Name: Charita Ramnathsing

Capstone I Final Submission

Date: April 27th, 2020

Mentor: Dipanjan Sarkar

Table of Contents

[Introduction 2](#_Toc38923569)

[Goal 2](#_Toc38923570)

[Data Set used 2](#_Toc38923571)

[General Analyses 3](#_Toc38923572)

[What is Latent Dirichlet Allocation? 3](#_Toc38923573)

[Data Cleaning 5](#_Toc38923574)

[Prepare Dictionary and Corpus 6](#_Toc38923575)

[LDA Mallet Model 6](#_Toc38923576)

[Feature Extraction for further visualization 8](#_Toc38923577)

[Visualization of the Topics and Distribution of Words 10](#_Toc38923578)

[Visualization With Topic Frequency and Word Count within a Theme 11](#_Toc38923579)

[Conclusion 11](#_Toc38923580)

# Introduction

Since 1946, Members of the United Nations gather together to debate over topics which are being held during session at the Headquarters in New York for the General Assembly. There is a specific preset of topics which will be further elaborated about during these sessions. Every member of the States united through the United Nations will discuss and work together on a wide array of international issues covered by the Charter of the United Nations. Expected topics can be:

* International Peace and Security
* Humanitarian and Disaster Relief
* Economic Growth and Sustainability
* Disarmament
* Drugs, Crime and Terrorism
* Etc.

The spoken topics are public and can also be found on Kaggle [[1]](#footnote-1)in an excerpt of the years from 1970 – 2016. As well as the history as the development of the Debates are interesting to look at and would be helpful if we could know more about the underlying intentions or politics of the Members of these States.

# Goal

My goal of this Capstone is to predict what the themes are being mentioned according to the Charter of the United Nations but also which topics are being highlighted per country to ultimately see where the Nations can find a platform for collaboration. Even though the goal is to predict the preset themes of the United Nations, there is possibly no conclusion being made about the sentiment of these analyses in the topics.

# Data Set used

On Kaggle there is a dataset available which can be completely used for the model. The main columns in this dataset are:

* **Session**: Every year the UN indicates a session by a number and a theme. Currently session 75 will be held from September 15 – 30, 2020 and the theme will be related to the Sustainable Development Goal: Strengthen the means of implementation and revitalize the global partnership for sustainable development. *The dataset though is from session 25 till session 70.*
* **Year**: the year that the session was held
* **Country**: the country of the represented Member States which the speaker is talking on behalf of.
* **Text**: The transcript presentation which the speaker has given.

The main column being used will be ‘Text’ because for making a good Topic Model it is important to make the ‘noise’ the model as little as possible, and that will be done through packages like NLTK, SpaCy and genism.

# General Analyses

A screenshot of a cell phone

Description automatically generated

With the data given at Kaggle we can start our journey to clean, pre-process and prepare a Dictionary and Corpus for the Model. The chosen model for this specific problem is LDA Mallet which is a Java based package for statistical natural language processing. There are more package which can be used but after looking at the pre-results of the prediction, I came to the conclusion that for this specific ‘prediction’ of the topics of the Debates it gives me the most accurate topics back. Most of the models I have seen used in natural language processing are based on LDA, Latent Dirichlet Allocation and so is LDA Mallet.

# What is Latent Dirichlet Allocation?

LDA is a generative probabilistic model of a corpus of documents made up of words and/or phrases. It is based on the mixed-membership model [[2]](#footnote-2), and this model is an extension of mixture models to grouped data. As Prof. David M. Blei indicated about ‘grouped data’, each data point (topic) is itself a collection of data and each collection can belong to multiple groups. LDA was applied in [machine learning](https://en.wikipedia.org/wiki/Machine_learning) by [David Blei](https://en.wikipedia.org/wiki/David_Blei), [Andrew Ng](https://en.wikipedia.org/wiki/Andrew_Ng) and [Michael I. Jordan](https://en.wikipedia.org/wiki/Michael_I._Jordan) in 2003. [[3]](#footnote-3)

The LDA is a three-level hierarchical Bayesian model, in which each item of a collection 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. In the context of text modeling, the topic probabilities provide an explicit representation of a document.

A screenshot of a cell phone

Description automatically generated

This model is being used in many predictions related to data and unsupervised learning. The basic ideas are:

* Data are grouped, and each group *x*i is a collection of xij, where *j* ∈ {1,…,ni}
* Each group (collection of words) is modeled with a mixture model
* The mixture components (recurrent patterns of observed words) are shared across groups
* The mixture proportions vary from group to group

In general, this algorithm has been adapted to all kinds of other data- images, computer code, music data, recommendation data, and others. It is a model of high-dimensional discrete data. I don’t expect you to understand this underlaying model of the LDA, but in general it can make more sense through the above basic idea.

Eventually we would like to have a dictionary where I can find my words and the frequencies of it, and I would like to have a corpus from which I can make a distribution of frequencies made out of the dictionary. So let’s start cleaning the data for making the dictionary and its distribution over the document, also called the ‘term document frequency’.

# Data Cleaning

Before using the data for the corpus we need to clean, tokenize, stem and lemmatize it. All these techniques can be seen as word normalization to make the text ready for the model. The cleaning part is related to taking the ‘Nan’ values out and try to make the data as simple as possible for ultimately taking out the noise in it and therewith easier to lemmatize and tokenize it. Stemming can be seen as bringing a word to its original form and in this way, words with the same writing and the same meaning in a sentence will be ‘stemmed’ as one word. Consigned, consigning, consignment can be stemmed to ‘consign’ for example. The process of tokenization can simply be seen as making separate words which can be used again as individual words but also taking out words which cannot be useful anymore. The lemmatization technique makes it possible to group together the inflected forms of a word, so they can be analyzed as a single item. I have also used Part Of Speech-tagging to make the text even more readable and clean for the model. Through this tagging technique we can mark up a word in a corpus to a corresponding part of a speech tag, based on its context and definition. In general tags being used are: ‘NOUN’, ‘VERB’, ‘ADJECTIVE’ and ‘ADVERB’.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

# Prepare Dictionary and Corpus

It is time to make the cleaned text ready for the dictionary and the Bag-of-Words used as a corpus. The Bag-of-Words is a representation of text cleaned text that describes the occurrence of words and its id within a document, comparable to the ‘mixture-components shared across groups’ from the above ‘mixed-membership model’. The dictionary is being made through the genism corpus with ‘Dictionary’. After the dictionary is made we can take out the extremes, with the words which are being mentioned less than 10 times over the corpus and more than 0.6 . The reason for that is to extract the least used and the most used words.

A screenshot of a cell phone

Description automatically generated

# LDA Mallet Model

Now the model will be able to read the input for the algorithm for topic modeling. We have a dictionary and a corpus for the model:

A screenshot of a cell phone

Description automatically generated

For use of the LDA Mallet , a specific installation needs to be made which we can find back in the ‘MALLET\_PATH’.

The MALLET topic modeling toolkit contains efficient, sampling-based implementations of Latent Dirichlet Allocation, Pachinko Allocation, and Hierarchical LDA. The latter models are out of the scope of this project but are recommendable for further research. As the description says, it includes an efficient implementation of Limited Memory BFGS, and many other optimization methods.

As seen in the screenshot above, I have chosen to apply 8 topics to come out of the model. The reason for that is based on a CoherenceModel used on the model and by simply looking at the number of topics used during the sessions of The UN Delegates.

A screenshot of a cell phone

Description automatically generated

A picture containing map, white, man, hanging

Description automatically generated

A screenshot of a social media post

Description automatically generated

When we look at these different scores we can conclude that the model can be seen as a good model in NLP.

# Feature Extraction for further visualization

We can get now the specified topic which are returned back by the model, in ‘lda\_mallet.show\_topic’ and work with it to build a DataFrame.

A screenshot of a cell phone

Description automatically generated

Further information of the model and the specific data therewith can be taken out through feature extraction. Through the code we can get the information about the most dominant topic per document, but also the percentage used of the specific topic in the specific document from the corpus.

A screenshot of a social media post

Description automatically generated

# Visualization of the Topics and Distribution of Words

A screenshot of text

Description automatically generated

# Visualization With Topic Frequency and Word Count within a Theme

A picture containing drawing

Description automatically generated

# Conclusion

As per document and thereby per session with the specific country, we can see that in general the mentioned topics are followed as planned per agenda of the SDG (Sustainable Development Goals). The SDG’s are coherent to the yearly debates which can also be seen in the visualization of the topics mentioned above. If we look at the ‘Word Count’ and the calculated ‘Weights’ within each theme, we can say that these indicators can only be seen in a topic as a distribution with the total amount of ‘1’. Meaning, if we look at theme 3, we see that ‘law’ and ‘freedom’ go hand-in-hand because to maintain ‘law’ the countries have to respect ‘freedom’ as well. But as with text, unlikely as with numbers, the interpretation is subjective.

1. https://www.kaggle.com/unitednations/un-general-debates [↑](#footnote-ref-1)
2. http://www.cs.columbia.edu/~blei/fogm/2015F/notes/mixed-membership.pdf [↑](#footnote-ref-2)
3. Blei, David M.; Ng, Andrew Y.; [Jordan, Michael I](https://en.wikipedia.org/wiki/Michael_I._Jordan) (January 2003). Lafferty, John (ed.). ["Latent Dirichlet Allocation"](https://web.archive.org/web/20120501152722/http:/jmlr.csail.mit.edu/papers/v3/blei03a.html). [Journal of Machine Learning Research](https://en.wikipedia.org/wiki/Journal_of_Machine_Learning_Research). **3** (4–5): pp. 993–1022. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1162/jmlr.2003.3.4-5.993](https://doi.org/10.1162%2Fjmlr.2003.3.4-5.993). Archived from [the original](http://jmlr.csail.mit.edu/papers/v3/blei03a.html) on 2012-05-01. Retrieved 2006-12-19. [↑](#footnote-ref-3)