Topic modeling for Australian Legal cases document

Background

Legal documents are generally written in complex legal language. Often, the task of interpretation too hard to understand legal documents and creating summaries is also laborious process, which requires considerable time and, therefore, money. Often, law firms have to hire additional support staff to assist with these tasks. One way to minimize the effort and cost is developing NLP in Machine learning algorithm, which creates summaries of legal texts. The following project shows the development of a topic modelling algorithm used to summarise full text legal documents using Australian Legal case data set from UCI machine learning repository. The data set contains around 4000 cases from the Federal court of Australia (FCA) for the year of 2006, 2007, 2008, and 2009.

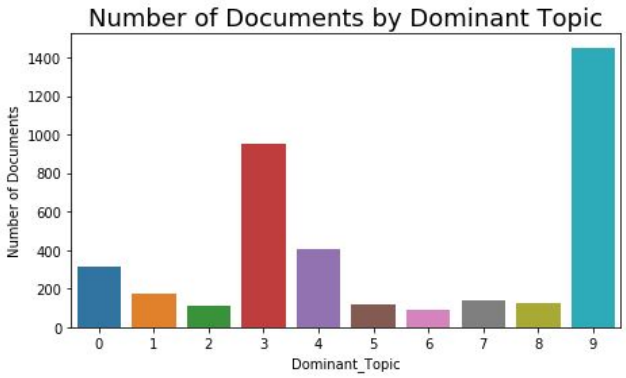
Data Exploration

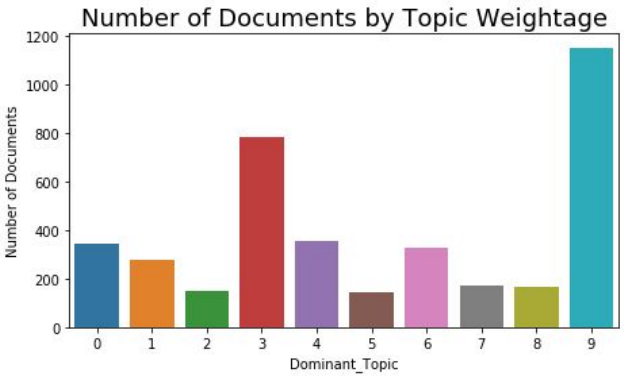
The data files are in the XML format. Each case is represented with an individual XML file. Beautiful soup packages are used to parse fields such as case name, link and sentences from XML files and to convert into pandas dataframe.

Text preprocessing

The text data can be used in machine learning algorithms by converting it into numbers. Ideally, a few techniques (bag of words, TF-IDF, Word2Vec, and GloVe etc.) could be tested to insure the most optimal accuracy; however, considering the time and computational resource constraints of this assignment, the following project is limited to the Bag of Words and TF-IDF techniques. These techniques are a frequency based approach, which converts text documents of the corpus to n-dimensional vectors. To prepare the data for the topic modelling process, information regarding counsel, solicitors, date info, non-english words, symbols, characters, numerical numbers, extract spaces, and most common English stop words are removed. Also, NLTK lemmatization is used to remove suffixes and converting each word into its original form.

Topic Modelling





Topic modeling is the one of statistical modeling algorithms used to summarize a large text. Different statistical techniques such as LDA, LSA, Non-Negative Matrix factorization, etc are used for topic modeling. LDA technique is used for topic modeling of the legal text data in this example. For the purpose of LDA, documents are considered as a collection of topics in a certain proportion and each topic is the collection of keywords in certain proportions. LDA technique is used to rearrange the topics distribution within each document and analyze keyword distribution within each topic to obtain a relevant composition of Dirichlet distribution. In this example, the Gensim library for NLP is used to implement LDA with both, Bag of Words and TF-IDF vectors. The figures above illustrate top 9 topics from the 4000 legal documents, where each topic is associated with its own set of words (see the fig. on the previous page) though some of them are common for a few topics (this could be improved with further tuning of parameters and more detailed text processing). For instance, topic-4 is associated with words related to visa, fraud, immigration, etc, which suggest the documents of this group are related to immigration cases. The distribution plot of the number of documents by dominant topics suggests that topics 3 and 9 appear in most cases. Finally, the plot of topic clusters below created using Bokeh library shows that the topics are clearly divided among these legal documents. The next steps for this project could be implementation of a more abstract approach for text embedding using Word2Vec or Glove for topic modeling and text summarization of legal text data.