Home

Part 14: Step by Step Guide to Master NLP – Basics of Topic Modelling

CHIRAG GOYAL — June 25, 2021

Advanced Algorithm NLP Project Python Text

This article was published as a part of the Data Science Blogathon

Introduction

This article is part of an ongoing blog series on Natural Language Processing (NLP). In this series, we completed our discussion on the entity extraction technique "Named Entity Recognition (NER)". But at that time, we didn't discuss another popular entity extraction technique called Topic Modelling. So, in continuation of that article, we will discuss Topic modelling in this article.

In this article, we will discuss firstly some of the basic concepts related to Topic Modelling. Probably, from the next articles, we will discuss some of the popular techniques to implement Topic Modelling.

This is part-14 of the blog series on the Step by Step Guide to Natural Language Processing.

Table of Contents

1. What is meant by Topics in NLP?

2. What is Topic Modelling?

3. What is the Importance of Topic Modelling?

4. What are the Objectives of Topic Modelling?

5. How does Topic modelling Works?

6. What are the Popular Techniques to implement Topic Modelling?

Topics in Natural Language Processing

In Natural Language Processing, the word topic means a set of words that **"go together"**. These are the words that come to our minds when we think a bit about that topic.

For Example, if we think sports, then some such words are athlete, soccer, and stadium, etc.

In simple words, Topics are defined as a repeating group of words that are statistically significant to a corpus.

Now, a question comes to mind:

What is the meaning of Statistical Significance here?

It means that the few words are occurring together in the same documents and they have similar ranges of TF-IDF values which

All these points are related to the statistical significance which implies that the group of words is important to the corpus.

Some more examples of topics are:

- A group of words containing the terms such as games, team, hockey, play, etc. This group is mostly about sports topics.
- Another topic group containing the words such as space, NASA, earth, henry, launch, etc. represents space-related topics.

Let's go one step ahead and discuss what exactly is the Topic model.

What is meant by Topic Model?

A topic model is defined as a system that automatically discovers topics occurring in a collection of documents or corpus. Then we may use the trained model for the following purposes:

- To find which of these topics occur in new documents.
- To pick out which portions of a document cover which topics.

For Example, Consider the website "Wikipedia"

This website contains millions of documents that cover hundreds of thousands of topics. So, it would be great if these could be discovered automatically and additionally a finer map of which documents cover which of the topics. These things become very useful for those peoples who want to explore Wikipedia.

We could also find some emerging topics since documents get written about them. In some settings (such as news) where new documents are constantly being formed and recency matters, this would help us to detect the trending topics.

What is Topic Modelling?

Topic modelling is an automatic process that aims to find the hidden topics embedded in the text data. This process is an unsupervised technique which means that we don't have to provide a labeled dataset to topic modelling algorithms and topics are identified automatically by the model.

Topic modelling can also be thought of as a form of text mining approach to obtain recurring patterns of words in the textual material.

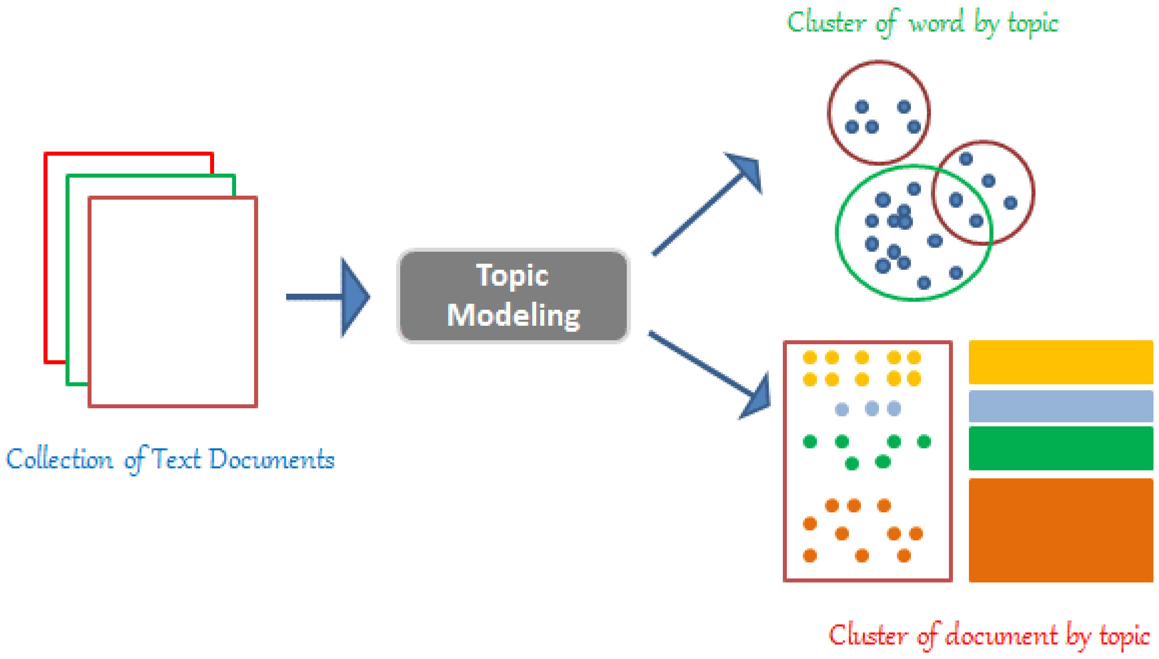

Image Source: Google Images

Therefore, Topic Modelling is the process of dividing a corpus of documents into the following two things:

- A list containing all the topics that are covered by the documents in the corpus.
- Grouped several sets of documents from the corpus-based on the topics they cover.

Here, the underlying assumption is that every document consists of a statistical mixture of topics, which means statistical distribution of topics that can be obtained by "summing up" all of the distributions for all the topics covered.

So, as a concluding step – Topic modelling tries to figure out which topics are present in the documents of the corpus and how strong the presence of those topics in documents are.

Importance of Topic Modelling

As we have known that every document from the corpus we read can be thought of as consisting of many topics all stacked upon one another.

Large amounts of data are collected every day. As more information becomes available, it becomes a tedious task to find what we are looking for. So, we require some sort of tools and techniques to organize, search and understand vast quantities of information.

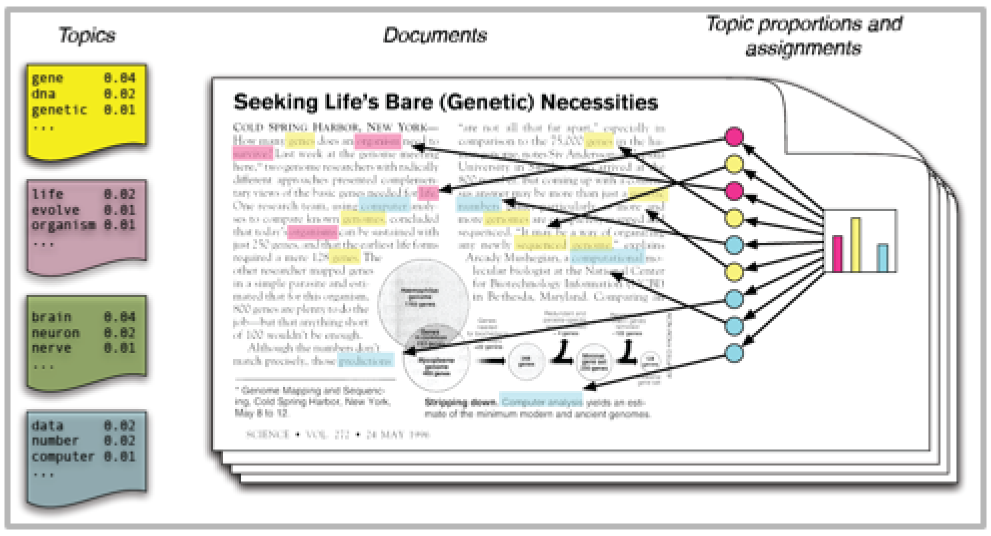

Image Source: Google Images

Therefore, Topic modelling helps us to organize, understand and summarize large collections of textual information.

Topic modelling helps in many ways such as:

- To extract hidden topical patterns that are present across the collection of documents.
- Annotation of all the documents according to these topics.
- With the help of annotations, we can organize, search and summarize texts.

Let's see the following example to discuss the power of Topic Modelling:

As we have discussed Topic modelling is recognizing the words from the topics present in the document or the corpus of data. This technique is useful, as extracting the words from a document takes more time and is much more complex than extracting them from topics present in the document.

For Example, Say we have 1000 documents and 500 words present in each document of the corpus. So while processing it, we require 500*1000 = 500000 threads.

But when you divide the document into certain topics then if we have overall 5 topics present in it, then the processing is just 5*500 words = 2500 threads.

So, we have observed that the processing of the corpus becomes easy if we go ahead with Topic Modelling instead of processing the entire documents.

This is how the topic modelling solves the problem and also helps in visualizing things better.

Objectives of Topic Modelling

Consider a corpus in which a set of documents are provided, these documents are about different types of text. The main goal

1. What are the most important topics?

This can also be called a **Topic term distribution**.

2. What are the topics which are assigned to every document?

This can also be called a **Document to topic distribution**.

Topic Modelling tries to find the latent structure in a text corpus that:

- Resembles "topics" (also "concepts")
- Best summarizes the collection
- Is based on Statistical Patterns
- Are obscured by synonyms, homonyms, stopwords,…
- May overlap

Now based on your understanding let's see the following questions:

Is Topic Modelling similar to Clustering?

Yes, Topic modelling is similar to clustering but with a slightly different "mindset":

- In clustering, the focus is on the data points/documents.
- In topic modelling, the focus is on the topics/cluster themselves.

Topic modelling allows us to cut through the noise (deal with the high dimensionality of text data) and identify the signal (the main topics) of our text data.

Extracting topics with the help of NLP gets us that much closer to generating something useful in the same way that dimensionality reduction techniques work like PCA, SVD, etc. that help us on the numerical side of the data science world.

Now, a question comes to mind:

Is Topic Modelling similar to Principal Component Analysis (PCA), a Dimensionality Reduction Technique?

Topic modelling is the practice of using a quantitative algorithm to tease out the key topics that a body of the text is about. It shares a lot of similarities with dimensionality reduction techniques such as PCA, which identifies the key quantitative trends (that explain the most variance) within your features.

So, the outputs of PCA are a way of summarizing our features. **For Example,** it allows us to go from something like 1000 features to 10 summary features. These 10 summary features are basically considered as the topics.

In NLP, this type of role is played by the Topic modelling algorithms that work almost exactly the same way. We want to distill our total corpus of documents having 1,00,000 features (distinct words) into 7 topics (decided arbitrarily). And once we have the topics along with what they consist of in our hand, we can transform each document that is present in our corpus from a noisy bag of words to a clean portfolio of topic loadings.

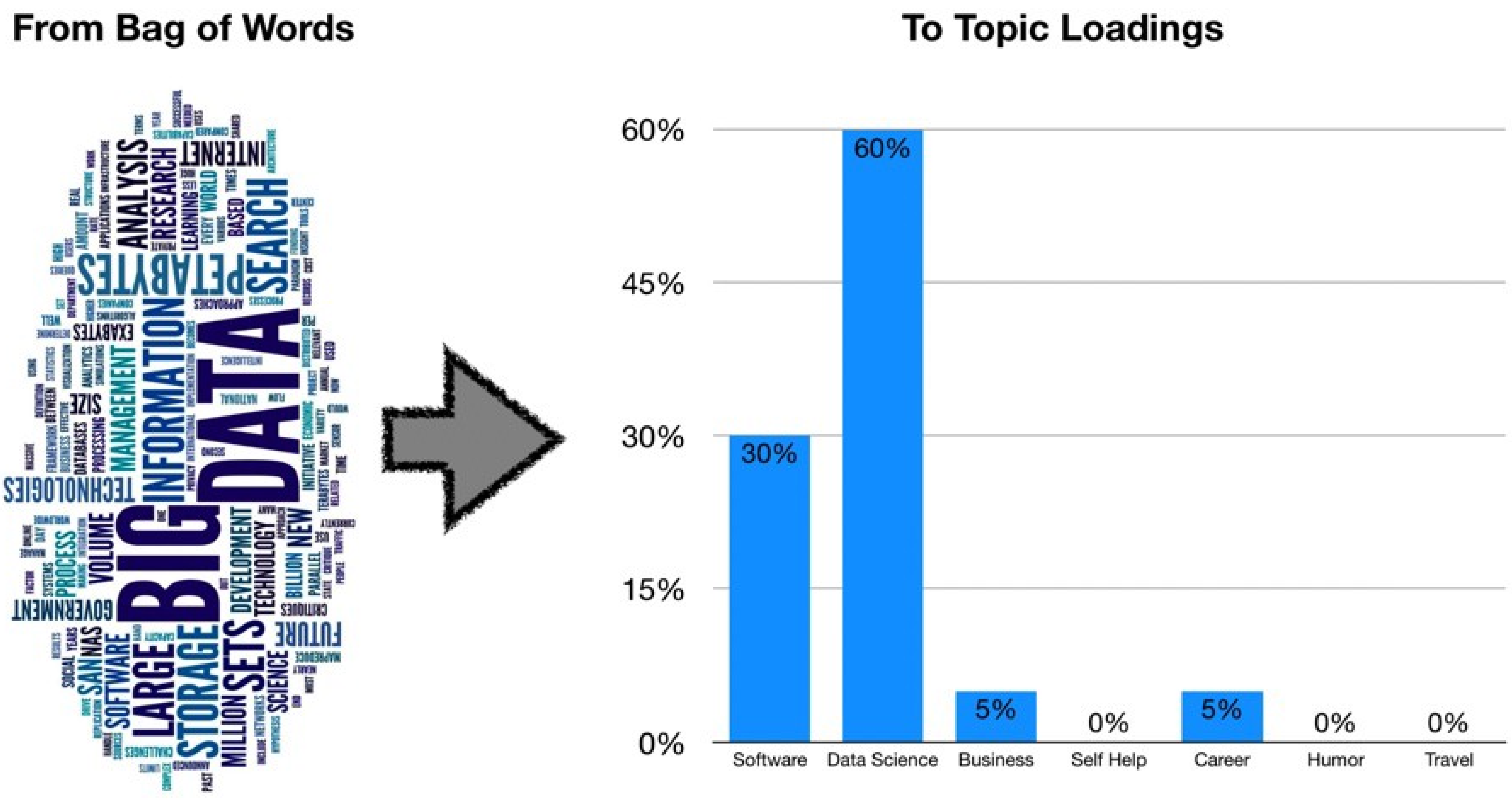

Image Source: Google Images

Homework Problem

Do you think we can use Topic Modelling algorithms for doing Feature Selection while doing Feature Engineering for our problem statement? Why or why not?

How Does Topic Modelling Works?

The working of topic modelling is not at all difficult. Topic modelling involves counting words and grouping similar word patterns to infer topics within unstructured data.

For Example, Imagine you are a manager of a software company and you want to know what customers are saying about particular features of your product. Instead of spending our valuable time going through heaps of feedback, in an attempt to find which texts are talking about your topics of interest, you could analyze them with the help of topic modelling algorithms.

Therefore, by detecting patterns such as word frequency and distance between words, a topic model clusters feedback that is similar, and words and expressions that appear most often. With this information, you can quickly deduce what each set of texts are talking about. Remember, this technique is 'unsupervised' in nature which means that no training is required.

Popular Techniques of Topic Modelling

Some of the popular techniques to implement topic modelling are as follows:

- LDA – Latent Dirichlet Allocation
- Non-negative Matrix Factorization
- LSA – Latent Semantic Allocation
- PLSA – Probabilistic Latent Semantic Analysis
- lda2vec – deep learning model
- tBERT – Topic BERT

Now, the question that comes to mind is:

What kind of Outputs we can get after applying Topic Modelling Algorithms?

For Example, consider the following corpus i.e, set of documents

```
Document-1: I had a peanut butter sandwich for breakfast
Document-2: I like to eat almonds, peanuts and walnuts
Document-3: My neighbour got a little dog yesterday
Document-4: Cats and dogs are mortal enemies
Document-5: You mustn’t feed peanuts to your dog
```

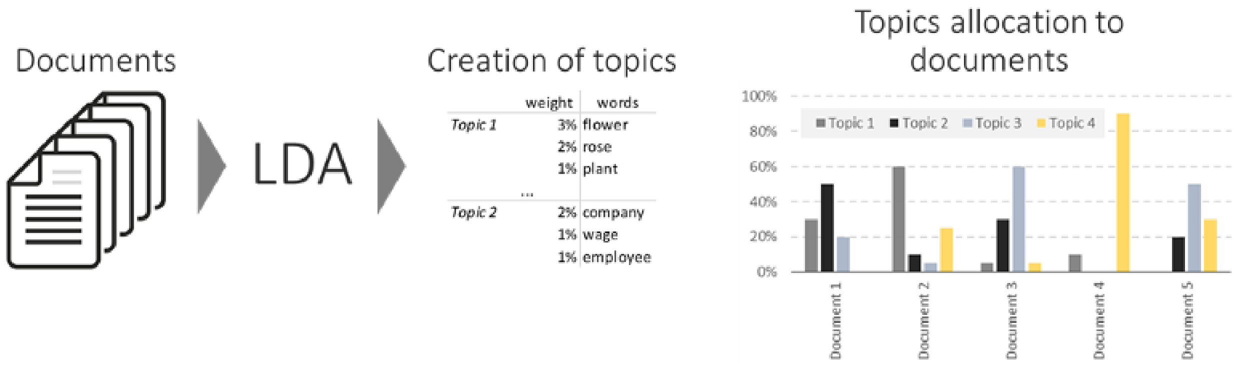

Image Source: Google Images

Creation of Topics

The algorithm discovers the different topics that the documents represent and how much of each topic is present in a document.

Topic 1: 30% peanuts, 15% almonds, 10% breakfast……(you can interpret that this topic deals with food).

Topic 2: 20% dogs, 10% cats, 5% peanuts…………(you can interpret that this topic deals with pet or animals)

Topics allocation to Documents

Apart from this, topic modelling algorithms would produce another output that will annotate the document with the topics present in the data. For example,

Document 1 and 2: 100% topic 1

Document 3: 100% topic 2

Document 4 and 5: 70% topic 1 and 30% topic 2

So, this is the kind of output that we will get after applying any technique.

This ends our Part-14 of the Blog Series on Natural Language Processing!

Other Blog Posts by Me

You can also check my previous blog posts.

[Previous Data Science Blog posts.]

LinkedIn

Here is [my Linkedin profile] in case you want to connect with me. I’ll be happy to be connected with you.

Email

For any queries, you can mail me on [Gmail].

End Notes