Big Data: Text Summarization

Team 5

New Zealand Earthquakes Summary

Rohit Kumar Chandaluri: rohitchandaluri@vt.edu
William Edmisten: wce@vt.edu
Alex Bochel: balex96@vt.edu
Jun Lee: jlee123@vt.edu

CS 5984/4984 - Fall 2018
Instructor: Edward A. Fox

Virginia Tech
Blacksburg VA, 24061
5 December 2018

Table of Contents

Table of Figures

Table of Tables

Abstract

The purpose of this Big Data project was to create a computer generated text summary of a major earthquake event in New Zealand. The summary was to be created from a large webpage dataset supplied for our team. This dataset contained 280MB of data. Our team used basic and advanced machine learning techniques in order to create the computer generated summary. The research behind finding an optimal way to create such summaries is important because it allows us to analyze large sets of textual information and to identify the most important parts. It takes a human a long time to write an accurate summary and may even be impossible with the number of documents in our dataset. The use of computers to do this automatically drastically increases the rate at which important information can be extracted from a set of data.

The process our team followed to achieve our results is as follows. First, we extracted the most frequently appearing words in our dataset. Our second step was to examine these words and to tag them with their part of speech. The next step our team took was to find and examine the most frequent named entities. Our team then improved our set of important words through TF-IDF vectorization. The prior steps were then repeated with the improved set of words. Next our team focused on creating an extractive summary. Once we completed this step, we used templating to create our final summary.

Our team had many interesting findings throughout this process. Our discoveries were as follows. We learned how to effectively use Zeppelin notebooks as a tool for prototyping code. We discovered an efficient way to run our large datasets using the Hadoop cluster along with PySpark. We discovered how to effectively clean our dataset prior to running our programs with it. We also discovered how to create the extractive summary using a template along with our important named entities. Our final result was achieved using the templating method together with abstractive summarization.

Our final result included a successful generation of an extractive summary using the templating system. This result was readable and accurate according to the dataset that we were given. We also achieved decent results from the extractive summary technique. These techniques provided mostly readable summaries but still included some noise. Since our templated summary was very specific it is the most coherent and contains only relevant information.

1. Introduction

This report contains a detailed account of our semester long project to create an accurate summary of our dataset concerning an earthquake in New Zealand. The following sections explain how our group used multiple datasets to create our final summary. Our first dataset is a small set of articles about Local Elections. Our second and primary dataset is concerning an earthquake in New Zealand. Our final summary with this dataset is compared with a gold standard summary created by another team in the class. The report also contains the gold standard summary that we wrote for Team 6 regarding the Stoneman Douglas High School shooting in Florida. The report contains sections regarding our methods as well as our results. These sections include information about our design, implementation, and timeline, as well as important discoveries. Our deliverables for each task can be found in the Approach, Design, and Implementation section. Important trials and unsuccessful runs can be found in our Timeline section. We also include enough information for developers and users to benefit from our research by continuing our project or simply using it.

2. Literature Review

Natural Language Processing (NLP) is a growing area of research and application that seeks to understand how humans comprehend language for the purpose of developing software that can mimic that ability [1]. The study of NLP began in the 1950s and has since developed substantially. NLP is a challenging subject, because of the complex nature of human communication. Words in the dictionary have a definite meaning; however, words strung together in sentences can alter meaning significantly, based on context. NLP must be able to extract meaning from these sentences. These meanings are commonly known as “semantics”. NLP has to be able to identify the relationship between text units using their part of speech and also identify abstract meanings from metaphors and more [2]. The importance of the part of speech of words has been taken into account for this project. Our team has tagged each important word throughout the entire document set with its part of speech using the Natural Language Toolkit (NLTK). NLTK was a Python library created in 2001 and has been a popular tool for NLP. NLTK strives for simplicity, consistency, extensibility, and modularity [3]. These features led this toolkit to be recommended by our professor and are ultimately what we chose to use.

Three of the summarization techniques that our team has used include extractive, abstractive, and template-based. An extractive summary is a summary that looks for sentences with important words and takes the whole sentence and strings it together with other important sentences. The template-based summary utilizes a template that we create manually and fill in automatically using the important named entities and specific words of a certain part of speech. The abstractive summary was generated by a pointer generator model and K-means clustering algorithms.

Processing big data can take a lot of time and computing power, but there are solutions to work around this. Using a computational cluster to run and store programs allows for more computing power. Our team chose to use a Hadoop cluster for much of this work. Hadoop is a software framework that can be installed on a Linux cluster to allow large scale data analytics. Hadoop uses a robust file system known as the Hadoop Distributed File System. This system is similar to Google’s file system and also features fault tolerance [5]. The system has allowed us to process thousands of documents in seconds. However, a cluster is not the only tool that can expedite processing. Using a special framework with Python allows for parallel processing of data. PySpark (a Python framework for Apache Spark) breaks up data into separate “RDD” (Resilient Distributed Dataset) files that can be processed in parallel. These RDD files are manipulated through functional programming and have a unique fault tolerance. If an RDD file fails it will automatically recover and the system will continue processing files. This method of batch processing allows for rapid organization and analysis of the data [4].

3. Approach and Evaluation

In order to achieve the best results in the shortest amount of time, our team elected to use certain third party tools. We made this decision so that we could learn a few tools very well. The tools we used were the following:

- Natural Language Processing Toolkit (NLTK)
- Python 2.7
- Zeppelin Notebooks
- Hadoop Cluster
- PySpark

Our team used NLTK to determine important information about words and sentences in our dataset. This information includes the part of speech of words, named entities, important topics, discriminating features, and word synsets. This library provided many built in functions that simplified our code. Our language of choice was Python version 2.7. The version we decided to use was constrained by the version available on the cluster we used. Our team initially used Zeppelin on our local machines to begin experimental development. Once we completed the first five tasks locally we decided to move to the cluster and no longer used Zeppelin. We used the Hadoop cluster processing for the final results. The cluster provided us with an environment that could process large volumes of data very quickly. Our last third party tool was PySpark. There was a large learning curve that was required in order to effectively learn how to use this tool, but it ultimately optimized the runtime of our code.

3.1 Pre-Processing: Cleaning the Data

We looked into multiple approaches for cleaning the data: deep-learning document classification, rule-based classification, and Justext, as suggested by many teams in the class. The Justext tool is used to remove HTML tags from the documents and return only pure sentences. Unfortunately, it didn't work well for our corpus data. It did not remove all noise like Javascript, random punctuation, and website boilerplate. We instead used OpenNLP, which did solve those problems. To further clean data to get more relevant documents, we used a rudimentary rule-based classifier that was based on the results of the most frequent words unit. This rule-based classifier simply checked if the documents contained the string ‘quake’ and a relevant location (‘zealand’, ‘kaikoura’, or ‘christchurch’). We also looked at which domains contributed to a large number of documents, and manually checked the quality of these documents. Some of these domains produced only noise, so we filtered them out. This left us with around 3,500 relevant documents out of the 10,000 original documents in the corpus.

3.2 Task 1: Most Frequent/Important Words

3.2.1 Approach, Design and Implementation

Our first task was to create a table of the most frequent and important words. To do this we elected to look through our entire dataset and use Python functions to count each word. To do this we had to tokenize each word. This process took a lot of time and was very difficult to run. We employed the use of PySpark to create RDD variable types which allowed us to process this information quicker. This approach gave us our final count but the results did not provide accurate information. The skewed data was caused by NLTK picking up pieces of text that were not words. This ranged from individual letters to pieces of HTML, Javascript, and punctuation. To remedy this, we further cleaned our dataset and added our own stopwords to the tokenizer from NLTK. Switching from JustText to OpenNLP also helped. Our next step was to start to instead focus on word importance, instead of just frequency. Our next step was to use TF-IDF to determine important words. After we successfully ran TF-IDF on the dataset we were left with a more accurate depiction of what the most important words in the dataset were.

Our initial code was written in the local Zeppelin Environment. This allowed us to tweak our code and process the small dataset. This benefitted us because we were able to experiment with separate paragraphs in the notebook to see the results of each step in the process. We moved into the Hadoop cluster once it was time to run our code on the large dataset. Our results are in the table below:

New	town	would	ground	Key
earthquake	Monday	19	areas	new
Zealand	hit	Sep	Google	road
quake	2016	Twitter	country	Read
said	Share	city	reports	hours
Kaikoura	buildings	14	3	large
people	read	reported	use	news
Christchurch	around	may	Zealand's	many
Wellington	2011	north	7.8	Image
1	minutes	first	region	km

Island	earthquakes	caused	local	capital
South	near	says	Defence	small
struck	time	November	along	fault
magnitude	could	News	powerful	Earthquake
damage	area	felt	evacuated	quakes
one	water	killed	NEW	still
tsunami	miles	roads	years	John
two	aftershocks	North	like	Canterbury
also	Facebook	residents	morning	major
2018	coast	2	ZEALAND	away

Table 3.1 Sample of Most Frequent / Important words

3.2.2 Testing, Evaluation and Assessment

We determined the accuracy of this section manually. We judged each number and word in this set based off of information read online about our events. We used Google to search for the name of our event along with some of the words we did not recognize, such as “Christchurch”. This helped us to understand the event better and to determine if certain places were relevant. We concluded that this section yielded results that were good enough to allow us to move on to the next task.

3.3 Task 2: A set of WordNet Synsets

3.3.1 Approach, Design and Implementation

For task 2 we utilized the NLTK library and the Hadoop cluster. The process of finding these words was very simple. We took the results of the TF-IDF most important words and passed them into a NLTK function which returns a set of synonyms (synsets), using an internal database of English synonyms. This gave us a set of synsets for our most important words.

A sample of the final results can be seen below. The table is read top to bottom, and then left to right. The synsets are grouped by text color corresponding to each input word. For example, if throw, toss, peg, and chuck were all one sysnet, they would be uniform in text color:

Synset('new.a.01')	Synset('christchurch.n.01')	Synset('smitten.s.01')
Synset('fresh.s.04')	Synset('wellington.n.01')	Synset('magnitude.n.01')
Synset('raw.s.12')	Synset('wellington.n.02')	Synset('orderofmagnitude.n.02')
Synset('new.s.04')	Synset('hessian_boot.n.01')	Synset('magnitude.n.03')
Synset('new.s.05')	Synset('one.n.01')	Synset('damage.n.01')
Synset('new.a.06')	Synset('one.s.01')]	Synset('damage.n.02')
Synset('newfangled.s.01')	Synset('island.n.01')	Synset('damage.n.03')
Synset('new.s.08')	Synset('island.n.02')	Synset('price.n.02')
Synset('modern.s.05')	Synset('south.n.01')	Synset('wrong.n.02')
Synset('new.s.10')	Synset('confederacy.n.01')	Synset('damage.v.01')
Synset('new.s.11')	Synset('south.n.03')	Synset('damage.v.02')
Synset('newly.r.01')	Synset('south.n.04')	Synset('one.n.01')
Synset('earthquake.n.01')	Synset('south.n.05')	Synset('one.n.02')
Synset('earthquake.n.02')	Synset('south.a.01')	Synset('one.s.01')
Synset('zealand.n.01')	Synset('south.r.01')	Synset('one.s.02')
Synset('earthquake.n.01')	Synset('strike.v.01')	Synset('one.s.03')
Synset('quiver.v.01')	Synset('affect.v.05')	Synset('one.s.04')
Synset('tremor.v.01')]	Synset('hit.v.02')	Synset('one.s.05')
Synset('state.v.01')	Synset('strike.v.04')	Synset('one.s.06')
Synset('allege.v.01')	Synset('strike.v.05')	Synset('matchless.s.01')
Synset('suppose.v.01')	Synset('hit.v.05')	[Synset('tsunami.n.01')
Synset('read.v.02')	Synset('strike.v.07')	Synset('two.n.01')

Synset('order.v.01')	Synset('fall.v.08')	Synset('deuce.n.04')
Synset('pronounce.v.01')	Synset('come_to.v.03')	Synset('two.s.01')
Synset('say.v.07')	Synset('strike.v.10')	Synset('besides.r.02')
Synset('say.v.08')	Synset('strike.v.11')	Synset('town.n.01')
Synset('say.v.09')	Synset('fall_upon.v.01')	Synset('town.n.02')
Synset('say.v.10')	Synset('strike.v.13')	Synset('township.n.01')
Synset('say.v.11')	Synset('strike.v.14')	Synset('town.n.04')
Synset('aforesaid.s.01')	Synset('hit.v.09')	Synset('monday.n.01')
Synset('people.n.01')	Synset('hit.v.12')	Synset('hit.n.01')
Synset('citizenry.n.01')	Synset('assume.v.05')	Synset('hit.n.02')
Synset('people.n.03')	Synset('mint.v.01')	Synset('hit.n.03')
Synset('multitude.n.03')	Synset('strickle.v.02')	Synset('collision.n.01')
Synset('people.v.01')	Synset('strike.v.20')	Synset('hit.n.05')
Synset('people.v.02')	Synset('strike.v.21')	Synset('hit.n.06')

Table 3.2 Sample of WordNet Synsets

3.3.2 Testing, Evaluation and Assessment

This task was evaluated manually by simply checking each synset and determining if they were accurate. We only evaluated the sets seen above since there were too many to handle by brute force, and assumed they were a representative sample. The results for this task are accurate and encompasses a variety of words.

3.4 Task 3: A set of POS Constrained Words

3.4.1 Approach, Design and Implementation

Tagging the part of speech of each word in the document was a simple task using the NLTK library. Since each important word was already tokenized from task 1 we did not have to tokenize the words. We simply tagged all of the most important words and then separated them by part of speech.

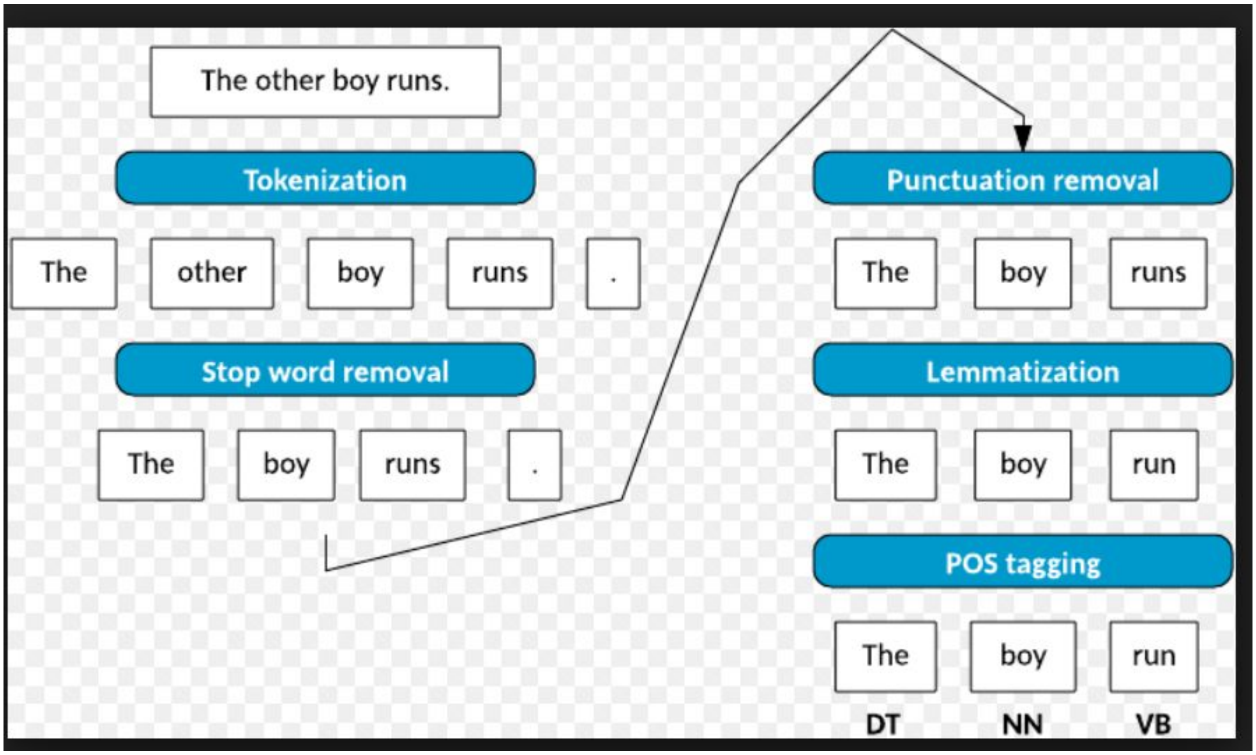


Figure 3.3 POS Tagging Concept Diagram
Source: www.ibm.com **[7]**

We began this task using Zeppelin similarly to how we began task 1. We then moved to Hadoop. The code to retrieve the parts of speech is a few lines following the code for TF-IDF. The results of the most important words by part of speech are as follows:

Nouns

New	time	Google
earthquake	area	country
Zealand	water	reports
quake	miles	use
Kaikoura	aftershocks	Zealand's
people	Facebook	region
Christchurch	coast	Defence
Wellington	Sep	years
Island	Twitter	morning
South	city	ZEALAND
damage	November	Key
tsunami	News	road
town	roads	Read
Monday	North	hours
Share	residents	news
buildings	ground	Image
minutes	areas	

Table 3.4 Sample of Most Frequent / Important nouns

Verbs:

said	shaking	found
struck	email	see
hit	following	expected
read	warning	went
earthquakes	closed	set
reported	told	metres
north	stranded	make
caused	east	seen
says	help	reserved
felt	take	recorded
killed	died	used
evacuated	say	inland
km	triggered	move
according	work	took
including	affected	know
left	collapsed	confirmed
damaged	cut	

Table 3.5 Sample of Most Frequent / Important Verbs

3.4.2 Testing, Evaluation and Assessment

The results for the part of speech tagging are fairly accurate and helpful. The nouns section was most accurate. This section presented us with the cities that we knew were important as well as other important nouns related to earthquakes like “ground” and “damage”. We compared these results with our TF-IDF most important words and determined that this set is a good extension of the TF-IDF set. The only issues with the set is the inclusion of social media words, included in website boilerplate. Although this noise is pervasive, it has not negatively impacted our final summaries.

The results for the verbs section is also good. Unlike the nouns section, this section did not have problems with including words like “Facebook” and “Twitter”. There are a few instances of words that are not verbs like “km” but this is a rare occurrence in our set. The majority of these words are important for a story about an earthquake. We decided that these results were sufficient and moved on to the next task.

3.5 Task 4: A set of Words and Stems that are Discriminating Features

3.5.1 Approach, Design and Implementation

To get the set of words and stems that are discriminating features in our dataset we tried to run lemmatization and stemming on the list of most important frequent words that we obtained from task 1. The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. We used the WordNetLemmatizer from the NLTK stem Wordnet library to apply lemmatization and stemming on the words.

The results we obtained from this task helped us to build a rule-based classifier to find the relevant web pages that are relevant to our topic and remove irrelevant and noisy webpages. Below are the top 25 words that resulted.

new	zealand	earthquake	kaikoura	people
read	share	sep	wellington	christchurch
news	south	island	2018	image
damage	minutes	tsunami	facebook	twitter
magnitude	water	monday	minute	caption

Table 3.6 Sample of Discriminating Features

3.5.2 Testing, Evaluation and Assessment

This set of words helped us to clean our data. When we used this set to clean the data it worked well. This indicates that the set was accurate. The questionable parts of this set are the words that have to do with social media like “facebook” and “twitter”. While many articles that we ended up using include these words they do not necessarily help us to clean our data. For that reason the set is mostly accurate, but it does have a few issues such as the social media words.

3.6 Task 5: A set of Important Named Entities

3.6.1 Approach, Design and Implementation

Tagging the named entities in the document was a simple task using the NLTK library. Our first approach was similar to how we handled the part of speech tagging. Since each important word was already tokenized from task 1 we did not tokenize the words. We simply tagged all of the most important words and then printed out only the important words that had been tagged. This approach was very inaccurate. The reason for the inaccuracy is described in the evaluation section of this task.

This task was not done with Zeppelin. We started this task on Hadoop by copying and pasting the script from the original TF-IDF most important words file and adding a small section at the end to handle the named entity recognition. Table 3.7 is our final named entity recognition result figure.

('GPE', u'New')	('ORGANIZATION', u'NEWS')
('PERSON', u'Zealand')	('ORGANIZATION', u'SHARE')
('PERSON', u'Kaikoura South')	('ORGANIZATION', u'USGS')
('ORGANIZATION', u'Christchurch')	('PERSON', u'Tsunami')
('PERSON', u'Wellington')	('PERSON', u'March')
('PERSON', u'Island Facebook')	('ORGANIZATION', u'GeoNet')
('PERSON', u'Twitter')	('GPE', u'Australian')
('GPE', u'North')	('PERSON', u'Survey Government')
('PERSON', u'Google')	('GPE', u'Search')
('ORGANIZATION', u'ZEALAND')	('GPE', u'U.S.')
('PERSON', u'John')	('GPE', u'Auckland')
('PERSON', u'Pacific')	('PERSON', u'Islands')
('ORGANIZATION', u'Canterbury')	('PERSON', u'Herald Andrew')
('PERSON', u'Retrieved')	('ORGANIZATION', u'University')
('GSP', u'US')	('ORGANIZATION', u'HMNZS')
('ORGANIZATION', u'NZ')	('ORGANIZATION', u'Rights')
('PERSON', u'Email')	('PERSON', u'Health National')
('PERSON', u'Emergency')	('PERSON', u'Cheviot')
('PERSON', u'People World')	('GPE', u'United')
('PERSON', u'Please Topics')	('PERSON', u'Aftershocks')
('ORGANIZATION', u'Local')	

Table 3.7 Sample of Most Frequent / Important Named Entities

3.6.2 Testing, Evaluation and Assessment

This task resulted in an almost completely wrong set of tagged entities. Our group categorized the top 40 named entities. To test the results, we compared our set to sets from other groups. There were many instances where the other groups also had very inaccurate data. We all came to the conclusion that the NLTK NER (named entity recognition) was out of date and that Spacy should instead be used. Unfortunately, Spacy was incompatible with the Hadoop cluster, and due to time constraints we could not explore its viability for this task.

As we were evaluating our results, we realized that a reason for the poor performance is because we were tagging the words directly from the TF-IDF list. The NER needs to tag words in their original contexts for best results. Since we did not run the NER on full sentences, the NER treated our list of most important words as one large sentence. We did not use NER for our template-based summary, but instead used REGEX, since this task did not give good results. For that reason, we moved on from this section without ideal results.

3.7. Task 6: A set of Important Topics

3.7.1 Approach, Design and Implementation

Latent Dirichlet allocation (LDA) was used to get the most important topics from a set of documents. LDA tries to cluster sets of important topics and documents depending on their textual content. We used LDAModel from the PySpark clustering library to apply LDA on the dataset. We observed that LDA did not give consistent results for each run. The results changed every run, because LDA is a stochastic process. After further research, we learned about LSA (Latent Semantic Analysis), a semi-supervised learning version of LDA. We were able to get consistent and more accurate results by applying this method instead. We used Gensim’s LSA library to run LSA on our dataset. Below are the top 10 results from running LSA on our dataset..

- (1, u'0.504*"volcano" + 0.222*"nyamuragira" + 0.183*"volcanoes" + 0.153*"kermadec" + 0.147*"volcanodiscovery"')
- (2, u'0.213*"est" + -0.212*";" + -0.195*"1931-02-02" + -0.195*"176.9" + -0.195*"-39.5"')
- (3, u'0.300*"est" + 0.155*"id" + 0.127*"kilometers" + 0.086*")" + -0.085*"school"')
- (4, u'-0.237*"cows" + 0.179*"she" + 0.158*"her" + -0.151*"cattle" + -0.145*"grass"')
- (5, u'0.131*"cattle" + 0.126*"cows" + 0.116*"grass" + 0.110*"she" + 0.108*"million"')
- (6, u'0.118*"cattle" + -0.117*"kaikoura" + 0.100*"million" + -0.097*"hide" + 0.095*"son"')
- (7, u'0.203*"cattle" + 0.164*"million" + -0.131*"abc" + 0.108*"outnumber" + 0.105*"ravines"')
- (8, u'0.118*"son" + 0.116*"usar" + 0.112*"school" + 0.107*"taitapanui" + 0.097*"rescue"')
- (9, u'-0.258*"hide" + -0.205*"shakes" + 0.192*"cows" + -0.153*"caption" + -0.148*"photos"')
- (10, u'-0.128*"photo" + 0.118*"spared" + 0.117*"largely" + 0.106*"cracked" + -0.102*"kilometres"')

Table 3.8 Top-10 LSA Result

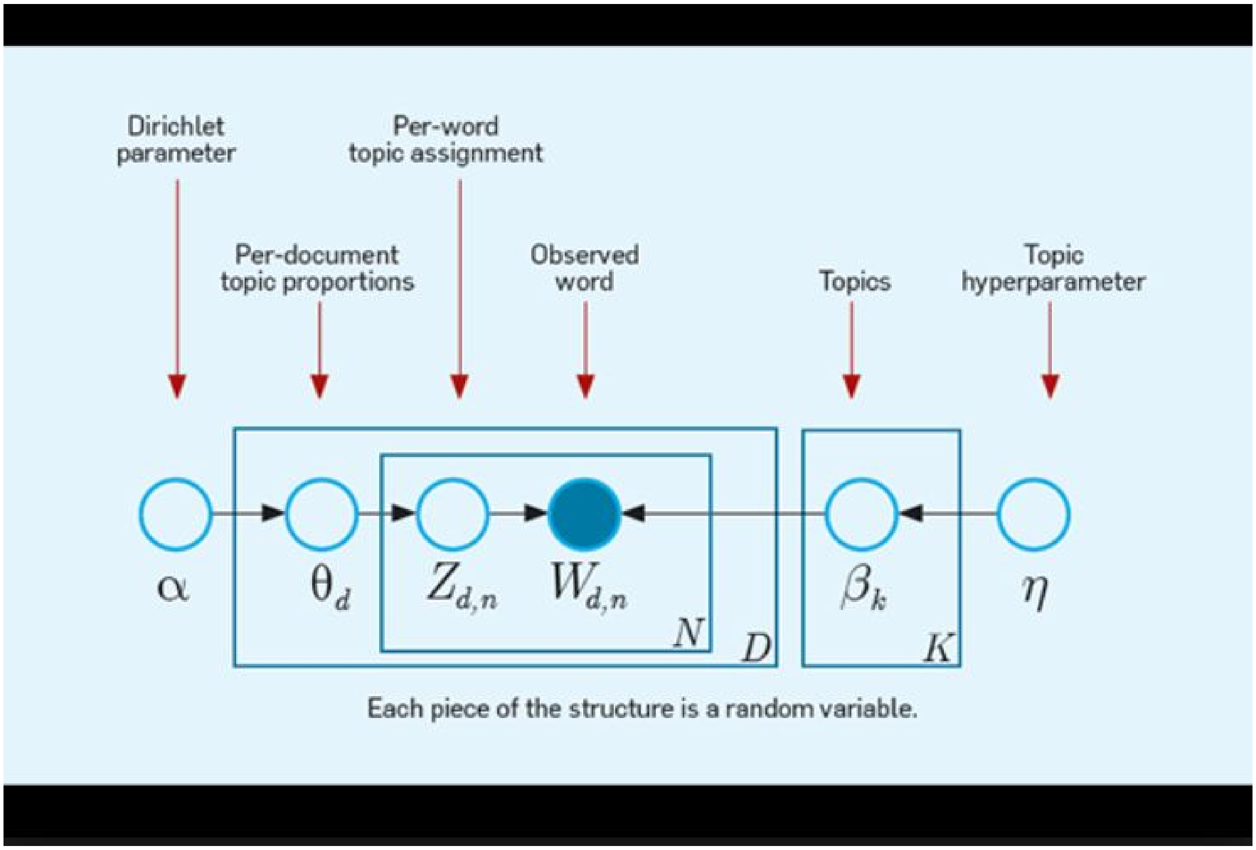

Dirichlet parameter
Per-word topic assignment
Per-document topic proportions
Observed word
Topics
Topic hyperparameter
α
θ_d
Z_{d,n}
W_{d,n}
N
D
β_k
K
η
Each piece of the structure is a random variable.

Figure 3.9 LSA Architecture
Source:.mypatentideas **[8]**

3.7.2 Testing, Evaluation and Assessment

The most important topics from Figure 3.9 include “volcano”, magnitudes of the earthquake, and names of the areas that were affected by the earthquake, like Kaikoura.

With LSA, we were able to learn that a volcano was also a part of the story. The earthquakes were the main focus of the articles but the volcanic eruption was also included in many of the documents.

We know that animals were also affected in the incidents due to further research on the subject. Cattle were lost because they were scared by the earthquake, and fled. We did not think that this part was very important because we only found a few documents that mentioned cattle and other animals.

Determining important topics with LDA and LSA helped us to create the template for task 9.

3.8. Task 7: Extractive Summary

3.8.1 Approach, Design and Implementation

We first implemented extractive summarization in a way that generated a summary for each article. This implementation utilized several steps of processing. First, we made each article its own dataframe. Second, we vectorized our dataset (utilizing the word2vec vectorization model). Lastly, we performed K-means clustering on the vectorized dataset from the previous step.

Our first step was to test the creation of dataframes on Zeppelin. However, we soon realized that some of the data contained empty content, which would cause issues with the clustering. At this point, our dataset was not completely clean, but we were not sure if we should postpone Task 7 until the cleaning was done. We decided to bring a temporary and impromptu solution of excluding all data frames of length equal to zero, in order to eliminate all the empty content article data. After this process we reorganized the document identification numbers, so that skipping in document identification number would not occur. The temporary solution was eventually reverted after the dataset was cleaned in a better way.

The next step was to vectorize the dataframe using word2vec vectorization model. This process was also tested in Zeppelin. We were eventually able to process this step without a single error, with our initial trial. The data frame was vectorized into 100 different vectors. We decided to convert each articles into 100 different vectors because of two reasons: according to study [10], most news articles are 500 to 800 words long; furthermore, our team agreed that usually it takes 5 to 8 words in English to convey a meaningful intention, regardless of if the words are in complete sentence. For that reason, we decided that for each article, there would be 100 meaningful parts of the article that we could convert into vectors and use in sentiment analysis.

After vectorization of the dataframe, we performed K-means clustering on our vectorized data. Clustering was done to generate 14 different clusters. The number of clusters were evaluated manually from 10 clusters to 20 to ensure that there are enough but not too many clusters to manually evaluate. When clustered in 14 different clusters, the all clustering result showed meaningful results. After fixing several syntax errors, we were able to get the centroid of each cluster, and the cluster labeling of each data point. Below is a sample of the results:

```
cluster center: 0 [-0.33174712  0.16968764  0.04645146  0.00244272  0.17423478  0.20103509
 0.08701435 -0.13080927 -0.01344027  0.22148678  0.0267731   0.17126637
 0.34163354  0.41269081 -0.16585926  0.55762091 -0.170854    0.34544742
 -0.24331773 -0.18159574 ...]
```

```
cluster center: 1 [-0.24931036  0.10121604  0.26656181  0.09524505  0.09476678  0.13907774
 -0.08238463  0.05369252 -0.05076714 -0.04704999  0.00883639 -0.02974392
 0.0042771   0.18168466 -0.24237925  0.22924463 -0.03366167  0.05963572
 -0.04540208 -0.17032995 ...]
```

```
cluster center: 2 [-0.47721872  0.193357   -0.20449835 -0.02787738  0.3004131   0.0013685
 -0.48886219 -0.32749994  0.48163639  0.08443132  0.20456542  0.16196493
 0.16338574 -0.21483108 -0.07351796 -0.06848809 -0.07894303 -0.01732902
 0.3554174   0.0884236  ... ]
```

```
cluster center: 3 [-0.14903758  0.23009507 -0.02384007 -0.18501541  0.04241503 -0.01847301
 -0.01589317 -0.13695922 -0.21778593  0.02634022  0.00757874  0.01757698
 0.08642961 -0.12720847 -0.04214456  0.0157517   0.02875143 -0.14802479
 0.09332923  0.16654796 ...]
```

```
cluster center: 4 [-0.21774166  0.17642005  0.02879819 -0.12483088  0.11046432  0.024135
 -0.08847094 -0.16666011 -0.27051458 -0.00337241  0.07640682 -0.00885124
 0.08786321 -0.08589822 -0.05491974  0.10186201  0.0598614  -0.06094049
 0.10924481  0.20247147 ...]
```

```
cluster center: 5 [-0.26556359 -0.11454757  0.49909203  0.30217224 -0.21129491  0.16741061
 -0.0125498  -0.39489343  0.13385775  0.27243051  0.14625433 -0.56190334
 -0.15261192  0.08937157  0.01218023  0.1220165   0.1893623  -0.48109283
 0.48827487 -0.40765118 ...]
```

```
cluster center: 6 [-0.161914    0.168821    0.03466178 -0.11787472 -0.03336642 -0.05471725
 -0.00181498 -0.0678705  -0.06950483  0.01349893 -0.04389722  0.0180956
 0.04043251 -0.06420383 -0.08237478 -0.03702432 -0.01788802 -0.16959141
 0.07600028  0.06235147 ...]
```

```
cluster center: 7 [-0.18436872  0.09367294  0.17108481 -0.00760946 -0.04704215 -0.00026561
 -0.0421621  -0.08381818  0.03456628  0.08297118 -0.01240274 -0.001102
 -0.04802505  0.03839754 -0.11702575 -0.04630364  0.00888154 -0.14755599
 0.08776715 -0.03605982 ...]
```

```
cluster center: 8 [-0.15963501  0.08939594 -0.26205246 -0.09068812 -0.07885027  0.13727086
 0.17055028 -0.02870683 -0.00506692  0.1155675   0.03801569 -0.10284286
 -0.23084999  0.36762295 -0.27020185  0.26466081 -0.25992597 -0.29280077
 -0.01831414 -0.31079328 ...]
```

```
cluster center: 9 [-2.04148467e-01  1.01824925e-01  7.04155686e-02 -6.41963653e-02
 6.60137993e-02 -1.95071975e-04 -8.73483235e-02 -1.29859602e-01
 -1.44177709e-01  7.39477637e-02 -1.84589261e-03 -1.09912879e-03
 8.05271042e-02 -3.71206639e-02 -8.99079624e-02 -3.48988258e-02
 -1.78513644e-02 -1.28647416e-01  8.18929022e-02  7.92211959e-02 ...]
```

```
cluster center: 10 [-0.93928464 -0.18821755 -0.82569991 -0.1552187   0.35770851  0.34518262
 -0.18235469 -0.28182655 -0.00679889  0.22031128 -0.68397829 -0.08712118
 -0.0294348   0.76481655  0.51694861 -0.07313353  0.61560249 -1.11792553
 0.19220868  0.16870537 ...]
```

```
cluster center: 11 [-0.17919838  0.07949853  0.20193391  0.40160835 -0.14501242  0.04385389
0.01730105 -0.28750777  0.00763054 -0.00574381  0.2436829   0.02424969
-0.16265562  0.18534832 -0.59574148 -0.03197009  0.05268378 -0.07036813
0.08325853  0.21466945 ...]

cluster center: 12 [-0.32993109  0.09940839 -0.02053637 -0.20773166 -0.11149118 -0.09761358
0.02360816  0.04216427 -0.02151035 -0.20981508  0.05188498  0.20656385
-0.14359228 -0.02069697 -0.07156611  0.00494207 -0.10737145 -0.25253983
0.05417769 -0.04491447 ...]

cluster center: 13 [-0.28667035  0.1800112   0.3140906   0.24198655 -0.17246435  0.05435806
-0.24618434 -0.48830457  0.10170453  0.09391689 -0.04452073 -0.2453267
-0.1908772   0.07105022  0.19731812 -0.01701054  0.00455986 -0.23196473
0.20533645 -0.02299293 ...]
```

Table 3.10 Cluster Center Result

```
doc_id: 0  cluster_label: 3
doc_id: 1  cluster_label: 7
doc_id: 2  cluster_label: 5
doc_id: 3  cluster_label: 5
doc_id: 4  cluster_label: 6
doc_id: 5  cluster_label: 12
doc_id: 6  cluster_label: 10
doc_id: 7  cluster_label: 5
doc_id: 8  cluster_label: 4
doc_id: 9  cluster_label: 3
doc_id: 10  cluster_label: 6
doc_id: 11  cluster_label: 3
doc_id: 12  cluster_label: 4
doc_id: 13  cluster_label: 3
doc_id: 14  cluster_label: 9
……
…...
```

Table 3.11 Sample (First 14) of Data Point Cluster Label

The representative article for each cluster was chosen by calculating the articles closest to each cluster centroid, using Euclidean distance. Clustering was performed with 14 different clusters, but only 10 were chosen for final result, since the 4 were either not relevant to the topic, or they were noise from the data set. The following are samples of the representative articles:

At least two people have been killed and thousands of landslides have occurred across the country, cutting off some towns from the rest of the country. Emergency response teams flew by helicopter to the region at the epicentre of the tremor, 57 miles northeast of Christchurch. New Zealand Prime Minister John Key told Civil Defence Minister Gerry Brownlee, according to Brownlee's Twitter account: It's just utter devastation, I just don't know... that's months of work. In the midst of the devastation, however, a few fortuitous cows, left stranded on a patch of earth, were spotted surrounded by disrupted earth and landslides. The three animals were noticed by a helicopter flying near Kaikoura, the footage from which shows how astonishingly lucky they were to survive. …...
The quake hit 196 km (122 miles) northwest of Auckland Island, at a depth of 10 km (6.2 miles). account. Create a free website or blog at WordPress.com. Privacy & Cookies: This site uses cookies. By continuing to use this website, you agree to their use.
A tsunami warning that led to mass evacuations after the original quake was downgraded after large swells hit New Zealand's capital Wellington, in the North Island, and Christchurch. Wellington was a virtual ghost town with workers ordered to stay away while the city council assessed the risk to buildings, several of which were damaged by the tremor. Severe weather with 140 km per hour (85 mph), gale-force winds was forecast for the area. Hundreds of aftershocks, the strongest a 6.2 quake at about 1.45pm local time (0045 GMT), rattled the South Pacific country …...
HUGE 5.5 magnitude earthquake has hit central New Zealand just weeks after the country was rocked by two massive quakes. GEONET The quake hit central New Zealand at around 3.30pm local time More than 7,000 people claim the earth beneath them shook as the quake hit the town of Kaikoura, the same area that was hit by a 7.8 magnitude earthquake last month. Local people have reported feeling aftershocks following the main earthquake, which struck at around 3.30pm local time (2.30am GMT). TWITTER New Zealanders took to social media after feeling the tremors The earthquake was felt in major cities including the capital Wellington and even Christchurch, almost 500km away.
Nov 21, 2016 by : Camila Domonoske Just over a week ago, a magnitude 7.8 earthquake shook New Zealand. Now, scientists investigating the damage have recorded mesmerizing footage of massive cracks that opened up in the earth. The videos were posted by GNS Science. One shows the Kekerengu Fault rupture, while the other reveals the Papatea Fault. Both faults lie near the coast in the northeastern corner of New Zealand's South Island.
The earthquake - the epicentre of which is close to Kaikoura, a small town near Christchurch - resulted in one man dying after his house collapsed in the quake. And McCaw, a commercial helicopter pilot now following his retirement from rugby, was one of the helicopter pilots who helped fly Urban Search and Rescue teams out to the house (as well as other places), according to the 'New Zealand Herald'. These latest actions from McCaw will only further endear him to the people of the country amidst the destructive chaos of the earthquake.
Therefore, they show the seismicity from 1964 until the earthquake occurrence. Regional deadly earthquakes from 1500 to 2000 Date Long.
Therefore, they show the seismicity from 1964 until the earthquake occurrence.
It is potentially the largest source of earthquake and tsunami hazards in New Zealand, but scientists say there is still much to learn about it.
New Zealand earthquakes caught on video 00:42 Story highlights Two people have died in series of earthquakes Small tsunami hits country's eastern coast, although warning has since been lifted (CNN)A series of powerful earthquakes jolted New Zealand's South Island Monday, triggering a tsunami and sending aftershocks across the country that left at least two dead, officials said. The first event, a 7.8-magnitude quake, struck just after midnight Monday near the coastal community of Kaikoura, some 93 kilometers (55 miles) northeast of the city of Christchurch, the US Geological Survey reported. …...

Table 3.12 Representative Articles from Clusters

3.8.2 Testing, Evaluation and Assessment

The results from the extractive summary are not great. The method gave us a readable summary that is full of the identified important words from TF-IDF. Upon manual assessment of the summary, we determined that a large problem is the end of many articles include noise like information about cookies or privacy.

To determine the accuracy of our summaries we examined 3 different ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics for comparing the extractive summary to the gold standard written for our dataset. ROUGE-1 examines the proportion of overlapping words (1-grams) between both summaries. ROUGE-2 similarly looks at overlapping 2-grams. ROUGE-SU4 looks at similar 4-grams with gaps.

The ROGUE evaluation score of the extractive summaries were not very good, and were the lowest metrics of all the summaries.

ROUGE-SU4	ROUGE-2	ROUGE-1
.01	0.0	.05556

Table 3.13 Rogue Score for Extractive Summary

3.9. Task 8: Set of Values for each Semantic Slot

3.9.1 Approach, Design and Implementation

In order to define which set of semantic slots are needed for our readable summary (which is in fact, task 9), we first had to define the template of a readable summary. Before we started to investigate how to obtain values for each slots, the format of the readable summary template was the first step. Since this is more closely related to task 9, the details will be discussed in the section for task 9.

After we outlined the readable summary template, we defined the set of semantic values needed as follows: magnitude, death count, injury count, location, epicenter, date, and time.

Meaningful progress was made for this task only after the mixed-discussion during class session, where members of different teams gathered to discuss about their approaches for each tasks. There, we were able to learn that using regex searching is a possible solution to successfully search and point out the values for each semantic slots. Also, while utilizing regex searches, we learned that the search should be done on n-grams of the data we have, since the context significantly matters on semantic searching. For the purpose of this project, we decided to use bi-grams, as the method seemed sufficient enough for the regex searching we are performing.

The first step was to create bi-grams of the data we had in PySpark's dataframe. The Pyspark library has a feature where it turns string data into a string array of bi-grams. The initial plan was to convert our dataframe of articles into a dataframe of bi-grams, but since PySpark's library transformed the articles (which are in string) into bi-grams (which are in string array), we had to change our plan. From here on, we decided to utilize the dataframes of string array that contains the bi-gram of each articles for task 8.

Table 3.14 is the sample bi-gram output of one of the articles:

Original	**Bi-gram**
[The, earthquake, hit, at, 4.40am, on, Friday,, about, 80, miles, (130km), north-east, of, the, East, Cape, community, ……]	[The earthquake, earthquake hit, hit at, at 4.40am, 4.40am on, on Friday,, Friday, about, about 80, 80 miles, miles (130km), (130km) north-east, north-east of, of the, the East, East Cape, Cape community, …...]
[It, was, at, a, depth, of, 13km,, 35km, north, of, Wairoa,, and, hit, at, 8.21pm., …...]	[It was, was at, at a, a depth, depth of, of 13km,, 13km, 35km, 35km north, north of, of Wairoa,, Wairoa, and, and hit, hit at, at 8.21pm., …...]
[Therefore,, they, show, the, seismicity, from, 1964, until, the, earthquake, occurrence.]	[Therefore, they, they show, show the, the seismicity, seismicity from, from 1964, 1964 until, until the, the earthquake, earthquake occurrence.]

Table 3.14 Sample of Bi-gram From the Articles

We decided to perform regex search on each article separately and obtain the set of semantic values for each article. In order produce the values for each articles, we first combined the bi-gram array into a one string value, joined by space character. Next, we mapped the bi-gram dataframe into another dataframe with columns for each semantic value by feeding the bi-gram string into separately defined regex search functions, for each semantic value. The regex search functions searched for the most frequent values that meet the criteria for each semantic value within the article. If there were no matching search results, number values were set to be 0, and string values were set to be the empty string.

There was a special condition for epicenter regex search – only the articles that mentioned the word 'epicenter' were used for epicenter regex search – since 'epicenter' has exactly the same characteristics as 'location'.

Below are the regex formats we used for regex search functions:

<table>
<tr><th>*Slot*</th><th>*Regex*</th></tr>
<tr><td>Magnitudes</td><td>'magnitude (?:of)?[0-9]\.?[0-9]*'
'[0-9]\.?[0-9]* magnitude'
'magnitude-[0-9]\.?[0-9]*'
'[0-9]\.?[0-9]*-magnitude'</td></tr>
<tr><td>Deaths</td><td>'(?:[0-9]+ deaths)'
'(?:[0-9]+ were killed)'
'killed (?:\w+){0,2}[0-9]+ people'
'(?:[0-9]+ casualties)'</td></tr>
<tr><td>Injuries</td><td>'[0-9]+ injur(?:(?:ies)|(?:ed))'
'(?:[0-9]+ were injured)'
'injured (?:\w+){0,2}[0-9]+ people'
'(?:[0-9]+ injuries)'</td></tr>
<tr><td>Location</td><td>"((in|at)\s([A-Z][a-zA-Z]{4,}|[A-Z][a-zA-Z]{2,}\s[A-Z][a-zA-Z]{3,})|\s+[A-Z][a-zA-Z]{3,},\s[A-Z][a-zA-Z]{2,}\s[A-Z][a-zA-Z]{3,})"</td></tr>
<tr><td>Epicenter</td><td>"((in|at)\s([A-Z][a-zA-Z]{4,}|[A-Z][a-zA-Z]{2,}\s[A-Z][a-zA-Z]{3,})|\s+[A-Z][a-zA-Z]{3,},\s[A-Z][a-zA-Z]{2,}\s[A-Z][a-zA-Z]{3,})"

% Only performed on articles that mentions the word “Epicenter”</td></tr>
<tr><td>Date</td><td>*Year:*
'(?:1|2)[0-9]{3}'

Month:
"(?:January|February|March|April|May|June|July|August|September|October|November|December)"

Day:
'(?:January|February|March|April|May|June|July|August|September|October|November|December),?\s([0-9]{,2})[^0-9]'</td></tr>
<tr><td>Time</td><td>*Time:*
'[0-9]{1,2}:[0-9][0-9](?::[0-9][0-9])?(?:\s?[apAP]\.?[mM]\.?)?'</td></tr>
</table>

Table 3.15 Regex Format Used for Regex Search

Finally, we counted the semantic values for the entire dataframe (all the articles), and ranked them by frequency. Empty results were excluded manually.

<table>
<tr><td><b><i>Magnitudes / Count</i></b></td><td>|7.5 | 395|
7.8	340
7.1	174
6.3	55
6.4	36|</td></tr>
<tr><td><b><i>Deaths / Count</i></b></td><td>|185.0	761
256.0	14|</td></tr>
<tr><td><b><i>Injuries / Count</i></b></td><td>|20.0	18 |</td></tr>
<tr><td><b><i>Location / Count</i></b></td><td>|Wellington	577
Kaikoura	306
Christchurch	177
September	130
Share	80
Sorry	70
February	65
Darwin	59 |</td></tr>
<tr><td><b><i>Epicenter / Count</i></b></td><td>|Wellington	204
Christchurch	51
Kaikoura	31
SYDNEY	17 |</td></tr>
<tr><td><b><i>Date / Count</i></b></td><td>|[,November,2016]	516
[,September,2011]	121
[,February,2011]	97
[,November,2011]	89
[,September,2018]	54 |</td></tr>
<tr><td><b><i>Time / Count</i></b></td><td>|11:02	111
00:19	61
1:50	57
4:52	52
12:02	48
5:55	45 |</td></tr>
</table>

Table 3.16 Regex Search Result in Rank

3.9.2 Testing, Evaluation and Assessment

Overall, the result was quite satisfying, as we manually inspected the outputs for each semantic slot. However, the result wasn't good for immediate use, and needed some human interpretation for actual use. For example, the most frequent 'date' was a 'date' without the number for 'day', so we had to investigate what the actual date would be for the event.

Furthermore, while composing this solution, we realized that if we were able to perform regex search for each sentence, not article, the accuracy of ‘epicenter’ search result would increase significantly. Regex search for ‘epicenter’ on every article that mentioned the word “epicenter” inevitably increased the noise in the ‘epicenter’ search result.

Also, regarding the magnitude of the earthquake, we realized that many of the articles say that the magnitude of the earthquake is 7.5. However, the actually magnitude was 7.8, which is ranked the second in the above chart. This is because the news articles corrected themselves with more accurate source of information, over time.

3.10. Task 9: Readable Summary with Slots and Values

3.10.1 Approach, Design and Implementation

Defining the template of the readable summary, which is part of task 9, was indeed processed before task 8.

In order to formulate a template that would summarize our topic (earthquake event), we observed several news articles that reported earthquake events. Further, on October 11, by Michael Horning, we were introduced to a professional document which we were able to reference with regards to earthquake reporting format. Based on the knowledge from the book *Melvin Mencher’s News Reporting and Writing*, we were able to produce the following template:

```
            # 0 = day         4 = time          7 = epicenter      10 = aftershock
            # 1 = month       5 = magnitude     8 = death          11 = tsunami
            # 2 = year        6 = location      9 = injury         12 = landslide

            summary = 'On {0} {1}, {2} at {4}, a earthquake with magnitude of {5} struck {6}.\
            The epicenter of the earthquake was located at {7}. The earthquake caused \
            number of casualties, reported to be {8} deaths and {9} injuries. There \
            {10} aftershocks reported, and there {11} tsunami caused by the \
            earthquake reported.  Reports mentioned there {12} landslides caused by \
            this earthquake.'.format(values['day'], values[month], values[year], values[time],
values[magnitude], values[location], values[epicenter], values[death], values[injury],
values[aftershock], values[tsunami], values[landslide])
```

Table 3.17 Readable Summary Template

The details of the landslide, aftershock, and tsunami was determined based on very simple criteria: if each word showed up in the data articles we had, for more than 30% of the articles, we considered that the specific event happened.

Transition from task 8 to task 9 would be very simple if the results of task 8 are returned in simple Python (or any other modern programming language) values. However, we faced a challenge in feeding each value into the template summary, since the results from task 8 are organized within a PySpark dataframe. For now, we are manually feeding the top results for each slot to the readable summary report template which we created.

> On {November , 2016} at {11:02}, a earthquake with magnitude of {7.5} struck {Wellington}. The epicenter of the earthquake was located at {Wellington}. The earthquake caused number of casualties, reported to be {185} deaths and {20} injuries. There {were} aftershocks reported, and there {was} tsunami caused by the earthquake reported. Reports mentioned there {were} landslides caused by this earthquake.

Table 3.18 Readable Summary with Slots and Values

3.10.2 Testing, Evaluation and Assessment

The result of the readable summary with slots and values was examined manually. It shows a very satisfying result, as it shows very closely correct information regarding the earthquake event which we are summarizing.

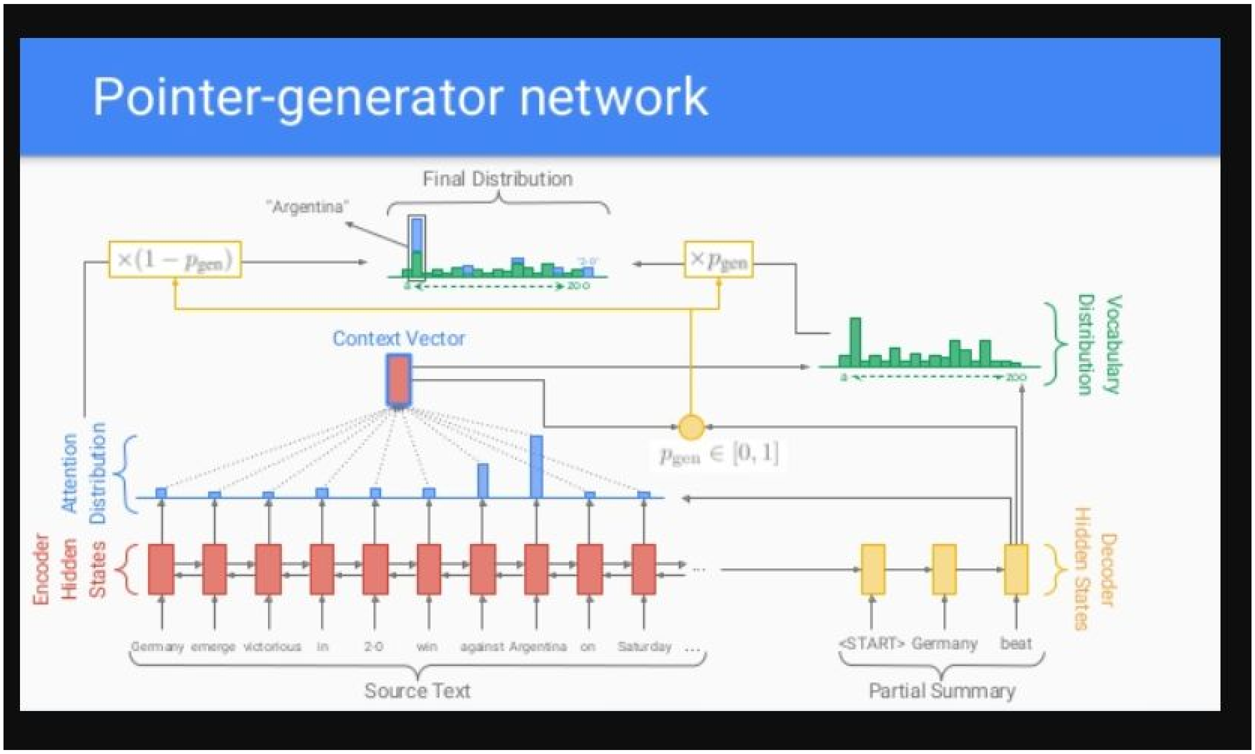
ROUGE-SU4	ROUGE-2	ROUGE-1
.035	.02857	.16667

Table 3.19 Rogue Score for Template-based Summary

3.11. Task 10: Readable Abstractive Summary with Deep Learning

3.11.1 Approach, Design and Implementation

After researching the topic and receiving suggestions from fellow students, we decided to use the TensorFlow pointer generator model to generate an abstractive summary for our dataset. The code repository for the pointer generator network also included a pre-trained model, that was trained on a dataset of CNN and DailyMail news articles. We first had to encode our dataset in a format that would match the pre-trained model. This was accomplished by using code from another repository, linked to by the first one. We then ran the encoded dataset with the pre-trained pointer generator model. The dataset contained about 3000 text files, so we used the Cascades cluster to speed up the process. After this step, we received 3000 new files, each being a summary of an input file from our dataset. Then we used K-means clustering to form 15 clusters of the summaries, based on similarity. We then took the closest file to the centroid of each cluster and appended these summaries (15 in total) into one larger summary. The abstractive summary is found below figure 3.20.



Pointer-generator network

Figure 3.20 Pointer Generator Model Architecture
Source: www.slideshare.net**[9]**

the earthquake hit at 4.40 am on friday , about 80 miles -lrb- 130 km -rrb- north-east of the east cape community of te araroa .
the 7.5-magnitude earthquake in the south island killed at least two people and triggered a tsunami alert for the entire east coast . seven aftershocks have been registered .
civil defence evacuated low-lying coastal areas after a tsunami warning was issued.
mountjoy suspects that the unique nature of the earthquake is behind the eerie exposure of the sea bed in the coastal town of kaikoura .
the effects can be seen all over kaikoura , with some areas rising a metre above the ground .
earthquake geologist nicola litchfield from gns science told michael daly from stuff that the uplift would have happened during the 90 seconds to 2 minutes the quake .

a state of emergency was declared in the town of kaikoura , a whale-watching destination and home to 2,000 people that has been completely cut off .

quake hit just after midnight -lrb- 11:02 gmt on sunday -rrb- , some 95 km -lrb- 59 miles -rrb- from christchurch .
officials said the first waves may not be the largest , with tsunami activity possible for several hours .

aap by jerico mandybur2016-11-14 utc after earthquakes with a magnitude of up to 7.5 struck new zealand 's upper south island just after midnight on monday .
two people were confirmed dead and extensive damage was done to buildings and roads .
with more than 250 aftershocks recorded and tsunami warnings sending people to higher ground , social media was flooded with videos and images as people captured the damage done to their businesses and properties .

four air force helicopters have been airlifting people out of kaikoura on the south island after battling strong winds and heavy rain earlier .
the town , northeast of christchurch , has been cut off by landslides triggered by the quakes .
it was then hit by poor weather on tuesday , with heavy rain and flooding .

prime minister john key media conference the ministry of civil defence and emergency management confirmed there had been two deaths from the earthquake .
one person died in a house that collapsed in kaikoura , and a second person died at a house in mt lyford , north of christchurch .
mr key said a defence force helicopter would be flying to kaikoura and another one would be available in wellington .nearly 500 evacuees came into christchurch early on wednesday morning .

6,000 people expected at free health clinic in fort worth your best pathway to health is a non-profit organization that exists to serve the physical needs of the under-served .emergency response teams flew by helicopter to the region at the epicenter of the tremor .

according to the united states geological survey .australian disaster management expert paul seinfort the most expensive cost will be the repairs to the railroad tunnels , several of which have caved in .
the tremor also created fractures in major roadways such as state highway one , brought buildings to the ground and forced frightened kiwis to flock to their local supermarkets to stock up on food , drinks and essential supplies .

tsunami follows powerful tremor sunday , november 13 , 2016 0 awoniyi omomayowa omomayowa a tsunami has hit after an earthquake struck new zealand 's south island .
tsunami waves a gauge at kaikoura , 181 km -lrb- 112 miles -rrb- north of christchurch .the first wave has arrived in the north eastern coast of the south island and it may arrive in eastern coast of the north and south islands .

three new zealand cows whose predicament captured the interest of many people around the world .
the farmer , who was not named by newshub , said the cows were desperate for water after they were rescued .
he said the quake fault line ran right beneath his farm , which had been relatively flat before the earthquake .

3.11.2 Testing, Evaluation and Assessment

Most of the important topics from LDA are covered in summary like earthquake, magnitude of earthquake, tsunami and cattle cows. Unfortunately, the summary is not readable. The punctuation is erratic and the sentences do not make grammatical sense. Even though this summary contains many keywords the poor grammar restricts us from making this summary our flagship summary. The ROUGE scores are as follows:

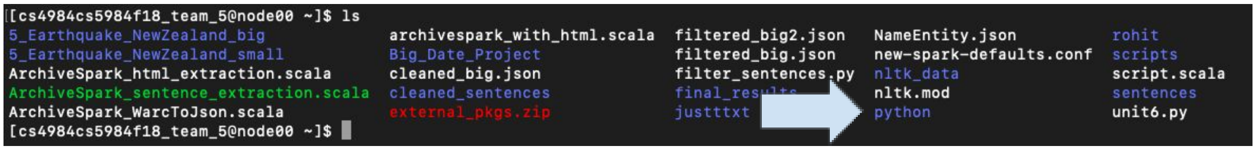
ROUGE-SU4	ROUGE-2	ROUGE-1
.07	.02857	.25

Table 3.21 Rogue score for Pointer Generator Model Summary

4. User's Manual

As stated in previous sections, our group utilized the Hadoop cluster to run our programs. In order to run our scripts one must log into the cluster by following the steps below. The data remains in the cluster and can still be accessed by the scripts. Each script is named based off of the task number and short keyword description. The tasks can be found above in the Approach and Evaluation section.

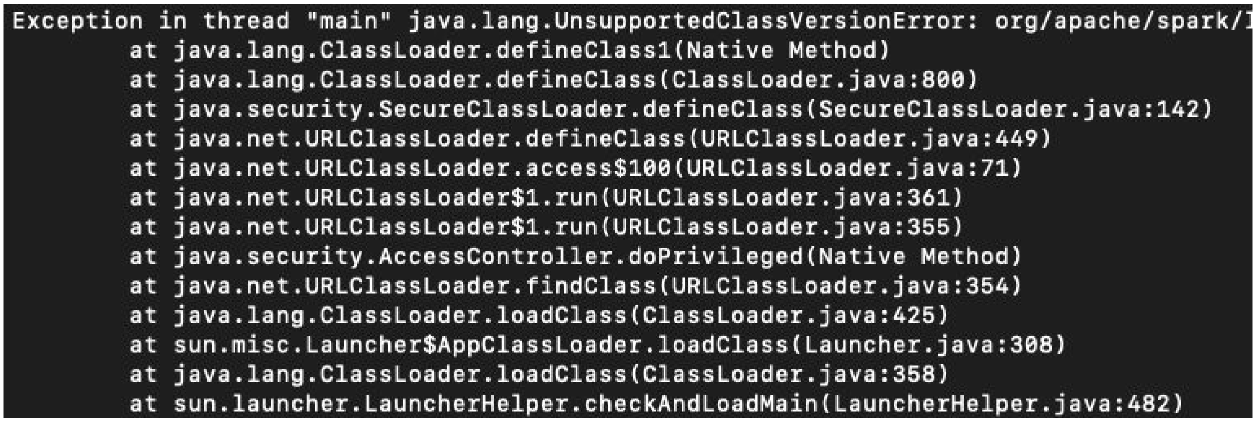
1. Login to the Hadoop cluster using the following command in the terminal:
   a. **ssh cs4984cs5984f18_team_5@hadoop.dlib.vt.edu -p 2222**
   b. Enter password: fox@vt.edu
2. Locate the folder "python". This folder contains the scripts for each deliverable in the project. Do not cd into it.



```
[cs4984cs5984f18_team_5@node00 ~]$ ls
5_Earthquake_NewZealand_big          archivespark_with_html.scala  filtered_big2.json     NameEntity.json            rohit
5_Earthquake_NewZealand_small        Big_Date_Project              filtered_big.json      new-spark-defaults.conf    scripts
ArchiveSpark_html_extraction.scala   cleaned_big.json              filter_sentences.py    nltk_data                  script.scala
ArchiveSpark_sentence_extraction.scala  cleaned_sentences          final_results          nltk.mod                   sentences
ArchiveSpark_WarcToJson.scala        external_pkgs.zip             justttxt               python                     unit6.py
[cs4984cs5984f18_team_5@node00 ~]$
```

Figure 4.1 Python Directory

3. From the home folder run the following command:
   a. **export JAVA_HOME='/usr/java/jdk1.8.0_171'**
   b. This command specifies the location of the JDK.
   c. If you receive the following error, you did not complete this step correctly.

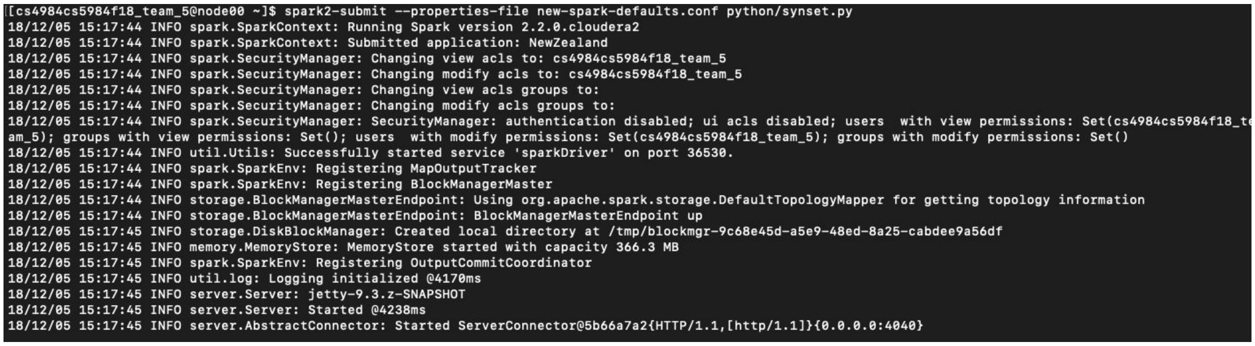


```
Exception in thread "main" java.lang.UnsupportedClassVersionError: org/apache/spark/
        at java.lang.ClassLoader.defineClass1(Native Method)
        at java.lang.ClassLoader.defineClass(ClassLoader.java:800)
        at java.security.SecureClassLoader.defineClass(SecureClassLoader.java:142)
        at java.net.URLClassLoader.defineClass(URLClassLoader.java:449)
        at java.net.URLClassLoader.access$100(URLClassLoader.java:71)
        at java.net.URLClassLoader$1.run(URLClassLoader.java:361)
        at java.net.URLClassLoader$1.run(URLClassLoader.java:355)
        at java.security.AccessController.doPrivileged(Native Method)
        at java.net.URLClassLoader.findClass(URLClassLoader.java:354)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:425)
        at sun.misc.Launcher$AppClassLoader.loadClass(Launcher.java:308)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:358)
        at sun.launcher.LauncherHelper.checkAndLoadMain(LauncherHelper.java:482)
```

Figure 4.2 Java Export Error

4. Run the following command from the terminal:
   a. **spark2-submit --properties-file new-spark-defaults.conf python/*name of file***
   b. Note: *name of file* is the script in which you want to run

c. The following should print from the console followed by status data if Spark started to run.



```
[[cs4984cs5984f18_team_5@node00 ~]$ spark2-submit --properties-file new-spark-defaults.conf python/synset.py
18/12/05 15:17:44 INFO spark.SparkContext: Running Spark version 2.2.0.cloudera2
18/12/05 15:17:44 INFO spark.SparkContext: Submitted application: NewZealand
18/12/05 15:17:44 INFO spark.SecurityManager: Changing view acls to: cs4984cs5984f18_team_5
18/12/05 15:17:44 INFO spark.SecurityManager: Changing modify acls to: cs4984cs5984f18_team_5
18/12/05 15:17:44 INFO spark.SecurityManager: Changing view acls groups to:
18/12/05 15:17:44 INFO spark.SecurityManager: Changing modify acls groups to:
18/12/05 15:17:44 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(cs4984cs5984f18_te
am_5); groups with view permissions: Set(); users  with modify permissions: Set(cs4984cs5984f18_team_5); groups with modify permissions: Set()
18/12/05 15:17:44 INFO util.Utils: Successfully started service 'sparkDriver' on port 36530.
18/12/05 15:17:44 INFO spark.SparkEnv: Registering MapOutputTracker
18/12/05 15:17:44 INFO spark.SparkEnv: Registering BlockManagerMaster
18/12/05 15:17:44 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
18/12/05 15:17:44 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
18/12/05 15:17:45 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-9c68e45d-a5e9-48ed-8a25-cabdee9a56df
18/12/05 15:17:45 INFO memory.MemoryStore: MemoryStore started with capacity 366.3 MB
18/12/05 15:17:45 INFO spark.SparkEnv: Registering OutputCommitCoordinator
18/12/05 15:17:45 INFO util.log: Logging initialized @4170ms
18/12/05 15:17:45 INFO server.Server: jetty-9.3.z-SNAPSHOT
18/12/05 15:17:45 INFO server.Server: Started @4238ms
18/12/05 15:17:45 INFO server.AbstractConnector: Started ServerConnector@5b66a7a2{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
```

Figure 4.3 PySpark Initiation Log

5. The results of each file will print to the console when Spark is finished running.

5. Developer's Manual

During our development process our team elected to use the following tools.

- **Natural Language Processing Toolkit (NLTK)**
- **Python 2.7**
- Zeppelin Notebooks
- **Hadoop Cluster**
- **PySpark**

In order to understand how to pick up from where our team left off it is important to understand which of these technologies is relevant. The items in bold are described in detail in the Approach and Evaluation section of the report. Information regarding running the files and logging into the hadoop system can be found in the above section, User Manual.

5.1 Data Flow

The data flow for this system is very simple because each script is independent of every other script. This does present reusability problems but as far as understanding flow, this architecture is the most simple. Each script takes the dataset and operates on it. Work is never outsourced from any script file; it is all contained within each task script.

The reason we took this approach is for simplicity and to allow each of us to work on parts of the system without causing problems for the other group members. Since we were unable to use version control in the Hadoop cluster, we did not want to create dependencies for specific files in our system. To avoid this we built our scripts as standalone objects even though we did not take advantage of abstracting out classes and other parts of our code.

5.1 File Inventory

This section provides a list of every file used and its purpose. Since each file serves a very particular purpose and has no code sharing, the inventory is very simple.

- cleaned_big.json
    - This file is the large cleaned dataset about earthquakes. It is imported into each script in order to be processed.
- Pos_tagger_important_words.py
    - This file takes in the cleaned_big.json file and outputs the most important words and their part of speech.
- Synset_most_freq.py

    - This file takes in the cleaned_big.json file and outputs the word synsets from the dataset.
- named_entity_recognizer_important_words.py
    - This file takes in the cleaned_big.json file and outputs the most important words that have been marked as a named entity.
- TF-IDF_new.py
    - This file takes in the cleaned_big.json and outputs the most important words as based on the TF-IDF method.
- clustering_word2vec.py
    - This file takes in the cleaned_big.json and creates multiple examples of an extractive summary.
- Templating.py
    - This shows the ranked values for the semantic slots that will be used for our abstractive summaries, using regex.

6. Lessons Learned

6.1 Timeline of Summaries

- October 24: First edition of extractive summary
  - [...significant, increase, number, devastating, earthquakes, striking, around, globe, 2018., claims, made, US, researchers, Roger, Bilham, Rebecca, Bendick, found, Earth's, rotation, decreased, slightly….]
  - [damaged, store, Wellington,, around, 440, km, Christchurch., Cracks, appeared, roads, around, Centre, Port,, Wellington,, earthquake, based, around, Cheviot, South, Island…]
- October 30: Second edition of extractive summary
  - At least two people were killed and thousands of landslides were triggered, cutting off some towns from the rest of the country. Emergency response teams flew by helicopter to the region at the epicentre of the tremor, which struck just after midnight some 91 km (57 miles) northeast of Christchurch in the South Island, amid reports of injuries and collapsed buildings. Read more Two killed in New Zealand earthquake Powerlines and telecommunications were down, with huge cracks in roads, land slips and other damage to infrastructure making it hard to reach the worst-affected areas. A tsunami warning that led to mass evacuations after the original quake was downgraded after large swells hit New Zealand's capital Wellington, in the North Island, and Christchurch. Wellington was a virtual ghost town with workers ordered to stay away while the city council assessed the risk to buildings, several of which were damaged by the tremor. Severe weather with 140 km per hour (85 mph), gale-force winds was forecast for the area. Hundreds of aftershocks, the strongest a 6.2 quake at about 1.45pm local time (0045 GMT), rattled the South Pacific country, fraying nerves in an area where memories of a deadly 2011 quake are still fresh. Christchurch, the largest city on New Zealand's ruggedly beautiful South Island, is still recovering from the 6.3 quake in 2011 that killed 185 people. New Zealand's Civil Defence declared a state of emergency for the Kaikoura region, centred on a tourist town about 150 km (90 miles) northeast of Christchurch, soon after Monday's large aftershock. Kaikoura, a popular spot for whale watching, bore the brunt of the quake. Up to 100,000 landslides were triggered in the region by the shock, the New Zealand Herald reported, some blocking roads and railways, cutting off the Kaikoura from the rest of the country. The Navy and Air Force have begun evacuating people stranded in the town. There was at least one death in area, as a large farmhouse collapsed with three people inside. Margaret Edgar, aged 100, was pulled alive from the ruins along with her daughter-in-law, but her son, Louis, was killed. Nearby, three cows were pictured huddled on an island of grass, the surrounding land having collapsed after the tremors. Three cows are stranded on

an island of grass in a paddock that had been ripped apart following an earthquake near Kaikoura, New Zealand (Newshub via AP) Urban Search and Rescue (USAR) said a 20-person rescue team and two sniffer dogs had arrived in the town. A second team was on standby in Christchurch, the USAR said in a statement. Hours after the quake, officials said a slip dam caused by the quakes that had blocked the Clarence River north of the town had breached, sending a wall of water downstream. New Zealand's Geonet measured Monday's first quake at magnitude 7.5, while the U.S. Geological Survey put it at 7.8. The quakes and aftershocks rattled buildings and woke residents across the country, hundreds of kilometres from the epicentre. Government research unit GNS Science later said the overnight quake appeared to have been two simultaneous quakes which together lasted more than two minutes. Around 90 percent of the world's earthquakes occur within this region. Stock exchange operator NZX Ltd said markets were trading, although many offices in the capital were closed. The New Zealand dollar initially fell to a one-month low before mostly recovering. Fonterra, the world's biggest dairy exporter, said some its farms were without power and would likely have to dump milk. Key postponed a trip to Argentina, where he had planned to hold a series of trade meetings ahead of the Asia-Pacific Economic Cooperation (APEC) leaders' summit in Peru this week, as he met disaster officials. Two other people were pulled alive from the same building. Shortly after the quake, residents heeding tsunami warnings in Wellington caused gridlock on the roads to Mount Victoria, a hill with a lookout over the low-lying coastal city. Yes, submit this vote Cancel You must be logged in to vote. Yes, flag this comment Cancel This comment has been flagged. This comment has been flagged. Yes, delete this comment Cancel This comment has been deleted. This comment has been deleted.

- October 6: Third Edition of Extractive Summary
    - Laced with seaweed and the purplish layers of rock, the raised platform is the raw product of the category 7.8 earthquake that struck the region on 13 November. Locals reported that the sound of water rushing off the new beach was louder than the earthquake itself, as gallons of seawater were suddenly displaced by the tectonic movement. The local fishing industry is expected to be severely impacted and there have been a number of landslides and rockfalls along the shore. (Image: Anna Redmund) The dramatic change, known as coastal uplift, occurred when at least four different faults broke at once. Although uplift in New Zealand is normal for large earthquakes near the coast, the Kaikoura quake offers geologists a chance to observe how several different faults can join together in a large earthquake. Palm oil is omnipresent in global consumption. But in many circles it is considered the scourge of the natural world, for the deforestation and habita... Join the thousands following us on Twitter, Facebook and Instagram and stay informed about the world. About Geographical Published in the UK since 1935, Geographical is the official magazine of the Royal Geographical Society (with IBG). All Rights Reserved.

- November 26: First Edition of Abstractive Summary with Template
  - On {November , 2016} at {11:02}, a earthquake with magnitude of {7.5} struck {Wellington}. The epicenter of the earthquake was located at {Wellington}. The earthquake caused number of casualties, reported to be {185} deaths and {20} injuries. There {were} aftershocks reported, and there {was} tsunami caused by the earthquake reported. Reports mentioned there {were} landslides caused by this earthquake.
- November 27: First Edition of Deep Learning Summary
  - the earthquake hit at 4.40 am on friday , about 80 miles -lrb- 130 km -rrb- north-east of the east cape community of te araroa .
    the 7.5-magnitude earthquake in the south island killed at least two people and triggered a tsunami alert for the entire east coast . seven aftershocks have been registered .
    civil defence evacuated low-lying coastal areas after a tsunami warning was issued.
    mountjoy suspects that the unique nature of the earthquake is behind the eerie exposure of the sea bed in the coastal town of kaikoura .
    the effects can be seen all over kaikoura , with some areas rising a metre above the ground .
    earthquake geologist nicola litchfield from gns science told michael daly from stuff that the uplift would have happened during the 90 seconds to 2 minutes the quake .

    a state of emergency was declared in the town of kaikoura , a whale-watching destination and home to 2,000 people that has been completely cut off .

    quake hit just after midnight -lrb- 11:02 gmt on sunday -rrb- , some 95 km -lrb- 59 miles -rrb- from christchurch .
    officials said the first waves may not be the largest , with tsunami activity possible for several hours .

    aap by jerico mandybur2016-11-14 utc after earthquakes with a magnitude of up to 7.5 struck new zealand 's upper south island just after midnight on monday .
    two people were confirmed dead and extensive damage was done to buildings and roads .
    with more than 250 aftershocks recorded and tsunami warnings sending people to higher ground , social media was flooded with videos and images as people captured the damage done to their businesses and properties .

    four air force helicopters have been airlifting people out of kaikoura on the south island after battling strong winds and heavy rain earlier .
    the town , northeast of christchurch , has been cut off by landslides triggered by the quakes .

it was then hit by poor weather on tuesday , with heavy rain and flooding .

prime minister john key media conference the ministry of civil defence and emergency management confirmed there had been two deaths from the earthquake .
one person died in a house that collapsed in kaikoura , and a second person died at a house in mt lyford , north of christchurch .
mr key said a defence force helicopter would be flying to kaikoura and another one would be available in wellington .nearly 500 evacuees came into christchurch early on wednesday morning .

6,000 people expected at free health clinic in fort worth your best pathway to health is a non-profit organization that exists to serve the physical needs of the under-served .emergency response teams flew by helicopter to the region at the epicenter of the tremor .

according to the united states geological survey .australian disaster management expert paul seinfort the most expensive cost will be the repairs to the railroad tunnels , several of which have caved in .
the tremor also created fractures in major roadways such as state highway one , brought buildings to the ground and forced frightened kiwis to flock to their local supermarkets to stock up on food , drinks and essential supplies .

tsunami follows powerful tremor sunday , november 13 , 2016 0 awoniyi omomayowa omomayowa a tsunami has hit after an earthquake struck new zealand 's south island .
tsunami waves a gauge at kaikoura , 181 km -lrb- 112 miles -rrb- north of christchurch .the first wave has arrived in the north eastern coast of the south island and it may arrive in eastern coast of the north and south islands .

three new zealand cows whose predicament captured the interest of many people around the world .
the farmer , who was not named by newshub , said the cows were desperate for water after they were rescued .
he said the quake fault line ran right beneath his farm , which had been relatively flat before the earthquake .

6.2 Challenges Faced

Our team faced a multitude of issues throughout the course of the semester. Some of these high-level issues include using third party software, cleaning unclean data, and choosing how we were going to implement our solutions.

One of the initial problems that we ran into was choosing how to implement our solution. There were many different clusters and frameworks that were presented to us at the beginning of the semester. Figuring out which technologies to use proved very difficult because none of the team members had any understanding of them previously. We did our best to learn about each technology before choosing but once we chose it was hard to switch because of the time we dedicated to something that did not work as well as we had hoped. If we had simplified the pros and cons would have made this decision much easier. The information that we were presented with about each technology did not help with our decision and we failed to look at it in a way that did. The information that was given was not practical. The information given revolved more around the history of the technologies and servers and how they are built rather than what kinds of third party software we were able to use on them and their limitations. Ultimately, we are happy with the decisions we made and have been able to make progress regardless of this challenge.

Integrating third party software proved to be difficult and created some barriers throughout the course of the semester. Our first instance of this problem was in Zeppelin. We were given small tutorials to help us with some of the tasks. However, in order to use the different packages in the notebook we had to install those packages on Zeppelin. Unfortunately, that was not possible. When we switched to Hadoop we found ourselves with the same problem. Not only did Hadoop restrict the rare third party packages like Spacy, but it also did not have universal libraries like git installed. Without a version control our team had a lot of trouble testing code and modifying it. We resorted to making duplicate files and relying on Hadoop to show each team member the most up to date version of the scripts. This method was not ideal but we were able to work past it by organizing our Python files well and familiarizing ourselves with each other's work.

The final large challenge we faced was cleaning our data. While some tasks were completed early on in the semester, cleaning our data was an ongoing task that never fully finished. When we first started using our dataset we noticed that there was a lot of Javascript, advertisements, and banners in the data. We also discovered many foreign pages and pages that did not even include information about New Zealand or earthquakes. As we improved the cleanliness of our data our lower level tasks improved. When these tasks improved our high level tasks like the extractive summary improved tremendously. We were able to work through the issues in data cleanliness largely due to the effort that multiple team members put into the project throughout the semester. Information about this process can be found in the *Approach, Design and Implementation* section above.

6.3 Solutions Developed

As explained in previous sections our group developed a solution that creates an accurate readable summary using the templating abstractive summary technique. This method of summarization proved to be our best and will be our final deliverable. Smaller task numbers provide data but none are as comprehensive or accurate as our final summary using the templating system. We also developed a decent solution using the extractive summary and deep learning techniques. These results were not ideal but still provided valuable insight into the concept of summarization.

6.4 Future Work

This project has the potential to be expanded greatly. There are three main areas of improvements/expansions that our group has identified. If another team were to take over our project at this current point the following suggestions should be taken into heavy consideration. To improve this project a team could create a more fluid class structure with all of the different scripts, utilize more up to date methods, and expand the template system beyond just earthquakes to be used in any dataset.

Our current suite of Python files is very jumbled and there is a lot of repeating code. In order to make the system easier to use it would be beneficial to separate redundant code into reusable classes. In each of our scripts we tokenize the words. This could be done once and then each script could then utilize that class to make that happen instead of calling multiple lines of NLTK code. Using parameters would also be very helpful. We could specify how many words (pieces of data) that we want to print out on any given run. It would also be beneficial to have a master script/class that runs each task simultaneously. Refactoring the scripts to classes that are easy to read would decrease onboarding time for new developers and make it easier for a user to run our system.

During our project we ran into problems with out of date software. The NLTK library is widely used and well known but the library does not have up to date POS tagging and named entity recognition (NER). Towards the end of the semester we found that some groups noticed this earlier and switched to using Spacy. Spacy worked well for them but we could not use it on the Hadoop cluster. For this reason it was already too late to migrate our system to another host and therefore we did not use the more accurate library. Moving to a new cluster would be beneficial for future work because it would allow for more accurate placement of values in semantic slots. We currently use purely regex. Using both regex and NER would increase the accuracy of our system.

The most ambitious future work goal is to expand the templating system to work with all types of natural disaster datasets instead of only one. This would be a great accomplishment because the system would be much more useful if it could summarize multiple types of datasets rather than just earthquakes. A high level approach to doing this is as follows. We would create a very generic template. Currently our template includes words like “earthquake”. We would replace those words with slots so that not only is our “death toll” and locations sots filled in but also the type of disaster. Depending on the type of disaster our template would rearrange itself and change some of the sentences. This process should be built on top of our existing functionality and deliverables.

7. Gold Standards

7.1 Gold Standard for Team 6

In order to help each group evaluate their summaries each group was assigned another group to write a gold standard for. We wrote a summary for Team 6. Their topic was on the Parkland Shooting. We used numerous sources to develop this summary and all are listed below the summary itself. The summary can be found below:

On February 14, 2018, gunman Nikolas Cruz killed seventeen people at Marjory Stoneman Douglas High School in Parkland, Florida. Fourteen students and three administrators were killed in the shooting. Seventeen people were also injured non-fatally in the shooting. Cruz began the attack at 2:21 p.m. and left the premises at 2:28 p.m. He carried out the attack without intervention with police officers, despite the fact that there was an armed deputy outside within two minutes of the start of the shooting.

The seventeen victims killed in the attack include: Joaquin Oliver, Aaron Feis, Martin Duque, Meadow Pollack, Alyssa Alhadeef, Jaime Guttenberg, Alaina Petty, Cara Loughran, Nicholas Dworet, Gina Montalto, Scott Beigel, Chris Hixon, Luke Hoyer, Helana Ramsay, Alex Schater, Peter Wang, and Carmen Schentrup. Many of these victims sacrificed their lives to help their peers. Aaron Feis, a football coach and security guard at the school, shielded students with his body to protect them from bullets. Peter Wang was fatally shot while holding the door open to allow his peers to escape the building.

The law enforcement response was complicated by several factors. For one, the emergency response calls from cell phones within the school were routed to the Coral Springs Fire Department. On the other hand, calls from parents were directed to the Broward Sheriff's Office. Another problem was that the police radio system became overloaded, forcing officers to use hand signals to communicate. Two sheriff's deputies did arrive at the scene during the shooting, but did not enter the building to intervene.

Other problems prevented internal mitigation of the attack. For instance, teachers were unable to lock their doors from within the classroom. They were forced to go outside the room with a key to lock the door. Another issue was that the gun smoke set off fire alarms in the building, causing confusion. Teachers and students were unsure whether the event was a fire, which required evacuation, or a shooting, which required seeking cover indoors.

Nikolas Cruz is currently being charged with seventeen counts of attempted murder and seventeen counts of first-degree, premeditated murder. Cruz could be sentenced to death if he is convicted. So far, he has entered a not-guilty plea in March. Cruz is also facing allegations for attacking a jail guard. Cruz was given an order to stop dragging his shoes on the floor by a guard. Upon hearing this, he became aggressive and repeatedly punched the guard in the head, and grabbed his stun gun. The stun gun was discharged in the altercation, after which the guard regained control of his weapon and punched Cruz in the face. Cruz later withdrew to a seat and was apprehended into custody.

This shooting sparked political discussion for gun control and mental health. Student led organizations as well as politicians called for gun control after the shooting. The “Never Again” movement sought to help push for these types of changes. Another topic of interest was mental health. Nikolaus Cruz showed many signs of being troubled but no serious action was ever taken to find him help. Mainly Democrats focused on the gun laws while Republicans focused on mental health.

Nikolas was adopted at a young age and never knew his birth mother. His birth mother was a drug addict and was often violent. His adoptive mother was known as being thoughtful and disciplined. Cruz took after his birth mother. He often had violent outbursts. He had an unhealthy obsession with guns and would sometimes shoot and torture animals. Cruz even released a video in which he described how he would carry out a school shooting. These signs were not taken seriously enough to stop the shooting from occurring. His mother told the press that she indulged him in his violent video games because they would calm his mood. This also could have led to his eventual shooting.

Cruz was able to purchase his guns legally. He instantly passed a background check and was not stopped due to his mental illness. This has been a big point made when discussing the issue of gun control. Three weeks after the shooting, Florida Governor Rick Scott signed into law the Marjory Stoneman Douglas High School Public Safety Act. This act raises the age in which people can buy guns and allows judges to bar someone from purchasing a gun if they display an act of violence. The bill also allows for the voluntary arming of school officials and a school “guardian” program in which an officer is assigned to a school facility. Advocates of this bill hope that it will make the loose gun laws in Florida more strict. Currently, buying a gun does not require a permit, private gun sales do not require a background check, gun ownership does not require the owner to register their gun. With the new restrictions, there is hope to stop the occurrence of mass shootings in the United States of America.

Citations

https://www.nytimes.com/2018/04/24/us/parkland-shooting-reconstruction.html
https://www.cbsnews.com/pictures/florida-school-shooting-victims-identified-by-authorities/14/
https://www.nytimes.com/2018/11/14/us/nikolas-cruz-parkland-shooting-charges.html
https://www.independent.co.uk/news/world/americas/nikolas-cruz-parkland-shooting-court-trial-school-shooter-jail-guard-stun-gun-a8634371.html
https://www.miamiherald.com/news/local/community/broward/article216909390.html
https://www.newyorker.com/news/news-desk/how-the-survivors-of-parkland-began-the-n

ever-again-movement
https://www.ajc.com/news/national/florida-gun-laws-how-have-they-changed-after-the-parkland-shooting/BIhOP1bppQJjV7NI1F7ZXI/

7.2 Our Golden Standard

In order to aid in our evaluation of our data, Team 4 wrote a gold standard for our summaries. The summary can be found below.

On Tuesday February 22, 2011 at 12:51 pm, an earthquake of magnitude 6.3 hit Christchurch, New Zealand, a South Island city of nearly 400,000 people. The country's deadliest natural disaster in 80 years killed 185 people from more than 20 countries and injured several thousand, 164 seriously. The earthquake epicenter was near Lyttelton, just 10 kilometers south-east of Christchurch's central business district at a depth of 5 kilometers. The earthquake, caused by a hidden fault, occurred more than five months after the 4 September 2010 earthquake, but is considered to be one of the more than 11,000 aftershocks of the earlier quake. In the ten minutes after the February 22 quake, there were 10 aftershocks of magnitude 4 or more. At 1:04 a magnitude 5.8, and at 2:50pm a magnitude 5.9, quake followed. The earthquakes stemmed from the deformation along regional plate boundaries where the Pacific and Indo-Australian tectonic plates push against one another. Although not as powerful as the magnitude 7.1 earthquake on 4 September 2010 which happened at night and was the first devastating earthquake of the current decade, causing some 3 billion pounds damage, this Canterbury region earthquake occurred on a shallow fault line that was close to the city, and there was very high ground acceleration, so the shaking was particularly destructive. It was lunchtime and many people were on the city streets. 115 people died in the Canterbury Television (CTV) building, 18 in the Pyne Gould Corporation (PGC) building, 36 in the central city (including 8 on buses crushed by crumbling walls), and 12 in the suburbs (including from falling rocks in the Redcliffs, Sumner, and Port Hills). The Chief Coroner determined that another four deaths were directly associated with the earthquake.

The earthquake brought down many buildings damaged in the previous September, especially older brick and mortar buildings. Up to 100,000 buildings were damaged and about 10,000 buildings were deemed unsalvageable, needing to be demolished. Heritage buildings suffered heavy damage, including the Provincial Council Chambers, Lyttelton's Timeball Station, the Anglican Christchurch Cathedral, and the Catholic Cathedral of the Blessed Sacrament. More than half of the buildings in the central business district have since been demolished, including the city's tallest building, the Hotel Grand Chancellor. Over a quarter of the buildings in the central business district were demolished.

Liquefaction was much more extensive than in the September 2010 earthquake, which is known as the Greendale, Darfield, Rolleston, September or Canterbury Earthquake. During the

2010 and 2011 Canterbury earthquakes, over 400,000 tons of silt came to the surface. Many roads, footpaths, schools and houses were flooded with silt. Properties and streets were buried in thick layers of silt, and water and sewage from broken pipes flooded streets. House foundations cracked and buckled, wrecking many homes. Despite the damage to homes, there were few serious injuries in residential houses in liquefaction areas. However, several thousand homes had to be demolished, and a large area of eastern Christchurch will probably never be reoccupied.

The government activated the National Crisis Management Centre and declared a national state of emergency the day after the quake. Authorities quickly cordoned off Christchurch's central business district. Christchurch's CBD remained cordoned off until June 2013. Power companies restored electricity to 75 percent of the city within three days, but re-establishing water supplies and sewerage systems took much longer. Rescue crews came to help from all over the world, especially Japan, the United States, the United Kingdom, Taiwan, and Australia.

In the weeks following the earthquake about 70,000 people were believed to have left the city due to uninhabitable homes, lack of basic services, and continuing aftershocks. Roads, bridges, power lines, cell phone towers, and ordinary phone lines were damaged. The water and sewage pipes were badly damaged. Many people needed to use portable or chemical toilets, and got their water from tankers for months after the quake. Timaru's population swelled by 20% and thousands of pupils registered at schools in other cities and towns. Many returned to Christchurch as conditions improved.

Several earthquakes happened during the following years but the next devastating and geological complex earthquake occurred in 2016 near the tourist town Kaikoura.

On November 14, 2016 at 12:02 am, a 7.8 magnitude earthquake, the second strongest quake since European settlement, struck 15 kilometers northeast of Culverden and 60 kilometers southwest of Kaikoura, lasting for nearly 2 minutes. There were 2 casualties with another 57 people treated for injuries. One of the casualties was a man who had been crushed by his collapsing home outside of Kaikoura. Two of his family members were rescued from the rubble. The other casualty was a woman who had a head injury during the earthquake. The earthquake also significantly damaged many roads, buildings, and bridges. Prime Minister John Key estimated that costs to fix all the infrastructure was in the billions.

Most significantly, the earthquake, which brought down between 80,000 and 100,000 landslides, effectively cut off all roads from Kaikoura to the rest of the country. By November 16, 200 people had been airlifted out of Kaikoura, while another 1,000 waited for evacuation. Many of the people who were stranded in Kaikoura were tourists, which elicited other nations to help in the evacuation efforts. The US volunteered the use of two navy helicopters. China sent helicopters to evacuate Chinese tourists.

Restoration of roads between Kaikoura and the rest of the country took a couple of weeks. On November 30, the designated Kaikoura Emergency Access Road was reopened to civilian drivers with permits. By the end of December 2016, the Kaikoura Emergency Access Road and State Highway 1 south of Kaikoura were completely reopened. The Main Rail North line was also severely damaged by the earthquake. It was partially restored by September 2017, but passenger service will not resume until December 2018. Large areas of the South Island's northeast coast shifted upward and northward; about 20 kilometers of the Marlborough coast was raised between 0.5 and 2 meters.

The tsunami following the 2016 earthquake had a maximum height of 7 meters at Goose Bay. After the earthquake, the tide level dropped 2.5 meters over 25 minutes. The tide level then rose 4 meters over the next 15 minutes. The tsunami caused some additional destruction, but the impact was minimized because the tsunami occurred at mid to low tide. One building was heavily damaged. The tsunami also scattered marine plants and animals across the Oaro River flood plain. The tsunami caused no additional casualties or injuries.

8. Acknowledgments

We would first like to thank Professor Fox for his management of the class and feedback throughout the semester. His goal for the class was for students to learn. Not only did we learn, but we accomplished a lot through our project. This has been very valuable to all of the group members.

We would also like to thank Liuqing Li for his work as the GTA. He was able to help our group consistently achieve results when we were blocked. His knowledge and ability to apply his skills were very useful throughout the semester.

Lastly, we would like to thank all of the special guest speakers who attended our classes. They helped us widen our understanding of the different approaches that we could take for our project and our summaries.

9. References

[1] G. Chowdhury, "Natural language processing", Annual Review of Information Science and Technology, vol. 37, no. 1, pp. 51-89, 2005.

[2] P. Nadkarni, L. Ohno-Machado and W. Chapman, "Natural language processing: an introduction", Journal of the American Medical Informatics Association, vol. 18, no. 5, pp. 544-551, 2011.

[3] S. Bird, E. Klein and E. Loper, "NLTK Book", Nltk.org, 2018. [Online]. Available: http://www.nltk.org/book/. [Accessed: 24- Oct- 2018]

[4] I. Matei Zaharia, "Apache Spark: A Unified Engine For Big Data Processing", Cacm.acm.org, 2018. [Online]. Available: https://cacm.acm.org/magazines/2016/11/209116-apache-spark/fulltext. [Accessed: 01- Nov-2018]

[5] R. Taylor, "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics", BMC Bioinformatics, vol. 11, no. 12, p. S1, 2010.

[6] Mencher, Melvin. *Melvin Menchers News Reporting and Writing*. McGraw-Hill, 2011.

[7] M. Tim Jones, “An introduction to natural language processing”. Available: https://www.ibm.com/developerworks/library/cc-cognitive-natural-language-processing/index.html [Accessed: 5 - Dec - 2018]

[8] Available: https://www.mypatentideas.com [Accessed: 5 - Dec - 2018]

[9] Abigail See, “Get To The Point: Summarization with Pointer-Generator Networks”. Available: https://www.slideshare.net/aclanthology/abigail-see-2017-get-to-the-point-summarization-with-pointergenerator-networks [Accessed: 5 - Dec - 2018]

[10] Corcoran, Liam. “What's The Average Word Count of Viral Stories?” NewsWhip, 12 Dec. 2013, www.newswhip.com/2013/12/article-length/.