Information Organization and Retrieval
Info 202 - Implementation Project

Research Methodology Comparison

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

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

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

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.

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.

Part 3 : Results and Insights

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.

As per the calculations, 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.

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 that 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 same emotions by LDA whereas a human probably put such text under the same umbrella.