

Mathematics and Big Data

Natural Language Processing - Part 3

Sundus Zafar

Department of Mathematics
Autonomous University of Barcelona

1 NLP: Topic Modelling

- Introduction
- Topic Modelling : Applications
- Latent Dirichlet Allocation (LDA)
- Latent Semantic Analysis : LSA
- References

NLP: TOPIC MODELLING

NLP: Topic Modelling

What is topic Modelling?

- An unsupervised technique to discover topics across various documents.
- It is a type of statistical modeling for discovering the abstract topics that occur in a collection of documents.

Here we will learn two algorithms widely used in topic modeling LDA and LSA.

Topic Modelling : Applications

Where do we use topic modelling?

The screenshot shows a Google search for "big data". The search bar is at the top, with the Google logo on the left and camera, voice, and search icons on the right. Below the search bar, there are tabs for "All", "Images", "News", "Videos", "Maps", and "More". A red box highlights a row of topic-related icons: "infographic", "cloud", "analytics", "machine learning", and "artificial intelligence". Two red arrows point from the word "Topics" (written in red) to the "machine learning" and "artificial intelligence" icons. Below the icons, there are three image thumbnails. The first thumbnail shows a network diagram with the caption "Four interesting ideas that harness big ...". The second thumbnail shows a hand holding a smartphone with various icons around it, with the caption "Big Data: ¿Qué es, cómo funciona y por ...". The third thumbnail shows a circular data visualization with the caption "El big data brinda nuevas oportunidades ...".

Google

big data

Topics

All Images News Videos Maps More

infographic cloud analytics machine learning artificial intelligence

Four interesting ideas that harness big ...

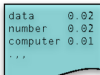
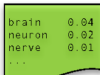
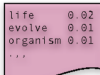
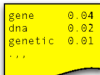
Big Data: ¿Qué es, cómo funciona y por ...

El big data brinda nuevas oportunidades ...

Topic Modelling : Applications

Which topic modelling examples you find every day?

Topics



Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,⁹ two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that **yeast's genome** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a single parasite and estimated that for this organism, 822 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, these **predictions**

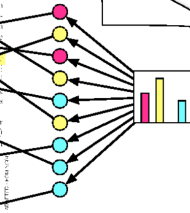
* Genome Mapping and Sequencing. Cold Spring Harbor, New York. May 8 to 12.

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Sir Alexander Hayes, a University of Sussex biologist, arrived at the 800 number. But coming up with a consensus answer may be more than just a **mathing** numbers. "Since particularly so many and more **genes** are completely unsequenced," he says, "it may be a way of organizing any newly **sequenced genomes**," explains Anady Muchugin, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

Topic proportions and assignments



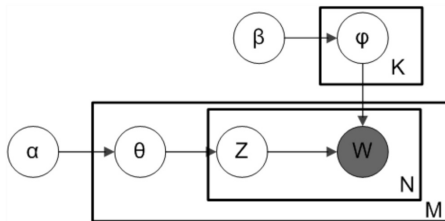
Latent Dirichlet Allocation (LDA)

LDA is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

LDA: Algorithm

- Assume k topics.
- Distribute k topics across document d .
- For each word w in document d assign correct topic k .
- Assign probability to word w for topic k .
- Repeat above steps for a number of times.

Latent Dirichlet Allocation (LDA)



Smoothed LDA from https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

Above is what is known as a plate diagram of an LDA model where:

α is the per-document topic distributions,

β is the per-topic word distribution,

θ is the topic distribution for document m ,

φ is the word distribution for topic k ,

z is the topic for the n -th word in document m , and

w is the specific word

Latent Semantic Analysis (LSA)

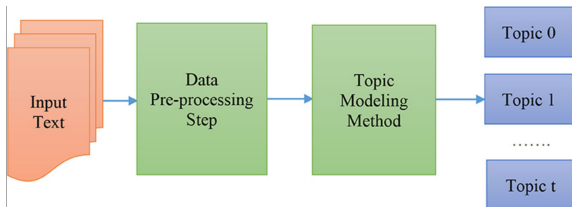
Latent semantic Analysis is used to leverage the context around the words to capture the hidden concepts. We use DTM and SVD to find vectors for every document and term in our corpus.

Algorithm

Let m be the number of text documents, n be the number of unique terms in the document d .

- To extract k topics from the text data, introduce number k .
- Generate an $m \times n$ document term matrix with TF-IDF scores.
- Reduce the dimension of DTM A to k dimensions using singular-value decomposition(SVD).
- Decompose the matrix A into matrix U, S and V^T .

Latent Semantic Analysis (LSA) : DTM



Latent Semantic Analysis (LSA) : DTM

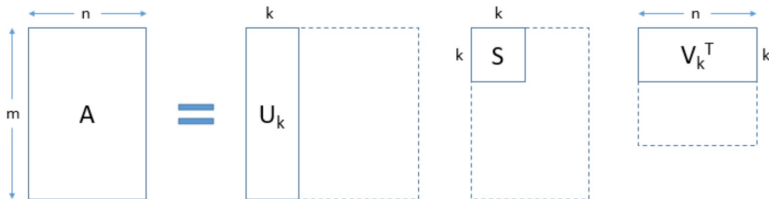
Document term Matrix

		Terms				
Documents		T1	T2	T3	...	Tn
	D1	0.2	0.1	0.5	...	0.1
	D2	0.1	0.3	0.4	...	0.3
	D3	0.3	0.1	0.1	...	0.5

	Dm	0.2	0.1	0.2	...	0.1

LSA : Singular Value Decompsition (SVD)

$$A = USV^T$$



Topic Modeling: LDA with Python

```
>>> from sklearn.decomposition import LatentDirichletAllocation

>>> LDA = LatentDirichletAllocation(n_components=7, random_state=42)

>>> LDA.fit(dtm) # dtm is Document Term Matrix

>>> LDA.components_

>>> single_topic = LDA.components_[0]
```

The detailed code for LDA and LSA provided in Practice 4 on CV

References

- ① Sentiment analysis and Opinion mining, Bing Liu.
- ② <https://aws.amazon.com/blogs/machine-learning/detect-sentiment-from-customer-reviews-using-amazon-comprehend/>
- ③ Applications of Topic Models, Jordan Boyd-Graber, Yuening Hu, David Mimno
- ④ Applied Natural Language processing with Python, Taweh Beysolow II.
- ⑤ Text Analytics with Python: A practical Real World Approach, Dipanjan Sarkar.