Topic Modeling PubMed's Articles Using LDA and BERT

Introduction

In science, we tend to superspecialize. This has given us Science that is efficient, standardized, and rigorous, leading to many technological advances in the recent decades. Yet, overspecialization can lead to 'thought and knowledge silos' that make it difficult for scientists to know about and critically evaluate scientific work outside of their micro-fields. With this in mind, I looked at the field of 'stem cell therapies' and used machine learning to find the strongest predictors of an article's influence level and hidden topics within the field.

Background

Scientific fields nowadays have many micro-topics that are highly specialized. Exploring the literature will present tens and hundreds of thousands of papers, which must be filtered for relevance. Review articles can be a good start, although there can be many of them and each one could be quite a long read. Take 'Covid-19' for example, a quick search on PubMed, a google for medical and life sciences publications, gives 158K query results (Covid-19 has only been around for just over a year at the time of writing!). What is the main consensus in the field? What kind of subtopics exist within the field? Even as somebody working in a similar field, it is not easy to answer these questions. This led me to my project:

`Can I use machine learning to elucidate key subtopics within a biomedical / life sciences field?`

This is a great task to apply data science techniques because we can leverage the computer to go through the tens and hundreds of thousands of scientific articles for us to find hidden topics within them. The resulting topics can aid in understanding the macro trends of the field and narrowing down the search. It can also lead to unexpected findings which may inspire interdisciplinary thinking and collaboration.

Data

For this task, I collected my own dataset by scraping information about 9000+ articles from a PubMed search query for the term '*stem cell therapies*'. Each entry included the article ID, title, type of publication, abstract, journal title, number of authors, author affiliations, number of citations, keywords, and number of references used. Since the information was extracted from the web page's html, the data had to be extensively cleaned. I removed duplicates, extracted extra information, filled in missing values, and reformatted the text data.

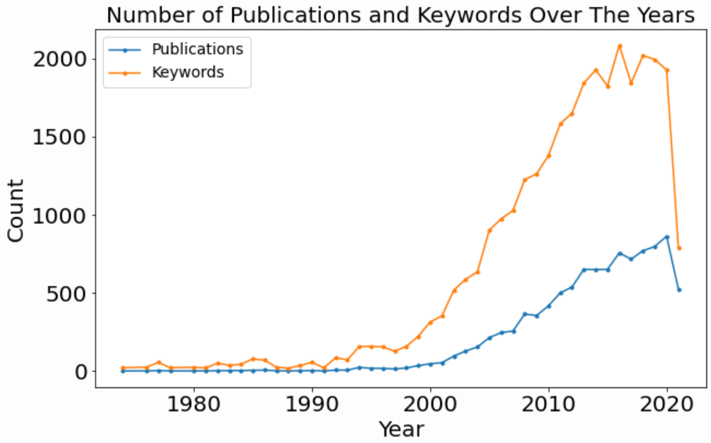

Fig 1. Number of Publications and Unique Keywords Between 1975-2021

From the initial analysis, I saw that the number of unique keywords from all of the articles published within a year has seen a dramatic increase since the 2000s at a rate exceeding that of the increase in the number of publications in the same time frame (**Fig. 1**). My main task was to reduce this down to 10-100 topics.

With this data, I performed the following analyses: a classification model to identify key predictors of an article's influence, followed by a Latent Dirichlet Allocation (LDA) model and a Bidirectional Encoder Representations from Transformers (BERT)-based model to cluster the articles into subtopics.

Models

1. **Classification**

   To understand the strongest predictors of an article's influence, I created three classes based on the number of citations to target my classification models. I preprocessed the data by count vectorizing the text, imputing missing values and normalizing the numeric data, and one hot encoding the categorical data. I tested Support Vector Classifier, Logistic Regression, K Neighbors Classifier, and Decision Tree Classifier leveraging the Gridsearch Pipeline to tune the hyperparameters. The resulting models were evaluated through their accuracy, precision, and recall scores.

2. **LDA Topic Modeling**

   To generate topics from the articles, I used Term Frequency Inverse Document Frequency (TF-IDF) on the titles + abstracts and fitted an LDA model on the resulting vectorized text to generate 10 topics. To tune the model, I adjusted some thresholds for the TF-IDF calculations, added stop words, and stemmed the vocabulary. Each iteration was evaluated through the most important keywords from the resulting topics.

3. **BERTopic Topic Modeling**

   For the BERT-based model, all of the text data besides the keywords were incorporated into one block of text that retained most of the original formatting. The resulting text was transformed into a vector with 768 dimensions using a pretrained model called 'allenai-specter', reduced to 5 dimensions and clustered into 48 groups. For each group, TF-IDF was performed to extract the top 10 words. I leveraged the GPU available in Google Colaboratory and reduced the model fitting time by ~15-fold.

4. **Topic Model Evaluation**

   To compare between the two topic models, I looked at the average number of unique keywords from 100 randomly sampled articles within each topic. For a further qualitative analysis, I chose three subtopics which were similar between the two models and extracted the top cited papers from both to look at how related they were to the main themes.

Findings

The best classification model was able to predict an article's influence level by 76.5%. The model identified the strongest predictors for the influence level to be mainly the journal titles, which is in line with the current tendency for scientists to judge a publication by the impact factor of the journal. However, the model was confused about moderate and strong influencers categorizing many of them as bad, most likely due to the heavy skew in the distribution of citations. Log-transforming the citations before binning could improve the results.

The LDA topic model showed promise after stop words were removed, minimum document frequency was set, and the number of features were reduced in TF-IDF. In some of the ten topics generated, different organ types were clustered in separate groups and disease types corresponded to organ types. This is in line with the field where the specializations have generally branched by specific organ types. The two groups which were the easiest to identify were heart- and blood-related topics, which were also easily identified as clusters when visualized. This could be improved by trying a different vectorizer, as well as tweaking the parameters for the vectorizer and the LDA model.

With BERTopic, the resulting clusters were sensible and well separated, mostly by the organ type. Interestingly there were topics specific for imaging and engineering. Within the organs, the biggest topic by number was the heart. For the brain and blood, there were subtopics mainly by disease type indicating relatively further advancements in these organ types. However, it is important to note that there were many outliers identified. I hope to reduce this by fine tuning the parameters for dimension reduction and clustering.

Direct comparison of topics generated by LDA and BERTopic showed that BERTopics were more sensible when evaluated by the average number of unique keywords in each topic. For similar topics, the articles in BERTopics were more related to the overall themes. The pretrained model with the bidirectional embedding seemed to solidify the model's contextual understanding of texts for better clustering. We must, however, consider the fact that the inputs were different for the two models and LDA was forced to separate the topics into 10 groups, while HBDSCAN was allowed to optimize the number of BERTopics. I would like to do a fairer comparison by fine-tuning both models to generate a similar number of topics from similar inputs. One caveat to 'allenai-specter' used in BERTopic is the fact that it was trained on citation graphs of scientific publications. As well as the fact that I added the journal title into the texts. This may end up biasing the groupings by the citations or journal, similarly to the current landscape of scientists filtering articles by a journal's reputation. Furthermore, I want to improve the quantitative measure by finding a way to normalize it to the number of documents found in each topic. I plan to incorporate other measures like perplexity and cohesion, as well.

Summary

With this project, I looked at the strongest predictors for an article's level of influence using a logistic regression model and identified key topics through topic modeling with LDA and BERTopic from the 9000+ articles scraped from PubMed about the field of 'stem cell therapies'. The many superspecializations within the field resulted in 2000+ unique keywords at the current moment. From there, I was able to cluster the articles using their title, abstract, journal title, publication year, and author information and boil it down to 48 unique subtopics using BERTopic. Returning to my original question, can I use machine learning to elucidate key subtopics within a biomedical / life sciences field, though not perfect, I believe the answer is a resounding yes. The groupings showed that the field is specialized mainly by organ type. There were also multidisciplinary topics like imaging and engineering, as well as further specialization for certain organs by disease type. I hope that these findings can help mitigate overspecialization by making the identification of subtopics within the field easier and lead to more interdisciplinary thinking and collaboration.

Future Steps

In the immediate future, I will work on fine-tuning the topic models and the evaluation metrics. In the near future, I would like to find a way to simplify the titles and abstracts from jargon to English. Down the line, I want to leverage a bigger dataset like SciDocs to investigate topic relations between larger topics. One day, I would love to realize this analysis into a usable web app to support researchers. My vision is to have a querying platform ("`FieldView`") for scientific papers, which will allow users to contextualize their query in relation to other topics within and surrounding the field, and navigate to a recommended reading list with some unexpected insights.

References

Latent Dirichlet Allocation
https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24
https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know

BERTopic
https://maartengr.github.io/BERTopic/index.html
https://github.com/MaartenGr/BERTopic
https://arxiv.org/pdf/2004.07180.pdf

Superspecialization in Science
https://theconversation.com/scientists-tend-to-superspecialize-but-there-are-ways-they-can-change-51644
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3993417/