Topic Modeling of Research Papers, Visualization and Correlations of topics

Sri Divya Pagadala Srinidhi vaddempudi Vennela Kilaru Sowmya Golla Manoj Koduri

ABSTRACT

Many techniques are used to obtain topic models. This paper aims to demonstrate the implementation of LDA: a widely used topic modeling technique and how to visualize the obtained topics. We use the topic model Latent Dirichlet Allocation (LDA) from genism package for extracting the topics from the abstracts and titles of different research papers present in the almetric data. We show the correlation among different topics extracted from both the abstracts and titles. We have done different visualizations like frequency distribution of word counts in documents, word counts of Top N keywords in each topic, word counts of topic keywords, sentence chart colored by topic, dominant topics in the documents, pyLDAVis interactive chart visualization.

KEYWORDS

Topic Models, Visualization, Correlation.

1. INTRODUCTION

Knowing what people are talking about and understanding their problems and opinions is highly valuable to businesses, administrators, political campaigns and it's hard to manually read through such large volumes and compile the topics. Thus, is required an automated algorithm that can read through the text documents and automatically output the topics discussed. One of the primary applications of natural language processing is to automatically extract what topics people are discussing from large volumes of text. Some examples of large text could be feeds from social media, customer reviews of hotels, movies, user feedbacks, news stories, e-mails of customer complaints etc.

Topic Modelling is a technique to extract the hidden topics from large volumes of text. Latent Dirichlet Allocation (LDA) is a popular algorithm for topic modelling with excellent implementations in the Python's Gensim package. We have used the Latent Dirichlet Allocation (LDA) from Gensim package. We have extracted the volume and percentage contribution of each topic to get an idea of how important a topic is. The challenge, however, is how to extract good quality of topics that are clear, segregated and meaningful. This depends heavily on the quality of text pre-processing and the strategy of finding the optimal number of topics.

Recent growth in text data affords an opportunity to study and analyze language at an unprecedented scale. The size of text corpora, however, often exceeds the limit of what a person can read and process. While statistical topic models have the potential to aid large-scale exploration, a review of the literature reveals a scarcity of real-world analyses involving topic models. So, visualization can be done to analyze the topics that come from the topic model.

The LDA model assumes that the words of each document arise from a mixture of topics, each of which is the distribution of the vocabulary. A limitation of LDA is the inability to model topic correlation even though, for example, a document about sports is more likely to be also about health than international finance. We have built a topic model and then found the correlation of the topics and then visualized the data.

2. DATA PREPROCESSING

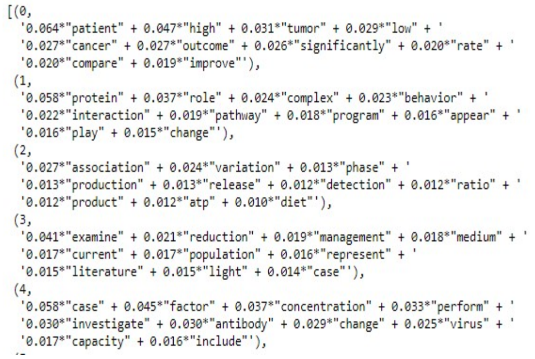
Here we have chosen 15774 files which is of 9GB data from the almetric data and used LDA to extract the naturally discussed topics. We have extracted the titles and abstracts from the selected files and converted to the data frames. Several steps are done for the pre-processing of data. After importing the data, the first step is tokenizing the sentences and cleaning them. The first step is about removing the emails, new line characters, single quotes, special character's using regular expressions and finally split the sentences into a list of words using gensim's simple_preprocess(). Setting the deacc=True option removes punctuations. The next step is to remove the stop words and do lemmatization. Stop words are extracted from NLTK package. Lemmatize each word to its root form, keeping only nouns, adjectives, verbs and adverbs. We keep only these POS tags because they are the ones contributing the most to the meaning of the sentences. Here, we used spacy for lemmatization. We keep only these POS tags because they are the ones contributing the most to the meaning of the sentences. Here, we used spacy for lemmatization. Lemmatization is the process of converting a word to its base form. The difference between stemming and lemmatization is, lemmatization considers the context and converts the word to its meaningful base form, whereas stemming just removes the last few characters, often leading to incorrect meanings and spelling errors. The third step is creating bigram and trigram models. Bigrams are two words frequently occurring together in the document. Trigrams are 3 words frequently occurring. Gensim's Phrases model is used to build and implement the bigrams, trigrams. The 2 important arguments to phrases are min_count and threshold. The higher the values of these params, the harder to form bigrams.

3. TOPIC MODEL, CORRELATION

The following are key factors to obtain good segregation topics:

- The quality of text processing.
- The variety of topics the text talks about.
- The choice of topic modeling algorithm.
- The number of topics fed to the algorithm.
- The algorithms tuning parameters.

The Topic Model is built using LdaModel() in Gensim's package. The two main inputs to the LDA topic model are the dictionary and the corpus. So, we created the dictionary and corpus. Gensim creates a unique id for each word in the document. The produced corpus is a mapping of (word_id, word frequency). In addition to the corpus and dictionary, you need to provide the number of topics as well. We have built LDA model with 20 different topics where each topic is a combination of keywords and each keyword contributes a certain weightage to the topic. Figure 1 represents the topics obtained after building an LDA model. It shows the topic number and all the keywords present in the topic. Looking at these keywords, we can guess what this topic could be? You may summarise it as health from topics 1. Model perplexity and topic coherence provide a convenient measure to judge how good a given topic model is. Topic coherence score, in particular, has been more helpful. We got the perplexity of -10.5653 and the coherence score of 0.340711 for our model.



```
[(0,
  '0.064*"patient" + 0.047*"high" + 0.031*"tumor" + 0.029*"low" + '
  '0.027*"cancer" + 0.027*"outcome" + 0.026*"significantly" + 0.020*"rate" + '
  '0.020*"compare" + 0.019*"improve"'),
 (1,
  '0.058*"protein" + 0.037*"role" + 0.024*"complex" + 0.023*"behavior" + '
  '0.022*"interaction" + 0.019*"pathway" + 0.018*"program" + 0.016*"appear" + '
  '0.016*"play" + 0.015*"change"'),
 (2,
  '0.027*"association" + 0.024*"variation" + 0.013*"phase" + '
  '0.013*"production" + 0.013*"release" + 0.012*"detection" + 0.012*"ratio" + '
  '0.012*"product" + 0.012*"atp" + 0.010*"diet"'),
 (3,
  '0.041*"examine" + 0.021*"reduction" + 0.019*"management" + 0.018*"medium" + '
  '0.017*"current" + 0.017*"population" + 0.016*"represent" + '
  '0.015*"literature" + 0.015*"light" + 0.014*"case"'),
 (4,
  '0.058*"case" + 0.045*"factor" + 0.037*"concentration" + 0.033*"perform" + '
  '0.030*"investigate" + 0.030*"antibody" + 0.029*"change" + 0.025*"virus" + '
  '0.017*"capacity" + 0.016*"include"'),
```

Fig. 1: *Topics obtained from the LDA model.*

Now that the LDA model is built, the next step is to examine the produced topics and the associated keywords. There is no better tool than pyLDAvis package's interactive chart and is designed to work well with jupyter notebooks. We have successfully built a good-looking topic model.

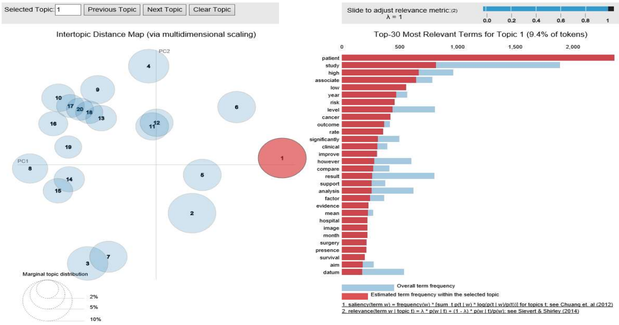
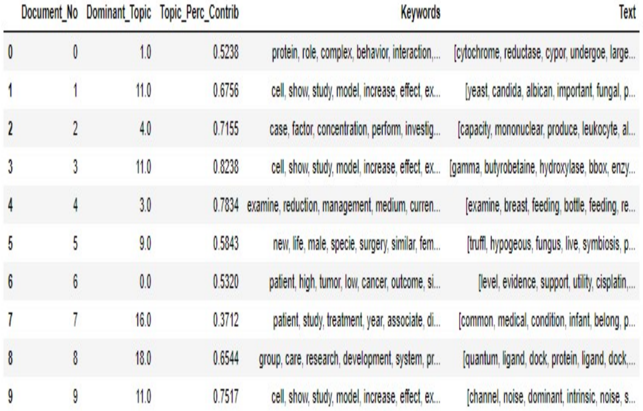

Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
4
9
10
17
20
18
13
16
11
12
6
19
1
8
14
15
5
2
3
7
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (9.4% of tokens)
0
500
1,000
1,500
2,000
patient
study
high
associate
low
year
risk
level
cancer
outcome
rate
significantly
clinical
improve
however
compare
result
support
analysis
factor
evidence
mean
hospital
image
month
surgery
presence
survival
aim
datum
Overall term frequency
Estimated term frequency within the selected topic
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)

Fig. 2: *Visualization of obtained topics*

Each bubble on the left-hand side plot represents a topic. The larger the bubble, the more prevalent is that topic. A good topic model will have big, non-overlapping bubbles scattered throughout the chart instead of being clustered in one quadrant. A model with too many topics, will typically have many overlaps, small sized bubbles clustered in one region of the chart.

In LDA models, each document is composed of multiple topics. But, typically only one of the topics is dominant. To find that, we find the topic number that has the highest percentage contribution in that document. We built a topic model for 20 topics. Figure 3 shows the dominant topic in each document and the percentage of dominant topic present in a document.



	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	1.0	0.5238	protein, role, complex, behavior, interaction,...	[cytochrome, reductase, cypor, undergoe, large...
1	1	11.0	0.6756	cell, show, study, model, increase, effect, ex...	[yeast, candida, albican, important, fungal, p...
2	2	4.0	0.7155	case, factor, concentration, perform, investig...	[capacity, mononuclear, produce, leukocyte, al...
3	3	11.0	0.8238	cell, show, study, model, increase, effect, ex...	[gamma, butyrobetaine, hydroxylase, bbox, enzy...
4	4	3.0	0.7834	examine, reduction, management, medium, curren...	[examine, breast, feeding, bottle, feeding, re...
5	5	9.0	0.5843	new, life, male, specie, surgery, similar, fem...	[truffl, hypogeous, fungus, live, symbiosis, p...
6	6	0.0	0.5320	patient, high, tumor, low, cancer, outcome, si...	[level, evidence, support, utility, cisplatin,...
7	7	16.0	0.3712	patient, study, treatment, year, associate, di...	[common, medical, condition, infant, belong, p...
8	8	18.0	0.6544	group, care, research, development, system, pr...	[quantum, ligand, dock, protein, ligand, dock,...
9	9	11.0	0.7517	cell, show, study, model, increase, effect, ex...	[channel, noise, dominant, intrinsic, noise, s...

Fig. 3: *Dominant topics in each document.*

The LDA model assumes that the words of each document arise from a mixture of topics, each of which is a distribution over the vocabulary. A limitation of LDA is the inability to model topic correlation. From the dataset of research papers, we have taken abstracts and titles and extracted topics from both and performed correlation for those topics. Figure 4 shows the correlation between the topics obtained from the abstracts.

Figure 5 shows the correlation between the topics obtained from the titles.

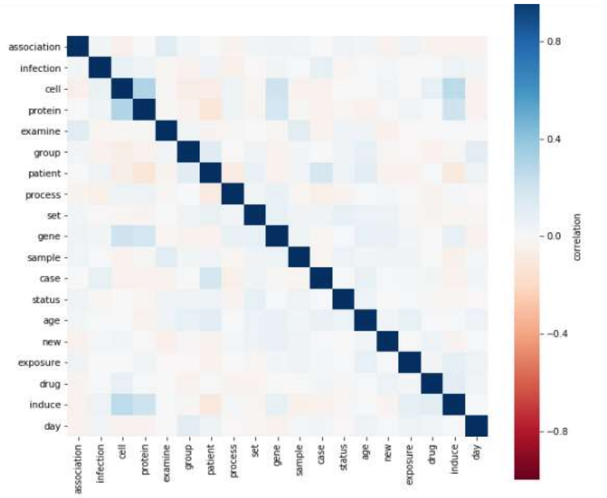


Fig. 4: *Correlation for topics obtained from Abstracts.*

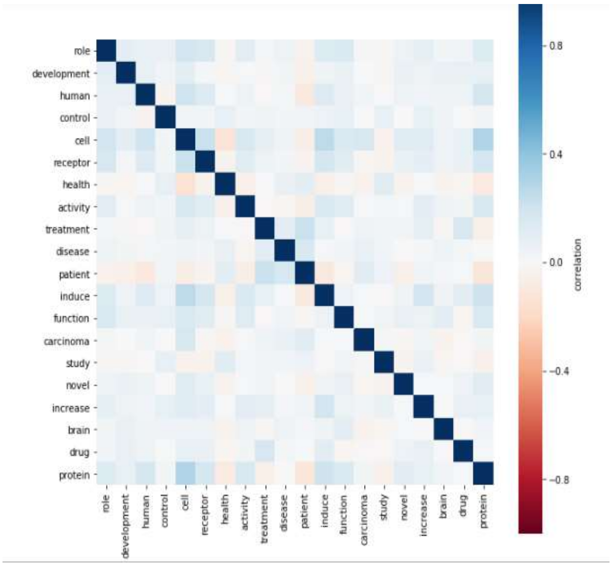
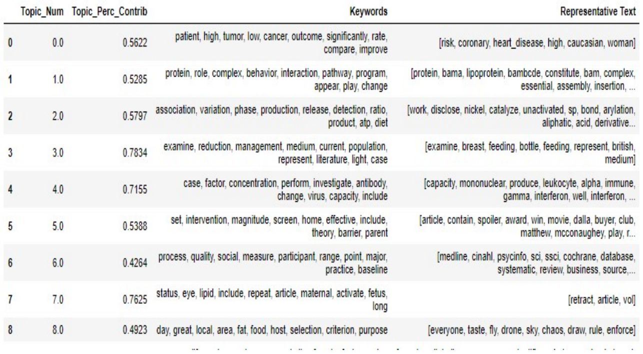


Fig. 5: *Correlation for topics obtained from Titles.*

4. VISUALIZATION

Sometimes just the topic keywords may not be enough to make sense of what a topic is about. So, to help with understanding the topic, you can find the sentences a given topic has contributed to the most and infer the topic by reading that document. Finally, we want to understand the volume and distribution of topics in order to judge how widely it was discussed. Figure 6 shows all the topics and their representative sentences. It also shows the topic numbers followed by its percentage contribution , keywords of each of the topics.



	Topic_Num	Topic_Perc_Contrib	Keywords	Representative Text
0	0.0	0.5622	patient, high, tumor, low, cancer, outcome, significantly, rate, compare, improve	[risk, coronary, heart_disease, high, caucasian, woman]
1	1.0	0.5285	protein, role, complex, behavior, interaction, pathway, program, appear, play, change	[protein, bama, lipoprotein, bambcde, constitute, bam, complex, essential, assembly, insertion, ...
2	2.0	0.5797	association, variation, phase, production, release, detection, ratio, product, atp, diet	[work, disclose, nickel, catalyze, unactivated, sp, bond, arylation, aliphatic, acid, derivative...
3	3.0	0.7834	examine, reduction, management, medium, current, population, represent, literature, light, case	[examine, breast, feeding, bottle, feeding, represent, british, medium]
4	4.0	0.7155	case, factor, concentration, perform, investigate, antibody, change, virus, capacity, include	[capacity, mononuclear, produce, leukocyte, alpha, immune, gamma, interferon, well, interferon, ...
5	5.0	0.5388	set, intervention, magnitude, screen, home, effective, include, theory, barrier, parent	[article, contain, spoiler, award, win, movie, dalla, buyer, club, matthew, mcconaughey, play, r...
6	6.0	0.4264	process, quality, social, measure, participant, range, point, major, practice, baseline	[medline, cinahl, psycinfo, sci, ssci, cochrane, database, systematic, review, business, source,...
7	7.0	0.7625	status, eye, lipid, include, repeat, article, maternal, activate, fetus, long	[retract, article, vol]
8	8.0	0.4923	day, great, local, area, fat, food, host, selection, criterion, purpose	[everyone, taste, fly, drone, sky, chaos, draw, rule, enforce]

Fig. 6: *Representative sentence for each topic.*

As we are working with a large number of texts, we also wanted to know how big the documents are i.e. the total number of words present in the document. In Figure 7, we can observe the number of documents and the number of words present in each document.

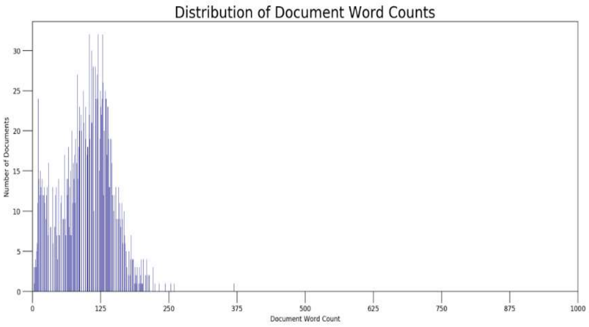


Fig. 7: *Distribution of Document Word Counts.*

The topic keywords for all the topics are known. We now made a word cloud with the size of the keywords proportional to the weights of those keywords. Different colors have been assigned to each topic and for the further visualizations, the color for a particular topic will be the same as allocated now. Figure 8 shows the word clusters.

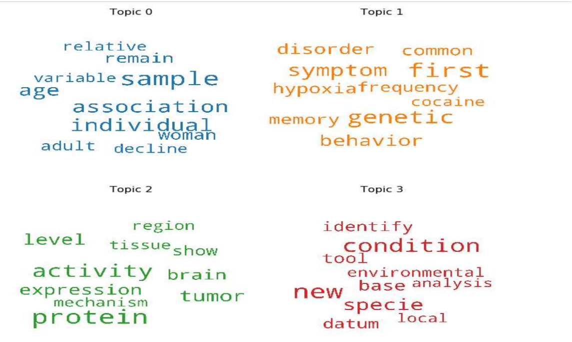


Fig. 8: *Word clouds of top keywords in each topic.*

When it comes to the topic keywords, there are two important factors, they are the weight of the keyword and the frequency of occurrence of a keyword in a document. We should keep an eye out on the words that occur in multiple topics and the ones whose relative frequency is more than the weight. Often such words turn out to be less important. Figure 9 is a result of adding several such words to the stop words list in the beginning and re-running the training process. It is a plot for both the word count and word weight.

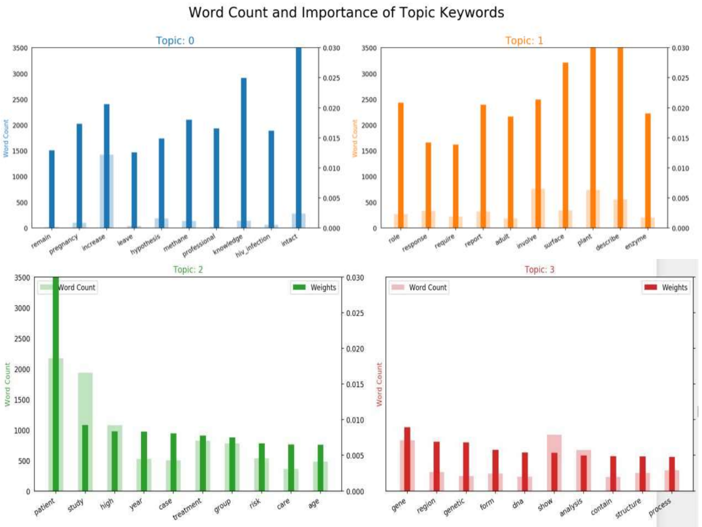


Fig. 9: *word count's and weights of keywords.*

Each word in the document is represented by one of the four topics extracted. we've colored each word in the given documents by the topic id it is attributed to. The color of the enclosing rectangle is the topic assigned to the document. Figure 10 shows the sentence topic colouring for the documents.

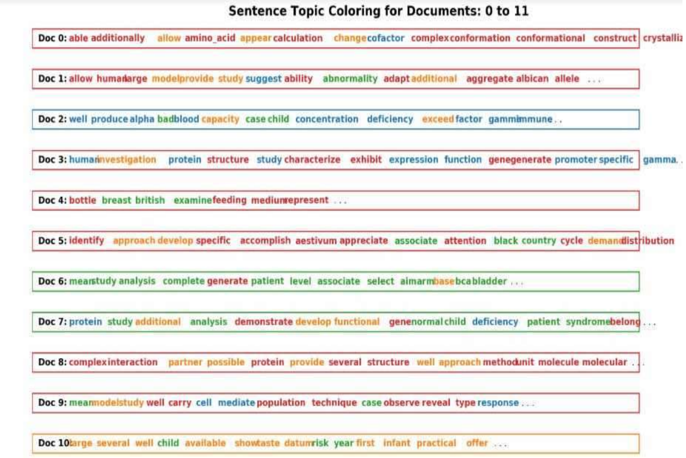


Fig. 10: *Sentence Topic Coloring for documents.*

We also computed the total number of documents attributed to each topic. Figure 11 shows the two plots; the first plot is the number of documents for each topic by assigning the document to the topic that has the most weight in that document and the second plot is the number of documents for each topic by summing up the actual weight contribution of each topic to respective documents.

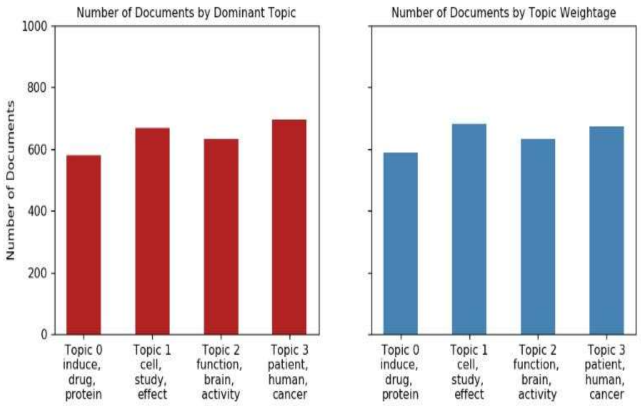


Fig. 11: *Number of Documents containing a particular topic*

5. CONCLUSION AND FUTURE WORK

In this paper, we built an LDA topic model and we have shown several preprocessing steps done before building the model i.e. steps like removing stop words, tokenization ,lemmatization. Then, after the model is built we found the topics manually by seeing the keywords obtained from the topic model. A limitation of LDA is its inability to find the topic correlation. We have found the correlation among topics by extracting both the abstracts and titles from the dataset. We have shown multiple ways to visualize the data. This model can be implemented using large amount of data than we used, and the results might vary for the correlation. This model can be improved by using Mallet's version of LDA algorithm.

REFERENCES:

1. https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24
2. https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/
3. https://www.aclweb.org/anthology/R09-1061
4. D. Hall, D. Jurafsky, and C. Manning. Studying the history of ideas using topic models. In Empirical Natural Language Processing Conference, 2008.
5. J. Han and M. Kamber. Data Mining: Concepts and Techniques. 2nd Edition. Morgan Kaufmann, San Francisco, 2006.