TEXT ANALYSIS FOR A COMPANY BID CENTER

CAPSTONE PROJECT SUMMARY FOR MS IN BUSINESS ANALYTICS

Hasan Mansoor Khan

Table of Contents

Introduction & Problem Setting

This Business Analytics project is focused on the text analysis of certain technical documents known as bids, tenders, requests for proposal (RFPs) or other public procurement notices in Europe. The client is a large firm which has dedicated Bid Center to study and examine large technical documents for identifying relevant projects based on business needs, product suitability and project feasibility in line with business opportunities and constraints. These documents are large in terms of length and contain excessive legal jargon and detailed descriptions of the project. The key is to identify the topic of the tender document which is in German language and identify if the relevant project advertised is of interest to the company or not. Currently many documents are manually analyzed which can cost the bid center in terms of lost opportunity due to time constraints and processing limitations. The aim of this text analytics project is due to utilize the vast potential of Natural Language Processing Techniques in analyzing the text documents for the client. The task is to devise an analytical strategy with the client to a develop an automated tool that takes the text of large documents as input and return more about that text such as common words, co-occurrence patterns or key underlying themes and topics. This text analytics project is therefore more centered around topic identification as opposed to the more commonly performed sentiment analysis. The major limitations observed are the availability of confidential and proprietary ownership of certain text documents which hampers the open analysis of these documents. Certain private and public portals exist which provide tender documents for either specific countries or the entire European region. The data provided can be searched based on Common Procurement Vocabulary or CPV codes. These prove instrument in acquiring a sample data base of text files based on specific words or keywords found in the description of these CPV codes. There more than 9000 CPV codes, specifically 9454 codes which all contain descriptions of the type of work or job they describe. Hence, once domain specific queries are made, the client & I agree on 36 specific CPV codes and their descriptions to proceed. Against these 36 specific codes certain PDF documents which are downloaded in PDF format and an illustration database is created for the purpose of detailed text analysis and project fit identification. The detailed process includes but is not limited to data processing, text preprocessing, feature engineering, model preparation and model application such as topic modelling or predictive analysis. The project is technically discussed in a separate technical discussion,

but this project summary includes a brief overview of the steps followed, key exploration techniques and topic modelling performed.

Feature Engineering & Text Preprocessing

One of the key components of data analysis, especially numerical data analysis is to identify certain variables and perform feature engineering. This can include complete transformation of key variables dependent on their distribution. However, in the field of text analysis the data processing pipeline is significantly different due to the nature of the data which is textual data. Text preprocessing is done in two specific stages: generic text preprocessing and specific preprocessing after exploring the data. To start with all the text files are examined and basic parameters are evaluated such as token length and language detection. A text file cannot be studied using Natural Language Processing techniques all the words are extracted from the text and specific tokens are created. The tokens are then uploaded in a data frame where there are two specific columns, the file name and the text containing tokens. The language detection tool identifies that the text my client and I are examining is in German language which will therefore require specific libraries to examine. Therefore, in the next steps the Spacy and NLTK libraries are specified that the text is in German. The first step in text preprocessing was to remove the stop words found in all the text files. Once this step is done, the text is stemmed so that each word is brought to its basic root word. Later, another column is added to include the process of text lemmatization which helps understanding the meaning or core of the word involved. The above mentioned three traditional steps are done for experimenting the libraries in this context of PDF tender documents. A further cleaning process is expected once exploratory data analysis is done. That is because each document must be further cleaned based on certain key characteristics.

Data Exploration & Analysis

Once the data is organized and a working environment has been established, one of key aspects of the text analysis capstone project is with respect to exploratory data analysis. This is the process where more statistical aspects of the data are figured out. In the context of text analysis, it is imperative that the text is first cleaned and tokenized. This is because the analysis as to shift to a certain aspect into quantifiable aspects that are ready for statistical analysis and modelling. For instance, one area of exploratory data analysis is the length of each document and its distribution. These basic steps help in application of further text analysis tools such as words correlation, co-occurrence, n-grams and so on. The token lengths are explored to check the distribution of the data and check for extreme values or potential errors. The text data is checked for token length for each document and visualized by box plots and bar and line charts. However, this just a count of tokens and acts more like a measure of data preparation rather than actual topic identification. The line chart captures this variation but as discussed, the overall length remains only measure of sanity check rather than a meaningful insight. The text documents are then studied for co-occurrence patterns which examines pair of words co-occurring simultaneously in the text data set. For example, certain key words appear frequently enough and n-grams such as “Slidescanner” and “Bilddatenamnagementsystem” are uncovered. Furthermore, n-grams, specifically two, three and four words based on keywords are visualized. The data exploration also includes most frequent words founds in a text and single document analysis which includes the creation of word clouds for specific document. The exploratory data analysis also shows a network graph where key topics and their relationships are displayed. However, more importantly bar charts and word clouds show the most common words or themes in the sample data analysis. Co-occurrence patterns are also visualized using tables sorted by frequency & scatter plots to better comprehend the two co-occurring words in the dataset.

Topic Modelling

Topic modelling in the realm of text analysis involves applying statistical models and helps identify underlying topics, clusters and clusters in each text body or dataset. This project begins with applying Term Frequency- Inverse Document Frequency (TF-IDF) analysis which identifies each token and its frequency via a value. In the resulting TF-IDF data frame each row remains a particular document, but each column of the data frame is a specific token. Therefore, there are 1923 variables or features each with a 0 to 1 value denoting if they are included or not in the text. The value increases correspondingly to the number of times the token emerges in the dataset and is then offset by the number of documents in the corpus. Furthermore, I proceed in applying a Non-Negative Matrix Factorization which is an attempt at topic categorization. Here, the dominant topics as expected turn out to be common themes in the sample dataset. This step is instrumental in creating a specific feature cleaning function for specific text preprocessing as outlined in the proceeding section. Then a part of speech tagging is conducted which emphasized the grammatical aspects of the text. Each token is identified for its POS such adverb, verb, noun, proper noun or even space. Finally, a Latent Dirichlet Allocation Analysis (LDA) is performed with an input of number topics as 5. Here the 5 common themes can be browsed interactively using an LDA visualization chart where circles on the left are representing topics & the size of the circles depicts prevalence of the topic. Then distance amongst each circle manifests the similarity in topics. The right-hand side show bar charts display the distribution or frequency of topics while hovering over a particular circle reveals the top keywords related to that specific topic. Amongst many other features of the LDA visualization the cell color intensity portrays the probability of the word belonging to that topic. The LDA visualization tool proved to be immensely insightful in topic categorization and key word analysis, but the results constrained by the nature of the sample data which is controlled. However, it proved that the models can scaled for application with real data and large data sets for topic categorization. In addition to LDA, the Naïve Bayes model has also been applied based on a prototype success index which is categorizing data into highly, moderately relevant, or irrelevant categories. The Naïve bayes is applied for illustration purposes (with nearing 71 percent for certain categories) and the client can apply a success index based on actual business needs and relevant supporting data to train the model along & predict whether a particular text is relevant or not.

Future Suggestions and Next Steps

Topic modelling has proven to be essential in text analysis as it allows for identifying underlying themes and topics based on the characteristics of the given text. Despite technical and legal limitations, the sample dataset proves that word counters, token frequency, co-occurrence patterns, keywords identification, NMF models, LDA models as well as Naïve Bayes predictive models are feasible and helpful for making informed business decision and maximizing the capacity of the Bid Team in securing the most suited business projects based on the products that the company is providing. A larger training data set and application of real-life large text data sets within the legal guidelines governing public procurement documents can provide a robust analytical model for the Bid Team which will be instrumental in achieving best LDA and Naïve Bayes scores. Significant generic feature engineering was performed in this text analysis project which included but was not limited to text preprocessing techniques. However, for future application with large data sets, it is highly recommended that domain specific customization will prove to boost the model in terms of efficiency. This is because based on domain specific knowledge, the bid team can modify the feature representation to reflect the properties of the text and adapt the training data accordingly. Lastly, apart from advanced topic modelling including real time topic modelling which is based on actual real time data and is aimed to enable timely analysis and identify upcoming or emerging topics which can in turn allow for proactive data management and analysis by the companies bid center.