Topic Modelling On News Articles

Documentation

- By Prasad Khedkar
(AlmaBetter Student)

Abstract:

Topic Modelling is extremely important technique in unsupervised machine learning field for doing analysis on large number of documentations.
Topic models have a wide range of applications like tag recommendation, text categorization, keyword extraction and similarity search in the broad fields of text mining, information retrieval, statistical language modeling.
The present project contains 2225 documents falling under 5 topics based on different news themes.
Models are built using mainly three different algorithms namely, Truncated SVD or LSA/LSI , LDA(Latent Dirichlet Allocation) & lastly NMF(Non- matrix factorization).
Among them LDA and NMF gave the best results.

Problem Statement:

- ***In the given problem, our main goal is to obtain the major topic/themes for given collection of BBC news which are in .txt format, using different algorithms.***

Dataset Information

- The data is unlabelled and unstructured

Features information:

For convenience we gave created dataframe of following :

- **News_text : Contains news in txt format**
- **News_theme : Topic which news belongs to.**

Introduction :

Topic Modelling is great technique in NLP , which comes under the unsupervised machine learning domain.
The main idea is to make the corpus of words from given set of documents & give us the clear-cut idea about the theme or section the document/file belongs to.

Steps performed :

-Data Cleaning

-NLP Text Processing

-Vectorizing the data

-Fitting different models

-Model Evaluation & Selection

Data Cleaning -

- In this step, data is cleaned by removing duplicates and null values for further processing.

NLP Text Processing –

- This is the most important process which deals with removing –
  - word/non-word characters
  -punctuations
  -numbers
  -html tags/urls
  present in the data to completely make it into text format(word format.)

Vectorizing the data :

- After text processing the data we get is unreadable by machine hence to make it readable, it is converted into vector matrix in which most of the entries are zero(sparse matrix)
- For this conversion, two popular vectorizers are used namely, CountVectorizer & TF-IDF vectorizer

- The CountVectorizer simply converts the text data into sparse matrix without giving advantage to unique terms or less frequent terms.
- Whereas the TF-IDF vectorizer gives more preference to the less frequent words appearing in the documents

Fitting the different Models :

Models have been trained based on following algorithms-

A. Truncated SVD
B. Latent Dirichlet Allocation(LDA)
C. Non-matrix factorization(NMF)

Truncated SVD :

The SVD algorithm is mainly used for reducing the dimensions in the dataset for simplifying the calculations.

Main idea is –
Any matrix (mxn) can be decomposed(or reduced) into 3 different matrices –

- *Unitary Matrix , U(m x m)*
- *Rectangular diagonal matrix, ∑(m x n)*
- *Complex unitary matrix, V*(n x n)*

Mathematically,

Original matrix = U.∑.V*

Truncated SVD is the type of SVD where number of columns is equal to the truncation,which is provided by us and everything will be removed.
For e.g., It rounds off the digits after the decimal , 99.98 can be truncated to 100.
As per the original documentation, this estimator does not centre the data before computing, meaning it can work with sparse matrix efficiently.
And since our vectorized data is full of such sparses , we can train this model with given vectorized sparse matrix, along with
truncated columns = no. of topics we want(5)

Latent Dirichlet Allocation:

This is the most popular & simplest topic modelling techniques present out there.
Main idea is- to find the topics/themes of the document based on words present in it.

It goes through two steps –

i. Find the words that belong to documents.

ii. The words (also probability) that belong to the topic, the thing in which we are interested

After this, we can train model by fitting the vectorized data to test the results.

Non-negative matrix factorization:

Just like SVD , this is also a factorization technique but instead of factorizing it into 3 matrices, NMF factorizes the matrix into two matrices i.e.,

Original Matrix(m x n) = W(m x k). H(k x n)

Th

The whole idea is –

NMF clusters the topics based on distribution hypothesis.

This distribution hypothesis is used for searching the words which are similar in meaning & context, occuring throughout the text/document and then group those similar words to give us topic(unlabelled)

Model performance :

Truncated SVD-

The resuts given by this algorithm-

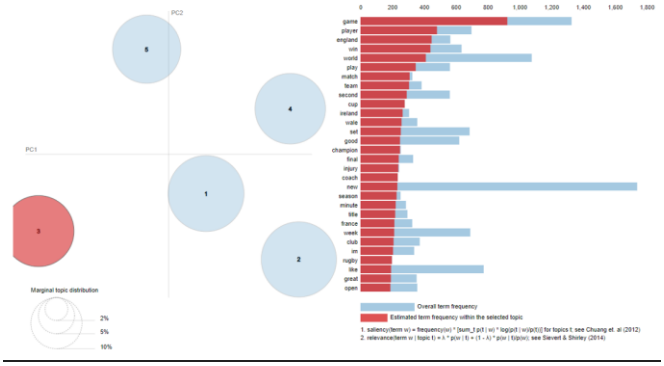
Top 20 for each topic according to SVD model -

	Top 20 words
Topic 1	[people, new, game, government, music, best, like, world, uk, way, think, party, good, service, company, labour, country, mobile, song, right]
Topic 2	[best, song, music, award, angel, robbie, film, game, urban, think, prize, artist, british, dont, stone, im, williams, album, brit, joss]
Topic 3	[best, song, labour, government, party, election, blair, award, tax, tory, music, minister, british, brown, angel, public, robbie, howard, britain, plan]
Topic 4	[game, england, win, party, wale, labour, play, roddick, best, election, team, world, ireland, blair, match, point, nadal, playing, cup, zealand]
Topic 5	[music, party, people, labour, game, election, mobile, urban, phone, tory, blair, ukip, kilroysilk, like, black, howard, joss, mp, campaign, thing]

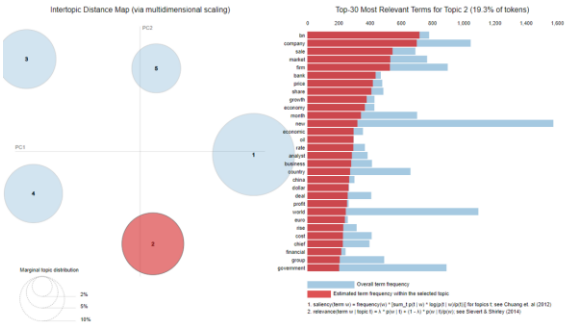
None of the top 20 words represent any topics –
[Business, Entertainment, Politics, Sports, Technology]
Gave us the poor outcome.

LDA & NMF Comparison –

Topic visualization for LDA –



Topic visualization for NMF –



Topics represented in LDA were adjacent to each other on PCA spacing representing the words may interconnect with the adjacent topic.

But the topics in NMF were farther from each others representing the unique words are used to for topic representation.

And is therefore more preferrable over LDA

Final Conclusion :

That's it!.
From start we load the raw data into dataframe, then cleaned the data by removing null & duplicate values present. Then we performed NLP text processing & removed unwanted data to further clean the data . Then we vectorized the data into machine readable matrix and trained different models by fitting the vectorized data.
And from human judgement (since there is no fix way to evaluate topic modellings), the final evaluation and assessment showed that NMF best suited for this topic modelling problem.

Future Scope :

- Neural networks along with NMF , for e.g giving output of NMF as input to neural networks can enhance the results and can give us the better representation.

- Algorithms like top2vec & BERTopic can also be implemented.

References :

Almabetter, towardsdatascience.com