Basics of Topic Modelling

* Topic modelling task is to infer the 'topics being talked about in a given set of Documents
* Topic modelling is a method for unsupervised classification of documents, similar to clustering on numeric data, which finds some natural groups of items (topics) even when we’re not sure what we’re looking for (unsupervised learning approach)
* TOPIC 1 (NLP) : statistics,model,learning, language,text, processing.
* For example, a document consisting of the terms 'MIT, Stanford, Cambridge' is probably about education, the best universities, research, or all of these 'topics'. These topics may not be explicitly present in the document, but they might still be there.
* Topic modelling is the art and science of identifying 'latent topics' in text. Let’s understand the basic idea of topic modelling.
* what is a Topic: Topic is it is the main idea or theme which is being discussed in the text or speech or whatever...
* Topic discovery: It becomes identifying in the text.
* why do we care about discovering topics? let us say a new product is launched, there are many reviews about the product.. topic discovery helps here we can say negative or positive by looking at all the reviews. you can identify what is going well or what is not going well.

There are various ways in which you can extract topics from text:

1. PLSA - Probabilistic Latent Semantic Analysis.

2. LDA - Latent Dirichlet Allocation.

**Techniques used for topic modelling such as Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA**).

Basic idea

* Topics and their discovery.
* Main idea being discussed in the text.
* Identifying the topics discussed in the text.
* This is Un supervised learning approach.
* The input to a topic model is the corpus of documents, for e.g. a set of customer reviews, research papers etc.
* There are two outputs of a topic model - 1. The distribution of topics in a document and 2. The distribution of words in a topic.

**Two major tasks in topic modelling are….**

* **Defining topic**
* Each term is a topic
* **Estimating coverage**
* simply count term frequencies
* What about synonyms…

**Example < disambiguation**

**The stars are out tonight at the RRR film audio ceremony...**

There is possibility for two topics

* Astronomy ) stars
* Movies( film

Words can also have different meanings

**Redefining a Topic: Distribution over a vocab… or terms**

* Consider topic - 'science'

Science (probabilities) or (weights)

High proportion)/experiment/lab/research/ low proportion)/art/humanities

**LDA** (latent dirichlet allocation)

It is one of the most popular topic modelling methods. Each document is made up of various words, and each topic also has various words belonging to it. The aim of LDA is to find topics a document belongs to, based on the words in it.

* In LDA, we assume that the document-topic and topic-term distributions are Dirichlet distributions (parameterized by some variables), and we want to infer these two distributions.
* Generative process: each document is a distribution of topics and each topic is a distribution of terms.

**Dirichlet Distribution:** Both document-topic and topic-term distributions come from a Dirichlet distribution. (Distribution of a Distribution is a Dirichlet Distribution).

LDA, Implementation.

* Choose the no of topics = k (ex : politics, sports, religion).
* Randomly assign each word in each document to one of the k topics.
* Once this assignment is over in the corpus.
* Now it will go through every word and its topic assignment in each document.
* it basically iterate through every word and its topic assignment in each document by looking at.
* how often the topic occurs in the doc.
* how often the word occurs in the topic.

(Multiple iterations of these steps)

* Based on this it will assign the word to a new topic.
* After some time algorithm will converge and topics will make sense…

(Library: genism package from genism import corpora)

(Spacy or NLTK: for pre-processing of text)

The input to a topic model is the corpus of documents.

for e.g. a set of customer reviews, research papers, books, articles etc.

There are two outputs of a topic model -

1. The distribution of topics in a document and

2. The distribution of words in a topic.

The matrix is the probability distribution of words in a topic, or matrix is the distribution of topics in a document.

**Topic is a distribution over terms**, i.e. each term has a certain 'weight' in each topic (which can be zero as well).

**Table of Content**

Latent Dirichlet Allocation for Topic Modelling

1. Preparing documents.
2. Cleaning and Pre-processing.
3. Given a corpus ( set of documents).
4. Preparing document term matrix (formatting data properly is imp that genism LDA understands).Once cleaning and pre-processing are finished, Feed your model with the data.
5. First, it needs corpus as a dictionary of id word mapping (identifiers for the words), also needs corpus as a term frequency matrix. so for each document and each word, this is the frequency you need to feed.
6. The (3, 7) represents the fact that the word with id=3 appears 7 times (frequency)in the document. (MATRIX WILL BE CREATED)
7. Build your model. (parameters are important inside your function – number of topics – we define upfront.
8. Choose several topics, start with 10 and improvise.
9. Running LDA model.
10. Print 10 topics, and look at the top 10 words and their corresponding probabilities.
11. Measure coherence score – if in a topic the words are very related and occurring together then that topic is coherent.
12. The words are similar, they are together and expected to be together.
13. The score is between 0 and 1.
14. 0 or below 0.5 is bad.
15. 1 0r nearby 1 is good.

**End notes**

* human involvement or judgement is essential.
* the big consideration is that LDA is more generalized than plsa, plsa has more parameters, it overfits when your data is small, it won't perform very well, regarding large data sets we can go with plsa or LDA. but mostly used one is LDA.
* probabilistic models are for interpretability.
* two methods are .....plsa is one specific instance or one specific formulation and LDA has fewer parameters, works better for small data sets. both are fine with large data sets.
* LDA is the most used one these days. this technique is highly recommended.

**APPLICATIONS**

1. **This is a very intresting approach to understanding the content or text in a document.**
2. **what do people like or dislike about the product.**
3. **we can optimize recommendation system with topic modelling. we can improve the accuracy of recommendation. (topic model based recommendation systems.**
4. **Using topic models in content based news recommender systems.**
5. **Automatic labelling of documents ( to multiple categories)**

**say there is a news article,if you would want to automatically categorize news article. wheather it is science or sports or business or travel or food.**

**you would want to automatically categorize based on the content in the document have a tag. this one is about travel or food etc...**

**Disadv**

PLSA is that it has a large number of parameters which grow linearly with the documents. Although estimating these parameters is not impossible, it is computationally very expensive.

**Adv**

LDA (Latent Dirichlet Distribution) is an alternative topic model that solves this problem. Unlike PLSA, LDA is a parametric model, i.e. you do not have to learn all the individual probabilities. Rather, you assume that the probabilities come from an underlying probability distribution (the 'Dirichlet' distribution) which you can model using a handful of parameters.

For example, the normal distribution is parameterized by only two parameters - the mean and the standard deviation. Modelling data using this distribution (such as the age of N people) means to estimate these two parameters.

Similarly, in LDA, we assume that the document-topic and topic-term distributions are Dirichlet distributions (parameterized by some variables), and we want to infer these two distributions.

**Algorithm of LDA, which is a generalised form of PLSA. in PLSA, It is assumed a generative process: each document is a distribution of topics and each topic is a distribution of terms.**

The generative process is assumed to be as follows:

For each term in a document, you first pick a topic from the document-topic distribution, then from the chosen topic, you pick a term from the topic-term distribution. You do this for all documents to create the corpus. Both document-topic and topic-term distributions come from a Dirichlet distribution.

Note that **alpha** is a parameter of the Dirichlet distribution which determines the document-topic distribution, while **eta** is the parameter which determines the topic-term distribution.

Parameter Estimation using **Gibbs Sampling.**

Gibbs sampling is a commonly used technique in LDA