Topic Modelling On News Articles

Harish Kollana

Data science trainees,

AlmaBetter, Bangalore.

Abstract:

The British Broadcasting Corporation (BBC) is the national broadcaster of the United Kingdom. Headquartered at Broadcasting House in London, it is the world's oldest national broadcaster, and the largest broadcaster in the world by number of employees, employing over 22,000 staff in total, of whom approximately 19,000 are in public-sector broadcasting.

Keywords: Lsa, Lda

1. Problem Statement

In this project your task is to identify major themes/topics across a collection of BBC news articles. You can use clustering algorithms such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) etc.

The dataset contains a set of news articles for each major segment consisting of business, entertainment, politics, sports and technology. You need to create an aggregate dataset of all the news articles and perform topic modeling on this dataset. Verify whether these topics correspond to the different tags available.

news: News Content

type: Type of News Category

2. Introduction

Our goal is to identify major themes/topics across a collection of BBC news articles by using clustering algorithms such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA).

3. Steps Involved:

- ### Data Cleaning:

  In Data Cleaning we dealt with stop words, punctuations, leammatization and unnecessary symbols from text of news articles.

- ### Explorative Data ANALYSIS:

  After cleaning the dataset, we created word clouds for every type of topic. This process helped us figuring out what are the most

popular words in that topic of articles.

- ## Topic Modelling:

  For the topic modelling we tried various Regression algorithms like:

  1. LSA - Latent Semantic Allocation
  2. LDA - Latent Dirichlet Allocation

4. Algorithms:

- ## LSA - Latent Semantic Allocation:

  Latent Semantic Analysis (LSA) involves creating structured data from a collection of unstructured texts. Before getting into the concept of LSA, let us have a quick intuitive understanding of the concept. When we write anything like text, the words are not chosen randomly from a vocabulary.

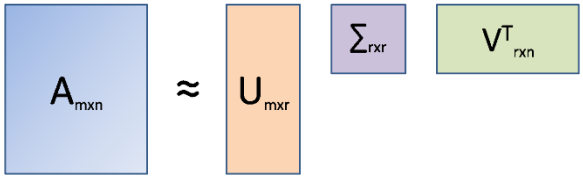
  Rather, we think about a theme (or topic) and then chose words such that we can express our thoughts to others in a more meaningful way. This theme or topic is usually considered as a latent dimension.

  It is latent because we can't see the dimension explicitly. Rather, we understand it only after going through the text. This means that most of the words are semantically linked to other words to express a theme. So, if words are occurring in a collection of documents with varying frequencies, it should indicate how different people try to express themselves using different words and different topics or themes.

  In other words, word frequencies in different documents play a key role in extracting the latent topics. LSA tries to extract the dimensions using a machine learning algorithm called Singular Value Decomposition or SVD.

  What is Singular Value Decomposition (SVD)?

  Singular Value Decomposition or SVD is essentially a matrix factorization technique. In this method, any matrix can be decomposed into three parts as shown below.



Here, A is the document-term matrix (documents in the rows(m), unique words in the columns(n), and frequencies at the intersections of documents and words). It is to be kept in mind that in LSA, the original document-term matrix is approximated by way of multiplying three other matrices, i.e., U, ∑ and VT. Here, r is the number of aspects or topics. Once we fix r (r<<n) and run SVD, the outcome that comes out is called Truncated SVD and LSA is essentially a truncated SVD only.

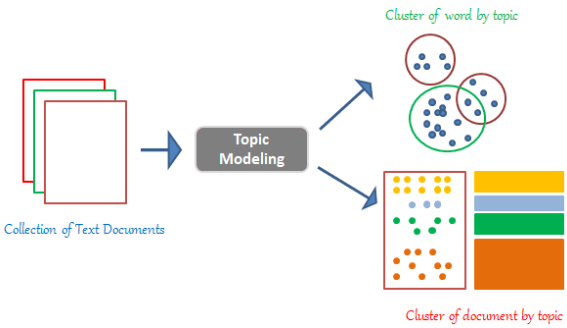
SVD is used in such situations because, unlike PCA, SVD does not require a correlation matrix or a covariance matrix to decompose. In that sense, SVD is free from any normality assumption of data (covariance calculation assumes a normal distribution of data). The U matrix is the document-aspect matrix, V is the word-aspect matrix, and ∑ is the diagonal matrix of the singular values. Similar to PCA, SVD also combines columns of the original matrix linearly to arrive at the U matrix. To arrive at the V matrix, SVD combines the rows of the original matrix linearly. Thus, from a sparse document-term matrix, it is possible to get a dense document-aspect matrix that can be used for either document clustering or document classification using available ML tools. The V matrix, on the other hand, is the word embedding matrix (i.e. each and every word is expressed by r floating-point numbers) and this matrix can be used in other sequential modeling tasks. However, for such tasks, Word2Vec and Glove vectors are available which are more popular.

- ## LDA - Latent Dirichlet Allocation:

Imagine walking into a bookstore to buy a book on world economics and not being able to figure out the section of the store that has this book, assuming the bookstore has simply stacked all types of books together. You then realize how important it is to divide the bookstore into different sections based on the type of book.

Topic Modelling is similar to dividing a bookstore based on the content of the books as it refers to the process of discovering themes in a text corpus and annotating the documents based on the identified topics.

When you need to segment, understand, and summarize a large collection of documents, topic modelling can be useful.



Latent Dirichlet Allocation (LDA) is one of the ways to implement Topic Modelling. It is a generative probabilistic model in which each document is assumed to be consisting of a different proportion of topics.

How does the LDA algorithm work?

The following steps are carried out in LDA to assign topics to each of the documents:

1) For each document, randomly initialize each word to a topic amongst the K topics where K is the number of pre-defined topics.

2) For each document d:

For each word w in the document, compute:

P(topic t| document d): Proportion of words in document d that are assigned to topic t

P(word w| topic t): Proportion of assignments to topic t across all documents from words that come from w

3) Reassign topic T' to word w with probability p(t'|d)*p(w|t') considering all other words and their topic assignments

The last step is repeated multiple times till we reach a steady state where the topic assignments do not change further. The proportion of topics for each document is then determined from these topic assignments.

Illustrative Example of LDA:

Let us say that we have the following 4 documents as the corpus and we wish to carry out topic modelling on these documents.

Document 1: We watch a lot of videos on YouTube.

Document 2: YouTube videos are very informative.

Document 3: Reading a technical blog makes me understand things easily.

Document 4: I prefer blogs to YouTube videos.

LDA modelling helps us in discovering topics in the above corpus and assigning topic mixtures for each of the documents. As an example, the model might output something as given below:

Topic 1: 40% videos, 60% YouTube

Topic 2: 95% blogs, 5% YouTube

Document 1 and 2 would then belong 100% to Topic 1. Document 3 would belong 100% to Topic 2. Document 4 would belong 80% to Topic 2 and 20% to Topic 1.

This assignment of topics to documents is carried out by LDA modelling using the steps that we discussed in the previous section. Let us now apply LDA to some text data and analyze the actual outputs in Python.

5. Conclusion:

In this Project we had analyzed the BBC articles using LDA and LSA Topic modelling techniques and found lsa seems more impactful on segregation of topics.

In future we can use one of model to predict the user input of text query to type of news. We can recommend news articles to the users by following these methods.

References-

1. Analyticsvindhya