UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

A Model for Processing Public Participation Feedback Using Topic Modeling

(A Case for Public Task Forces in Kenya)

By:

Stephen Nyaga Muita

P52/11224/2018

Supervised By: Dr. Lawrence Muchemi

A PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF MASTERS OF SCIENCE (COMPUTATIONAL INTELLIGENCE) IN THE SCHOOL OF COMPUTING AND INFORMATICS OF UNIVERSITY OF NAIROBI

July, 2020

Declaration

This research project is my original work and has not been submitted to any other university for academic award.

Sign…………………………. Date…………………………

STEPHEN NYAGA MUITA

P52/11224/2018

This project has been submitted with my approval as the appointed University supervisor.

Sign………………………….. Date………………………..

DR. LAWRENCE MUCHEMI

PROJECT SUPERVISOR

Dedication

I dedicate this project to my father, Samuel Muita for his unwavering encouragement in my academic journey. I also dedicate it to Martha Wangui who has constantly offered support and help during the entire Msc. program. Above all, I dedicate this project to God for His grace and strength throughout the study period.

Acknowledgement

I would like to express my heartfelt gratitude to Dr. Lawrence Muchemi for his valuable and insightful guidance during the entire period of this project. His patience, insight, suggestions and recommendations have been instrumental in the successful completion of this work. I also wish to express my gratitude to the teaching fraternity of the Msc (Computational Intelligence) at the School of Computing and Informatics, University of Nairobi for the high quality of input and teaching that inspired this research. Finally, I thank my family, colleagues and friends for their support during my studies.

Abstract

Governments acknowledge the need to involve citizens, through different platforms like public task forces, when public policies are drafted. Public task forces are expected to consider views from different stakeholders in the process of drafting their recommendations and policy proposals. They tend to receive large amounts of submissions and in most cases, the submissions are contained in massive documentations. The amount of material received is beyond the task forces' processing ability causing input from critical stakeholders being ignored completely leading to biased output. This provides a situation where topic modeling can be applied. Topic modeling is an unsupervised machine learning algorithm that can process large corpora of data by classifying them by identifying themes in those corpora. In our study, we set out to develop a model that task forces can use to process the received feedback. The model was validated by comparing the topics identified by the model against those identified by a human expert. An experiment was conducted where we built and trained an LDA topic model with 15 submissions. We then presented 7 submissions both to the trained model for processing and also to a human expert to manually identify the topics contained in those submissions. The topics generated by the model were compared to the topics identified by the human expert. The model generated topics that are similar to the topics identified by a human expert. Distinctive topics are contained in submissions. These topics are few and trying to extract a higher number of topics, will lead to more overlapping and nonsensical topics being generated. Extraction of topics contained in feedback from the public can be automated. This study contributes to practice by enabling task forces to objectively identify topics and themes covered by submissions which results into more acceptance of outputs and recommendations of task forces by the citizens.

Keywords: Text mining, topic modeling, text summarization, natural language processing, LDA

Table of Contents

List of Figures

List of Tables

Abbreviations

NLP - Natural Language Processing

API - Application Programming Interface

LDA - Latent Dirichlet Allocation

LSI - Latent Semantic Indexing

LSA - Latest Semantic Analysis

PLSA - Probabilistic Latent Semantic Analysis

PLSI - Probabilistic Latent Semantic Indexing

TF-IDF - Term Frequency–Inverse Document Frequency

SA - Sentiment Analysis

NER - Named Entity Recognition

SVD – Singular Value decomposition (SVD)

PDF – Portable Document Format

CHAPTER ONE

INTRODUCTION

1.1 Background

Most governments today are keen on involving citizens when formulating public policies. Different platforms exist where governments seek to engage the citizenry. One such platforms is public task forces. Citizens and other stakeholders have the opportunity to air their views and opinions through the task forces. "Governments also have the opportunity to listen, understand and adapt to social issues as they unfold" (Evangelopoulos and Visinescu, 2012). The output of the task forces is expected to be objective and lacking bias clearly reflecting the submitted views. Task forces are set up, for a set duration of time, to collect and collate views from different stakeholders and use the views to develop policies regarding a specific area or sector of governance. Stakeholders, (normally different interest groups), submit views in three different forms; printed documents, soft copy documents and oral forms. Where submissions are made orally, transcription is performed to convert the audio into written text.

The first step in processing feedback is reading the summary provided by a stakeholder, which could be either in written or oral form. The next step is to review the submissions in detail. The process of analyzing the submitted documents is cumbersome as, ideally, it would require the contents of the entire document to be reviewed. The aim of this step is to understand the issues that have been raised in the document. The results of processing a single document has to be recorded for later comparison against the results of reviewing other submissions. The results of the processing stage are then collated for final analysis. The task force then develops policies and recommendations that seek to address specific objectives using the results of the processing and analysis exercise. These policies, once implemented, then affect the stakeholders either directly or indirectly. It is common to find one submission addressing multiple themes. There exists many stakeholders including individuals, private companies, special interest groups and government bodies. Public task forces are comprised of different professionals and experts.

A key operating guideline for the task forces is that they should ensure that as many stakeholders as possible participate and give their input in a transparent manner. Some of the task forces are considered a success while others are deemed to not have achieved their set objectives. The recommendations and proposals of task forces often face public outrage with some stakeholders complaining about their views not being reflected in the final recommendations. Task forces normally find it difficult to justify their recommendations and what really informed the decisions. Part of the reason for this difficulty is the large amount of received submissions which proves very challenging to process fully. When policy makers are not able to collect and analyze all information that would be relevant, they tend to give close attention to a small range of issues while completely ignoring other issues. Due to this, they tend to make emotional and often quick decisions. This scenario also makes them easy to persuade and causes them to focus on stories and issues which take advantage of their emotions and biases reinforcing their beliefs. This phenomenon is referred to as "bounded rationality" (Cairney, 2015).

Public task forces operate within a set duration of time. The bulk of their work involves analyzing views from stakeholders. This is a very time consuming exercise because it requires one to be able to quickly grasp the content of the submissions. It is not uncommon to find a submission from a stakeholder spanning tens of pages. Synthesizing the content of such a large document becomes very difficult to accomplish given the time constraint. It is possible that some submissions never get reviewed at all due to their large size and some of them end up being reviewed partially. Leaving out submissions may lead to a biased decision or a biased policy recommendation by the task force.

With the growing level of attention being accorded task forces by the general public and other interest groups, it is highly likely that the amount of feedback submitted to the task forces will continue to grow both in terms of magnitude and scope. There is therefore a need to apply machine learning techniques to automatically process the submissions in an efficient manner to ensure that task forces cope with the large amount of submissions.

Text mining has attracted significant research interest in the recent past. Information is stored in both structured and unstructured forms. Majority of the world's data is stored in form of unstructured text. Further processing of such unstructured text using computers is a difficult and a challenging endeavor. Unstructured text usually contains valuable information that requires text mining techniques to extract that information. Text mining is also referred to as knowledge discovery from textual databases and leads to discovery of valuable information from textual data which otherwise remains hidden or secret by using different machine learning techniques (Ramanathan et al., 2013). Aggarwal and Zhai (2012, p.2) reported that text mining involves "going beyond information access to further help users analyze and digest information and facilitate decision making" at a low cost. Text mining techniques save time from collection of data and information processing perspectives. "Based on the principle of adaptation, text mining applications save time in terms of data collection and information processing. Thus, decision makers can have more time to adapt to the actual problem environment and conduct better outcome estimations" (Ngai et al., 2016).

Various techniques have been used to perform basic text mining. They are:

- **Information Retrieval**. Search engines recognize documents that are associated with a given set of words. Documents are first retrieved followed by information extraction where the user query is processed against the retrieved documents.
- **Information Extraction (IE).** This technique involves extracting useful information from text by identifying entities and establishing relationships that exist between these entities. Named entities are extracted such as names of people, organizations etc.
- **Categorization**. This is a form of supervised learning where documents are assigned to a predefined list of tags depending on the content of each document. Spam filtering in emails, topic genre identification and generation of document metadata are examples of how text categorization can be applied (Ramanathan et al., 2013).
- **Clustering**. This is an unsupervised technique where information from text is arranged into groups or clusters with each cluster being associated with related information pieces. Examples of application include pattern recognition, web search and document retrieval.
- **Summarization**. This technique involves compressing long text while still preserving a significant amount of information that was present in the original text.

Text mining techniques have been applied in academics by researchers to analyze journal and conference papers whose aim is to unearth patterns and trends in research. Text mining tools have also been used to identify trends in topics that exist within journal proceedings and how they change over time as well as categorizing journal papers into different genres (Dang, 2014).

Supervised and unsupervised text mining techniques have been applied to automatically process copyright applications.

Organizations are enhancing customer relations by using text mining to analyze complaints and support tickets. By converting the data into themes and semantic networks, an organization is able to get insights into the quality of feedback received. Sources of such data include web pages, support forums and customer support systems.

Governments are applying text mining to monitor and analyze, for security purposes, plain text sources such as blogs and emails. The main aim is to be able to detect threats to security by identifying semantic patterns in these plain text sources.

Public task forces have largely relied on physical examination of every submission. In most cases, the volume of the submissions is too large beyond the manual processing ability of the task force. Text mining techniques can be applied to automatically process the submissions in an efficient manner to ensure that task forces cope with the large amount of submissions. Various techniques could be applied such as detecting the hidden topics in the submissions, identifying the most relevant words and also identifying named entities contained in the submissions. Application of such techniques will greatly expedite processing of submissions and consequently reduce the time allocated for the overall operation of the task forces. By performing the automatic processing of the submitted views, the task forces will be able to make recommendations that are devoid of human bias or preference.

In this study, we developed a model that processes views from stakeholders efficiently and in an objective manner. We applied topic modeling text mining technique using Latent Dirichlet Allocation (LDA) algorithm to identify the topics that are hidden (latent) in the submitted views.

This document is organized into five main sections. The first section is the introduction that highlights the background to the problem, the statement of the problem, objectives and research questions as well as justification and scope of our work. Section two contains detailed literature related to our work. In this section we describe the various text mining techniques, the algorithms applied in text mining, the process model, previous work in the field of text mining and the existing research gaps. Section three describes the research methodology including how we collected data, how we designed and validated the model including the different experiments that we performed. Section four contains the results of our experiments while section five contains discussions, conclusions and recommendations for future work.

1.2 Problem Statement

Public task forces receive large amount of submissions in various forms which should all be processed and analyzed. The results of the analysis inform proposals and recommendations made by task forces. Currently, submissions are processed manually which is very time consuming. Some submissions are never processed, or are partially processed, either because of their bulkiness or due to a limitation of the allocated time. The amount of feedback that public task forces are likely to encounter is simply beyond their human processing capacity.

By omitting some submissions, task forces end up making proposals that do not incorporate input from all stakeholders. As a result of this, stakeholders have in the past raised their dissatisfaction with proposals made by task forces. The dissatisfaction that follows a task force’s report makes it difficult for the recommendations therein to be implemented. This has resulted to a large number of task forces’ reports to remain unimplemented despite money and resources having been spent on them. Manual processing of feedback is characterized by human bias or preference of the task force members leading to skewed proposals.

There is a need to establish an efficient means to process feedback from as many stakeholders as possible in an objective manner. Human bias need to be eliminated when processing feedback from stakeholders. To address the problem of partial and subjective processing, this study aimed at developing a model that public task forces can use when processing feedback. The model incorporates text mining technique to extract useful information contained in the submissions.

1.3 Research Question

We used this study to answer the following research question:

1. How can public task forces process the large amount of feedback in a consistent and objective manner?

1.4 Objectives of the Study

The objectives of our study are:

1. Develop a model to process feedback received by public task forces
2. Validate the developed model to ensure that feedback is processed effectively

1.5 Scope

Figure 1 shows the general framework of operations of public task forces. Our study focused on how public task forces can effectively process the large amount of feedback they receive during their tenure.

1.6 Significance of the Study

The output of this study would provide public task forces with a platform where large amount of stakeholders' views can be processed efficiently and objectively. The model also results to improved level of satisfaction with the proposals of task forces thereby enhancing adoption of their recommendations. This will be achieved because automated and objective processing of the views will be performed thereby leading to data driven proposals and recommendations.

1.7 Justification

There are various benefits for designing this model within the set out framework. Firstly, having views that are represented as raw data ensures that the risk of altering the original meaning has been eliminated. The task forces will use the views as a means of supporting and justifying why a certain proposal or recommendation was made. The views that the users submitted will be annotated with different types of tags, e.g topic tags, thus providing a means of querying views. For example, a user can query views that address a particular topic.

1.8 Contribution

This research contributes towards practice by:

- Enabling task forces automatically process submitted views using various text mining techniques.
- Improving the acceptance and adoption of recommendations of task forces.

1.9 Assumptions

Our research is based on the assumption that public task forces will continue to receive submissions in soft form.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

In this section, we describe how textual data is processed and analyzed within similar contexts and the different text mining techniques that are applied. We also describe the different text mining techniques and the algorithms applied. Lastly, we detail previous work done in text mining field and the research gap.

2.2 Process Flow Model

Figure 1 shows a high level description of the different entities and processes that relate to public task forces and how they interact. Our study addressed the feedback processing and analysis of submitted views.

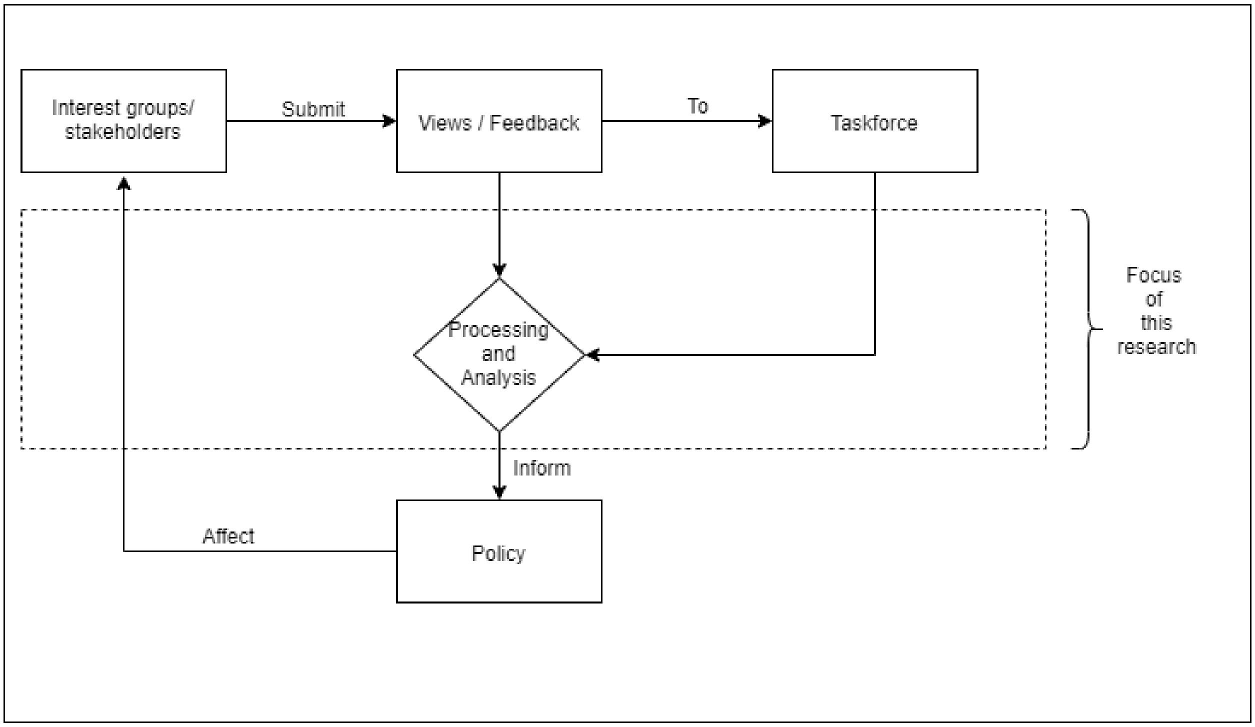


Figure 1: Process Flow Model

Feedback provided to public task forces is usually in form of written text. This unstructured data contains precious information that require to be extracted. The process of extracting this valuable information is what is referred to as text mining. In this section, we will discuss the various techniques that can be applied to perform text mining on general textual data, namely sentimental analysis, named entity recognition, term frequency – inverse document frequency and topic modeling.

2.3 Sentiment Analysis

Sentiment analysis (SA) assesses the emotional reaction or attitude of a person towards an issue. SA uses different tools, methods and techniques to understand people’s opinion about something. In other words, SA tries to mine people’s opinion and is usually considered a natural language processing task. Different levels of granularity have been applied when performing SA starting from document level, down to sentence level and lastly at the phrase level (Birjali et al., 2016, p. 2).

Research in the field of sentiment analysis in the recent past has focused on detecting polarity of unstructured data. The aim has mainly been to perform binary classification to categorize text as either positive or negative. Other research work has focused on analyzing intensity of emotion in a specified piece of text.

2.4 Named Entity Recognition (NER)

NER identifies and classifies named entities contained in unstructured text into pre-defined categories. Examples of categories include person names, locations, time, money, quantities etc. A named entity is a term that represent objects whose context is unique and informative within the text. A named entity represents real-world objects like places, organizations and persons. NER is also referred to as entity chunking/extraction. Consider the following example:

Peter bought 200 laptops from Apple in 2008.

A NER model will produce an output similar to the text below.

[Peter]$_{\textit{Person}}$ *bought 200 laptops from [Apple.]*$_{\textit{Organization}}$ *in [2008]*$_{\textit{Time}}$*.*

A person name, a company name and a temporal expression have been detected.

2.5 Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF counts the occurrence frequency of a term in a document and evaluates its relative importance to the collection of documents. Some words that appear more frequently within the set of documents may not be useful, for example ‘the’, ‘and’, ‘is’, ‘a’. However, some words that appear more frequently within a document may be useful in classifying the document.

The more frequently a word occurs within a document, the more important that word is considered to be. However, we need to offset this by how frequently that word occurs in the entire set of documents. In other words, we normalize the importance of that word according to how rampant that word is in the entire set of documents. TF-IDF assigns weights to words and measures relevance as opposed to frequency. In other words, instead of frequency being associated with a term, a weight is associated with that term instead. Search engines use TF-IDF to rank a particular document’s relevance to a given user search query. TF-IDF are also used when building classifier models, for example those that classify call center classifications.

TF-IDF is a combination of two terms, that is, normalized Term Frequency (TF) and Inverse Document Frequency (IDF).

TF(t) = Count of term t in a document / Total number of terms in the document
IDF(t) = log (Total count of documents / Number of documents with term t in it).

Thus: *TF-IDF = TF * IDF*

TF-IDF is beneficial in identifying sets of words that are discriminative for documents in a collection. Conversely, it has a disadvantage in that it completely ignores syntax information and possibly excludes words that have the same meaning (semantic information), such as hotel and motel.

2.6 Topic Modeling

Topic modeling is a technique applied to identify major themes in a large collection of documents or a large corpus of text. The general idea is, given a set of documents, one can infer what the major topics are, based on the words that are present in the documents. Topic modeling answers three basic questions: Given a document, what is the likelihood that a selected topic is covered in that document? And given a selected topic, what is the likelihood that a particular word would be used for that topic? Topic modeling algorithm is iterative and involves multiple iterations of taking guesses about what words are associated with what topics and the algorithm settles on the best set of combinations of words for topics and topics for documents (Merrett, 2015).

Topic modeling is an unsupervised machine learning approach that processes a set of documents to detect words patterns and easily clusters those patterns to best describe the documents. Topic modeling does not require predefined labels to be provided to the algorithm. This is the unsupervised learning process. It will automatically group recurring words and patterns into different clusters (Pascual, 2019). Topic modeling can be thought of as a form of dimensionality reduction where instead of representing text in word feature space, the text is represented it in its topic feature space. The three key elements of a topic model are a document, a topic, and a word (term). A document is a mixture of topics (Blei et al., 2003). A topic is a theme discussed in one or more documents and is represented as a probability distribution over words (Griffiths and Steyvers et al., 2007).

A topic model is used to discover hidden semantic structures in textual data. Assuming that a document is addressing a certain topic, then it is normal to expect some words to appear in that document more often than others. For example, for a document about tourism, hotel, accommodation and meals will appear more frequently. For a document about transport, car, bicycle and airplane are expected to appear in that document more than hotel, accommodation or meals. In both documents however, “is” and “the” are expected to occur frequently in both documents. The topics that are generated by a topic model are represented as clusters of words. A topic model is used to represent this notion in a mathematical framework which leads to discovery of hidden semantics based on word statistics within the set of documents (“Topic Model”, 2019).

A topic model is probabilistic because it applies a probability assumption when processing documents. The model assumes that documents are formed using a generative process of picking a probabilistic distribution over a set of topics. The main assumption is that to form a new document, a probabilistic distribution of topics is chosen. “Then for each word a random topic is chosen according to this distribution and a word is drawn from the topic” (Blei et al., 2003). Topic modeling aims to reverse this process of document generation using a statistical framework such that the topics that were used to generate the documents can be inferred (Amin, 2016).

The process of inferring topics involves counting of words and phrases that appear more often together and then grouping them to into clusters. A topic model outputs these clusters and by looking at them, one is able to deduce what each set of words or phrases is talking about (Pascual, 2019). Topic modeling process can be illustrated using the figure below:

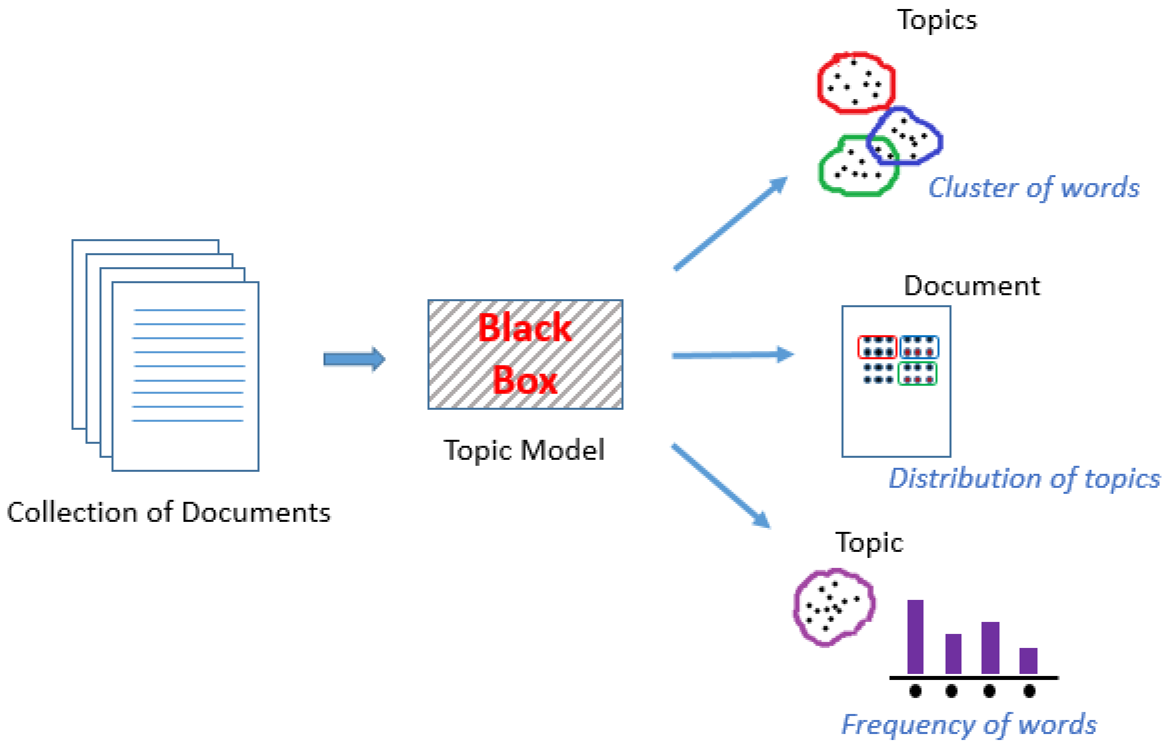


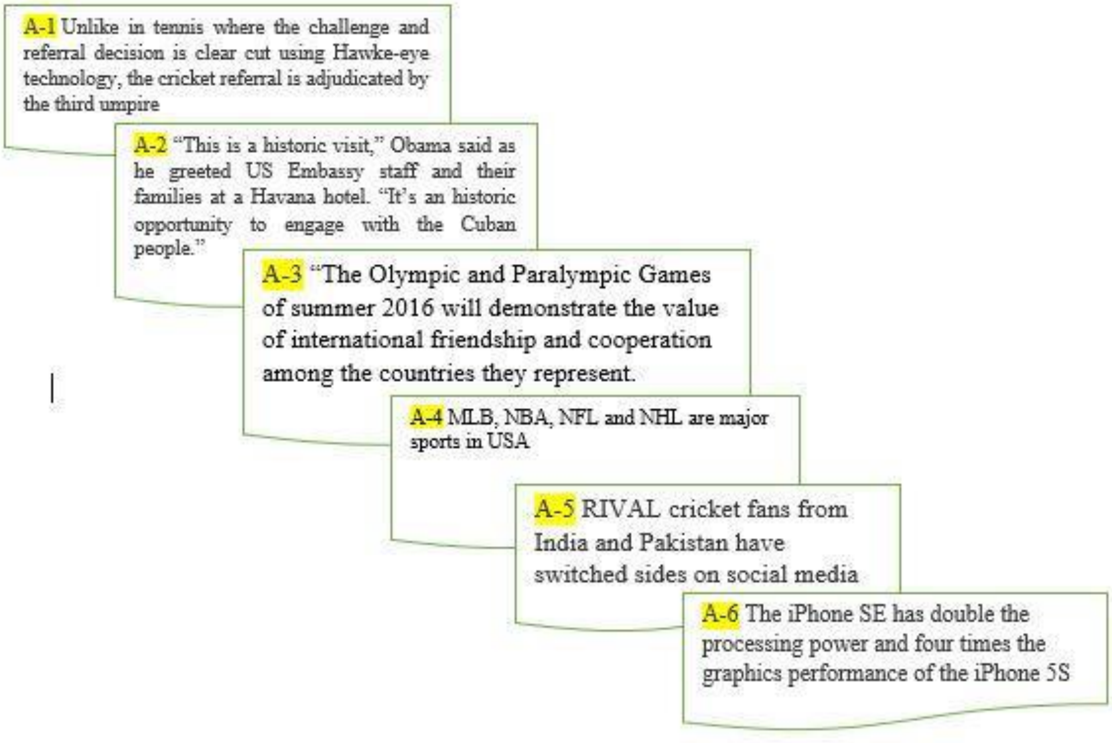
Image source: *https://www.analyticsvidhya.com/blog/2018/10/stepwise-guide-topic-modeling-latent-semantic-analysis/*

Figure 2: Topic modeling framework

We will first describe how topic modeling works then we will describe Latent Dirichlet Allocation topic modeling algorithm.

2.6.1 How does a topic model work?

A topic model accepts a collection of documents or textual data as the input to the model. The user then specifies a numeric parameter, **k**, that represents the number of topics that the set of documents should be classified into. The model then processes the documents and outputs a set of keywords for each for topic. By looking at the keywords for each cluster, the user is able to infer what topic is being addressed by these keywords. Each keyword associated with a topic has a weight which is a measure of how important that keyword is to a particular topic. Consider a set of articles shown in Figure 3.



A-1 Unlike in tennis where the challenge and referral decision is clear cut using Hawke-eye technology, the cricket referral is adjudicated by the third umpire

A-2 “This is a historic visit,” Obama said as he greeted US Embassy staff and their families at a Havana hotel. “It’s an historic opportunity to engage with the Cuban people.”

A-3 “The Olympic and Paralympic Games of summer 2016 will demonstrate the value of international friendship and cooperation among the countries they represent.

A-4 MLB, NBA, NFL and NHL are major sports in USA

A-5 RIVAL cricket fans from India and Pakistan have switched sides on social media

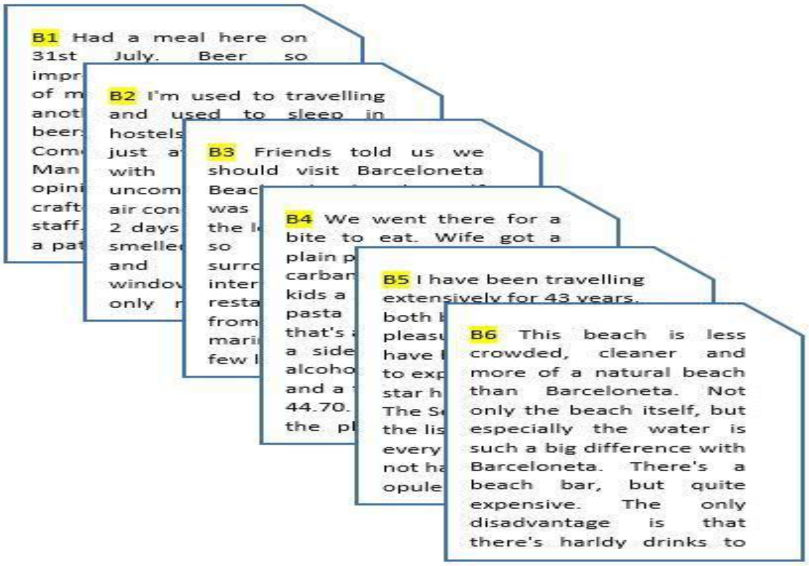
A-6 The iPhone SE has double the processing power and four times the graphics performance of the iPhone 5S

Image source: Veer Reddy Sathi and Jai Simha Ramanujapura (July 2016).

Figure 3: Sample news articles

If the user was to set the number of topics to be 3, then the topics contained in the articles may be broadly classified as Technology, Sports and foreign relations Looking at article A-4, we can conclude that the article is entirely about sports. Also, the contribution of article A-4 to the topic about sports is very high. For article A-3, the article refers to both sports and foreign relations. We can estimate that the article contributes in equal proportion to the topics of sports and foreign relations. Article A-5 on the other hand describes sports rivalry between India and Pakistan. It touches on social media which contributes to the topic on technology. It also mentions cricket which contributes to the topic on sports. By mentioning India and Pakistan and the rivalry thereof, the article also contributes to the foreign relations topic. The contribution of article A-5 can be in different portions to the three listed topics (Sathi, 2016). In the example above, the topic interpretability is high meaning that anyone looking at the topics can easily decipher them.

Figure 4 shows different reviews about hotels, restaurants and beaches from a website. We keep the number of topics to be 3 just like in the previous example.



B1 Had a meal here on
31st July. Beer so
impr
of m
anot
beer
Com
Man
opini
craft
staff.
a pat

B2 I'm used to travelling
and used to sleep in
hostels
just a
with
uncom
air con
2 days
smelle
and
windo
only r

B3 Friends told us we
should visit Barceloneta
Beac
was
the l
so
surro
inter
resta
from
mari
few l

B4 We went there for a
bite to eat. Wife got a
plain p
carban
kids a
pasta
that's a
a side
alcoho
and a
44.70.
the pl

B5 I have been travelling
extensively for 43 years.
both
pleasu
have
to exp
star h
The S
the lis
every
not ha
opule

B6 This beach is less
crowded, cleaner and
more of a natural beach
than Barceloneta. Not
only the beach itself, but
especially the water is
such a big difference with
Barceloneta. There's a
beach bar, but quite
expensive. The only
disadvantage is that
there's harldy drinks to

Image source: Veer Reddy Sathi and Jai Simha Ramanujapura (July 2016)

Figure 4: Sample reviews data

Topic_1	Topic_2	Topic_3
Spaghetti	Stay	Beach
Meal	Make	Walk
Plain	People	Beer
Hot	Cheap	Food
Staff	Sincerely	Barcelonata

Table 1: Topics with low topic interpretability

Table 1 shows the different topics inferred from the data. The three topics are not easily interpretable. Topic_1 is likely to be about food because of the words restaurant, meal and spaghetti and staff that made people happy. Topic_2 could be discussing accommodation but it is difficult to interpret the word sincerely because it could mean a salutation in a review or that the service offered was done so sincerely. Topic_3 could be interpreted as walking on the beach while having some food and beer. The difficulty in interpreting the topics could be due to the topic model parameters. Also the word Barcelonata is difficult to interpret as it refers to only one location of a beach. It is possible that the topics could have been clearer had the number of topics been higher than 3. The parameter k, i.e the number of topics, plays a critical role in determining how interpretable the topics generated by a topic model are (Sathi et al., 2016).

2.6.2 How is a topic model evaluated?

A topic model can be evaluated both quantitatively and qualitatively. Qualitative evaluation involves assessing the quality of the topics generated by the model. Do the generated topics represent the data? To qualitatively evaluate a model, the output of the model is presented to a human being for assessment on how well the topics identified represent the content of the documents. Visualizing the topics is also an effective way to evaluate topic models qualitatively. To quantitatively evaluate topic models, two main metrics are used; perplexity and coherence score

Perplexity

Perplexity, in the normal usage, refers to the inability to understand something. In topic modeling, it is a measure of how well a probability distribution predicts a collection of documents. A model with a lower perplexity score is considered better than one with a high perplexity score. A model with a low perplexity score is considered a generalized model that will be able to perform fairly well when presented with new unseen document. To calculate perplexity, the dataset is split into training and testing sets. In the topic modeling scenario, the test set is a set of documents that have not previously been seen by the topic model. “Perplexity measures the log-likelihood of a test set” (Sathi et al., 2016).

Perplexity measures how much a model is surprised to see new data. In other words, how probable is some unseen data when it is presented to a model that was earlier trained. It measures how well a model predicts a sample. Perplexity score is generally used to compare models with different values of k. Usually, the model with the lowest perplexity score is picked as it is considered more generalized. However, optimizing perplexity does not always yield to more human interpretable topics (Kapadia, 2019).

Topic Coherence

Topic coherence measures the degree of similarity between the words that are most relevant for each topic as generated by a topic model. It helps differentiate between topics that are interpretable semantically from those that are purely as an outcome of statistical inference. “A set of statements or facts is said to be coherent if they support each other” (Kapadia, 2019). Coherence score measures different topic models based on how human-interpretable the topics are. A model with a high coherence score is considered as better. There are different coherence measures as listed below. More details about these coherence measures are contained in Röder et al., 2015.

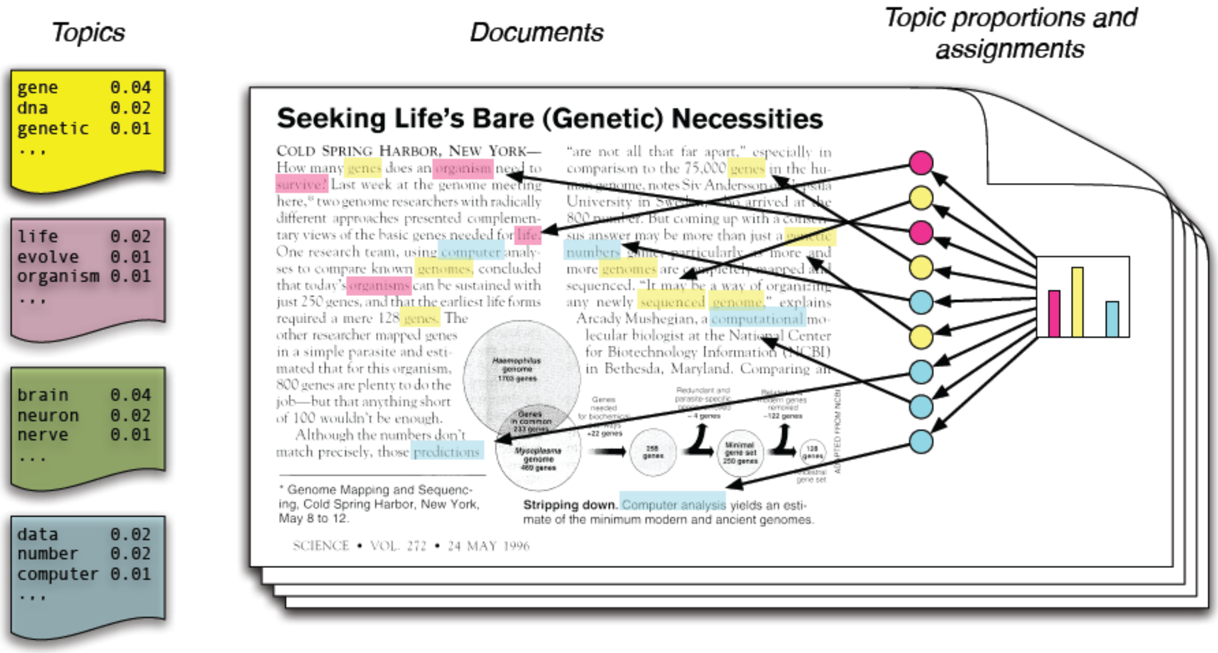
2.6.3 Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model that assumes a Dirichlet distribution prior over hidden topics. Details of the LDA algorithm are given by Blei et al., 2003. The main assumption underlying LDA, is that a document contains multiple themes. These themes in a document and the words that make up these themes are Dirichlet distributions. “LDA is a three-level hierarchical Bayesian model where each item of a collection of text is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities” (Blei et al., 2003). The probabilities of topics represent and describe the document (Daud et al., 2010). In addition, every topic is described by a probability distribution of words. LDA classifies a dataset into K topics. Each of the topics is represented by the most relevant N words. Described in another way, the documents in the dataset are a random combination of hidden topics and each topic is a random combination of a set of terms represented as Dirichlet distribution. LDA does not output explicit topics but rather it outputs a combination of top N keywords together with the relative importance of each keyword. The topics are latent meaning that they have to be inferred given the top N relevant keywords. LDA is an unsupervised machine learning technique that discovers the themes in a corpus. It is thus suited to even noisy data because of its unsupervised nature (Zhai et al., 2012).

LDA algorithm identifies words that are contained in a corpus of text and statistically ranks each word into a topic. One term can contribute to multiple topics albeit in varying levels of importance. That is, the importance of a term can vary from topic to topic. The output of a topic model is such that each word is represented together with the probability that the term is associated with specific topics. The probability is a measure of how important a word is to a particular topic. The higher the probability, the higher the word's importance to a topic. Additionally, LDA considers that a document is a collection of topics in differing proportions. In other words, each document can be represented as a probability distribution over a number of topics. This means that a single document can address multiple topics even though the contribution to the topics might vary. The proportion of one topic in a document is likely to be different from the proportion of the other topic in that same document. LDA algorithm takes as an input parameter the number of topics into which the documents should be categorized into. What LDA does is to rearrange the topic distributions for the documents and map the keywords distributions into topics until convergence to generate the best topic keyword combinations.

LDA is based on two main assumptions: First, that similar words are used for same or similar topics; Second, that a document or collection of documents discuss multiple topics which can be represented as a statistical distribution. LDA aims to summarize a document into a set of topics which describe the document as much as possible. LDA algorithm starts by assigning random topics to an arrangement of terms (n-grams). This is in consistency with the assumption that documents are generated using an arrangement of words and that topics are determined by these arrangements. LDA treats documents as a bag of words thus completely ignoring syntactical meaning. Each word can separately be linked to different topics with different measures of relevance represented by the probability of the word being associated with a particular topic. LDA determines the composition of topics within a document (Pascual, 2019).

Figure 5 shows the assumptions made by LDA about topics and documents.


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Image source: Blei D. M (2012) Probabilistic topic models. Communications of the ACM. 55(4), 77-84

Figure 5: Topic Model

LDA Assumptions

LDA algorithm, being a generative model, relies on some a priori assumptions:

- Order of words in the documents is not important.
- Position or order of a document within a dataset is not important.
- The number of topics is known in advance
- One word can contribute to multiple topics
- Each document is treated as a combination of k hidden (latent) topics
- Each topic can be represented as a distribution over a set of words

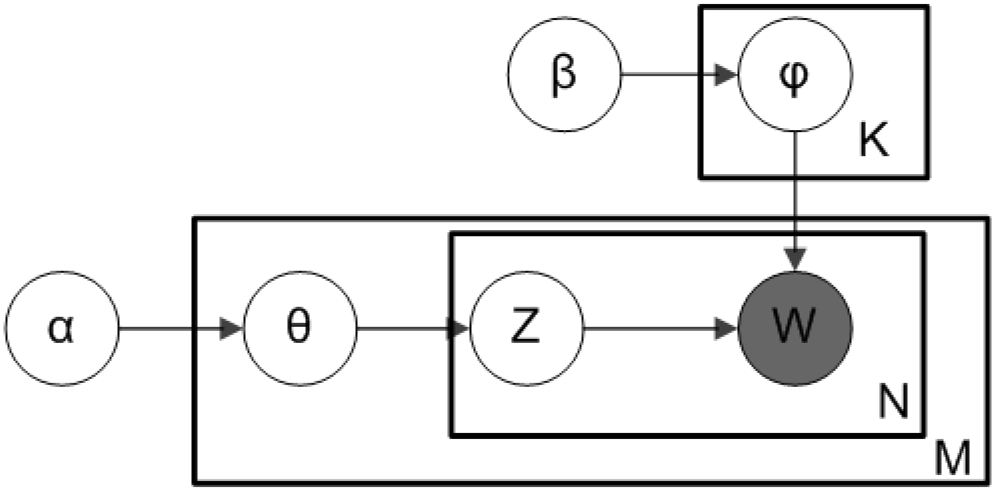


Figure 6: Graphical model for LDA

M: Count of documents in the collection

N: Total count of word tokens in the documents

K: Number of topics to infer from the documents

A: Hyper-parameter for per-document topic distribution

B: Hyper-parameter for per-topic word distribution

Φ: Word distribution for topics **k**

Θ: Topic distribution for the i^{th} document

$\mathbf{Z_{(i,\,j)}}$**:** Topic assignment for **w(i,j)** (**j**th word in **i**th document)

The outer box represents documents while the inner plates represent words contained in a document. **α** and **β** hyper-parameters are set during the single process of generating a corpus from the collection of documents.

<u>LDA Steps</u>

Assuming we have set the value of **k**, during training the LDA algorithm operates as below:

- STEP 1: Each word in each document is assigned randomly to one of the ***k*** topics.
- STEP 2: For each word ***w*** in document ***d*** and for each topic ***t***, compute:
    - ***p(topic t | document d):*** Proportion of terms in ***d*** currently associated to ***t.***
    - ***p(word w | topic t)***: Proportion of associations to ***t*** over all documents containing ***w***.

- STEP 3: Reassign ***w*** to a new topic ***t,*** based on ***p(word w | topic t) * p(topic t | document d).***
- STEP 4: Repeat step 2 and 3 until convergence.

2.7 Why topic modeling and LDA?

Given the characteristics of text data we will be processing, topic modeling is the text mining technique best suited to process public participation data. This is because of the following reasons:

- The content of the submissions are not structured in a standardized way. Different stakeholders make submissions in a format that is convenient to them
- Some submissions span multiple pages
- Topic modelling allows removing of stop words meaning commonly repeated words can be ignored and only leave words that are useful in inferring the theme of the document
- The submissions do not have labels which makes them a good candidate for unsupervised learning techniques like clustering

We chose LDA over LSA as the topic modeling algorithm because of the following reasons:

- LSA is slow on large corpora. The amount of text to be processed by our model will be very large.
- LSA is quicker to train its accuracy is lower compared to LDA.

2.8 Previous Work

In their work, Ujang Fahmi (2019) used a combination of “Topic Modelling, Social Network Analysis and Discourse Analysis” to extract discourse from data in the form of text. Their study was categorized as Corpus Assisted Discourse. Discourse analysis began with extracting topics from tweets under the #JogjaOraDidol hashtag from Twitter. They then applied LDA algorithm on the text that had undergone pre-processing stages. The text that has been cleaned later became bags of words without removing the identity (id) of the document (tweet) so that 179,134 terms were identified. The topics obtained were classified to form separate themes. Classification was based on the same reference data from two other data sources, namely media and interview results. In other words, discourse analysis was referred to as intertextuality, that is, a method of finding the context of a text based on other texts. Quinn et al. (2010) referred to this phenomenom as intrasemantic-validity where each topic obtained was examined qualitatively. In the context of using the #JogjaOraDidol hashtag on Twitter the researcher agreed with Koltsova et al. (2013) who said that social media (blogs and microblogs) can be utilized in the policy making process, especially in the agenda setting stage. In this agenda setting stage, policy makers need to know and have evidence about things that concern the public.

Tong and Zhang (2016) analyzed Wikipedia articles and built a topic model solution on searching, exploring and recommending articles with the perspective focused on topics. They also analyzed users' tweets thus setting up a topic model that provided a full research and analysis of the users' interests. Over 200,000 Wikipedia articles were downloaded as simplified Wikipedia English version database backups from Wikipedia Foundation. An LDA model was developed using R package called *topicmodels*. They computed the distribution coverage of each topic for each document which showed the extent of association of a document to each topic. To find the document the user was looking for, they used the topic distribution as a means of searching and exploring between topics. They applied Jensen-Shannon divergence to evaluate the distance between each document distribution. They also built a recommender system that would sort documents based on their similarities as represented by their distributions. The shorter the distance between the subject article and another article, the more related the two articles. They also applied LDA and topic modelling on Twitter users where a detailed model of Twitter users' personality and preference would be inferred. They used *twitteR* package to analyze 10,000 valid Twitter users. A valid Twitter user was deemed to be one whose profile is unprotected, has at least 100 tweets and the user must use English as a major language when tweeting. Each user tweet was treated as an article (document). They concluded that the model could be a used as a tool for conducting social and business research.

Evangelopoulos and Visinescu (2012) presented two case studies that analyzed how unstructured data could give politicians the ability to efficiently interpret feedback from citizens. The two case studies used E-democracy data. The first case study involved analyzing 902 SMS messages that had been sent to Barack Obama by African citizens. The second case analyzed 1,481 ideas submitted to the U.S. Department of Homeland Security. Factors that represented main suggestions for better and efficient government were extracted from these submitted ideas. To extract the articulated factors in both case studies, the authors employed a methodological twist of LSA which involved rotating the corresponding mathematical components. For both case studies, the authors demonstrated the usefulness of LSA in processing short and unstructured text and also provided a means of tying up the loose ends in political discourse between the citizens and their leaders.

2.9 Research Gap

The existing means of processing and analyzing submissions is cumbersome because the task force members have to examine their content by hand. Task forces tend to design their own means of analyzing data for the oral and written submissions. Transcription is normally done for oral submission. Written submissions are normally presented as either in hard or soft formats. The task force members have to read the physical documents to be able to synthesize the content therein. Task forces receive massive amounts of documents that are beyond their processing capability leading to some of the submissions never being reviewed or considered. This scenario is ripe for developing of an automated standardized means of processing and analyzing all submissions in an efficient manner. Text mining techniques when properly applied to public task forces' data processing and analysis will make a contribution to public policy making in ensuring that the citizen feedback is understood.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This section describes the activities and steps we undertook in collecting data, the techniques applied to analyze data, how the model was designed, the experiments that were set up and finally how we validated the model to ensure it provided a solution to the problem as described in the problem statement section earlier. All the activities were designed so as to achieve the earlier stated objectives that will also enable us to answer the research question.

3.2 Conceptual Framework

Number of topics, alpha and beta parameters are the independent variables that impact the quality and number of topics generated from the submissions. Alpha and beta parameters affect the sparsity of topics. Alpha hyper-parameter affects the topic distribution in each document while the beta hyper-parameter affects the word distribution for each topic. Parameter k, determines the quality and number of topics that will be generated by the topic model.

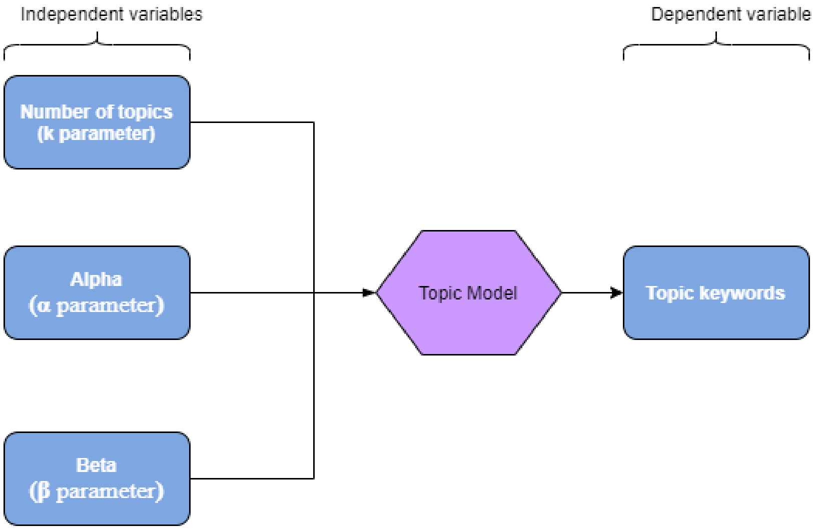


Figure 7: Conceptual Framework

3.3 Research Approach

We chose Experiment as the means to answer our research question. Figure 8 shows how the research methodology contributed to answering the research question by achieving the objectives of this study. We conducted a systematic literature study to identify the best method to process public feedback. We then conducted an experiment where we trained and optimized the model using a training dataset of fifteen (15) submissions. We then processed seven (7) new submissions through the model and also gave them to a human expert who identified the topics that were being addressed. The topics generated by the models were then compared to those identified by the human expert as part of the model validation step.

There are various ways of qualitatively evaluating topic models but we settled on the interpretability of topics as an evaluation method. We used topic coherence as a metric to evaluate the performance of the topic models quantitatively. Coherence scores for different topic models was calculated and the results used to determine the model that best picks the topics within the text. The experiment would help us answer the research question earlier stated on how public task forces process the large amount of feedback in a consistent and objective manner. Developing the model would help us achieve the first objective (*Developed a model to process feedback received by public task forces*) while engaging the expert enabled the researcher to achieve the second objective (*validate the developed model to ensure that feedback is processed effectively*).

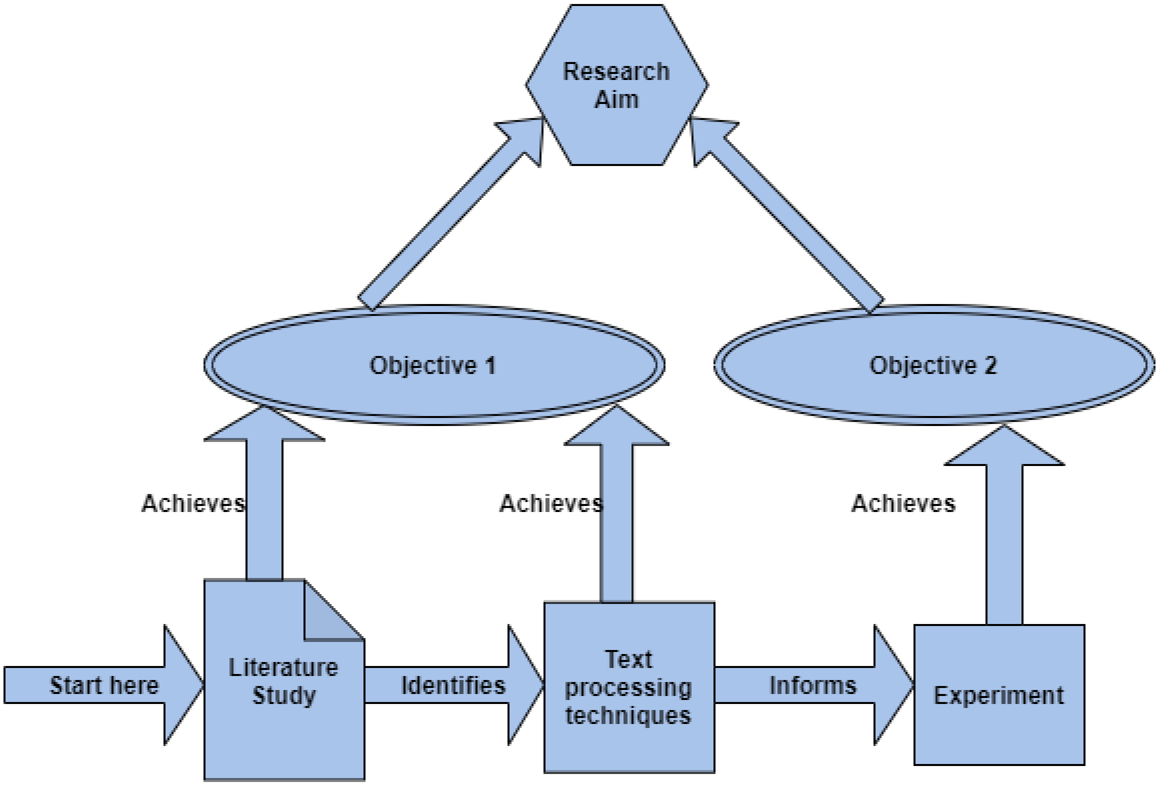


Figure 8: Research Approach

3.4 The Experiment

To be able to achieve both objectives, we conducted an experiment where we were able to manipulate the independent variables and consequently measure the effect on the dependent variable, that is, the topics. The experiment enabled us to assess the cause-effect relationship between the independent and dependent variables (Sathi et al., 2016).

In our work, we developed model to process public participation feedback where we used a human expert to evaluate the quality of the topic models. The variables that we vary are 1) the number of topics K, 2) alpha and 3) beta parameter. We verified and validated the topics generated from the submissions by assigning them to a human expert for qualitative evaluation.

The steps what we undertook to develop the model are highlighted below. Later, we describe each step in detail in the subsequent sub-sections.

1. Data collection
2. Exploratory analysis

3. Model design
    i. Data pre-processing
    ii. Base model design
    iii. Hyper-parameters tuning
    iv. Final model design
    v. Model validation

3.4.1 Data Collection

The study used publicly available submissions to a task force on ICT procurement in form of PDF and Microsoft Word documents. A total of twenty-two (22) submissions were used. Each document represents a submission by a single entity. The task force had listed specific questions that needed to be addressed in each submission. However, the submissions were not in any standard structure of writing. Some chose to address all the questions in prose form, some addressed each question at a time, others rephrased the questions, some addressed a subset of the questions and even lastly, some submissions went ahead to address issues that had not been raised by the questions. The lack of a standardized document structure is a key characteristic of submissions that are received by task forces. While the task force tries to structure the questions in a standard way, more often than not, they receive submissions in varied structures. There is no strict insistence on the format of the submissions as long as the submission addresses the issues that the task force is dealing with.

We split the dataset into training and testing sets. The first set, containing fifteen (15) submissions, was used to train the LDA model. The remaining seven (7) submissions were used to validate the model. To facilitate model validation, the seven (7) submissions were tagged by a procurement expert who was able to generate a list of topics (labels), that each of the submission was addressing.

3.4.2 Exploratory analysis

In this step, we performed exploratory data analysis where we identified the characteristics of the training data as well and computed the descriptive statistics. The purpose of this step was to get a basic overview and understanding of the unprocessed data that we were about to use to design and validate the model. We plotted a histogram for unigrams, bigrams and trigrams showing the frequencies for each n-gram. We compared the histograms before removing the stop words and the histograms after removing the stop words and lemmatizing the words. Lemmatization usually removes any inflection a word and returns the base form of a word that belongs to the language. It ignores any tense in the word. For example, if the word “ate” is presented, the lemmatized form of the word would be “eat”. Histograms enabled us to identify frequently occurring words that would be meaningless. An example of such words are the “country name”, “name of task force” and “date of submission”.

3.4.3 Model Design

This section describes in detail the steps that the authors followed in designing the model. We implemented the model using Gensim python package. As described in the literature review section, we used topic modeling as the text mining technique because we wanted to be able to automatically identify the topics that are contained in the submitted views. LDA topic modeling algorithm was applied. We will describe sequentially the different steps we undertook in designing the model.

3.4.3.1 Data Pre-processing

Having understood the data by exploring it, we found there were a number of data pre-processing activities that we needed to conduct. These are:

1. Remove newline characters
2. Remove all email addresses
3. Tokenize words and clean up. Sentences were split into their constituent words and removing all punctuations. All the words were converted into lowercase and those with fewer than three characters were removed.
4. Remove stop words. Stop words are commonly used word that have no value in describing the text since they appear so commonly in the text. Examples of common stop words are “the”, “a”, “an”, “in”. Gensim library has its own set of stop words but in addition we used stop words from the NLTK python library.
5. The next step was to lemmatize the remaining words. We then run the text into a spellchecker to remove words that have been misspelt. For further reduction, we then

removed any word occurring in more than a 75% of the total number of documents. This cut-off percentage was set through experimentation and the purpose was to remove words that are too common in the corpus and thus would be of little value in identifying different topics within the submissions.

6. Removed names of individuals or organizations who made the submissions as well as the task force name as part of anonymization.

3.4.3.2 Base Model Design

The next step was to train an LDA model using the default LDA algorithm settings. The purpose of this step was to give us a baseline against which we will compare the results after we tune the hyper-parameters of the model.

LDA algorithm consumes numeric values. So we needed to assign a unique numeric id to all of the words in the cleaned text. To achieve this, the cleaned and tokenized text was converted into bag of words by assigning a numeric value to each token and counting the number of occurences of that word in the corpus. A dictionary is used to represent this transformation where the dictionary key is the word and value is the number of instances of that word in the entire corpus. Consider the example below:

Dictionary: {'all': 0, 'cow': 1, 'eat': 2, 'grass': 3, 'every': 4, 'day': 5 .....

> id for word 'all' is 0, id for word 'cow' is 1 and so on.

Corpus: [(0, 3), (1, 5), (2, 9), (3, 7), (4, 3), (5, 2), (6, 5), (7, 1), (8, 21),...

> (0, 3) means that in the first document, word 'all' occurs three times. Word 'cow' occurs five times etc.

An LDA topic model takes the following inputs:

- Dictionary. Represents unique identifier for each token.
- Corpus. A representation of the unique word id and its frequency within a specific document.
- Alpha and beta as hyper-parameters. They both default to 1.0 divided by the value of **k.**

- Chunk_size. This determines how many documents are processed at a single pass during model training.
- Passes. Determines how frequently the model is trained on the entire set of data.

The perplexity and coherence score of the base model were computed. These values would act as baselines against which we evaluated the process of hyper-tuning the parameters and the effects on the model performance and output. We compared the coherence and perplexity scores of the base line model against those of the tuned models to determine the optimal value of the hyper-parameters.

3.4.3.3 Hyper-parameter tuning

Hyper-parameters generally are settings that are tuned before training of the algorithm. For the LDA model they are k, alpha and beta parameters. Model parameters, on the other hand, are the values learnt and set by a model during training. An example is the importance of each word within a given topic.

We focused on tuning three LDA model hyper-parameters. These are, the number of topics **k**, alpha **α** and beta **β** values. The sensitivity of the model towards a change of each of the hyper-parameter was evaluated. To tune each parameter, the values of the other two parameters were held constant.

To pick the optimal number of topics, we trained models with different values of k and evaluated their coherence scores. The model to pick is the one that gave the highest coherence score before flattening or before a consistent drop. The value of **k** for that picked model is the optimal number of topics.

To select the optimal values of alpha and beta, we picked the combination of the values of alpha and beta that yielded the highest coherence score for the selected optimal number of topics. These optimal values of k, alpha and beta yield an improvement over the baseline model score. Next, we trained our final (optimal) model using the optimal hyper-parameters.

3.4.3.4 Final Model Design

We then trained our LDA model using the optimal values of the hyper-parameters. We compared the coherence score of the base model against that of the optimal model. We noted an increase in the coherence score. This optimal model is what will process the unseen data (new submissions) to generate the topic keywords.

We also provided an interface where the user can alter the number of topics parameter. The user will specify the number of topics, represented by **k,** that the words will be sorted into. LDA will be run for different values of k until a good range of topics is found. The appropriateness of a topic range is assessed by evaluating the topics output for each value of k. For example, if assessing the generated topics reveals that they contain words and phrases that are likely to belong to more than one topic, then the algorithm is re-run, this time with a higher value of **k**. If by visualizing the topics, it is discovered that the topics are sparse or concentrated in one quadrant, then the algorithm is re-run with a lower value of **k** (Cvitanic et al., 2016).

3.4.3.5 Model Validation

The purpose of this step is to develop confidence that the model will perform as expected when it will be presented with previously unseen data input. In other words, the model should be able to infer the latent topics contained in a new submission that the model has never processed before. To facilitate this step, we engaged a human expert to label the seven (7) submissions that we had set aside. One by one, each submission was presented to the final model and the keywords generated by the model were compared with the topics that had been identified by the human expert. As an extra step, we developed an LDA model using *Mallet*. Mallet is an industrial toolkit for text mining and natural language processing. We also compared the output of the Gensim and Mallet LDA models.

3.4.4 Model Evaluation

Why and how to evaluate the model?

LDA is a probabilistic topic model which provides latent topic representation of the corpus using keyword-probability pairs. In LDA, the model undergoes an unsupervised training which makes it important to objectively assess how good or bad the model is. Conventionally, to determine if the model has learnt well about the corpus, qualitative and quantitative methods are used. The authors used perplexity and topic coherence as the measures to evaluate the model quantitatively and assessed topic interpretability to evaluate the model qualitatively.

Interactive Visualizing Topics

We used pyLDAvis Python package to visualize the LDA topic model as shown in Figure 12. PyLDAvis is an interactive chart that visually displays topics and related keywords that have been learned from a text corpus. The chart is web-based allowing user interaction and it displays information that has been extracted from an LDA model. The main elements of the chart are:

- Topic circles where size of each circle is proportional to the proportion of a specific topic in the entire corpus
- An interface that allows a user to vary some model parameters
- Red bars that horizontally show count of the times a particular term occurs in documents that contribute to the selected topic.
- Blue bars that horizontally show the total number of times a particular term occurs within the entire corpus.

CHAPTER FOUR

RESULTS

4.1 Introduction

This chapter presents the results of our analysis and the interpretation thereof. The purpose of this study was to develop a model for processing public participation feedback. More specifically, the study aimed at developing a model for processing of feedback received by public task forces in Kenya. This purpose was guided by the following research question.

- How can public task forces process the large amount of feedback in a consistent and objective manner?

The researcher used publicly available submissions to a public task force addressing procurement within the ICT sector. A total of twenty two (22) submissions were used as input to the model. The submissions were in Portable Document Format (PDF) and Microsoft Word formats with the file name being the institution that made the submission. For ethical reasons, the researchers undertook a step to anonymize the input data. This was done by renaming the files into a format of document_1.pdf, document_2.pdf etc. The other phase of anonymization was by replacing the name of the institution making the submission by organization A, organization B etc. For training and testing the model, fifteen (15) submissions were used. The remaining seven (7) submissions were used for validating the model.

For the model validation stage, seven (7) submissions were presented to a human expert for tagging. The tags assigned to each of these submissions represent the major topics addressed within the submissions. Later, these submissions were run through the model to identify the topics present in them. The output of the model, (topics as represented by keywords), was then compared with the tags earlier assigned by the human expert.

In this chapter, we will present the results of our analysis in line with addressing the above research question.

4.2 Environment

Here, we describe the software tools, hardware specifications and data formats within within which we conducted the experiment to perform topic modeling. The LDA topic model was developed using Gensim python package. Gensim has an inbuilt LDA algorithm used for Topic modeling. Gensim also comes with a Python wrapper for LDA using Mallet. Mallet is developed using JAVA. The wrapper is an interface between Gensim and Mallet that allows both packages to communicate and interact using data files on disk. This is only a python wrapper for Mallet LDA, and thus we needed to install original implementation first and then pass the path to Mallet binary to the wrapper. Gensim uses Variational Bayes sampling method while Mallet uses Gibb's sampling. Variational Bayes sampling method is faster while Gibbs Sampling is more precise.

Item	**Specification**
Processor	Intel® Pentium(R) i7 CPU G3250 @ 3.20GHz × 2
RAM	11.6 GB
Operating System	Ubuntu 18.04 LTS 64bit
Tools	1. Gensim – for topic modeling, 2. Jupyter Notebook - For live code, equations, visualizations and explanatory text 3. Python 3.7 4. Mallet 2.0.8
Programming Languages	Python
Input data format	- Portable document format (PDF) - Microsoft word documents
Input data format not supported	- Images - Nested tables
Java	JDK Version 11.0.6

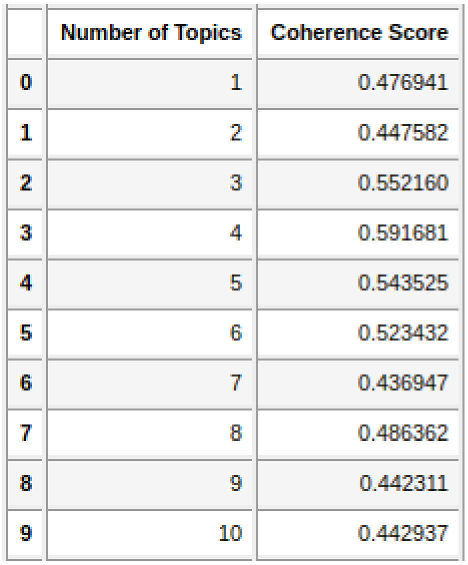
Table 2: Environment setup

4.3 Experiment Results

4.3.1 Quantitative evaluation

For each value of **k** (number of topics), a Gensim LDA model was trained using default hyper-parameter settings of alpha (α) and beta (β).

Table 3 shows coherence scores of the model for different values of **k**. The higher the coherence score, the more coherent the topics are and therefore the better the model. The highest score was achieved when the value of **k** was set as 4.



	Number of Topics	Coherence Score
0	1	0.476941
1	2	0.447582
2	3	0.552160
3	4	0.591681
4	5	0.543525
5	6	0.523432
6	7	0.436947
7	8	0.486362
8	9	0.442311
9	10	0.442937

Table 3: Coherence Scores for different values of K

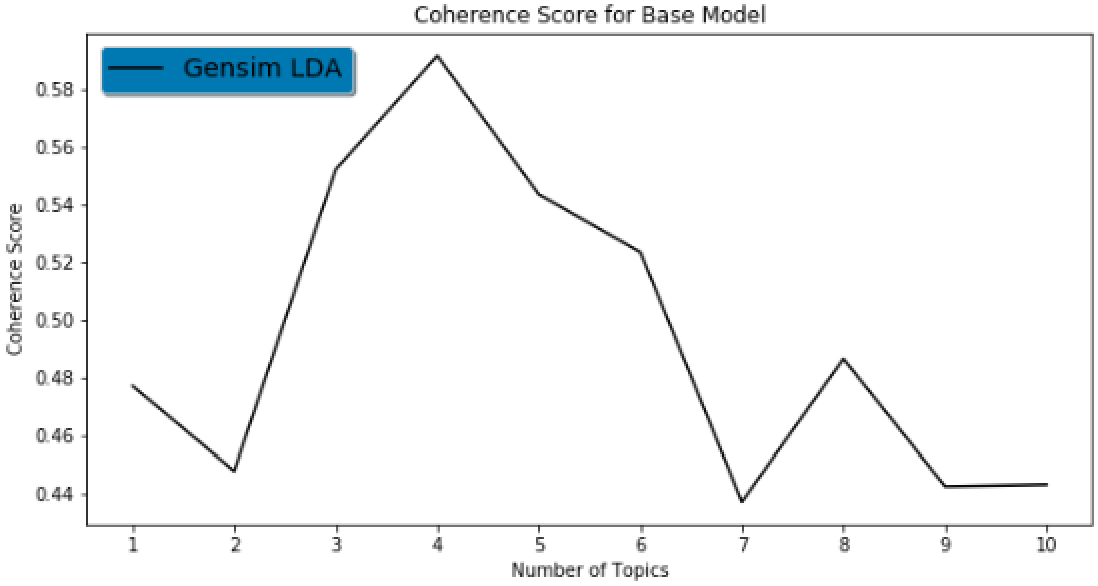


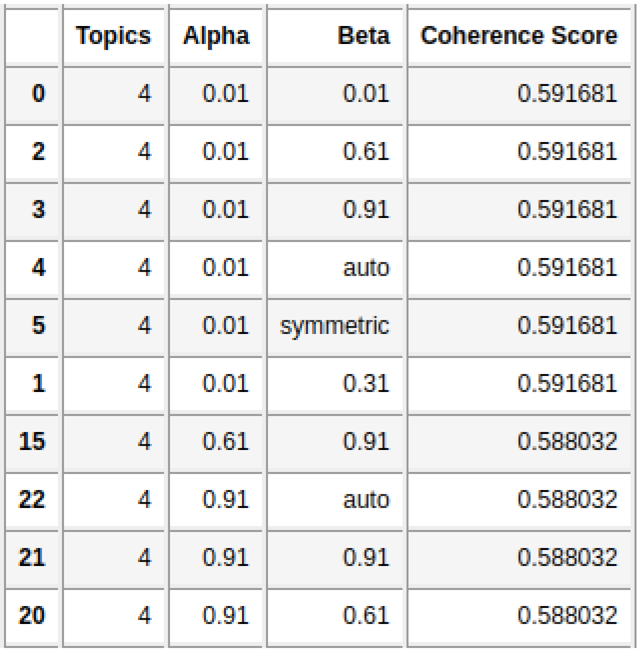
Figure 9: Gensim LDA Coherence Scores

The performance of the LDA model increase rapidly as the value of k changes from 2 to 4. As the value of k changes from 4 upwards, the performance starts to decrease. This is due to the fact that all topics are generally worse (less coherent) as the value of k increases. Another possible reason is that the corpus contains good topics, although few, and thus adjusting the value of k upwards leads to more nonsensical topics as depicted by the low coherence scores for higher values of k. Also considering the fact that task forces are set up to address specific issues, it is highly probable that submissions will be addressing a narrow range of topics. The results show that topic models produce good topics but adjusting the value of k upwards only leads to more incoherent topics and thus decrease the average coherence scores.

4.3.2 Hyper-parameters tuning

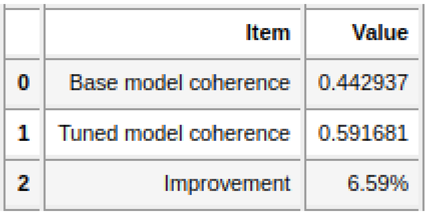
1

Table 4 shows combinations of values of alpha and beta hyper-parameters and the corresponding coherence score given the optimal number of k (4). The combination of alpha and beta that gave the best results for a k value of 4 are 0.01 and 0.01 respectively. Table 5 shows the coherence score of the base Gensim model and that of the model after tuning the hyper parameters. For each of the models, the value of k is set at 4, which is the optimal value identified earlier. An improvement of 6.59% was achieved through tuning alpha and beta parameters as shown in Table 5.



	Topics	Alpha	Beta	Coherence Score
0	4	0.01	0.01	0.591681
2	4	0.01	0.61	0.591681
3	4	0.01	0.91	0.591681
4	4	0.01	auto	0.591681
5	4	0.01	symmetric	0.591681
1	4	0.01	0.31	0.591681
15	4	0.61	0.91	0.588032
22	4	0.91	auto	0.588032
21	4	0.91	0.91	0.588032
20	4	0.91	0.61	0.588032

Table 4: Hyper-parameter tuning

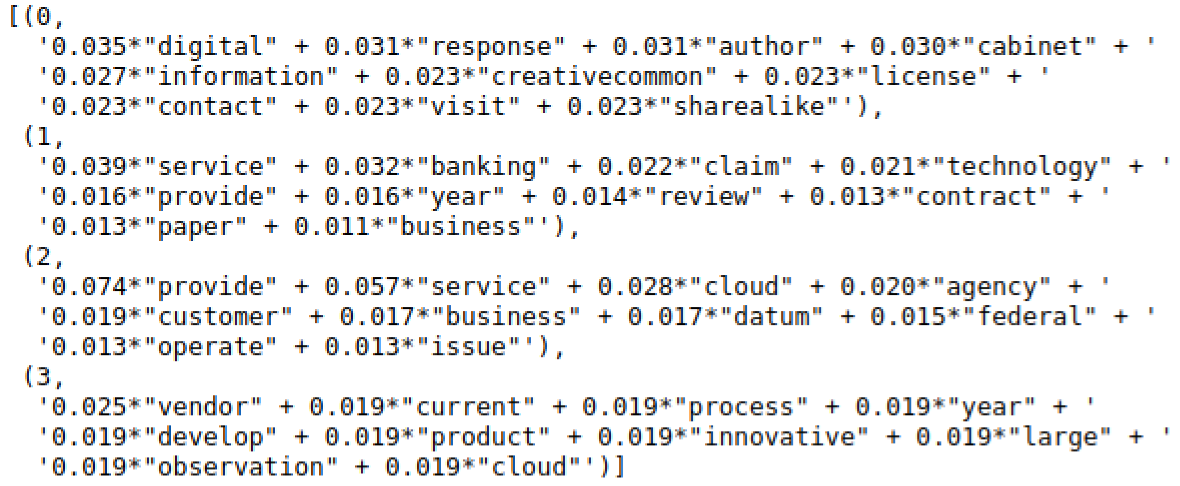


	Item	Value
0	Base model coherence	0.442937
1	Tuned model coherence	0.591681
2	Improvement	6.59%

Table 5: Tuned Gensim LDA model improvement

4.3.3 Top 4 topic keywords

Figure 10 shows the output of the Gensim LDA topic model. It shows the top 10 keywords for the top 4 topics and the weightage (importance) of each keyword.



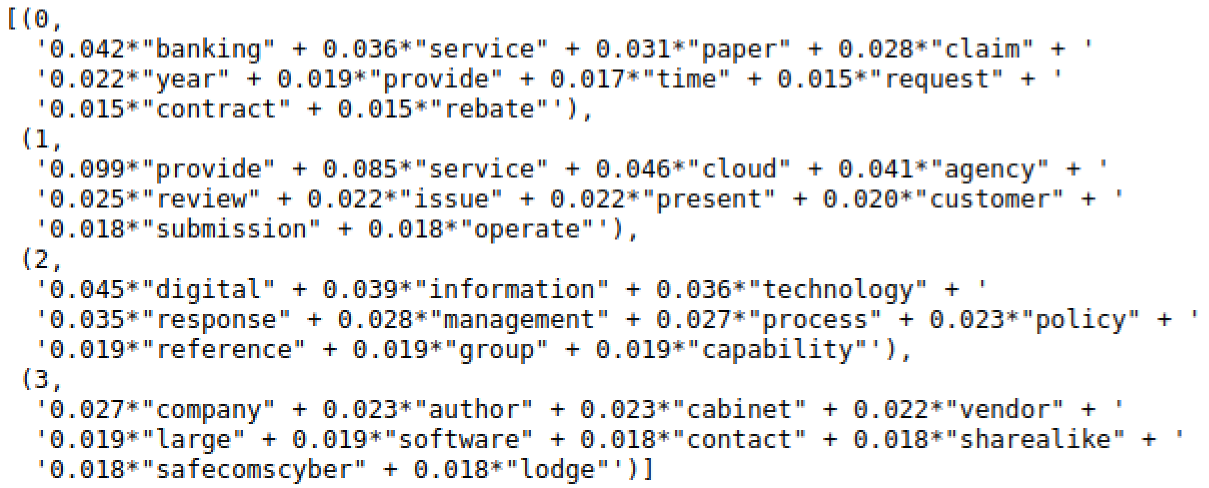
```
[(0,
  '0.035*"digital" + 0.031*"response" + 0.031*"author" + 0.030*"cabinet" + '
  '0.027*"information" + 0.023*"creativecommon" + 0.023*"license" + '
  '0.023*"contact" + 0.023*"visit" + 0.023*"sharealike"'),
 (1,
  '0.039*"service" + 0.032*"banking" + 0.022*"claim" + 0.021*"technology" + '
  '0.016*"provide" + 0.016*"year" + 0.014*"review" + 0.013*"contract" + '
  '0.013*"paper" + 0.011*"business"'),
 (2,
  '0.074*"provide" + 0.057*"service" + 0.028*"cloud" + 0.020*"agency" + '
  '0.019*"customer" + 0.017*"business" + 0.017*"datum" + 0.015*"federal" + '
  '0.013*"operate" + 0.013*"issue"'),
 (3,
  '0.025*"vendor" + 0.019*"current" + 0.019*"process" + 0.019*"year" + '
  '0.019*"develop" + 0.019*"product" + 0.019*"innovative" + 0.019*"large" + '
  '0.019*"observation" + 0.019*"cloud"')]
```

Figure 10: Topic keywords for k topics for Gensim LDA

Topic 0 is a represented as _"0.035*"digital" + 0.031*"response" + 0.031*"author" + 0.030*"cabinet" + "0.027*"information" + 0.023*"creativecommon" + 0.023*"license" + "0.023*"contact" + 0.023*"visit" + 0.023*"sharealike"'._

This implies that the top 10 keywords that are most relevant to this topic are: ‘digital’, ‘response’, ‘author’, ‘cabinet’, ‘information’, ‘creativecommon’, ‘license’, ‘contact’, ‘visit’ and ‘sharealike’, and the weight of ‘digital’ on topic 0 is 0.035. The weights are a reflection of how important or relevant a keyword is to that particular topic. The words are listed according to their topic-word probability, in other words, the first word listed for each topic is most probable while the second word is the one that is secondly most likely to be generated by the respective topic. Words that are less common are positioned further down the list. By analyzing these keywords, we can infer what topic is being referred to, in this case it is about ‘contract management’. Likewise, topic 2 seems to be discussing issues around customer issues. However, it can be clearly seen that the topics are not highly distinctive of each other. The reason is because all the articles address the same topic on ICT procurement. It can also be noted that the same word can contribute to multiple topics. For example, the word ‘provide’ contributes to topics 2 and 3 with importance of 0.016 and 0.074 respectively.

Figure 11 shows the output of the Mallet LDA topic model. It shows the top 10 keywords for the first 4 topics and the weightage of each keyword. We can infer, for example, that topic 0 is addressing about contract execution while topic 2 is addressing issues regarding IT processes and policy management. We observe that topics output by the Mallet model are more coherent compared to those output by the Gensim model.

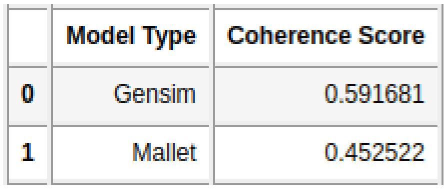


```
[(0,
  '0.042*"banking" + 0.036*"service" + 0.031*"paper" + 0.028*"claim" + '
  '0.022*"year" + 0.019*"provide" + 0.017*"time" + 0.015*"request" + '
  '0.015*"contract" + 0.015*"rebate"'),
 (1,
  '0.099*"provide" + 0.085*"service" + 0.046*"cloud" + 0.041*"agency" + '
  '0.025*"review" + 0.022*"issue" + 0.022*"present" + 0.020*"customer" + '
  '0.018*"submission" + 0.018*"operate"'),
 (2,
  '0.045*"digital" + 0.039*"information" + 0.036*"technology" + '
  '0.035*"response" + 0.028*"management" + 0.027*"process" + 0.023*"policy" + '
  '0.019*"reference" + 0.019*"group" + 0.019*"capability"'),
 (3,
  '0.027*"company" + 0.023*"author" + 0.023*"cabinet" + 0.022*"vendor" + '
  '0.019*"large" + 0.019*"software" + 0.018*"contact" + 0.018*"sharealike" + '
  '0.018*"safecomscyber" + 0.018*"lodge"')]
```

Figure 11: Topic keywords for k topics for Mallet LDA

4.3.4 Gensim versus Mallet LDA models

Table 6 shows the coherence score for Mallet and Gensim models. Gensim model performed better compared to Mallet LDA model given the same dataset.

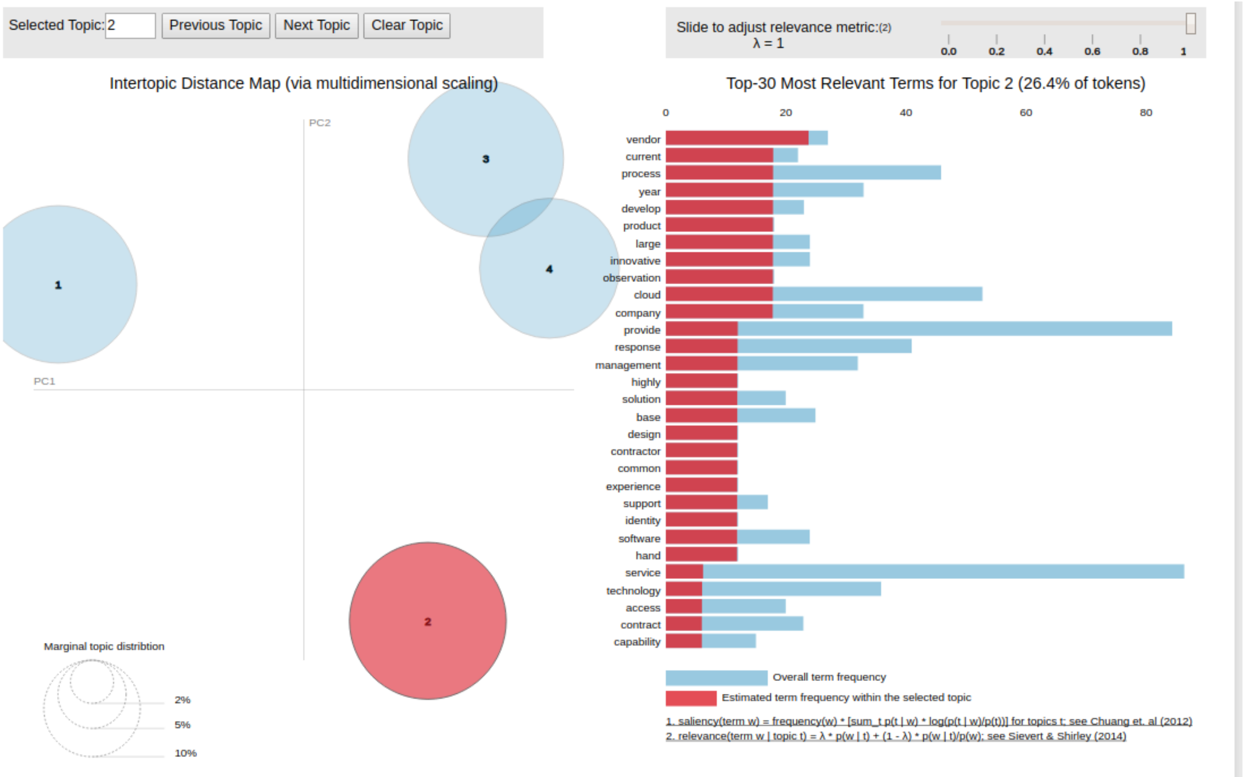


	Model Type	Coherence Score
0	Gensim	0.591681
1	Mallet	0.452522

Table 6: Gensim versus Mallet LDA models

4.3.5 Top 4 topics visualization

Figure 12 shows the visualization of the output of the Gensim LDA model. Topic 1 and topic 2 are highly distinctive of each other while topic 3 and topic 4 can be seen to be overlapping. Topics 3 and 4 are closely related because of the short inter-topic distance between them. The blue circles on the left side of the chart represents the topics. The size of the blue circle is directly proportional to the prevalence of that topic within the corpus. A fairly good topic model will typically have big and non-overlapping circles that are distributed in all the four quadrants. When a model has many small circles concentrated in one quadrant, that model is considered not good and sometimes reducing the value of k is likely to result into a better model. The top N terms for the selected topic are displayed on the right side of the chart. The bars colored blue show the frequency of the specific term within the entire corpus. The bars shaded red show how frequent the term is within the selected topic.



Selected Topic: 2

Slide to adjust relevance metric:(2) λ = 1

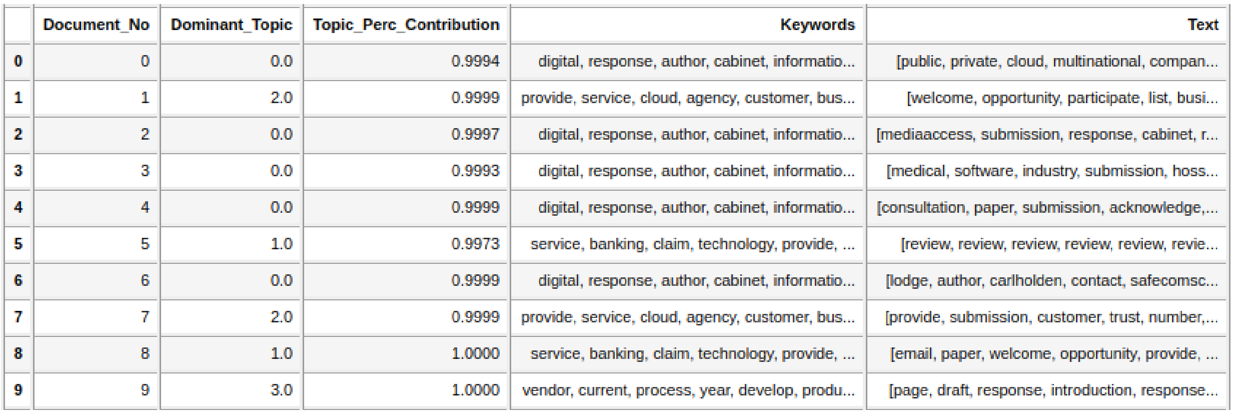
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 12: Gensim LDA model topic visualization

4.3.6 Dominant topic for each document

We want to determine the dominant topic for each of the document. In other words, to which topic does a particular document contribute the highest to? Identifying the dominant topic for each document can help understand the topic better by reading that particular document. This is in addition to analyzing the topic keywords which sometimes may not make it obvious what topic the keywords are addressing.

Table 7 shows the first 10 documents and the topic that each document contributes highest to in percentage. For instance, the dominant topic addressed by document number 5 is **topic 1.0** and the document contributes **0.9973 (99.73%)** to **topic 1.0**.

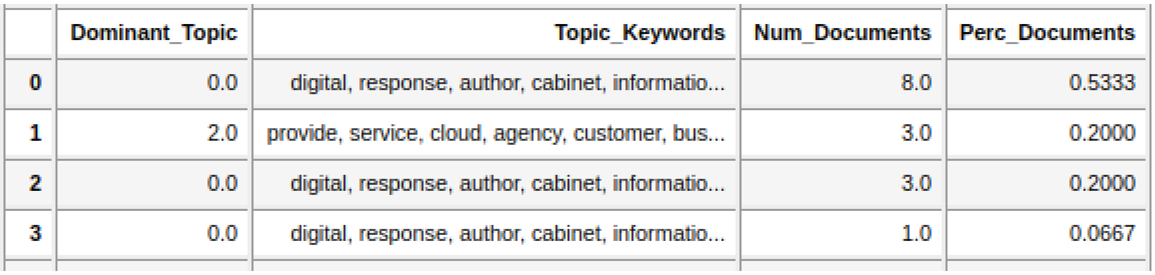


	Document_No	Dominant_Topic	Topic_Perc_Contribution	Keywords	Text
0	0	0.0	0.9994	digital, response, author, cabinet, informatio...	[public, private, cloud, multinational, compan...
1	1	2.0	0.9999	provide, service, cloud, agency, customer, bus...	[welcome, opportunity, participate, list, busi...
2	2	0.0	0.9997	digital, response, author, cabinet, informatio...	[mediaaccess, submission, response, cabinet, r...
3	3	0.0	0.9993	digital, response, author, cabinet, informatio...	[medical, software, industry, submission, hoss...
4	4	0.0	0.9999	digital, response, author, cabinet, informatio...	[consultation, paper, submission, acknowledge,...
5	5	1.0	0.9973	service, banking, claim, technology, provide, ...	[review, review, review, review, review, revie...
6	6	0.0	0.9999	digital, response, author, cabinet, informatio...	[lodge, author, carlholden, contact, safecomsc...
7	7	2.0	0.9999	provide, service, cloud, agency, customer, bus...	[provide, submission, customer, trust, number,...
8	8	1.0	1.0000	service, banking, claim, technology, provide, ...	[email, paper, welcome, opportunity, provide, ...
9	9	3.0	1.0000	vendor, current, process, year, develop, produ...	[page, draft, response, introduction, response...

Table 7: Dominant topic per document

4.3.7 Topic distribution across documents

Table 8 shows the volume and distribution of topics. Table 8 shows that topic 0 was discussed in 8 documents in total which represents 53.33% of the total number of documents. The total number of training documents is 15. Similarly, topic 2 was discussed in 3 documents representing 20% of the total number of documents.



	Dominant_Topic	Topic_Keywords	Num_Documents	Perc_Documents
0	0.0	digital, response, author, cabinet, informatio...	8.0	0.5333
1	2.0	provide, service, cloud, agency, customer, bus...	3.0	0.2000
2	0.0	digital, response, author, cabinet, informatio...	3.0	0.2000
3	0.0	digital, response, author, cabinet, informatio...	1.0	0.0667

Table 8: Topic distribution across documents

4.3.8 Most discussed topics in the documents

Figure 13 shows the number of documents that are attributable to each topic. The topic to be assigned to each document is the topic that is associated has the highest weight or contribution in that document. For example, if Document 1 contributes to Topic 0.0 and Topic 1.0 in proportions of 0.6 and 0.4 respectively, the topic to be selected here is Topic 0.0 as it has the highest contribution to Document 1.0. The x axis shows the 3 top-most keywords for a particular topic.

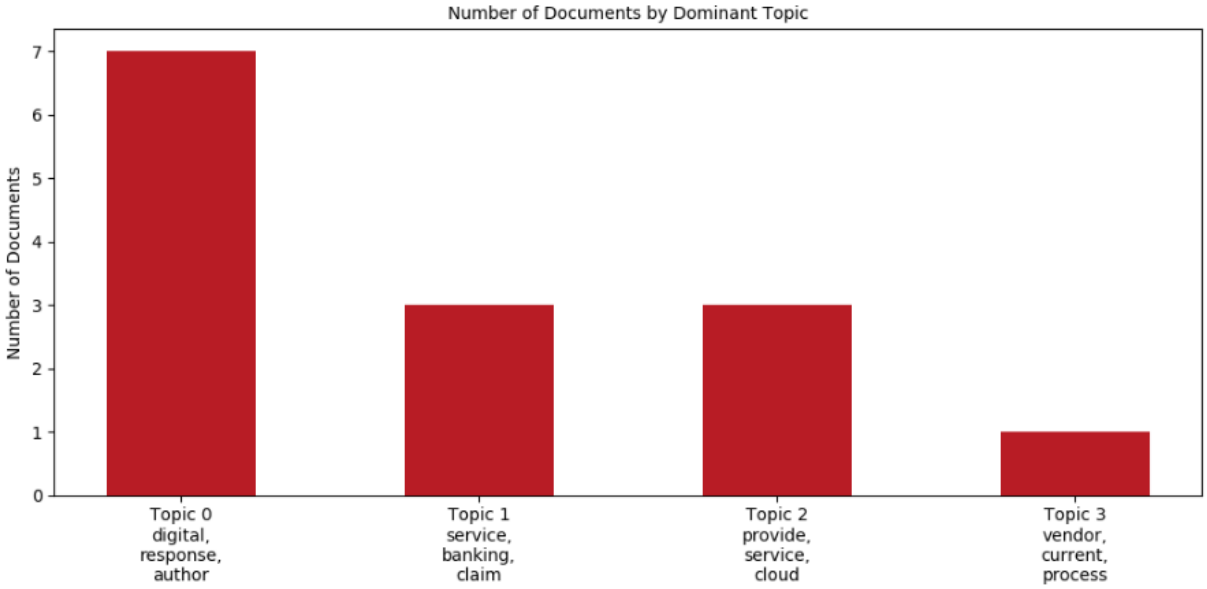


Figure 13: Document count by topic

Figure 14 shows the count of documents against each topic which is arrived at by adding up the weight contribution of each topic to the associated documents. The x axis shows the 3 top-most keywords for a particular topic.

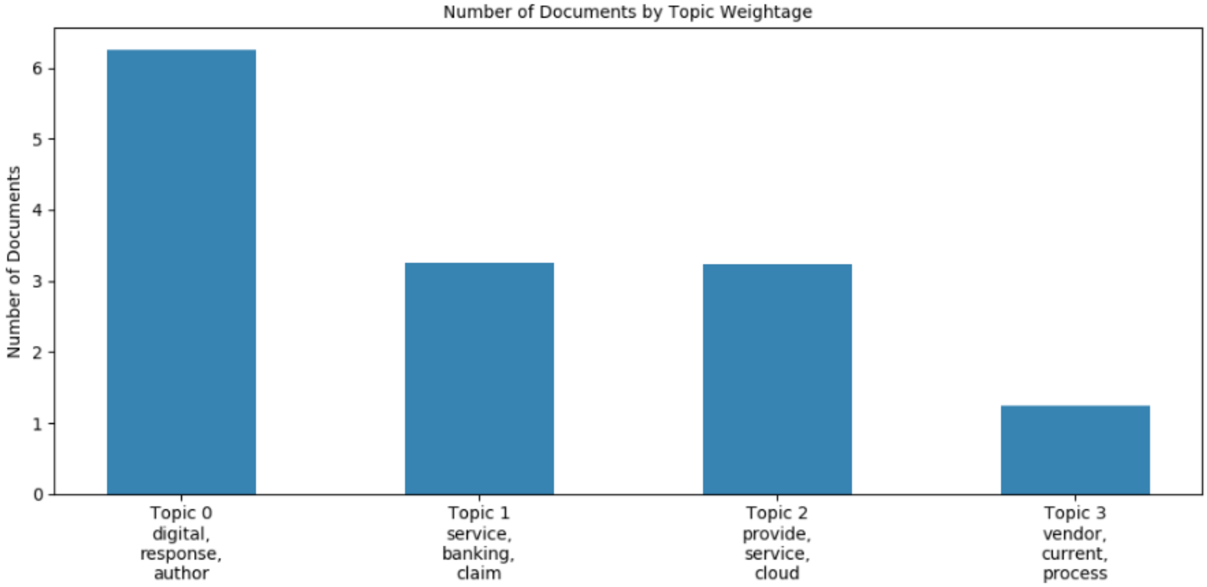


Figure 14: Document count by topic weightage

4.3.9 Word clouds per topic

Figure 15 shows the word clouds for each topic with the top 10 keywords displayed. The size of a word is proportional to the importance of that word in the respective topic.

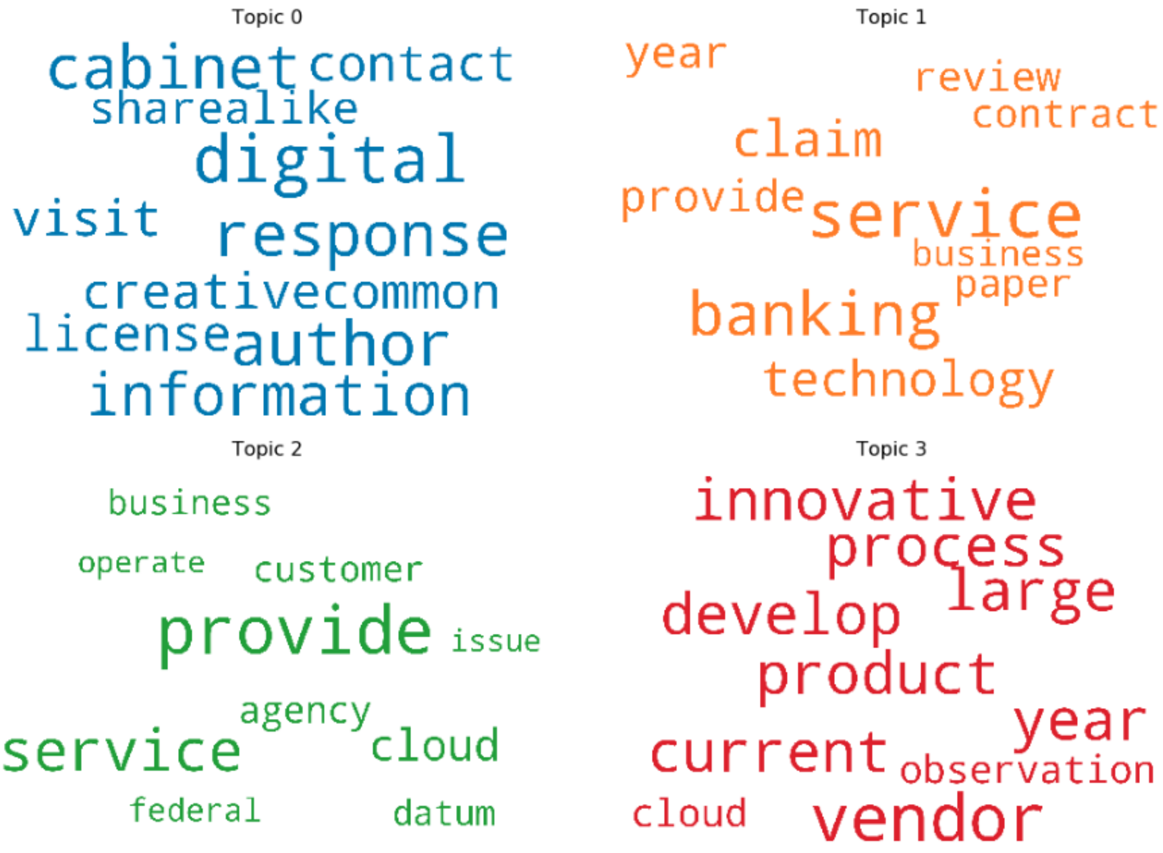


Figure 15: Word clouds of top 10 keywords per topic

4.3.10 Frequency and Importance of Topic Keywords

The weights of keywords and their frequency within the documents are important as we need to be able to assess words that have their relative frequency higher than their importance. Such words are usually less important. The count of top 10 words are plotted on the same chart as their respective weights as shown in Figure 16. As can be observed, all the keywords have their weights higher than the frequency.

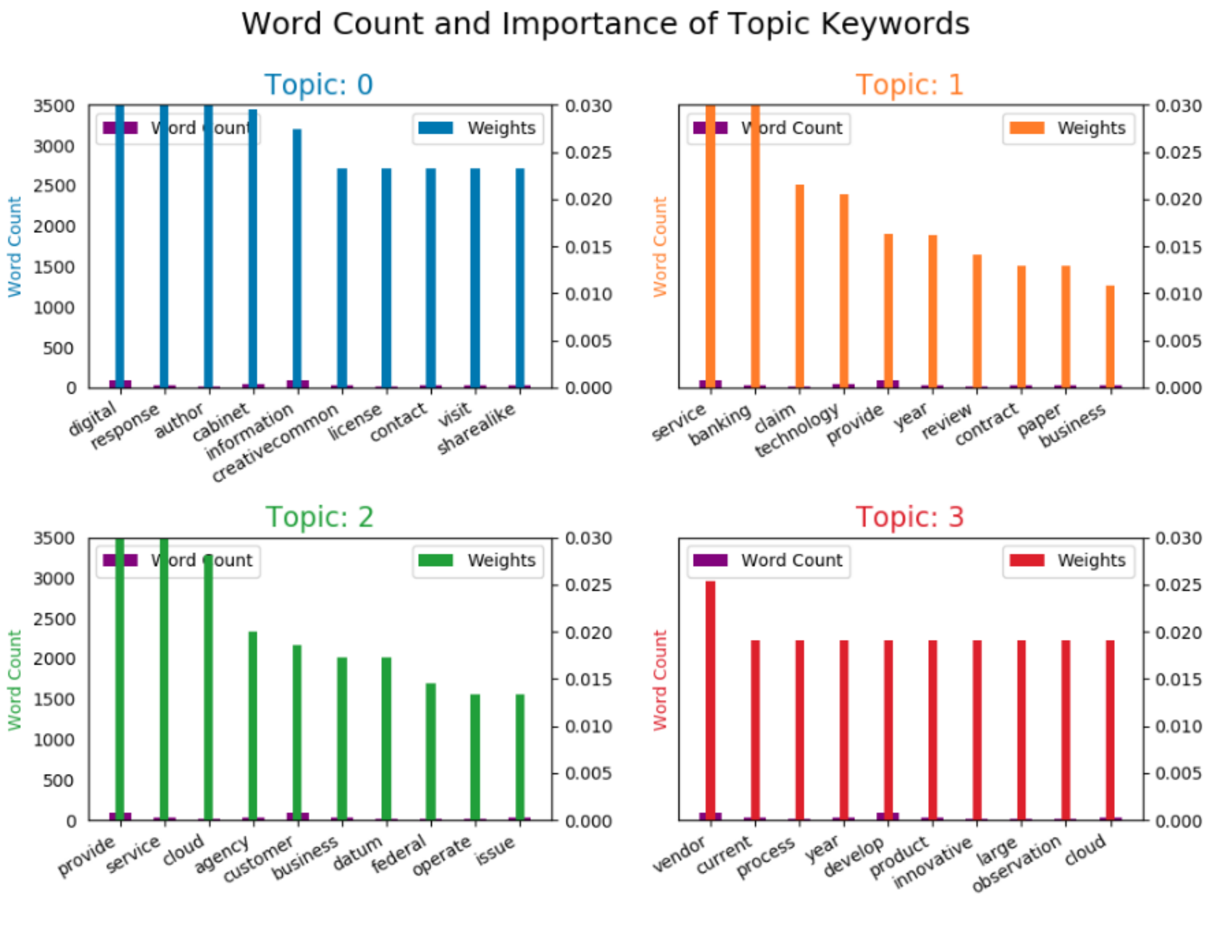


Figure 16: Word frequencies and importance of top 10 keywords per topic

4.3.11 Sentence topic coloring

Figure 17 shows the tokens for the first 10 documents plotted with different coloring. A word in a document can be associated with multiple topics. We associate different topics with distinguishing colors. For this chart, the topic to pick for a particular word is the topic to which the word has the highest importance. The color of the enclosing border for each document represents the color associated with the topic that the document contributes highest to. Similarly, each word is colored depending on the topic that it contributes highest to. As can be seen, most of the top 10 words for each document are attributable to the major topic for the document.

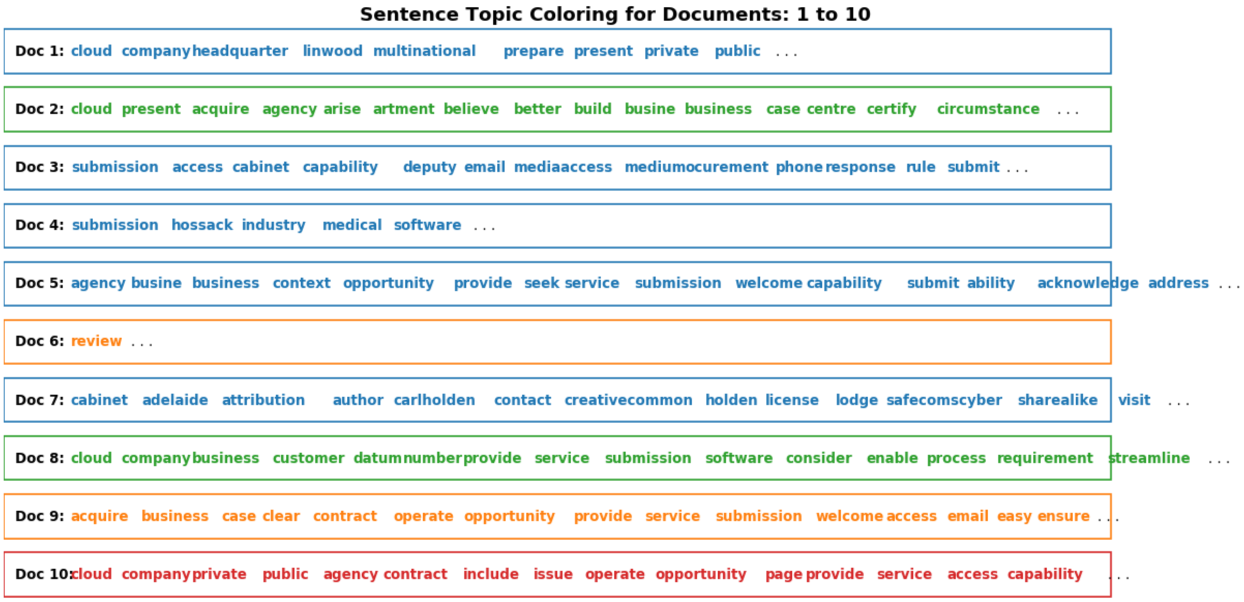


Figure 17: Sentence topic coloring

4.3.12 Frequency Distribution of Documents Lengths

We want to visualize the length of the documents that are associated with a particular topic. The length of a document is basically the count of words in that document. Figure 18 shows the size of documents in terms of word counts for topics 0 and 1. For example, for all the documents associated with topic 0, there is only one document whose length is between 200 and 400 words.

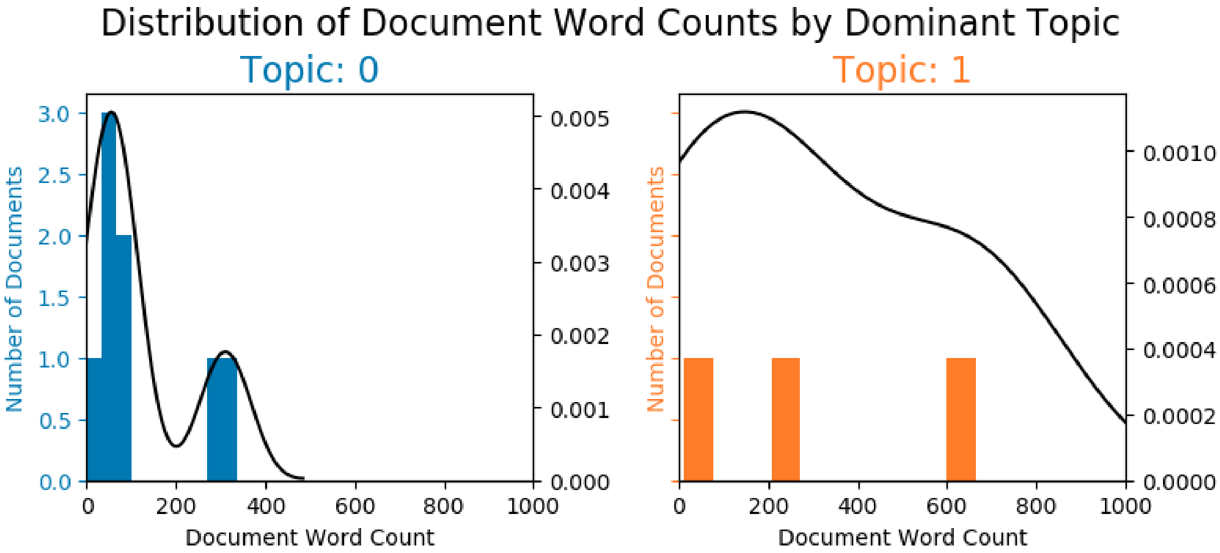
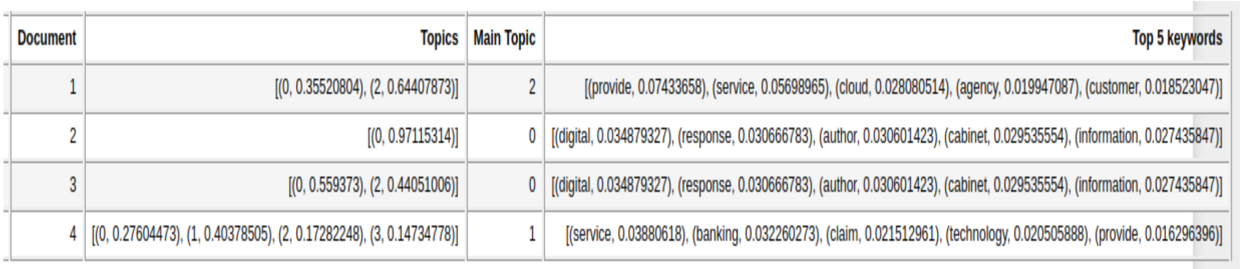


Figure 18: Distribution of word counts by dominant topic

4.3.13 Topic keyword assignment by human expert versus model prediction

Table 9 shows analysis of four (4) submissions as predicted by the trained LDA model. The first column shows the id of the document whose topics we are predicting. The second column shows the topics that each document is addressing. The first element on the tuple is the topic number. The second element of the tuple is the probability that a particular topic is being addressed in the document. The third column shows the main topic which is basically the topic with the highest probability as depicted in the second column. The fourth column shows the top 5 keywords associated with the main topic associated with the document. The first element of the tuple is the term, while the second item of the tuple is the importance of that term within the topic.

Table 10 shows the tags that were generated by the human expert on four (4) new validation submissions that were also presented to the model for prediction. The second column shows the tags that the human expert assigned to the document. The last column contains the remarks which reflect the degree to which the human expert agrees with the output of the model while considering the actual contents of the documents.



Document	Topics	Main Topic	Top 5 keywords
1	[(0, 0.35520804), (2, 0.64407873)]	2	[(provide, 0.07433658), (service, 0.05698965), (cloud, 0.028080514), (agency, 0.019947087), (customer, 0.018523047)]
2	[(0, 0.97115314)]	0	[(digital, 0.034879327), (response, 0.030666783), (author, 0.030601423), (cabinet, 0.029535554), (information, 0.027435847)]
3	[(0, 0.559373), (2, 0.44051006)]	0	[(digital, 0.034879327), (response, 0.030666783), (author, 0.030601423), (cabinet, 0.029535554), (information, 0.027435847)]
4	[(0, 0.27604473), (1, 0.40378505), (2, 0.17282248), (3, 0.14734778)]	1	[(service, 0.03880618), (banking, 0.032260273), (claim, 0.021512961), (technology, 0.020505888), (provide, 0.016296396)]

Table 9: Model prediction for new documents

We measured the human expert’s assessment based on the Likert scale on measuring agreement and the results are shown in Table 10. The assessment question presented to the expert was *“Do the top 10 keywords generated by the model accurately represent the content of the documents you reviewed?”*.

Document	Tags by human expert	Human expert assessment
1	Inclusion, rules, training, change, policy, awareness	Disagree
2	Innovation, government rules, culture, business, public service, departments	Disagree
3	Panel procurement, rates, relaxing rules, requirements	Disagree
4	Bias, favouring large corporations, quality of service, bid selection process	Agree

Table 10: Tags generated by human expert

4.4 Validity Threats

Every research is faced with risks and factors, which if not mitigated, are likely to adversely affect the research or even threaten the validity thereof (Wohlin et al., 2012). We will describe three types of threats, their impacts, and how we mitigated them.

4.4.1 Internal validity

This type of validity threat exists when there is some form of collaboration between dependent and independent variables (Wohlin et al., 2012). One threat to our experiment is the potential bias in the human expert. During the experiment, the researchers ensured that the human expert had no information that the tags he would generate would be compared with the keywords generated by our model. This ensured that the expert assigned the tags objectively. The researcher also ensured that the expert had no prior knowledge or access of the submission he was to tag. The task to be carried by the expert was clearly explained at the onset of the experiment. The expert was asked to complete tagging the documents in a single pass. We also asked the expert to perform the task continuously without taking breaks so as to avoid a situation where the expert would have to pause and think about the topics which would make the expert accustomed to the topics. By doing this, we were able to eliminate potential bias.

4.4.2 External validity

This type of validity measures the degree to which results of a research can be generalized outside of the environment within which it was conducted. In other words, can the experiment be repeated in another setting and still yield valid results? Most times, the design of the research and the people involved can affect its validity (Wohlin et al., 2012). It was difficult accessing actual submissions of a task force, and this was expected given the black box nature of the performance of public task forces. This limitation threatens the generalizability of our model outside procurement sector. However, the researchers have clearly described the design of our experiment including the software and hardware specifications in which it was conducted. This greatly mitigates the generalization threat. The same design approach can be extended to other task forces addressing other issues.

4.4.3 Construct validity

Construct validity assesses how much the objectives of the research are reflected in the actual setup of the experiment. In other words, does the environment of the experiment affect the results (Sathi et al., 2016). The entire research was under close supervision by the research supervisor who continuously suggested changes. The entire research and experiments steps and processes were setup and documented with the overall goal of achieving the research objectives.

CHAPTER FIVE

DISCUSSIONS, CONCLUSIONS AND RECOMMENDATIONS

This chapter concludes this report and contains five sections. We start by presenting a summary of the research work, we then present the discussion arising from the results of our analysis and how they are interpreted. We conclude by acknowledging the limitations of this study and finally giving recommendations for future work.

5.1 Summary

This study focused on processing of public participation feedback with focus on public task forces. Empirical literature indicates that text mining techniques have continued to generate a lot of interest from researchers. Much of the focus is on how to process large corpora of text and how to gain insights from it. However, little focus has been given to processing text which may only be containing a narrow range of themes. This gap inspired this study where we sought how to process submissions that are likely to be addressing a narrow range of topics. The literature review was concentrated on how themes and topics resident in submissions can be identified without the need to review the entire submission manually. Emphasis was given on topic modeling as a text mining technique with specific application of LDA as the text processing algorithm.

The experiment was conducted using twenty-two (22) submissions to a public task force on ICT procurement. Fifteen of these were used to train the model while the remaining seven submissions were used to validate that the model would perform as expected given new unseen data. The findings show that submissions address more than one topic. The topics are not so distinctive of each other and are likely to be overlapping. We also noted that when the value of k is high (more than 4 for our training data), the topics generated are no longer coherent. We also found out that Gensim LDA model performed better than Mallet LDA model given the same input parameters and data. As described in the sample model validation results, the keywords generated for 4 documents do not accurately represent the contents of the respective document for a majority of the documents. The human expert only agreed that the keywords correctly represent the contents for only one of the documents out of a total of four.

5.2 Discussion and interpretation of findings

As shown when visualizing a topic model in Figure 12, the blue circles on the left side of the chart represents the topics. The size of the blue circle is directly proportional to the prevalence of that topic within the corpus. A fairly good topic model will typically have big and non-overlapping circles that are distributed in all the four quadrants. When a model has many small circles concentrated in one quadrant, that model is considered not desirable and sometimes reducing the value of k is likely to result into a better model.

According to the results, given the top n words for each topic, it is evident that the topics are not easily identifiable. This is because the keywords all relate to the same general theme of ICT procurement. This closeness between keywords representing the topics explains why the optimal number of topics is low, that is, 4 topics. In addition, we observe that most of the submissions are classified into Topic 0. This is explainable because all submissions will naturally be addressing at least one main theme that a public task force is addressing.

A topic model requires a sufficiently large volume of text and documents for it to predict accurately. Corpus size is a function of the number of documents and the length of those documents. No proven guidelines were identified in the existing literature that specifies the ideal or minimum size of corpus that is required to train a model to predict accurately. However, some experimental studies suggest that an LDA that is trained with less than 100 documents produces results that have a low interpretability regardless of the size of the documents. The studies further seem to suggest that results of a topic model are reasonably interpretable when number of documents used to train the model is around 1000 (Nguyen, 2015).

The extent to which topic keywords represent the actual contents of a submission is relatively low. This is because of the small size of training data that was used. The model did not have sufficiently large enough data to properly train it so as to avoid overfitting. Thus, when it was presented with new data, the model was able to categorize the document into respective topics although the keywords are not very distinctive.

5.3 Conclusion

The purpose of our research was to demonstrate the value of topic modeling in processing public participation data. The study concluded that extraction of topics contained in feedback from the public can be automated. The findings show that distinctive topics are usually contained in submissions. However, those topics are few and increasing parameter k, will lead to more overlapping and nonsensical topics being generated.

The results also showed that one word could contribute to multiple topics. Inferring topics from the generated keywords requires knowledge in the subject area since topic models identify hidden (latent) topics within documents.

In light of the findings of the study, applying topic modeling when processing public feedback will lead to objective and comprehensive analysis of the submissions. The study demonstrates that topic models can be applied to processing any type of textual feedback whose volume is large and can be used as a collaborative tool in public policy making. By applying topic modeling to the large volume of public feedback, objective and effective processing is achieved, thus eliminating would be partiality and bias by capturing all the input from stakeholders. This greatly enhances the process of citizen oriented public policy making process.

5.4 Limitations of the study

This study had two limitations. Firstly, the relatively small size of training data that was used may lead to overfitting of the topic model. Overfitting causes a model to perform well only on data whose characteristics are similar to the dataset used to train the model. But when the same model is presented with data whose characteristics are quite different from the training dataset, the model predicts unreliably. In other words, the model cannot be generalized. An overfitted model picks noise in the data as features of the data.

Secondly, the time limitation associated with this study did not allow in-depth probing of the topic model in terms of how generalizable it was. The model was developed using small amount of submissions related to ICT procurement. We did not test the model to check how it would perform given submissions related to procurement in other sectors.

5.5 Recommendations for future work

In our research, we have developed an LDA model to extract the latent topics in submissions to public task forces. One of the issues that has emerged is the overlapping of topics between submissions. This is because public task forces are set up to address a limited range of issues. Thus, it is likely that different submissions will be handling similar topics. An interesting future work would be to identify sub-topics within a main topic. This would be for example looking at the responses given for a specific question.

In this study, 22 submissions were considered for this research. The same task can be extended to a much larger dataset to evaluate if the model performance would improve.

Lastly, due to the limitation of time allocation for this research, we did not perform a time-based analysis on the topics contained in submissions. It is possible that topics contained in earlier submissions are different from those contained in topics made later in the life of task forces. Performing such an analysis would help understand the relationship between external developments, like economic, policy and governance, with the issues being addressed by a public task force. For example, a change in a government policy on taxation regime midway during the life time of a task force would be reflected differently by submissions depending on the time a submission was made.

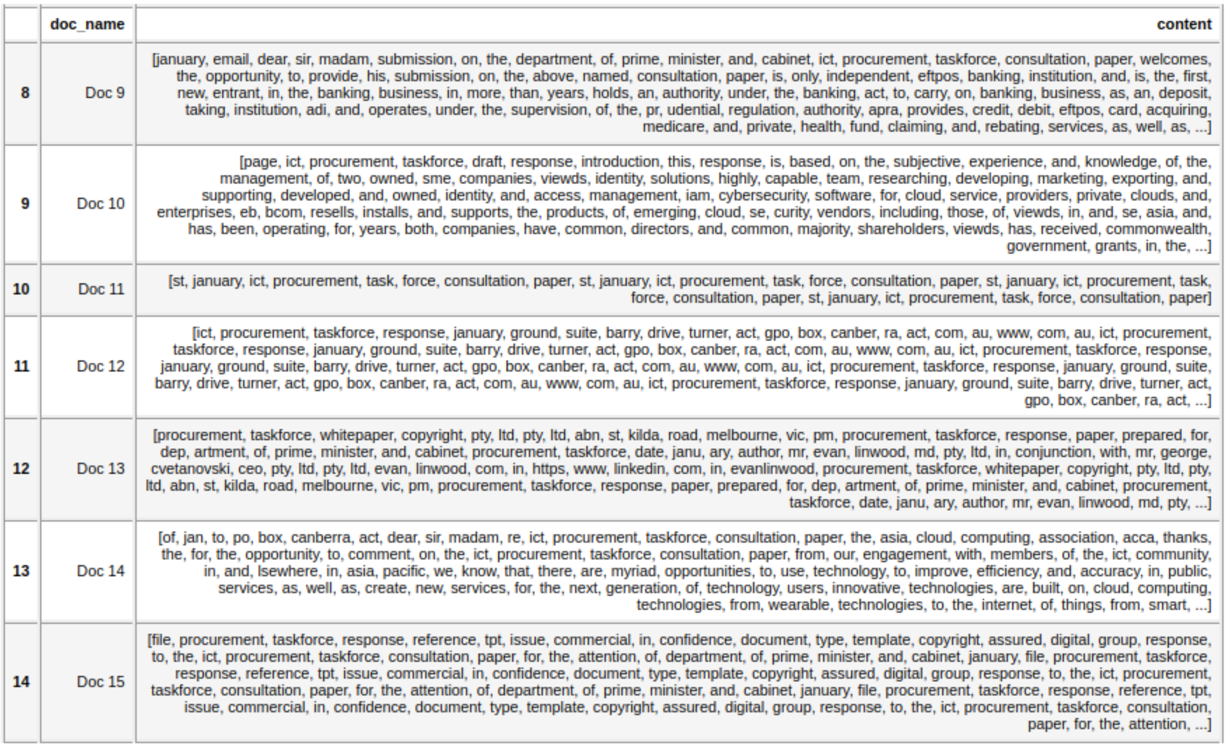
REFERENCES

1. Aggarwal, C. C., & Zhai, C. (2012). An introduction to text mining. In Aggarwal, C. C., & Zhai, C. (Eds.), Mining text data (1 – 10). Springer.
2. Ali Daud, Juanzi LI, Lizhu Zhou, Faqir Muhammad (June 2010). Knowledge discovery through directed probabilistic topic models: a survey
3. C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, Experimentation in software engineering. Springer Science & Business Media, 2012.
4. David M. Blei, Andrew Y. Ng and Michael I. Jordan (January 2003). Latent Dirichlet Allocation
5. Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Björn W Schuller. (2016). SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives. In COLING. 2666–2677.
6. Evangelopoulos, N., & Visinescu, L. L. (2012). Text-mining the voice of the people. *Commun. ACM*, *55*(2), 62-69.
7. Federico Pascual (September 2019). Introduction to Topic Modeling. Retrieved from https://monkeylearn.com/blog/introduction-to-topic-modeling/
8. Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, Dragomir R. Radev (2010). How to analyze political attention with minimal assumptions and costs. American Journal of Political Science, 54(1), 209–228. doi:10.1111/j.1540-5907.2009.00427.x
9. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. In the proceedings of the National Academy of Sciences, 101(Suppl 1):522
10. Marouane Birjali, Abderrahim Beni-Hssane,and Mohammed Erritali. (2016). Prediction of Suicidal Ideation in Twitter Data using Machine Learning algorithms. International Arab Conference on Information Technology (ACIT'2016).
11. Michael Röder, Andreas Both, Alexander Hinneburg (February 2015). Exploring the Space of Topic Coherence Measures.
12. Mohamed Amin (2016). A Topic Modeling Approach to Categorizing API Customer Value Propositions
13. Ngai, E. W., & Lee, P. T. Y. (2016, June). A Review of the literature on Applications of Text Mining in Policy Making. In PACIS (p. 343).

14. Nguyen, L. (2015). Topic modeling with more confidence: A theory and some algorithms. Paper presented at the Pacific-Asia Knowledge Discovery and Data Mining, Ho Chi Minh City.
15. Olessia Koltsova and Sergei Koltcov (June 2013) Mapping the Public Agenda with Topic Modeling
16. Paul Cairney (September 2015). Politics & Public Policy. Retrieved from https://paulcairney.wordpress.com/2015/09/22/key-theories-in-policymaking-how-to-explain-what-is-going-on-in-scotland/
17. Rebecca Merrett (May 2015). 5 Tools and Techniques for Text Analytics. Retrieved from https://www.cio.com/article/3498302/5-tools-and-techniques-for-text-analytics.html
18. Shashank Kapadia (August 2019). Evaluate Topic Models: Latent Dirichlet Allocation (LDA). Retrieved from https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0
19. Shilpa Dang (December 2014). Text Mining : Techniques and its Application
20. T.Rajasundari, P.Subathra, P.N.Kumar (July 2017). Performance analysis of topic modeling algorithms for news articles.
21. Thomas L. Griffiths and Mark Steyvers (2007). Topics in Semantic Representation
22. Toni Cvitanic, Bumsoo Lee, Hyeon Ik Song, Katherine Fu, and David Rosen (2016). LDA v. LSA: A Comparison of Two Computational Text Analysis Tools for the Functional Categorization of Patents
23. Ujang Fahmi (January 2019). Cultural Public Sphere: Tracking the Yogyakarta City Policy Agenda through the #JogjaOraDidol Hashtag on Twitter
24. Vallikannu Ramanathan, T. Meyyappan "Survey of Text Mining", International Conference on Technology and Business and Management, March 2013, pp. 508-514.
25. Veer Reddy Sathi and Jai Simha Ramanujapura (July 2016). A Quality Criteria Based Evaluation of Topic Models
26. Zhai, K., Boyd-Graber, J., Asadi, N., & Alkhouja, M. L. (2012, April). Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of the 21st international conference on World Wide Web* (pp. 879-888).
27. Zhou Tong and Haiyi Zhang (May 2016). A Text Mining Research Based on LDA Topic Modelling

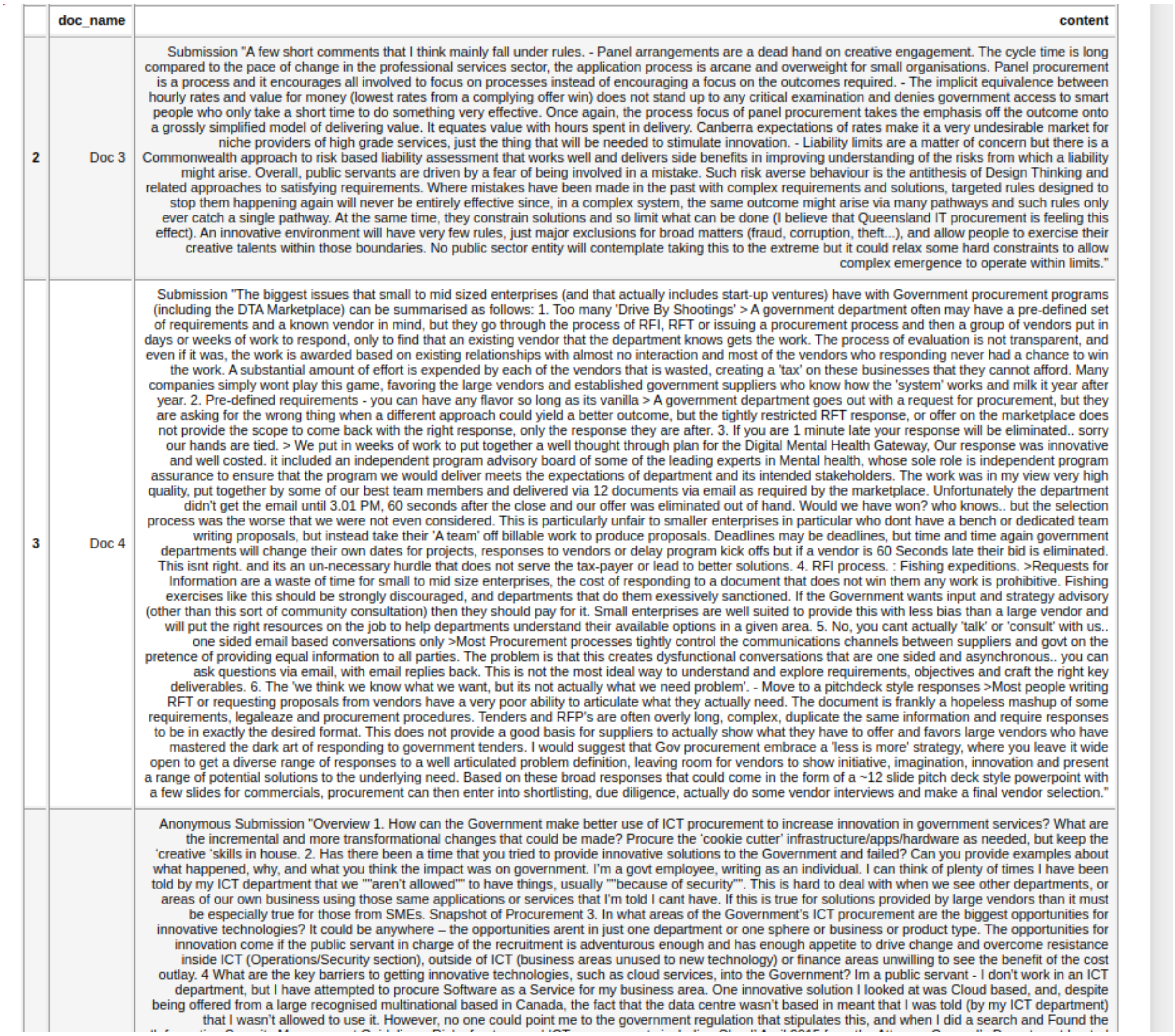
APPENDICES

7.1 Snapshot of data used to train LDA model



	doc_name	content
8	Doc 9	[january, email, dear, sir, madam, submission, on, the, department, of, prime, minister, and, cabinet, ict, procurement, taskforce, consultation, paper, welcomes, the, opportunity, to, provide, his, submission, on, the, above, named, consultation, paper, is, only, independent, eftpos, banking, institution, and, is, the, first, new, entrant, in, the, banking, business, in, more, than, years, holds, an, authority, under, the, banking, act, to, carry, on, banking, business, as, an, deposit, taking, institution, adi, and, operates, under, the, supervision, of, the, pr, udential, regulation, authority, apra, provides, credit, debit, eftpos, card, acquiring, medicare, and, private, health, fund, claiming, and, rebating, services, as, well, as, ...]
9	Doc 10	[page, ict, procurement, taskforce, draft, response, introduction, this, response, is, based, on, the, subjective, experience, and, knowledge, of, the, management, of, two, owned, sme, companies, viewds, identity, solutions, highly, capable, team, researching, developing, marketing, exporting, and, supporting, developed, and, owned, identity, and, access, management, iam, cybersecurity, software, for, cloud, service, providers, private, clouds, and, enterprises, eb, bcom, resells, installs, and, supports, the, products, of, emerging, cloud, se, curity, vendors, including, those, of, viewds, in, and, se, asia, and, has, been, operating, for, years, both, companies, have, common, directors, and, common, majority, shareholders, viewds, has, received, commonwealth, government, grants, in, the, ...]
10	Doc 11	[st, january, ict, procurement, task, force, consultation, paper, st, january, ict, procurement, task, force, consultation, paper, st, january, ict, procurement, task, force, consultation, paper, st, january, ict, procurement, task, force, consultation, paper]
11	Doc 12	[ict, procurement, taskforce, response, january, ground, suite, barry, drive, turner, act, gpo, box, canber, ra, act, com, au, www, com, au, ict, procurement, taskforce, response, january, ground, suite, barry, drive, turner, act, gpo, box, canber, ra, act, com, au, www, com, au, ict, procurement, taskforce, response, january, ground, suite, barry, drive, turner, act, gpo, box, canber, ra, act, com, au, www, com, au, ict, procurement, taskforce, response, january, ground, suite, barry, drive, turner, act, gpo, box, canber, ra, act, com, au, www, com, au, ict, procurement, taskforce, response, january, ground, suite, barry, drive, turner, act, gpo, box, canber, ra, act, ...]
12	Doc 13	[procurement, taskforce, whitepaper, copyright, pty, ltd, pty, ltd, abn, st, kilda, road, melbourne, vic, pm, procurement, taskforce, response, paper, prepared, for, dep, artment, of, prime, minister, and, cabinet, procurement, taskforce, date, janu, ary, author, mr, evan, linwood, md, pty, ltd, in, conjunction, with, mr, george, cvetanovski, ceo, pty, ltd, pty, ltd, evan, linwood, com, in, https, www, linkedin, com, in, evanlinwood, procurement, taskforce, whitepaper, copyright, pty, ltd, pty, ltd, abn, st, kilda, road, melbourne, vic, pm, procurement, taskforce, response, paper, prepared, for, dep, artment, of, prime, minister, and, cabinet, procurement, taskforce, date, janu, ary, author, mr, evan, linwood, md, pty, ...]
13	Doc 14	[of, jan, to, po, box, canberra, act, dear, sir, madam, re, ict, procurement, taskforce, consultation, paper, the, asia, cloud, computing, association, acca, thanks, the, for, the, opportunity, to, comment, on, the, ict, procurement, taskforce, consultation, paper, from, our, engagement, with, members, of, the, ict, community, in, and, lsewhere, in, asia, pacific, we, know, that, there, are, myriad, opportunities, to, use, technology, to, improve, efficiency, and, accuracy, in, public, services, as, well, as, create, new, services, for, the, next, generation, of, technology, users, innovative, technologies, are, built, on, cloud, computing, technologies, from, wearable, technologies, to, the, internet, of, things, from, smart, ...]
14	Doc 15	[file, procurement, taskforce, response, reference, tpt, issue, commercial, in, confidence, document, type, template, copyright, assured, digital, group, response, to, the, ict, procurement, taskforce, consultation, paper, for, the, attention, of, department, of, prime, minister, and, cabinet, january, file, procurement, taskforce, response, reference, tpt, issue, commercial, in, confidence, document, type, template, copyright, assured, digital, group, response, to, the, ict, procurement, taskforce, consultation, paper, for, the, attention, of, department, of, prime, minister, and, cabinet, january, file, procurement, taskforce, response, reference, tpt, issue, commercial, in, confidence, document, type, template, copyright, assured, digital, group, response, to, the, ict, procurement, taskforce, consultation, paper, for, the, attention, ...]

7.2 Snapshot of documents given to expert



	doc_name	content
2	Doc 3	Submission "A few short comments that I think mainly fall under rules. - Panel arrangements are a dead hand on creative engagement. The cycle time is long compared to the pace of change in the professional services sector, the application process is arcane and overweight for small organisations. Panel procurement is a process and it encourages all involved to focus on processes instead of encouraging a focus on the outcomes required. - The implicit equivalence between hourly rates and value for money (lowest rates from a complying offer win) does not stand up to any critical examination and denies government access to smart people who only take a short time to do something very effective. Once again, the process focus of panel procurement takes the emphasis off the outcome onto a grossly simplified model of delivering value. It equates value with hours spent in delivery. Canberra expectations of rates make it a very undesirable market for niche providers of high grade services, just the thing that will be needed to stimulate innovation. - Liability limits are a matter of concern but there is a Commonwealth approach to risk based liability assessment that works well and delivers side benefits in improving understanding of the risks from which a liability might arise. Overall, public servants are driven by a fear of being involved in a mistake. Such risk averse behaviour is the antithesis of Design Thinking and related approaches to satisfying requirements. Where mistakes have been made in the past with complex requirements and solutions, targeted rules designed to stop them happening again will never be entirely effective since, in a complex system, the same outcome might arise via many pathways and such rules only ever catch a single pathway. At the same time, they constrain solutions and so limit what can be done (I believe that Queensland IT procurement is feeling this effect). An innovative environment will have very few rules, just major exclusions for broad matters (fraud, corruption, theft...), and allow people to exercise their creative talents within those boundaries. No public sector entity will contemplate taking this to the extreme but it could relax some hard constraints to allow complex emergence to operate within limits."
3	Doc 4	Submission "The biggest issues that small to mid sized enterprises (and that actually includes start-up ventures) have with Government procurement programs (including the DTA Marketplace) can be summarised as follows: 1. Too many 'Drive By Shootings' > A government department often may have a pre-defined set of requirements and a known vendor in mind, but they go through the process of RFI, RFT or issuing a procurement process and then a group of vendors put in days or weeks of work to respond, only to find that an existing vendor that the department knows gets the work. The process of evaluation is not transparent, and even if it was, the work is awarded based on existing relationships with almost no interaction and most of the vendors who responding never had a chance to win the work. A substantial amount of effort is expended by each of the vendors that is wasted, creating a 'tax' on these businesses that they cannot afford. Many companies simply wont play this game, favoring the large vendors and established government suppliers who know how the 'system' works and milk it year after year. 2. Pre-defined requirements - you can have any flavor so long as its vanilla > A government department goes out with a request for procurement, but they are asking for the wrong thing when a different approach could yield a better outcome, but the tightly restricted RFT response, or offer on the marketplace does not provide the scope to come back with the right response, only the response they are after. 3. If you are 1 minute late your response will be eliminated.. sorry our hands are tied. > We put in weeks of work to put together a well thought through plan for the Digital Mental Health Gateway, Our response was innovative and well costed. it included an independent program advisory board of some of the leading experts in Mental health, whose sole role is independent program assurance to ensure that the program we would deliver meets the expectations of department and its intended stakeholders. The work was in my view very high quality, put together by some of our best team members and delivered via 12 documents via email as required by the marketplace. Unfortunately the department didn't get the email until 3.01 PM, 60 seconds after the close and our offer was eliminated out of hand. Would we have won? who knows.. but the selection process was the worse that we were not even considered. This is particularly unfair to smaller enterprises in particular who dont have a bench or dedicated team writing proposals, but instead take their 'A team' off billable work to produce proposals. Deadlines may be deadlines, but time and time again government departments will change their own dates for projects, responses to vendors or delay program kick offs but if a vendor is 60 Seconds late their bid is eliminated. This isnt right. and its an un-necessary hurdle that does not serve the tax-payer or lead to better solutions. 4. RFI process. : Fishing expeditions. >Requests for Information are a waste of time for small to mid size enterprises, the cost of responding to a document that does not win them any work is prohibitive. Fishing exercises like this should be strongly discouraged, and departments that do them exessively sanctioned. If the Government wants input and strategy advisory (other than this sort of community consultation) then they should pay for it. Small enterprises are well suited to provide this with less bias than a large vendor and will put the right resources on the job to help departments understand their available options in a given area. 5. No, you cant actually 'talk' or 'consult' with us.. one sided email based conversations only >Most Procurement processes tightly control the communications channels between suppliers and govt on the pretence of providing equal information to all parties. The problem is that this creates dysfunctional conversations that are one sided and asynchronous.. you can ask questions via email, with email replies back. This is not the most ideal way to understand and explore requirements, objectives and craft the right key deliverables. 6. The 'we think we know what we want, but its not actually what we need problem'. - Move to a pitchdeck style responses >Most people writing RFT or requesting proposals from vendors have a very poor ability to articulate what they actually need. The document is frankly a hopeless mashup of some requirements, legaleaze and procurement procedures. Tenders and RFP's are often overly long, complex, duplicate the same information and require responses to be in exactly the desired format. This does not provide a good basis for suppliers to actually show what they have to offer and favors large vendors who have mastered the dark art of responding to government tenders. I would suggest that Gov procurement embrace a 'less is more' strategy, where you leave it wide open to get a diverse range of responses to a well articulated problem definition, leaving room for vendors to show initiative, imagination, innovation and present a range of potential solutions to the underlying need. Based on these broad responses that could come in the form of a ~12 slide pitch deck style powerpoint with a few slides for commercials, procurement can then enter into shortlisting, due diligence, actually do some vendor interviews and make a final vendor selection."
		Anonymous Submission "Overview 1. How can the Government make better use of ICT procurement to increase innovation in government services? What are the incremental and more transformational changes that could be made? Procure the 'cookie cutter' infrastructure/apps/hardware as needed, but keep the 'creative 'skills in house. 2. Has there been a time that you tried to provide innovative solutions to the Government and failed? Can you provide examples about what happened, why, and what you think the impact was on government. I'm a govt employee, writing as an individual. I can think of plenty of times I have been told by my ICT department that we ""aren't allowed"" to have things, usually ""because of security"". This is hard to deal with when we see other departments, or areas of our own business using those same applications or services that I'm told I cant have. If this is true for solutions provided by large vendors than it must be especially true for those from SMEs. Snapshot of Procurement 3. In what areas of the Government's ICT procurement are the biggest opportunities for innovative technologies? It could be anywhere – the opportunities arent in just one department or one sphere or business or product type. The opportunities for innovation come if the public servant in charge of the recruitment is adventurous enough and has enough appetite to drive change and overcome resistance inside ICT (Operations/Security section), outside of ICT (business areas unused to new technology) or finance areas unwilling to see the benefit of the cost outlay. 4 What are the key barriers to getting innovative technologies, such as cloud services, into the Government? Im a public servant - I don't work in an ICT department, but I have attempted to procure Software as a Service for my business area. One innovative solution I looked at was Cloud based, and, despite being offered from a large recognised multinational based in Canada, the fact that the data centre wasn't based in meant that I was told (by my ICT department) that I wasn't allowed to use it. However, no one could point me to the government regulation that stipulates this, and when I did a search and Found the

7.3 Instructions issued to expert

Guidelines on tagging validation documents with associated topics

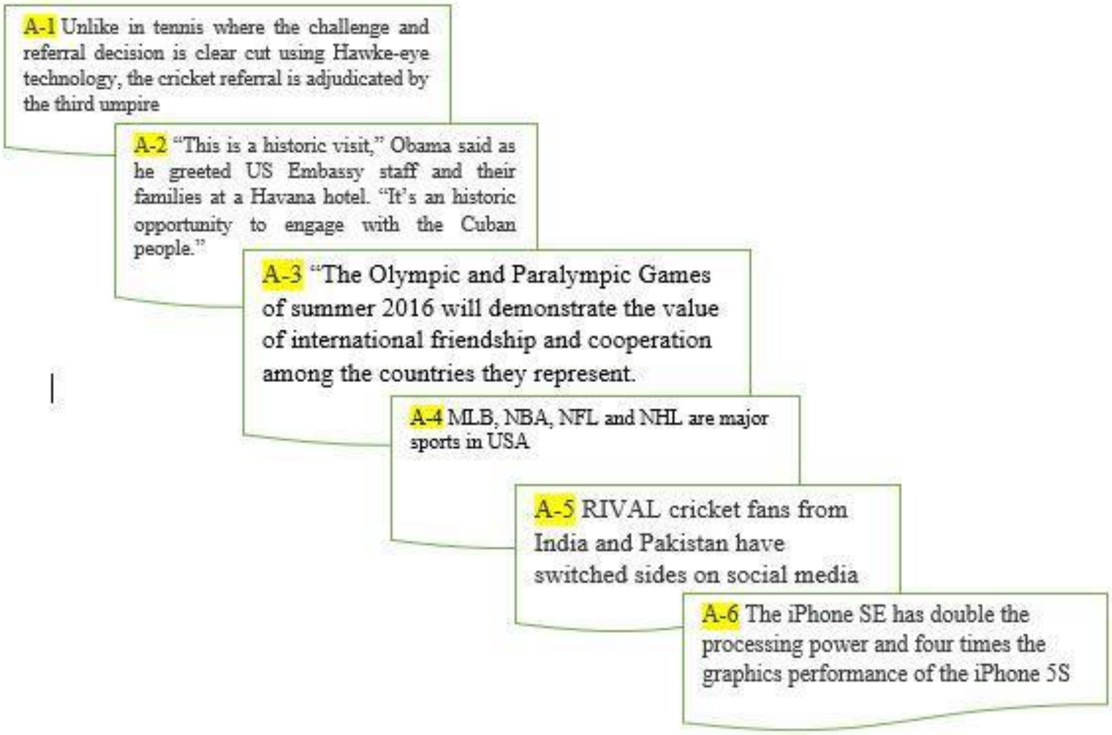
Introduction

The main objective of this experiment is to process public participation feedback by performing topic modeling.

Topic modeling refers to a natural language processing technique used to discover hidden semantic structures of text in a collection of documents. A topic modeling tool takes a single text (or corpus) and looks for patterns in the use of words. A topic to the computer is a list of words that occur in statistically meaningful ways.

Example

An example has been provided below to better understand what topic modeling is.



A-1 Unlike in tennis where the challenge and referral decision is clear cut using Hawke-eye technology, the cricket referral is adjudicated by the third umpire

A-2 "This is a historic visit," Obama said as he greeted US Embassy staff and their families at a Havana hotel. "It's an historic opportunity to engage with the Cuban people."

A-3 "The Olympic and Paralympic Games of summer 2016 will demonstrate the value of international friendship and cooperation among the countries they represent.

A-4 MLB, NBA, NFL and NHL are major sports in USA

A-5 RIVAL cricket fans from India and Pakistan have switched sides on social media

A-6 The iPhone SE has double the processing power and four times the graphics performance of the iPhone 5S

Considering the above set of documents, it is apparent that the articles are addressing topics about Sports, technology and relation between countries. In addition to the most frequently occurring words in each topic, we also get the proportion of each article in that topic. We can observe that article A-4 deals with sports entirely, the document-topic proportion of article A-4 is very high for topic 3. This means 99% of the article completely falls under topic 3. In the case of article A-3, where it explains about the Olympics and foreign relations, the document-topic proportions of article A-3 may be 0.5 for topic 3 and 0.5 for topic 2. Article A-5 describes the rivalry between cricket fans of two countries on social media. This article is a mixture of sports, technology and foreign relations. The document-topic proportions for this article may be similar for all the topics (0.4, 0.3 and 0.3).

The table below shows the topics that associated with each article.

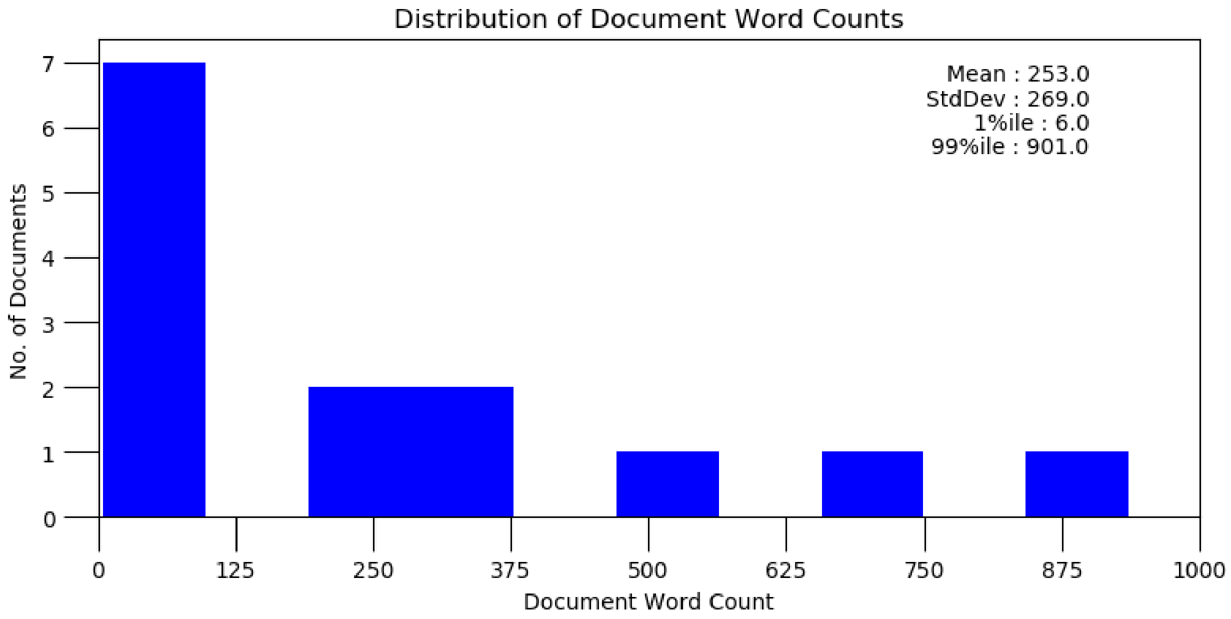
Article	**Topics**
A-1	Sport
A-2	Foreign Relations
A-3	Sport, Foreign Relations
A-4	Sport
A-5	Sport, Technology, Foreign Relations
A-6	Technology

Task to be performed

- You will be presented with seven documents that represent actual submission to a task force on ICT procurement.
- Your task is to carefully study the document and then list down the topics that each of the document is addressing.
- Each document may (or may not) have more than a single tag

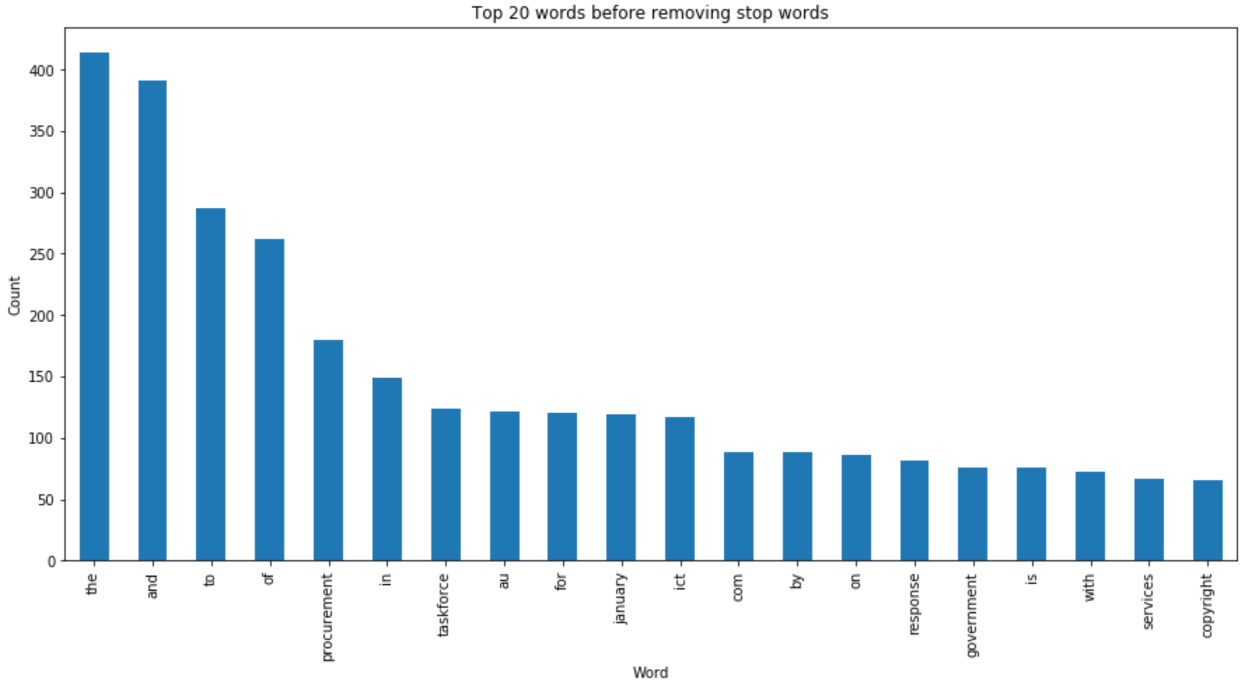
7.4 Distribution of document word counts

The histogram below shows how big the documents are as a whole in terms of word counts.



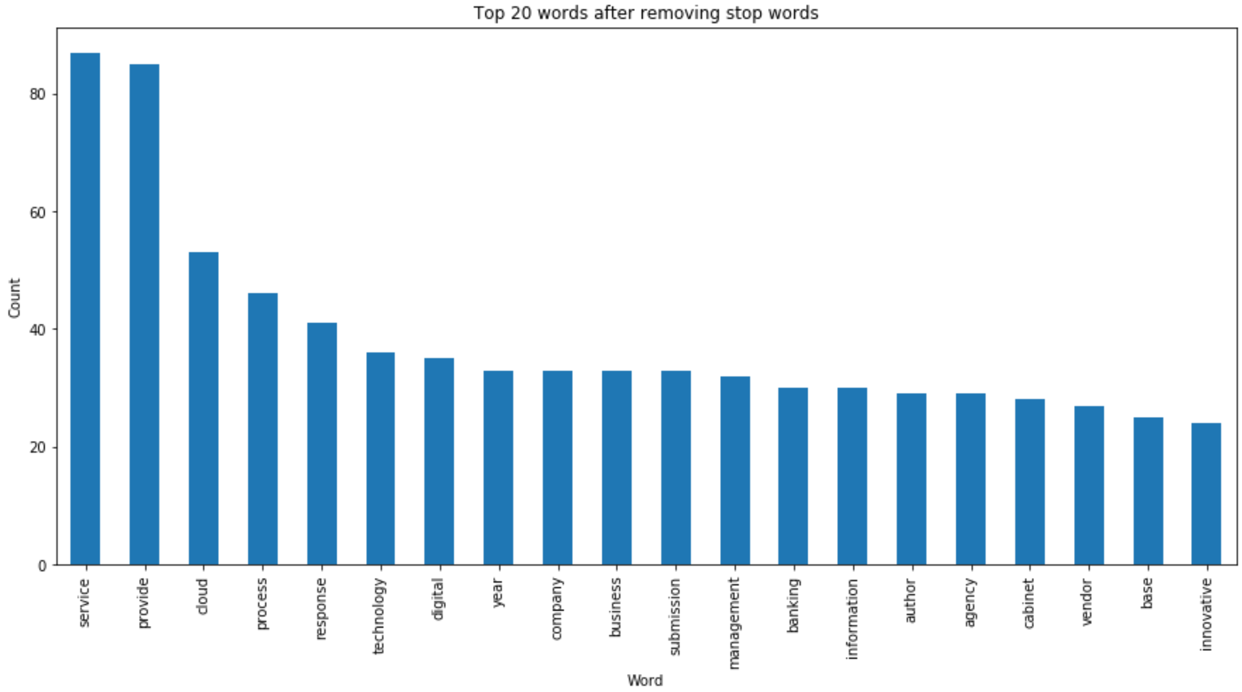
7.5 Top 20 words before removing stop words

The chart below shows the top 20 unigrams across the documents before removing stop words. It can be seen that unigram “the” is the most occurring word despite it being a stop word.



7.6 Top 20 words after removing stop words

The chart below shows the top 20 unigrams after removing stop words. After removing stop words, it can be seen that the terms are a bit distinctive semantically.


Top 20 words after removing stop words
Count
service
provide
cloud
process
response
technology
digital
year
company
business
submission
management
banking
information
author
agency
cabinet
vendor
base
innovative
Word