Topic modeling and unsupervised classification

NSF Research Awards Abstracts

Gustavo Gouvea

The goal of my code is to analyze a collection of NSF Research Awards Abstracts stored as XML files to identify topics in the documents and classify them.

First, I import necessary libraries like NumPy, pandas, scikit-learn, gensim, and NLTK for data manipulation, topic modeling, and natural language processing. I also download essential NLTK resources such as stopwords and wordnet.

To prepare the document data, I clean common phrases, remove punctuation, tokenize the text, and eliminate stopwords using my customized set (which I created by manually analyzing common words in the documents). Additionally, I lemmatize the words to ensure consistency.

Next, I perform topic modeling using Latent Dirichlet Allocation (LDA). I create a document-term matrix and apply LDA to extract topic probabilities for each document. By examining the most significant words associated with each topic, I gain insights into the underlying themes.

To classify the documents into clusters based on their topics, I employ k-means clustering. I specify the desired number of clusters (three in my case) and use the feature matrix derived from the topic probabilities.

The results include the identification of topics within the document collection and the classification of documents into clusters. I print the most significant words for each topic to gain further understanding of the identified themes. Finally, I export the classification results, including file names, cluster numbers, and document abstracts, to a CSV file named 'results.csv'.