Information Organization and Retrieval
Info 202 - Implementation Project

Research Methodology Comparison

AI/NLP (Topic Modeling) vs Social Sciences (Grounded Coding)

Github link - https://github.com/guptaanik41/Info202_Project_Fall22

Being an aspiring Data Scientist with a prior background in ML (mainly supervised), I am highly interested in learning unsupervised methods and this project gave me the right opportunity to explore that. I will be using NLP techniques, to create categories from a *News headlines dataset*. Primarily, my focus will be Latent Dirichlet Allocation (LDA), which is a topic modeling technique to extract topics from a large set of documents.Human grounded coding is a common method employed in the field of social sciences whereas NLP is an AI algorithm which gives computers the same power as humans to analyze text.

Its results will be compared to LDA using the Inter-Annotator agreement. The project will also provide insights about each method's distinct benefits and limitations, as well as the broader roles that these methods play in the respective domains where they are frequently employed. These discoveries will help us to better understand the trade-offs involved in selecting various approaches to textual data analysis. Thus, trying to highlight the research gap between the AI/NLP techniques and the research methods in human social sciences.

The Dataset consists of over a million news headlines that have been published over a period of 19 years. It has been taken from Kaggle from the link shared above. *A thorough analysis of the results will also provide us valuable insights into the major topics in news over the past decade and also how those topics evolved with time?*

So, we are comparing two different methods of analyzing textual data, namely -

1. Human Grounded Coding
2. Natural Language Processing (NLP)

Part 1 : Topic Modeling using LDA (from NLP)

Topic Modeling is an unsupervised ML algorithm used for clustering documents into different topics (buckets) based on the content they hold. Latent Derelicht Analysis, aka LDA technique has been employed for performing the same.

1. Importing *nltk* and *gensim* libraries along with other basic ones such as *pandas*, *numpy*, etc. for loading the dataset.

2. Data Pre-processing

- *Function used: gensim.utils.simple_preprocess()*
- Tokenization - Converting a document into sentences and the sentences into words which are further converted into lowercase tokens.
- Dropping words that are too short or too long, *min_len* = 4, *max_len* = 15.
- Discarding the stopwords present in the data
- Employing Stemming and Lemmatization techniques

3. Creating Word Vectors using the Bag of Words technique

    - Creating mapping of all the words, i.e. token with their respective integer ids.
    - Filtering the dictionary
        - *no_below* to filter out words appearing less than 15 times
        - *no_above* to filter out words appearing in more than 60% of all documents
        - *keep_only* the first 10000 most frequent tokens
    - Converting the above dictionary into a bag of words. Basically, for each document a dictionary is created which has various words and their respective frequency.

4. Creating Word Vectors using the TF-IDF technique as well.
5. Perform LDA based topic modeling using both Bag of words and TF-IDF techniques. Here, we have created 10 topics.
6. Each headline is assigned one of the topics from the 10 topics that we created above. This is done with the help of *'get_document_topics'* which returns the probability of a particular headline belonging to every topic. For a headline to belong to a particular topic, the corresponding probability should be maximum.

Note - Detailed comments have been provided in the code and so, the things have not been repeated here.

Assigning a specific topic code to each of the 10 topics created using LDA

After creating 10 topics using LDA, we have given a title code to each one of them. This will be useful when categorizing the headline into various topics, i.e. each headline will be given a title code.

Since the size of data is 1.2M, we have selected a generous dataset of 40 headlines. The sample is small enough for us to read and significant enough to perform NLP techniques.

Part 2 : Grounded Coding (from Sociology/Human Sciences)

Grounded coding is performed on the same data sample that we used above. Here, we are reading the text ourselves and then assigning one of the 10 title codes that we generated above using topic modeling.

Grounded coding is performed in two ways -

1. There are 10 topics (Topic0, Topic1, …, Topic9) generated by the LDA. Each of the topics is given a code. Grounded coding is then performed on the data using these codes.

2. The text is read multiple times and the codes are generated manually. Each headline in the dataset is then assigned a code.

Part 3 : Results and Insights

As we learned in week 7 of our class, Inter Annotator Score or Kappa tells us how well a given set of annotators (2 in our case - LDA and grounded coding) can make the decisions of classifying a certain document into the same category.

For this, the grounded coding performed in the first way is used, i.e. categorisation obtained by using the topic modeling codes. As per the calculations shown in the excel sheet, the value of Kappa is **78.5%.** So, these both nearly give us the same results.

1. Grounded theory is a bottom-up methodology and LDA based topic modeling is an unsupervised ML learning method. The former involves reading and re-reading again and again whereas the latter is basically an iterative process carried out many times.

2. Grounded theory is an open-minded approach, i.e. it doesn't involve any specific mathematical constraints or any stringent rules, whereas LDA is a math heavy scientific approach. There are various assumptions in LDA which are mathematically coded into the model.

3. In grounded theory (~1960s) , we start from the documents and don't come with any idea from the documents (*we are at the ground level*). The general procedure followed for grounded coding involves:

   > Reading data > form various concepts from the data > find interconnection between things, collect and group concepts into various categories > assign each category a respective code, i.e. generalizing into high level categories using codes. Here '>' indicates the next step.

4. In LDA (~1990s), we start with documents and we know that there are a certain number of topics present. We write an algorithm to repeatedly read these documents and build a connection between topics and text. The topics that are generated to us with the help of code are the concepts that we further use for analysis.

5. Grounded theory approach requires a lot of time and effort to read, re-read, form codes, and categorize various texts into respective categories. LDA is a computerized approach which significantly takes less time (*it may take good time if the textual data is HUGE*). This doesn't mean that LDA is easier. It takes good thinking to write code and work with large datasets but the point is to have a computer who assists us consistently.

6. Based on my understanding from the project, topic modeling can be used in areas of insight-driven analysis, where we are specially looking for some high level insights. This might explain the difference in the performance.

7. Humans are much better at identifying the commonalities in different texts. They are better at understanding emotions present in a text as compared to machines. Hence, we tend to observe different categories formed corresponding to the same emotions by LDA whereas a human probably put such text under the same umbrella.

8. There are no standard rules of performing grounded theory and hence everyone tends to achieve a different result. The case is different for topic modeling using LDA wherein we will get the same topic clusters provided the dataset and the hyperparameters are the same. Also, the grounded theory method which we employed in our previous assignments as well produces a large amount of data to be handled manually. Moreover, I observed that since grounded theory is the majority of manual work in making categories, we as humans tend to involve our personal bias into the approach. Hence, one should not bring any of his prior knowledge and perform the tasks objectively.

9. Having said that, grounded theory brings humans creativity into the picture while forming categories which any computational method lacks in analyzing textual data. It also allows us to go as much in-depth into a concept as possible. It also follows a systematic and defined way of data analysis as compared to other qualitative research methods. But it also opens the door for the possibility of high error since the majority of the work is done manually and quite exhaustive in nature.

10. As per my research, there are innovative ways to combine both grounded coding and topic modeling in order to leverage the advantages of both the techniques. This research paper - Combining topic Modeling and Grounded coding proposes an interesting way to do this.

    Steps followed in the paper-

    a. Raw coding - Perform LDA based topic modeling to generate topics.
    b. Expert coding - Experts naming the various topics and providing relevance points to each topic.
    c. Focus Coding - Engage with the codes formed and find any possible similarities, patterns within the codes.
    d. Theory Building - Combine various categories and compare them. Develop themes and build new knowledge.

I believe we can use the power of LDA to form the various broad categories and then use the concepts of Ground coding to categorize various documents/text into those topics. This technique can somehow reduce the time and the manual effort required to read and re-read the textual data to form various categories by Grounded coding.

References - https://dl.acm.org/doi/10.1002/asi.23786, https://guides.temple.edu/c.php?g=77914&p=505635, https://dongpng.github.io/attached/