**Introduction**

Earlier, email was used for merely exchanging communications and information. Now, it is used in many applications such as online shopping, accessing news, and accessing online services. This caused people to receive more and more mechanized emails. The average person receives more than one hundred emails a day (Campbell, 2022). Reading such a number of emails not only is unnecessary but also wastes a significant amount of time. In the United States, in 2021, users spent about five and a half hours daily checking emails and spent more time checking personal emails including newsletters, promotional and automated emails (Statista, 2022), indicating the person giving more time on unnecessary emails. Further, even for useful emails, a user might have his own order of priorities. It is, therefore, organizing incoming emails has become very important for the users not just to ensure that the important emails are read timely but also to handle all the emails efficiently.

Topic modeling, a natural language processing technique based on unsupervised machine learning algorithms, provides a way to organize emails, automatically, by creating topics based on the contents of the emails. We used the Latent Dirichlet Allocation algorithm to create clusters of emails for the [Kaggle dataset](https://www.kaggle.com/datasets/nitishabharathi/email-spam-dataset?select=lingSpam.csv) and got three well-separated clusters each containing set of different words. We analyzed these clusters and their distribution of words to label the name of each cluster.

**The overall process**

The figure below shows the process we adopted to model the topic for the email dataset.



Figure 1 Process applied for topic modeling

**Preprocessing**

The raw emails are preprocessed to remove stop words, non -alphabetic characters, and punctuations. Also, the emails are normalized and lemmatized.

**Model Development using Latent Dirichlet Allocation (LDA)**

We used gensim library to use Latent Dirichlet Allocation (LDA) algorithm to develop and train the model. The preprocessed emails are passed as input to the algorithm that tokenizes all the preprocessed emails and creates a document term matrix, based on the bag of words approach.

***LDA algorithm and its parameters***

LDA uses Dirichlet distributions for document-topic and word-topic distributions, allowing for the Bayesian inference over a three-level hierarchical model (Blei et.al, 2003, p.993) as shown in figure 2.



Figure 2 Graphical Representation of LDA model (Blei et.al., 2003, p.997)

The M represents the number of documents and the Ѳₘ is document topic distribution for document m. N is the number of words in the document, and Wₘₙ is the observed word for the mth document. So, Zₘₙ is a word topic assignment for Wₘₙ. K is the number of topics and 𝜑ₖ is the word distribution for the topic k. The hyperparameter ⍺ controls the mixture of topics for any document, and 𝝱 controls the distribution of words per topic[[1]](#footnote-1) (Naushan, 2020).

We set default parameters and K = 8 for the base model. The outputs of the model, which are document topic and word topic distribution matrices are evaluated using the coherence score and manual assessment. We got a decent coherence score of 0.52. However, output matrices contained overlapping key topic words. So, to improve the model, we iterate the process by extending stop words and removing the twenty most common words. Also, we optimized the coherence score by selecting the value of K = 3, in the combination of default values of other parameters.

**Output Assessment**

At an optimized coherence score of about 56 percent, we got three well-separated clusters of words as shown in figure no 3, in which each cluster represents a topic for the email. Further, each cluster contains an almost equal number of tokens, suggesting it is a good model. We further, assigned the label of each email, that was dropped earlier to build the LDA model, to see whether spam emails are clustered separately. Figure 4 shows that most of the spam emails are clustered in topic 1, hence this shows spam emails have the same tokens or words.

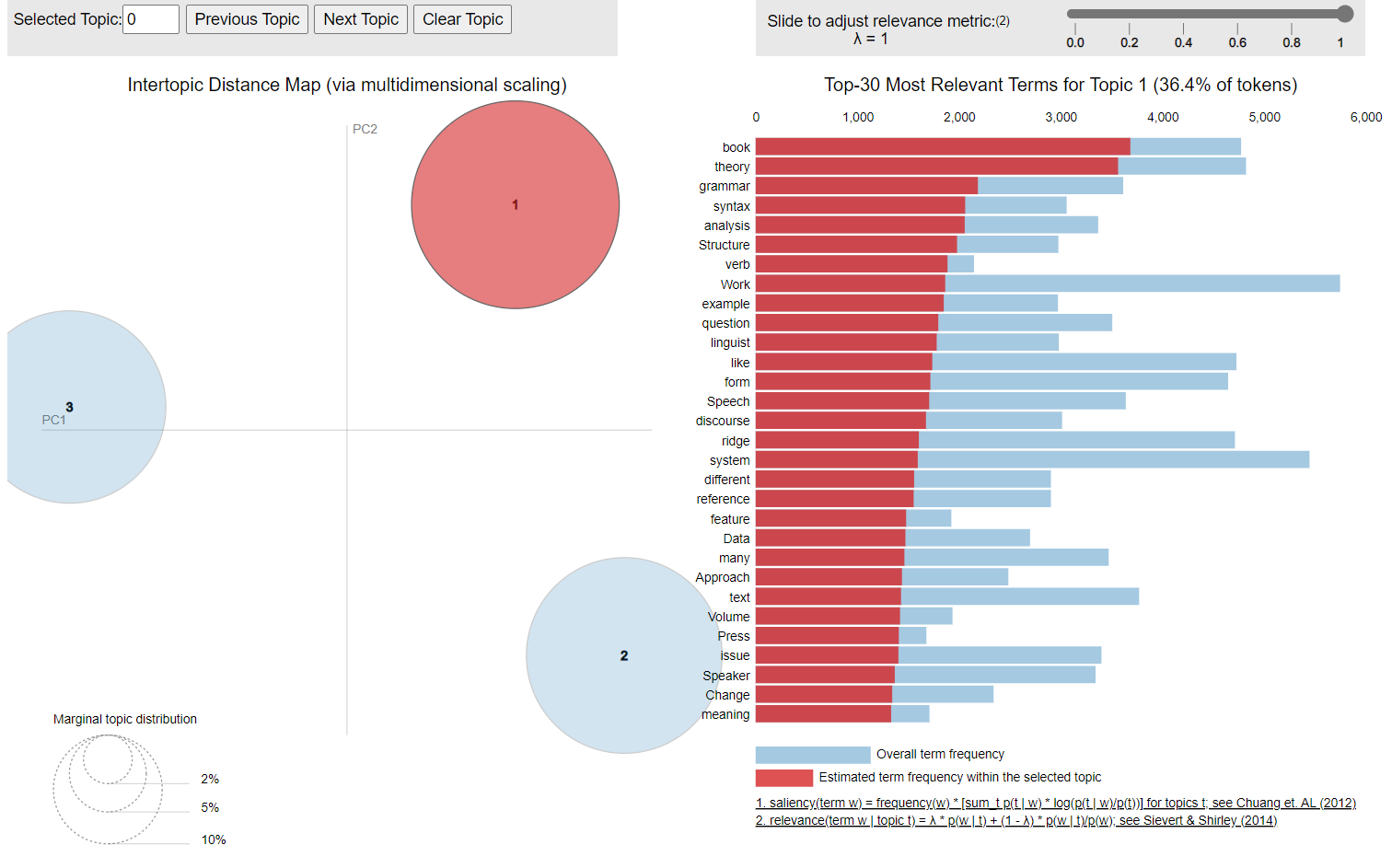
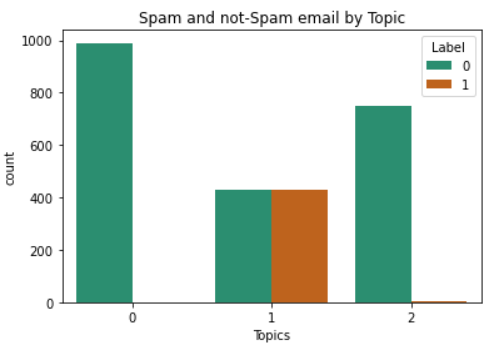


Figure 3 Distribution of spam and non-spam emails by topic

Figure 4 Clusters of words, an output visualization of LDA model

**Label Assignment**

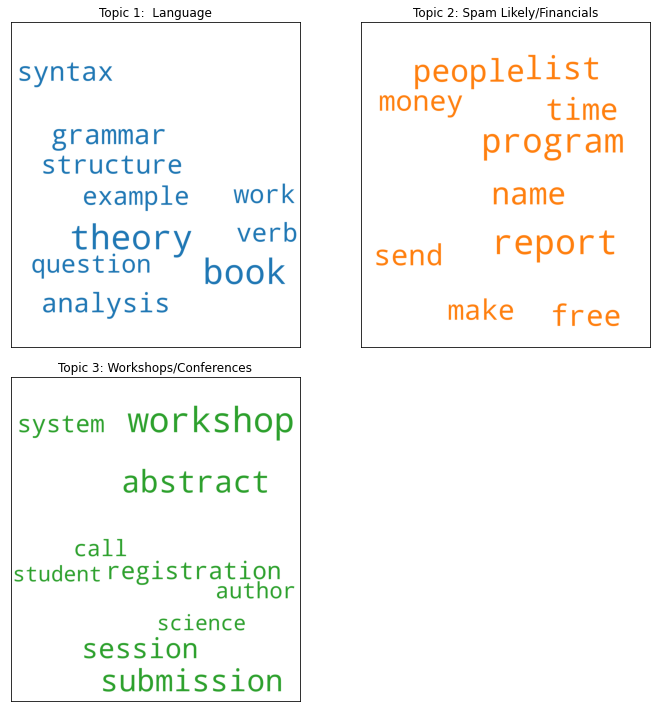
We assigned a label for each topic based on the keyword’s distributions for each topic as shown in figure 5. For example, in topic 1, keywords are syntax, grammar, structure, theory and etc., So, we label this topic as “Language”.

Figure 5 Topic keywords distributions and label assignment

**Conclusion:**

To enhance the user experience in management of their email inboxes, we addressed to model the topics for the emails.

Topic modeling provides a way to organize emails, automatically, by creating topics based on the contents of the emails.  We used the Latent Dirichlet Allocation algorithm and got three well-separated clusters each containing a set of different words. We labeled a name, which is a topic, for each cluster by analyzing its distribution of the words.

**References:**

Blog Vector. (2015). *LDA Alpha and Beta Parameters - The Intuition.* Retrieved from https://www.thoughtvector.io/blog/lda-alpha-and-beta-parameters-the-intuition/

Campbell, S. (2022, November 23). *How many emails does the average person receive per day in 2022?* Retrieved from The small business blog : https://thesmallbusinessblog.net/how-many-emails-does-the-average-person-receive-per-day/

David M.Blei, A. Y. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research 3*, 993 -1022. Retrieved from https://dl.acm.org/doi/pdf/10.5555/944919.944937

Naushan, H. (2020, December 2). *Topic Modeling with Latent Dirichlet Allocation.* Retrieved from towards data sience : https://towardsdatascience.com/topic-modeling-with-latent-dirichlet-allocation-e7ff75290f8

Statista. (2022, September 15). *Minutes spent daily checking emails in the U.S. 2016-2021, by type*. Retrieved from statista: https://www-statista-com.remote.baruch.cuny.edu/statistics/1332517/time-spent-checking-emails-us-users-daily/

1. In the symmetric distribution, the higher the value of these hyperparameters, the higher will be the density (Blog Vector, 2015). [↑](#footnote-ref-1)