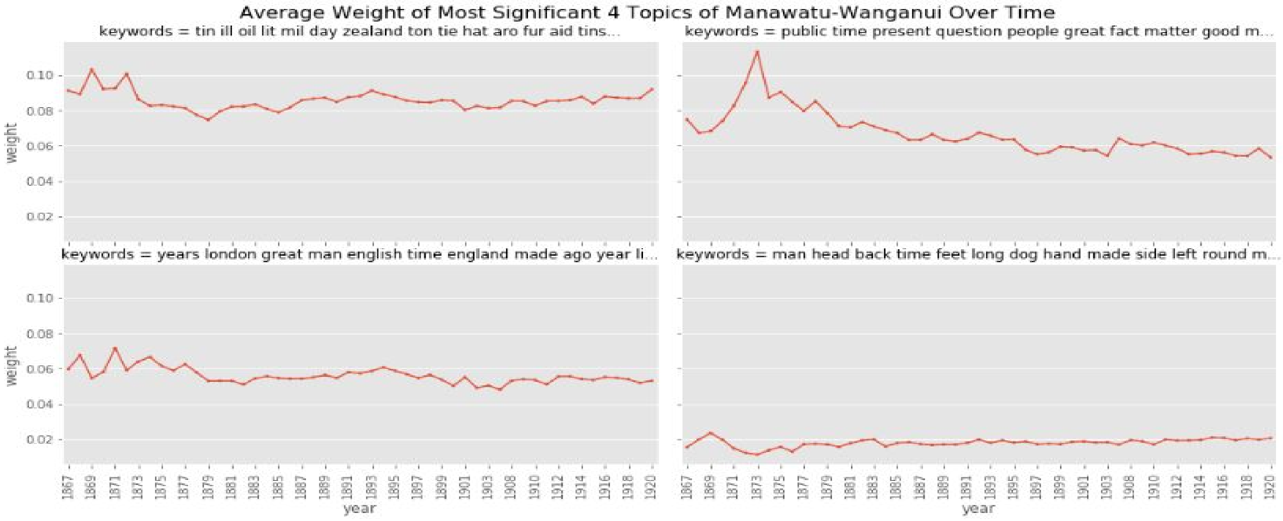


Fig. 25: Average weight of the most significant topics in Manawatu-Wanganui over time.

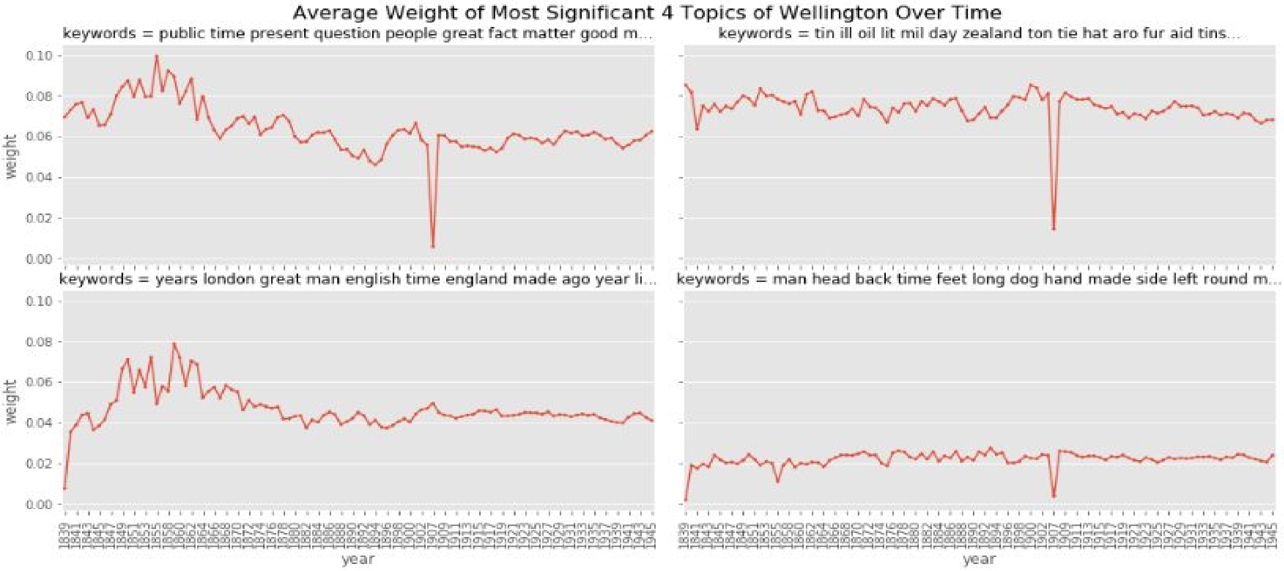


Fig. 26: Average weight of the most significant topics in Wellington over time.

In the figures 23-26, we have the four most significant topics for each region. The most significant topics are being repeated across the four regions. This is expected and it makes it simple to make meaningful comparisons. Although the topics are the same across regions, the trend of average weight differs. For example, if we were to look at the topic in the bottom-left corner of the plots. Wellington peaks more between 1845 and 1866 more than any other regions. In Otago, this topic reaches its peak in 1914 and 1918.

This shows that even though the regions have the same significant topics, periods in which the topic's average weight is highest severely differs. Figure 27 is another example of this if we choose topic 161 and compare them across regions. We can see that average weights are vastly different from each other, Canterbury and Manawatu-Wanganui trend downwards. While Otago and Wellington trend upward and show far more peaks troughs.

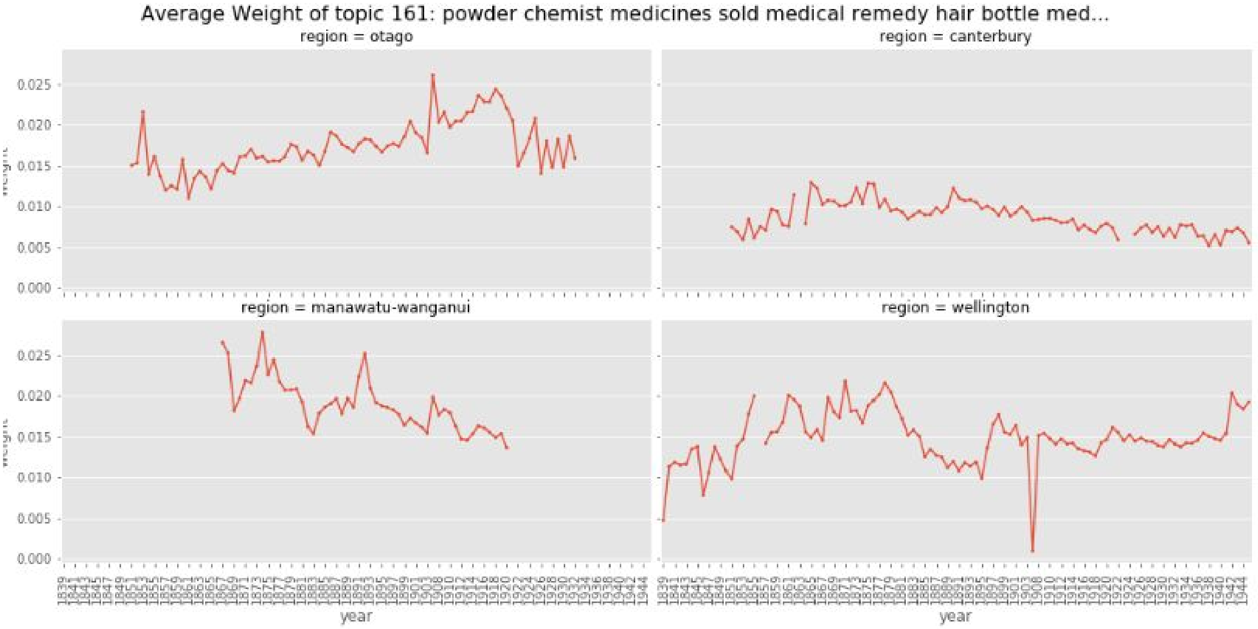


Fig. 27: Average weight by region for topic 161 over time.

Advertisements

Analysing the data set's advertisements was not one of our core research questions. In the brief for *Papers Past*, we state is an overdrive objective and could provide a source of further insight into the discourse of New Zealand.

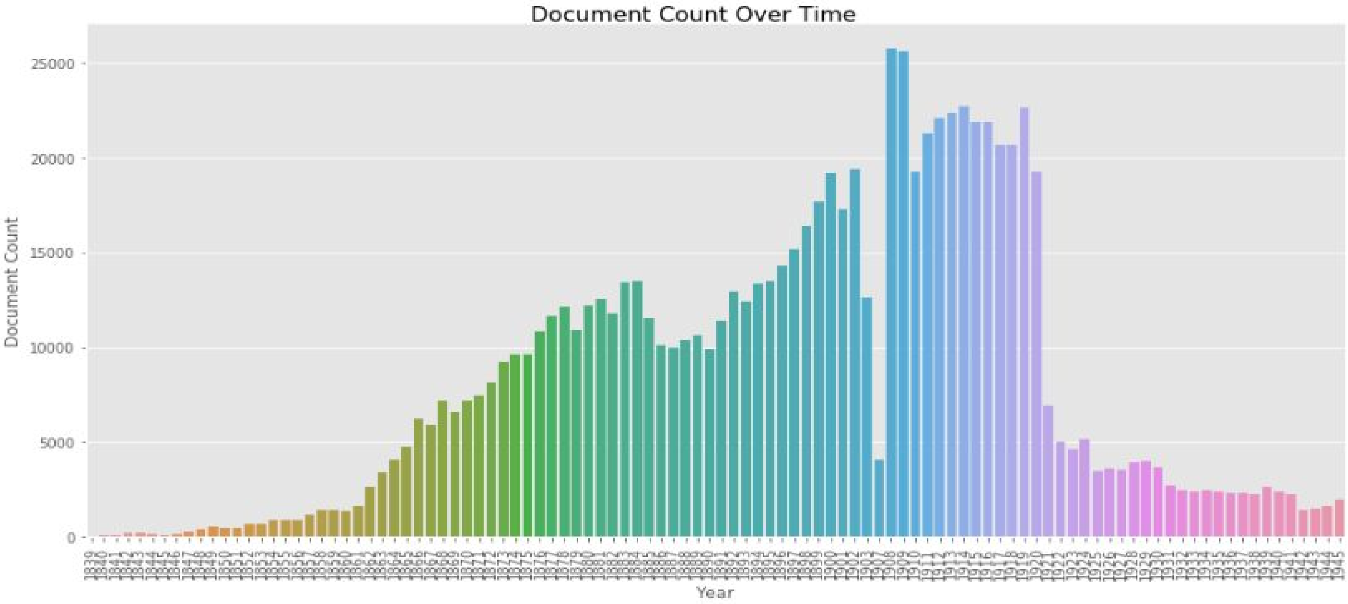


Fig. 28: Advertisement count over time.

The counts of advertisements of the time periods appear to match the total document count. This makes intuitive sense, more documents being published allows for more advertisements to published. So we would expect the two counts to correlate positively.

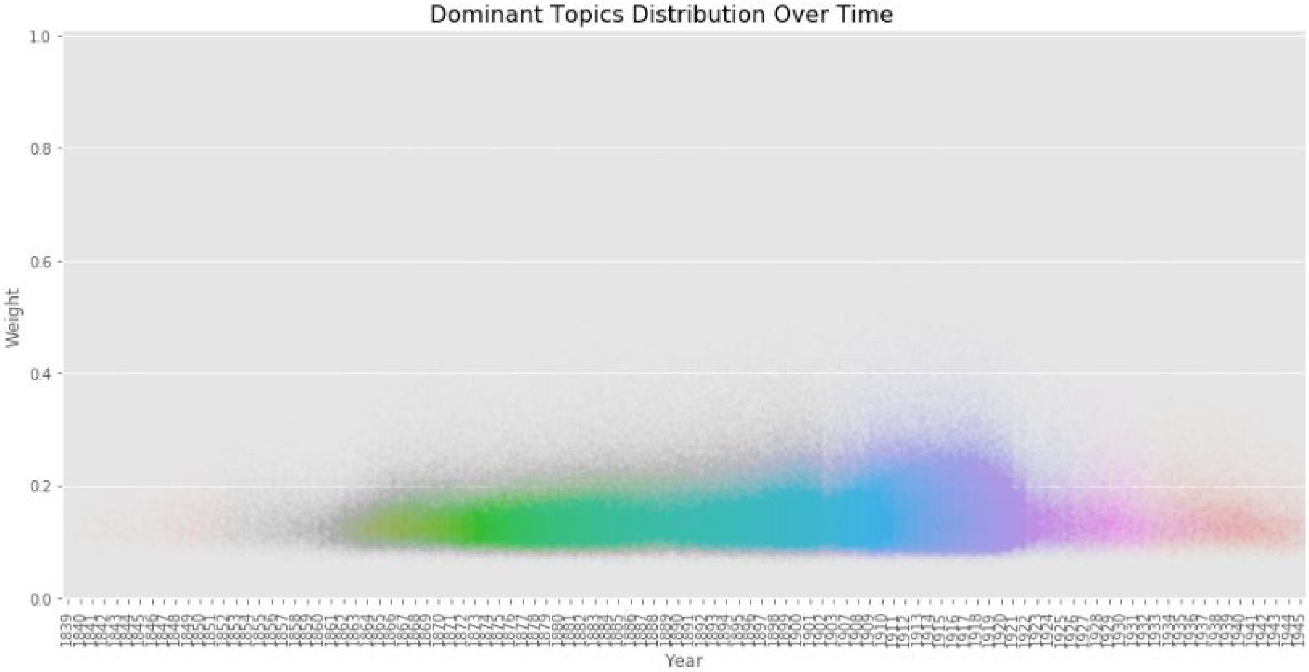


Fig. 29: Dominant advertisement topic distribution over time

The distribution two makes sense as it follows the document distribution in figure 29. It becomes denser in the same time period 1860 through to 1922. The obvious difference here is the weight of the advertisement in the document is significantly less than the dominant topic. Accounting for roughly 10% - 25% in most documents.

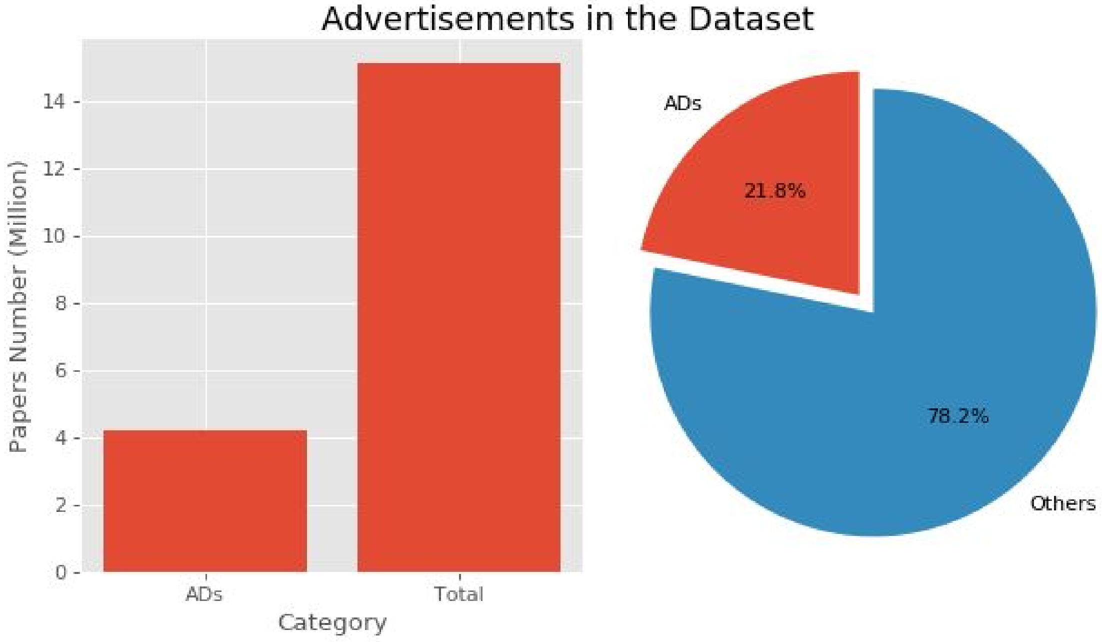
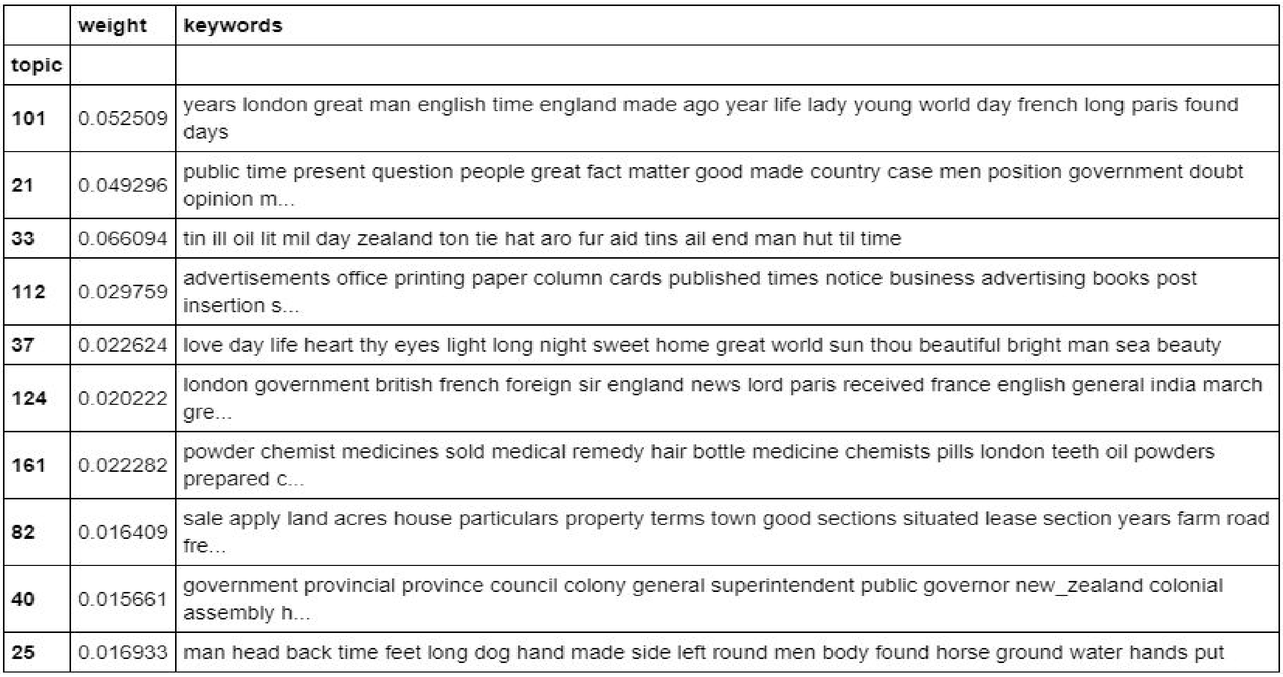


Fig. 30: An overview of advertisements in the dataset.

Figure 30 shows that advertisements account for 21.8% of the total dataset. In the clean dataset it advertisements account for 27.8%. This is important since when analyzing the advertisements we separate them into their own subset.



	weight	keywords
topic		
101	0.052509	years london great man english time england made ago year life lady young world day french long paris found days
21	0.049296	public time present question people great fact matter good made country case men position government doubt opinion m...
33	0.066094	tin ill oil lit mil day zealand ton tie hat aro fur aid tins ail end man hut til time
112	0.029759	advertisements office printing paper column cards published times notice business advertising books post insertion s...
37	0.022624	love day life heart thy eyes light long night sweet home great world sun thou beautiful bright man sea beauty
124	0.020222	london government british french foreign sir england news lord paris received france english general india march gre...
161	0.022282	powder chemist medicines sold medical remedy hair bottle medicine chemists pills london teeth oil powders prepared c...
82	0.016409	sale apply land acres house particulars property terms town good sections situated lease section years farm road fre...
40	0.015661	government provincial province council colony general superintendent public governor new_zealand colonial assembly h...
25	0.016933	man head back time feet long dog hand made side left round men body found horse ground water hands put

Fig. 31: The 10 most significant advertisement topics.

In figure 31 it shows the 10 most significant advertisement topics. It is not inherently obvious that these topics are related to advertisements without knowing beforehand. Topic 101 and 25, for example, seem to more general in nature. Topics 161 and 82 seem to be advertising pharmaceuticals and the sale/leasing of land, respectively.

In figure 31, topic 82 shows a very clear high period between 1839 and 1870. The advertisement for land sales/leases makes sense for this time period. This is when migration to New Zealand was at its height, as more people came from Europe. We mentioned that immigration would be an important event to use as a reference. The dates of the average weight seem to line up accurately with when the booms.

However, we found it difficult to make any inferences about the other topics and weight plots. This is for two reasons:

- We lack the historical knowledge of certain time periods and the daily life of New Zealand to draw any meaning from the topics.
- Topics don't appear to be directly related to any good or service that could be advertised. This could be a result of how we initially defined 'advertisements' against 'document'. Otherwise, it could be an issue with the LDA algorithm and how it topic modelled the advertisement subset.

We have received some usefulness out of modeling the advertisements but the outputs are not perfect. Further exploration would have to do in future work

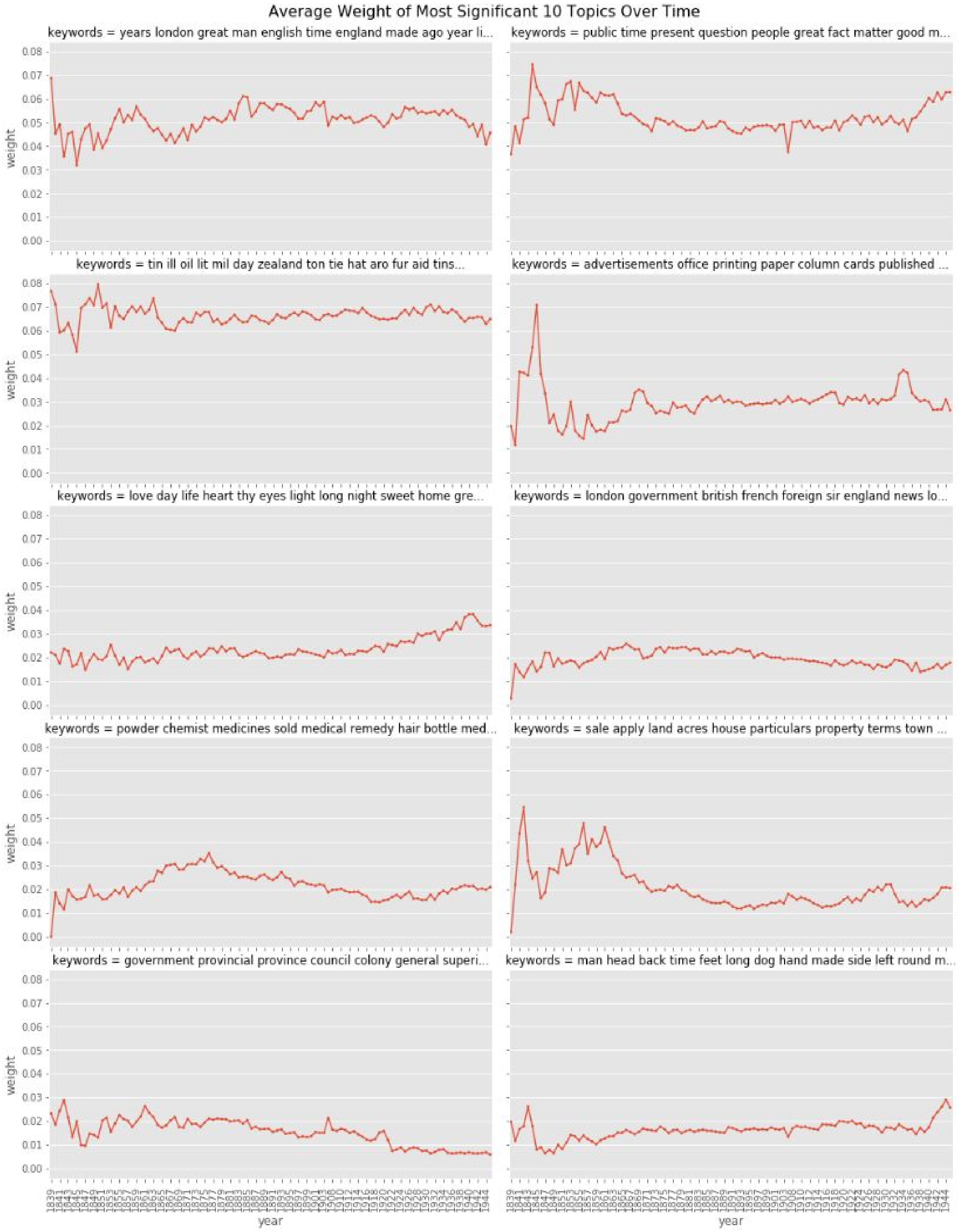


Fig. 32: Average weight of most significant 10 advertisement topics over time.

Application on Data Mining

We applied a simple data mining algorithm - linear regression on topic modeling result files to find with time goes there are correlations between topics or not.

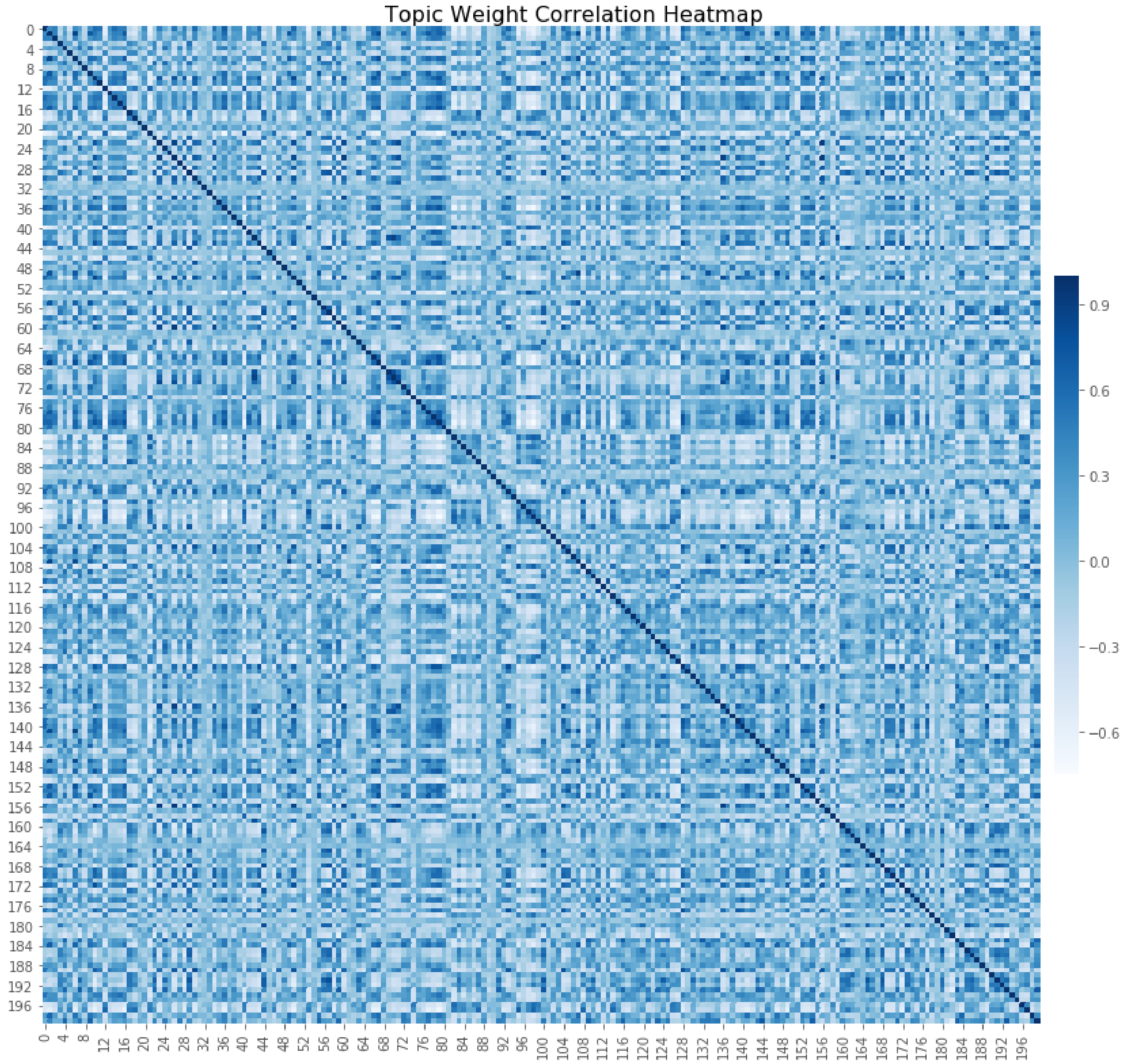
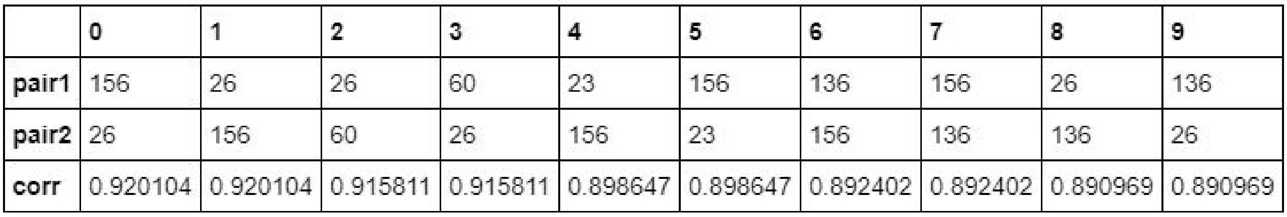


Fig. 33: Topic weight correlation heatmap

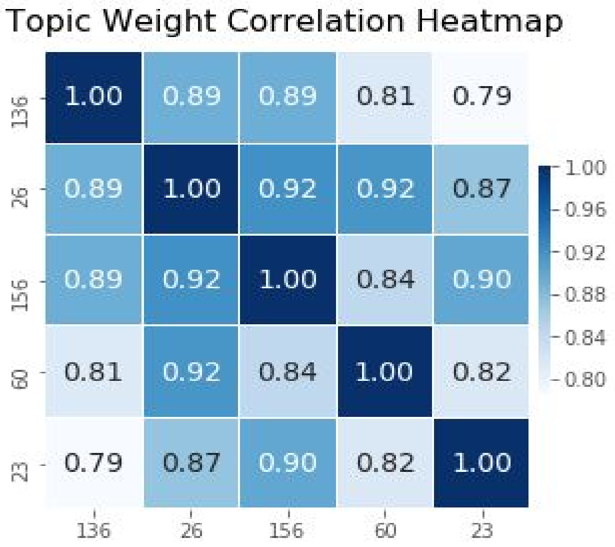
It is difficult to display the heatmap to faithfully show the correlations. This is due to the density of topics and the limitations of displaying it in an A4 sized report. Below are the top ten correlated pairs of topics.



	0	1	2	3	4	5	6	7	8	9
pair1	156	26	26	60	23	156	136	156	26	136
pair2	26	156	60	26	156	23	156	136	136	26
corr	0.920104	0.920104	0.915811	0.915811	0.898647	0.898647	0.892402	0.892402	0.890969	0.890969

Fig. 34: Ten most positively correlated topic pairs table.

From figure 34 you can see that topics; 136, 23, 60, 156, 26 are the most positively correlated.



	136	26	156	60	23
136	1.00	0.89	0.89	0.81	0.79
26	0.89	1.00	0.92	0.92	0.87
156	0.89	0.92	1.00	0.84	0.90
60	0.81	0.92	0.84	1.00	0.82
23	0.79	0.87	0.90	0.82	1.00

Fig. 35: Topic Weight Correlation Heatmap for topics; 136, 26, 156, 60, 23.

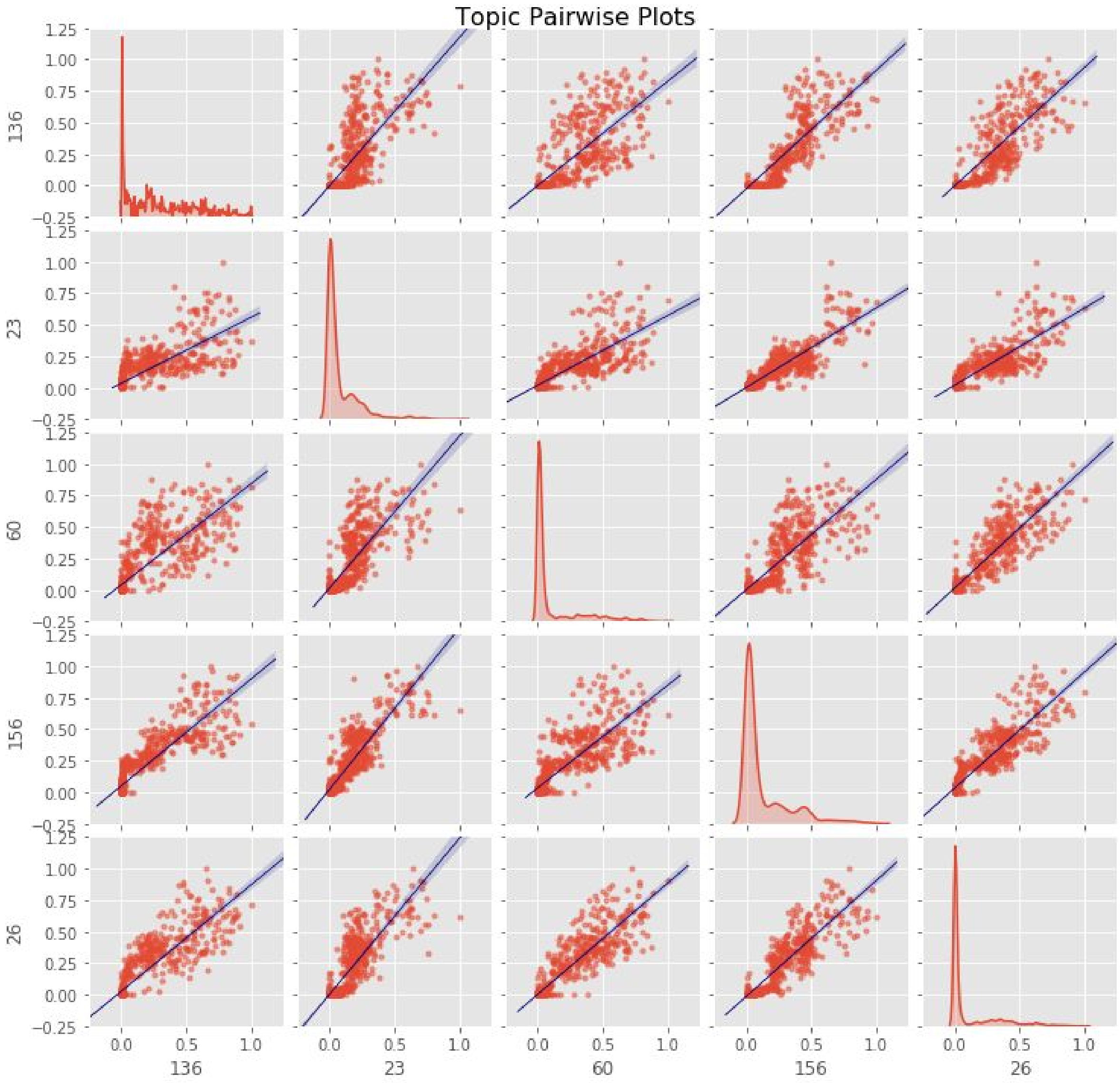
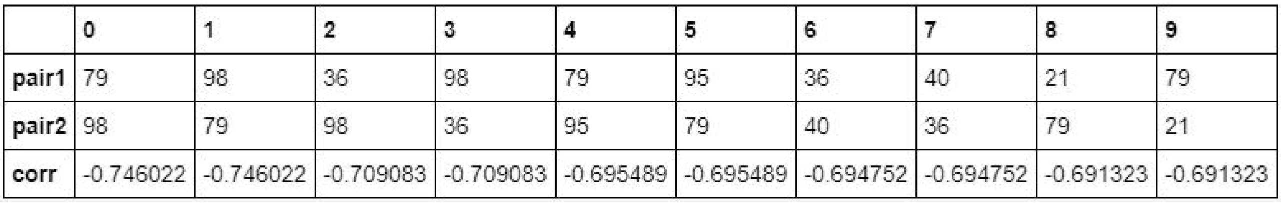


Fig. 36: Topic pairwise of the most positively correlated topics.

We can see that the relationships between the top 5 topics show a strong relationship.



	0	1	2	3	4	5	6	7	8	9
pair1	79	98	36	98	79	95	36	40	21	79
pair2	98	79	98	36	95	79	40	36	79	21
corr	-0.746022	-0.746022	-0.709083	-0.709083	-0.695489	-0.695489	-0.694752	-0.694752	-0.691323	-0.691323

Fig. 37: Ten most negatively correlated topic pairs table.

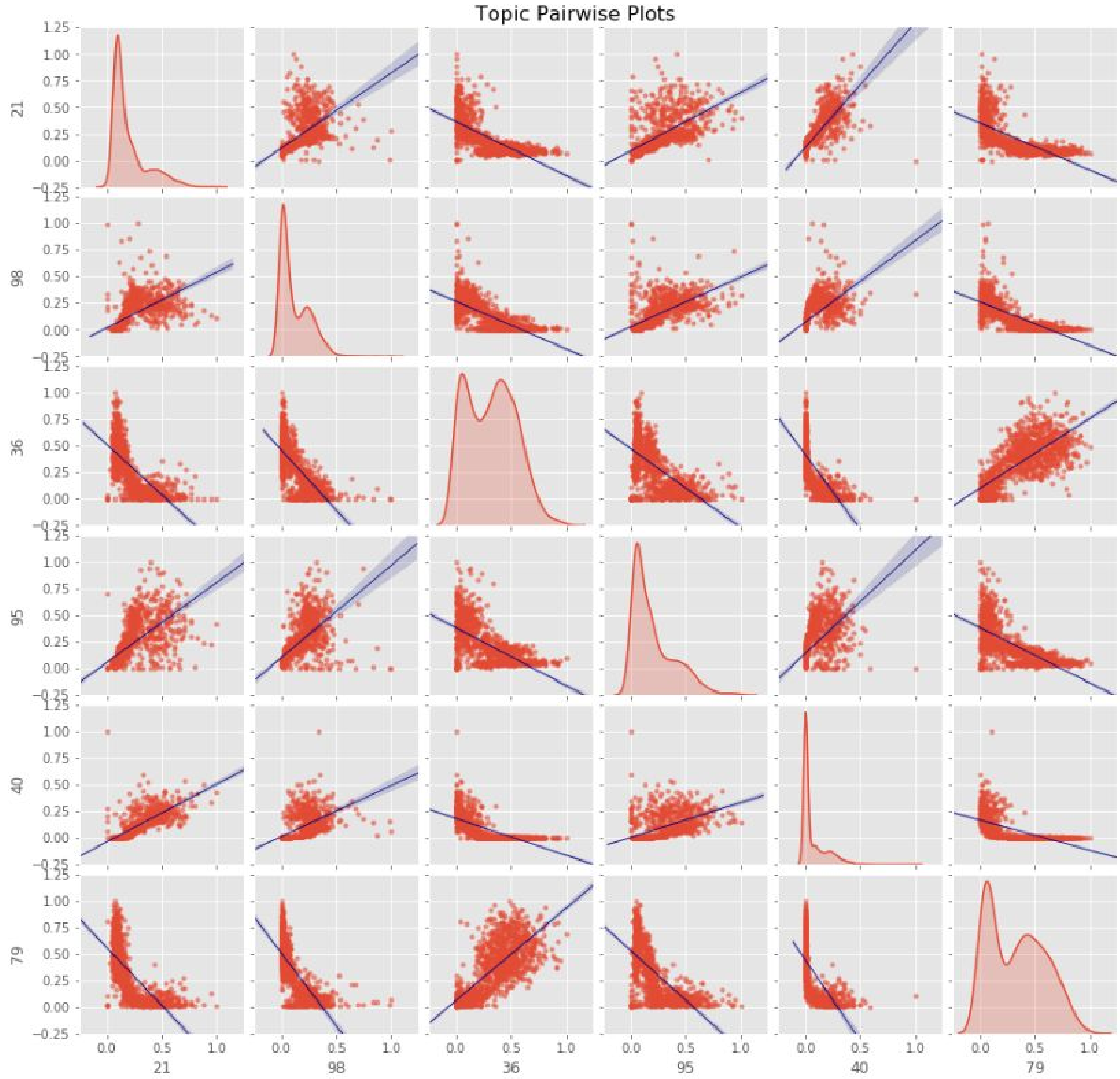
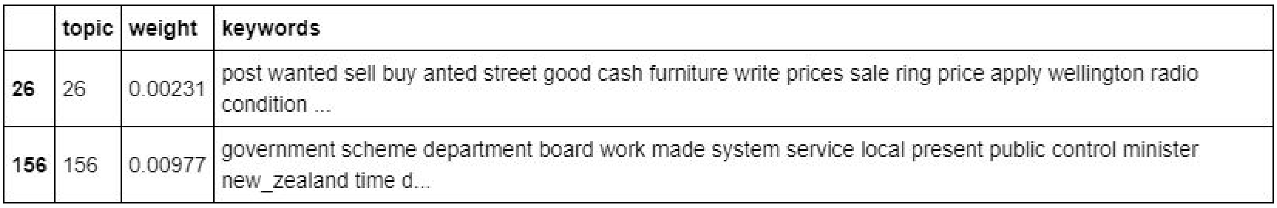


Fig. 37: Topic pairwise of the most negatively correlated topics.

From figure 37 you can see that topics; 21, 98, 36, 95, 40, 79 are the most negatively correlated. The top 5 most negative correlated topics show that pairwise plots that are more ambiguous and do not follow as similar relationships as the positively correlated topics did.



	topic	weight	keywords
26	26	0.00231	post wanted sell buy anted street good cash furniture write prices sale ring price apply wellington radio condition ...
156	156	0.00977	government scheme department board work made system service local present public control minister new_zealand time d...

Fig. 38: The most correlated topic pairs; 26 and 156.

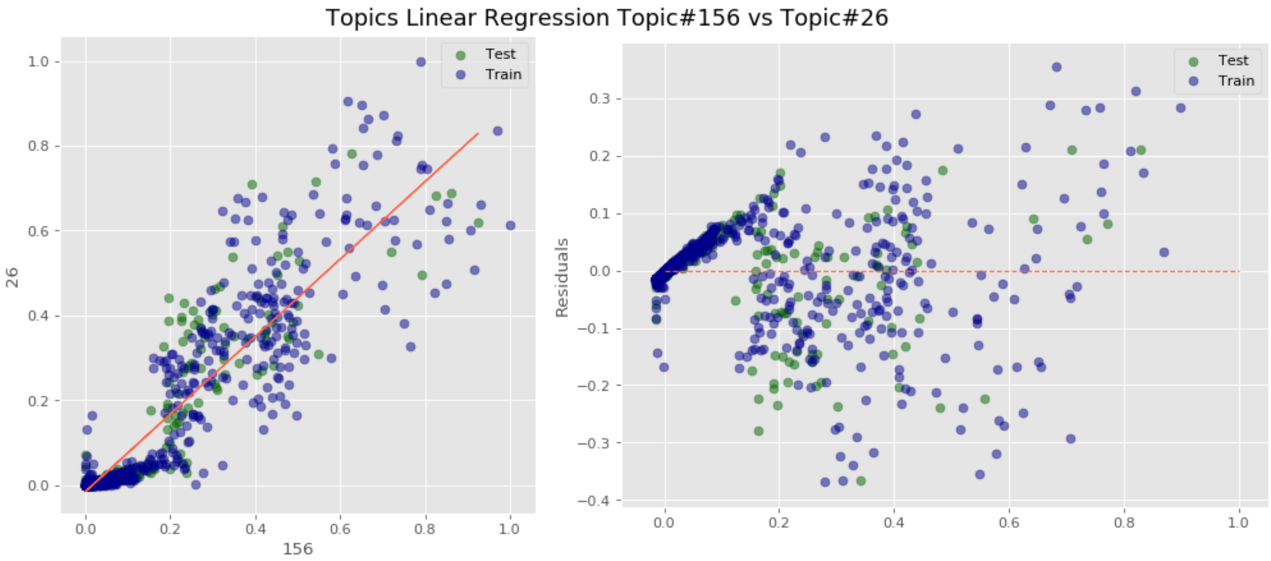


Fig. 39: Topic linear regression topic #156 vs Topic #26

Figure 39 shows the linear regression to find the correlation between the topics 156 and 26, the result is not as good as we expected. In the “Average Weight Topic over time” figure 40 (found in the appendix) there is a positive correlation between topic 44 and 100. While there is a negative correlation between topic 24 and 44. We cannot detect this correlation using linear regression, there are other methods of analyzing correlation, such as the Pearson Correlation. Other methods will have to be considered in future work.

Sentiment Analysis

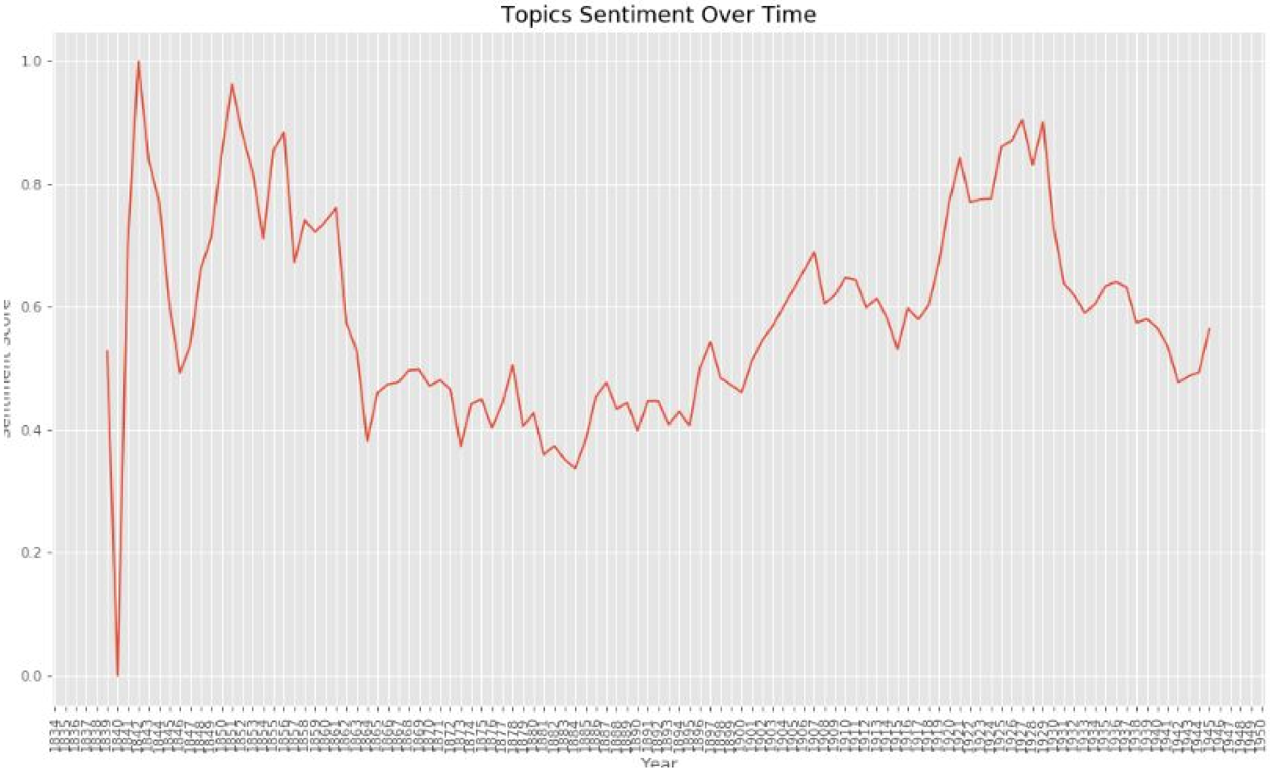


Fig. 40: Topics Sentiment over year

Sentiment Analysis is an incredibly interesting application of natural language processing. The idea is to attempt to extract opinions from the text, they are entirely subjective impressions and not facts. Sentiment analysis assumes that opinions take a binary opposition, for example, positive or negative. Sentiment Analysis allows us to understand how historical events shaped the country's opinion.

Very interesting, the social sentiment is in some level reflected by topics models. With our limited historical knowledge, in 1840 the treaty of Waitangi is signed and Europeans begin increasing immigration. This time period has a high sentiment score so topics from newspapers have a positive connotation. There is a decrease after 1860, which we know is when some of the New Zealand wars took place.

Some interesting points that appear to relate to known historical events. If we look at 1914, we can see a huge decrease in sentiment score. This is the start of World War 1 and we would expect that event to have a negative impact on society. Then the score increases when the war is drawing to a close and soldiers return.

After the first world war, the great depression hits, obviously a troubling time for the majority of concerns with market insecurity and mass unemployment. We would expect negative opinions to arise and the sentiment score clearly shows that decrease at around the same time. Then decreases further as the tensions between the great powers increases and the second world war break out.

Unfortunately, this highly speculative and we can only use our knowledge our well-known significant events to tie and infer meaning from our results. There could be further underlying reasons for the peaks and troughs, but the events seem to match up well with our reference events.

Sentiment analysis allows us to go further than topics. It brings a deeper insight into the opinions around topics. We could take this further with *Contextual Semantic Search* [20], this can identify what aspects of the topic are actually being discussed. It will return the distribution of positive, negative, and neutral mentions of topic aspect. Our work on sentiment analysis is brief, there are other methods and applications that were not attempted in this project and would be considered for future work.

[20] More information about Contextual Semantic Search found at Gupta, S. (2018, May 11)

Discussion

Pandas vs PySpark

Pandas run on Python, PySpark runs on Java and both can be used to manipulate data. At first, we preferred using Pandas rather than PySpark to wrangle data for its high efficiency and easy use, but when data size reached to over 10 GB, the speed of data loading, exporting, some manipulations are far slower than PySpark, especially compare to the speed of running on Spark/Hadoop cluster. Finally, we selected PySpark when we need to deal with big data frames, and used Pandas to process small data frames.

Topic Number

What is the “correct” topic number is a hot question in the topic modeling field? In the beginning, we were confused about the topic number and applied *Topic Coherence* method to quantitatively evaluate the effect of different topic number (shown below). It is similar to cross-validation, iteratively try the different topic number and calculate the topic coherence then decide what’s the best topic number. The problem is, if one iteration takes 1 hour, the method is acceptable, but if one iteration takes over 10 hours, the method is out of consideration for us.

We do not think of the number of topics is a natural characteristic of corpora. The topic number is not really combinations of multinomial distributions, so there is no "right" topic number. We think of the number of topics as the scale of a map of corpora. If we want a broad overview, we use a small topic number. If we want more detail, use a larger topic number. The right number is the value that produces meaningful results that allow us to accomplish our goal and there is a wide range of good values for us.

Text Preprocessing

Normally we will perform text preprocessing step before we fit a model to get a better result, such as tokenize, stopwords, lemmatization, stemming, bi-grams, TF, etc. MALLET provides tokenize, stopwords, function by default, the differences are mainly lemmatization and stemming. After many times experiments and read some literature, we consider them are not very useful in big dataset topic modeling. A well-trained topic model will cluster related words to one topic or related topics, lemma and stem words would help in small corpus but not improve big dataset topic modeling, it will help improve the weight of rare words but it’s not what we concern.

In a word, text preprocessing would not improve topic modeling clearly in the big data condition. We would not be entangled in with or without text preprocessing issue, just ignore it.

OCR quality and Spelling Correction

As for the result of topic modeling, the quality of corpora is the most important factor, with the high-quality corpora (long, typo-free text), even TF-IDF could do the topic modeling job, while if the corpora is low quality (short text or full of errors), there is no way to get good result. In our case, we have enough long texts but all documents are full of typos, which means if we have some methods to correct errors, the result would improve obviously.

Unfortunately, we did not have the time to develop a way to do it, after trying SymSpell package, we did not find it contributes to topic modeling and depreciated it. The replacement method would be useful which find some typo pattern and replace the pattern with correct letters, for example, OCR may take ‘mallet’ to ‘mallot’ or ‘malbt’, we use ‘mallet’ replace ‘mallot’ and ‘malbt’ will correct this pattern of error. But find patterns and generate replacement list may be complicated and time-consuming.

In this project we utilize the “extra stopwords” argument of MALLET to add “bad OCR” topics from last topic modeling, then the next topic modeling process will take those “bad OCR” words as “stop words” and ignore them, in this way, after several iterations of topic modeling, the “bad OCR” rarely appear in the topic list.

MALLET vs gensim

MALLET is a full solution for topic modeling and other fields, by default it has two methods to use: command mode and API mode, we did not use API mode because it works in JAVA development environment, so it appears we could only use command mode. In this mode, we have very limited space to perform text preprocessing using MALLET before topic modeling. However, gensim provide a wrapper of MALLET, we can perform MALLET in Python development environment. With the help of abundant NLP tools in Python environment, such as NLTK, spacy, pyLDAvis, etc., we could easily control every detail of the full process from tokenizing text to visualizing topic models, so we choose gensim to perform topic modeling at first and get impressive results.

With the progress of the project, the data size increased rapidly, the efficiency issue appeared, it took much more time comparing with MALLET. Finally, we gave up text preprocessing step, fed MALLET with raw text directly. The advantages are that the workflow because concise and high efficiency, and the results were not worse than the process with text processing, it turns out that general text processing is not so important in topic modeling.

Pruning Bag of Words

There is an important step is pruning which impacts the efficiency and accuracy of topic modeling by MALLET. MALLET will generate an import-file model which is the bag of words, as our understanding. The model generally contains billions level (or more) features/vocabularies which will be used to generate matrices with documents/topic number to calculate topics iteratively. With billions features the matrices are huge and would lead to out of memory issue if the memory is not big enough.

So we have to prune the model using MALLET, but if we trim it too much the accuracy of topics would go down, so there is a tradeoff in it. MALLET provide some arguments to downsize the model, we turned some of them and selected use “max-IDF” to reduce features, and to avoid use “min-IDF” which may filter out frequent but meaningful words.

The Metrics to Quantitatively Evaluate Topics

To evaluate and compare the distribution of topics, we use two basic metrics, in practice, we use both to find target topics:

The dominant topic counts

- Every document has a dominant topic represent the most portion of its contents. By summing the number of each dominant topic in a range of time, we see the variety and trend of each topic over time. It is easy to understand and fit for scatter plot, but in this way, it ignores non-dominant topics.

- Put simply, the goal to classify a document belonging to a particular topic and assign that topic. The dominant topic is defined as the topic that exceeds all other topic proportions. Documents will generally provide a mixture of topics, so the dominant topic is the primary topic.

The average weight of topics over time

- This metric is the average weight of a topic through all documents in a range of time, it avoids the issue above, reflect the topic variety and trend over time. However, it might give a frequent/normal topic (which has a small weight in each document) a high weight because the topic appears in a large number of documents when summing all weights of each topic and dividing by total weight of all topics, this normal topic may get a high weight.

Trained Topic Models vs Inferred Topic Models

In this project, although the result of the trained topic model is stable and reasonable, we are not satisfied with the results of inferred topic models. The MALLET seems not a profession at inferring topic models, e.g. MALLET could only use one core to infer topic model, which limited the usage of inferring, and the doc-weight matrix is not coincident with a trained model based on our experiment. If we have to use MALLET to perform topic modeling in the future, we would prefer using the training command of MALLET not inferring command. Overall, MALLET is a high-performance topic modeling tool, excellent at training topic models of large corpora, except the inferring weakness.

Future Work

- Spelling correction - We mentioned in the discussion about the quality of the OCR and we could improve our results if we removed the typos. When we tested out SymSpell, we found that it didn't improve results and wasn't efficient and took far too long. There are other methods to deal with the OCR error words or typos, such as matching the error pattern to replace typos with correct words. If we have a chance do a similar project, we would invest much more time in spelling correction field, it is totally worth it.

- Quantitatively tune the topic number - Many metric methods and tools could help us to quantitatively tune the topic number, such as Hierarchical Dirichlet process, dating and topic coherence.

- Correlation - Explore other algorithms to find the correlation of topics over time, in our experiment the linear regression is not suited for analyzing time series type of data. The range of time may affect the regression a lot, in a specific period time the correlation between topics may appear stronger than the whole time range. By applying regression on topics, we can evaluate the relationship between topics quantitatively, and have more chance to find (or even predict) interesting patterns or features in topics.

- Sentiment Analysis - Better methods to generate sentiment values, current work quite brief and not explorative. We have some preliminary results but there are difficult to accurately interpret and could certainly be improved. Applications such as contextual semantic search could be explored, to gain greater insight into the topics.

- Pronoun Tagging - In the brief, we mentioned that of our applications of Topic Modeling could be the detection of pronouns. It was intended to be an application for the website. The idea is that we could make a model to 'tag' place names and personal pronouns. This would allow articles to be linked by a particular name. For example, if you are interested in following the life of Julius Vogel, eighth premier of New Zealand. You could find one document mentioning him and the model would recommend more documents that mention Julius Vogel.

- Top topic contributors – This was considered to be another application for the website. Somewhere around the document, whether it be the sidebar or at the beginning. There would be a display of topics contribute to a large portion of the document. This would show the keywords from the topic and the most significant words.

Acknowledgments

This project was supported by the University of Canterbury. Chris Thomson, Ben Adams and James Williams provided immense help with both our technical and theoretical questions. As well as providing guidance and supervisory support throughout the course of this project.

References

Alves, T. (2017, September 17). 12 moments that shaped New Zealand's History. Retrieved from: https://theculturetrip.com/pacific/new-zealand/articles/12-moments-that-shaped-new-zealands-history/

A Gospel of Health and Salvation. (n.d.). Retrieved from http://jeriwieringa.com/portfolio/dissertation/

Bast, H., Buchhold, B., & Haussmann, E. (2016). Semantic Search on Text and Knowledge Bases. Foundations and Trends in Information Retrieval, pp. 119-271. *https://doi.org/*10.1561/1500000032.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM,55*(4), pp. 77. doi:10.1145/2133806.2133826

Blei, D. M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John, ed. Latent Dirichlet Allocation. *Journal of Machine Learning Research. 3 (4–5): pp. 993–1022.*

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D M. (2009) Reading Tea Leaves: How Humans Interpret Topic Models. Neural Information Processing Systems, pp. 288-296. https://www.researchgate.net/publication/221618226_Reading_Tea_Leaves_How_Humans_Interpret_Topic_Models

Gupta, S. (2018, May 11). Automated survey using contextual semantic search [Blog Post]. Retrieved from: https://dzone.com/articles/automated-survey-processing-using-contextual-seman

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. *Proceedings of the First Workshop on Social Media Analytics - SOMA 10*. doi:10.1145/1964858.1964870

Kompalli, S., Nayak, S., Setlur, S., & Govindaraju, V. (2005). Challenges in OCR of Devanagari documents. *Eighth International Conference on Document Analysis and Recognition (ICDAR05)*. doi:10.1109/icdar.2005.70

Krasnashchok, K., & Jouili, S. (2018). Improving Topic Quality by Promoting Named Entities in Topic Modeling. *ACL*., 247 - 251. https://www.semanticscholar.org/paper/Improving-Topic-Quality-by-Promoting-Named-Entities-Krasnashchok-Jouili/5145ca7c35142177955b6d66996b78b374b1e4ab

Jelodar, H., Wang, Y., Yuan, C,. Feng, X., Jiang, X., Li, Y., & Zhao, L. (2018) Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. Cornell University, 1-32. https://arxiv.org/pdf/1711.04305.pdf

Jockers, M L., & Mimno, D. (2013). Significant themes in 19th-century literature. *Science Direct, 751 - 768. https://doi.org/10.1016/j.poetic.2013.08.005*

Joshi, P. (2018, October 16). An NLP approach to mining online using topic modeling (with python codes) [Blog Post]. Retrieved from: https://www.analyticsvidhya.com/blog/2018/10/mining-online-reviews-topic-modeling-lda/

Li, S. (2018, March 31). Topic Modelling in Python with NLTK and Gensim [Blog Post]. Retrieved from: https://towardsdatascience.com/topic-modelling-in-python-with-nltk-and-gensim-4ef03213cd21

Li, S. (2018, May 31). Topic Modeling and Latent Dirichlet Allocation (LDA) in Python [Blog Post]. Retrieved from: https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24

Mallet: MAchine Learning for LanguagE Toolkit.(2018). *Topic Modeling*. Retrieved from: http://mallet.cs.umass.edu/topics.php

Mallet: MAchine Learning for LanguagE Toolkit.(2018). Optimization. Retrieved from: http://mallet.cs.umass.edu/optimization.php

Mining the Dispatch. (n.d.). Retrieved from http://dsl.richmond.edu/dispatch/

Mulligan, J. (2017, March 9). The history of immigration booms in NZ [Audio Podcast]. Retrieved from: https://www.radionz.co.nz/national/programmes/afternoons/audio/201836026/the-history-of-immigration-booms-in-nz

Nair, G. (2016, July). Text Mining 101: Topic Modeling [Blog Post]. Retrieved from: https://www.kdnuggets.com/2016/07/text-mining-101-topic-modeling.html

New Zealand History. (2017). *The war at home.* Retrieved from: https://nzhistory.govt.nz/war/first-world-war-overview/defending-our-shores

New Zealand History. (2014). *New Zealand and the Second World War.* Retrieved from: https://nzhistory.govt.nz/war/new-zealand-and-the-second-world-war-overview

Nikolenko, S.I. (2016). Topic Quality Metrics Based on Distributed Word Representations. *SIGIR*, 1029 - 1031. *https://doi.org/*10.1145/2911451.2914720

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM 15*. doi:10.1145/2684822.2685324

Sukhija, N., Tatineni, M., Brown, N., Moer, M. V., Rodriguez, P., & Callicott, S. (2016). Topic Modeling and Visualization for Big Data in Social Sciences. *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*. doi:10.1109/uic-atc-scalcom-cbdcom-iop-smartworld.2016.0183

Phillips, J. (2012, June 20). Story: South African War. Retrieved from: https://teara.govt.nz/en/south-african-war

Williams, T. & Betak, J. (2018). A Comparison of LSA and LDA for the Analysis of Railroad Accident Text. Scient Direct, 98-102. https://doi.org/10.1016/j.procs.2018.04.017

Yang, T., Torget, A., & Michalcea, R. (2011). Topic Modeling on historical newspapers. ACM Digital Library. 96-101. https://dl.acm.org/citation.cfm?id=2107649

Yildirim, I. (2012). Bayesian inference: Gibbs Sampling. Technical Note. University of Rochester. PP. 1-5. http://nlp.chonbuk.ac.kr/PGM/slides_other/GibbsSampling.pdf

Appendix

A Repos

https://github.com/xandercai/papers-past-topic-modeling

Directory instruction:

```
papers-past-topic-modeling
├── 1-loading               # part 1. Data Loading
├── 2-wrangling             # part 2. Data Wrangling
├── 3-exploring             # part 3. Data Exploring
├── 4-preprocessing         # part 4. Text Preprocessing
├── 5-modeling              # part 5. Topic Modeling
├── 6-analyzing             # part 6. Analysis and Visualization
├── 7-applying              # part 7. Applications
├── models/train            # topic keywords file
└── previous                # solutions of week 1 to week 7
```

B Annual average weight over time of each topic

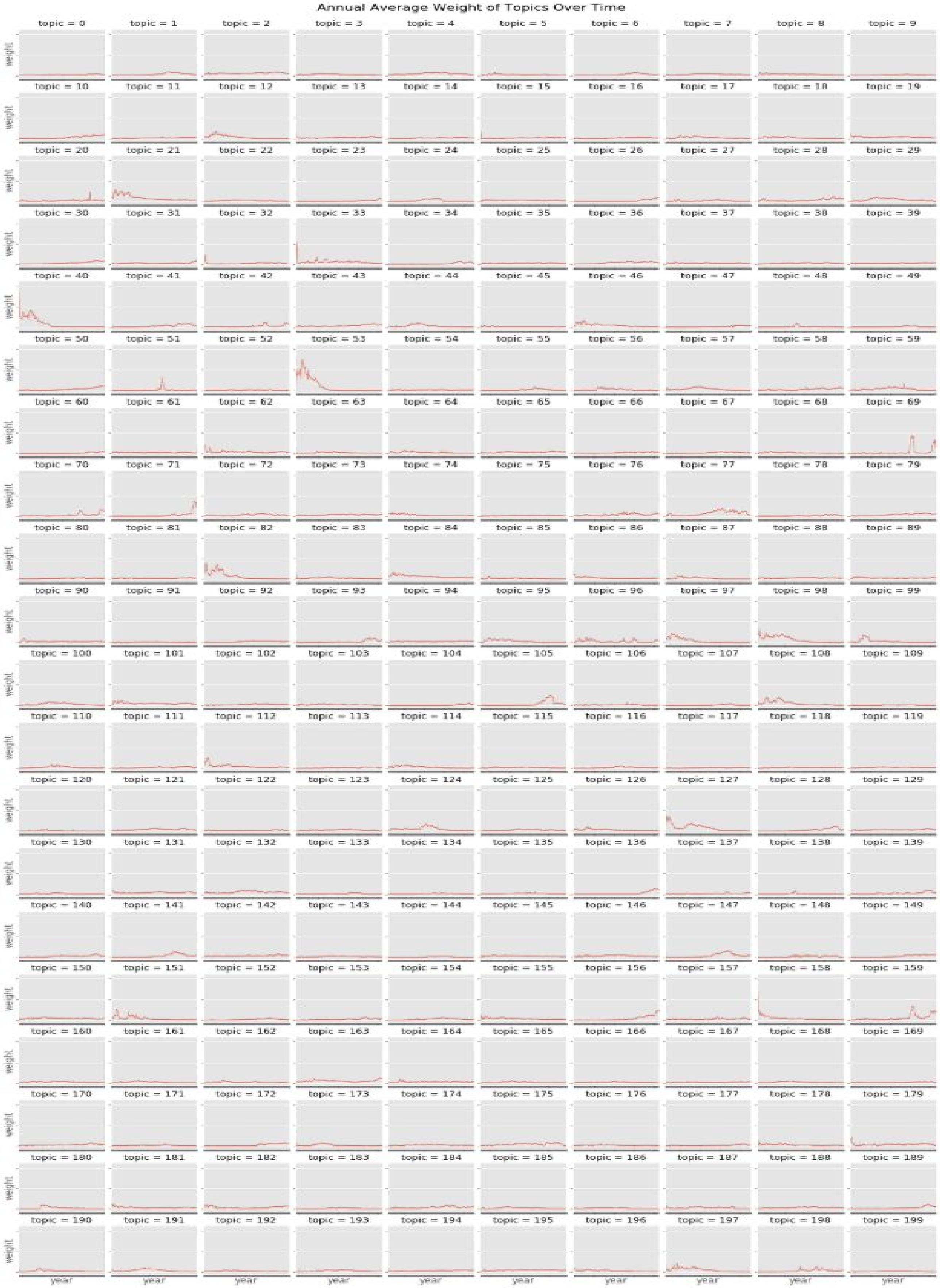


Fig 41: Average weights for all topics over time