Mathematics and Big Data

Natural Language Processing - Part 3

Sundus Zafar

Department of Mathematics
Autonomous University of Barcelona

NLP: TOPIC MODELLING

NLP: Topic Modelling

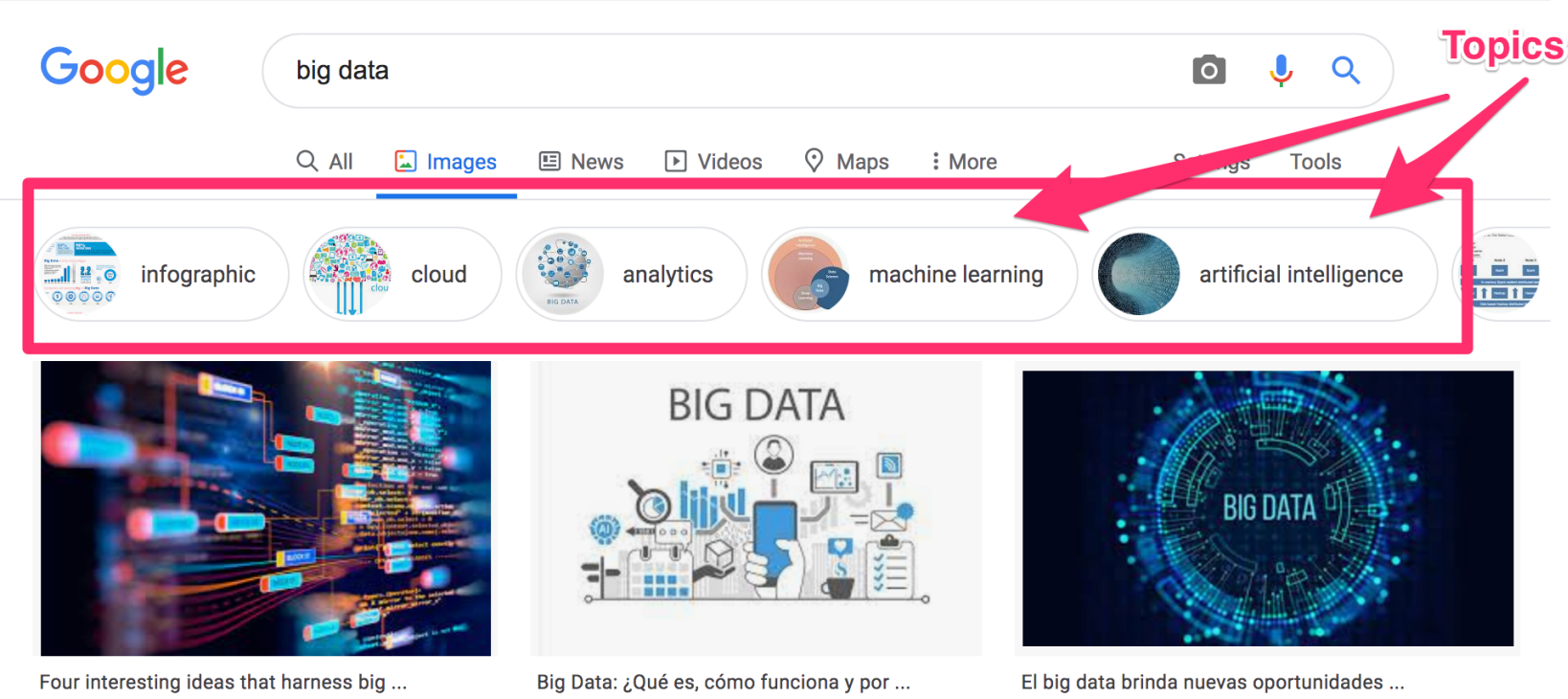
What is topic Modelling?

- An unsupervised technique to discover topics across various documents.
- It is a type of statistiacl 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.

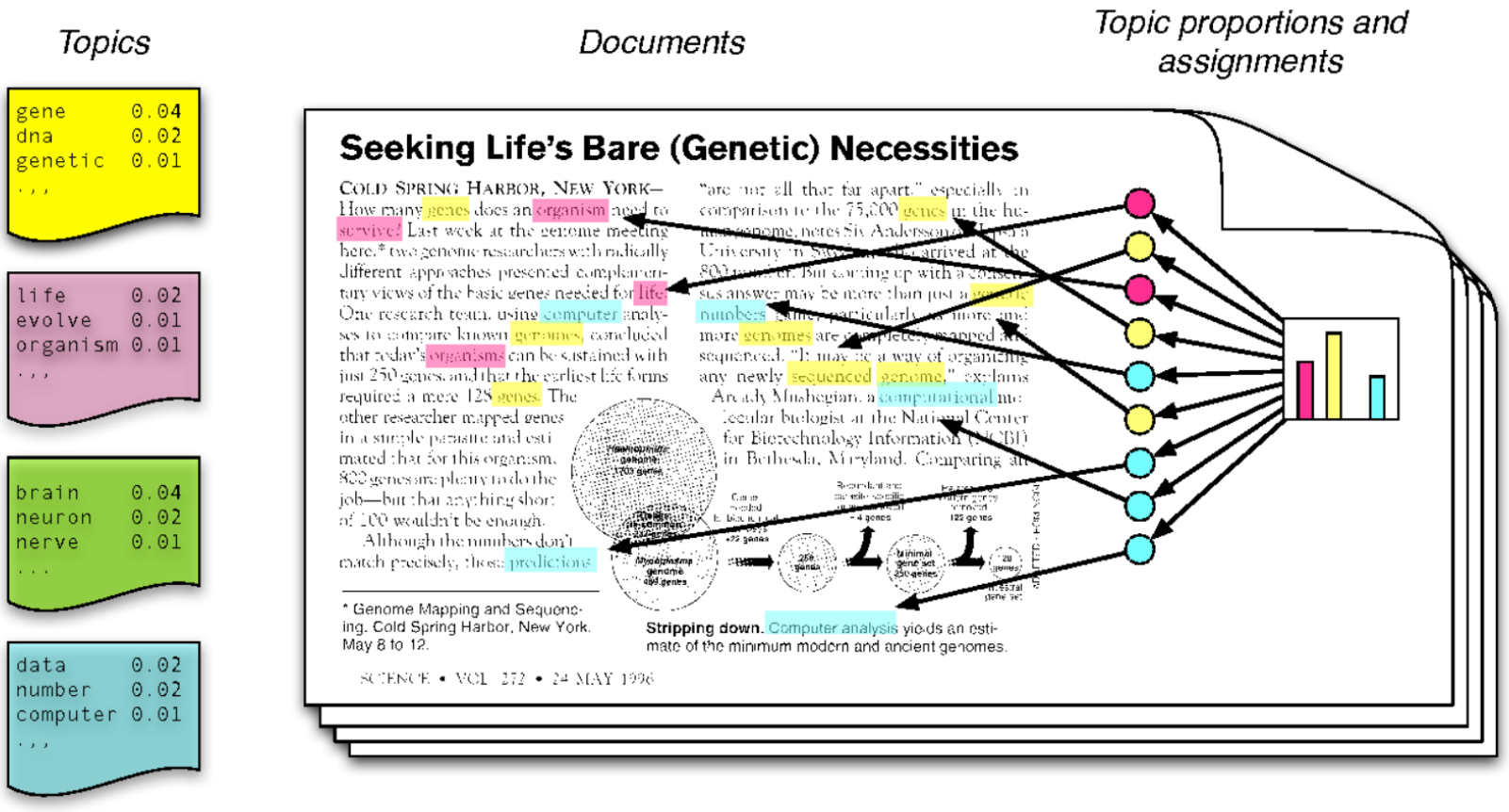
Topic Modelling : Applications

Where do we use topic modelling?



Topic Modelling : Applications

Which topic modelling examles you find every day?



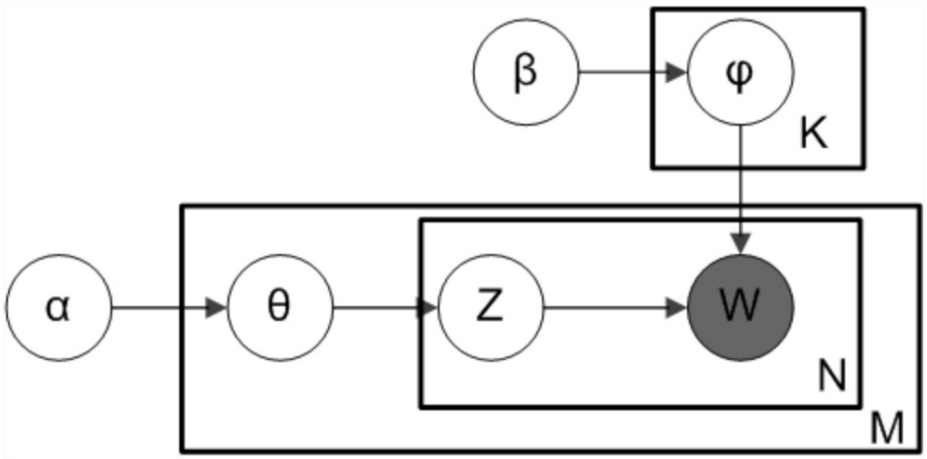
Latent Dirichlet Allocation (LDA)

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 probablity 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)

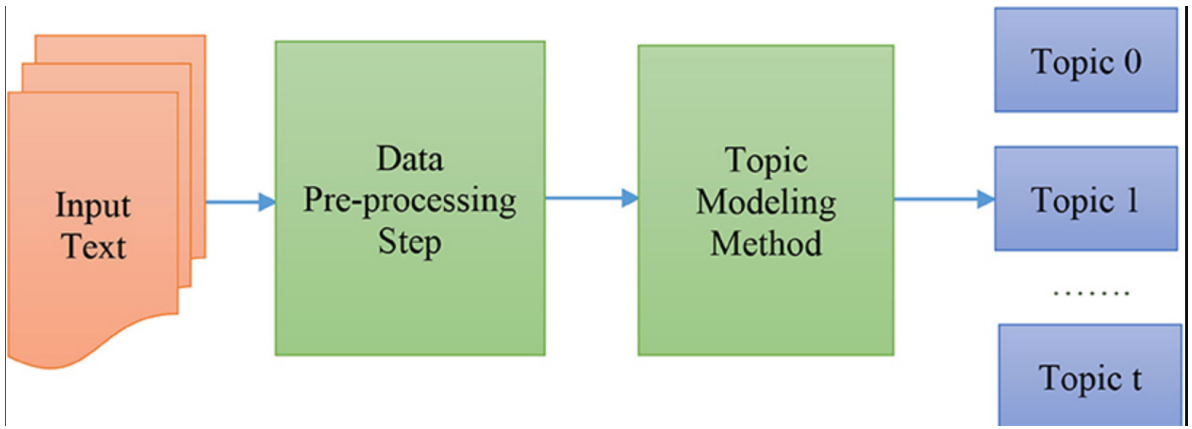
Latent semnatic Analysis is used to leverage the context aroudn the words to capture the hidden concepts. We use DTM and SVD to find vestors for every document and term in our corpus.

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

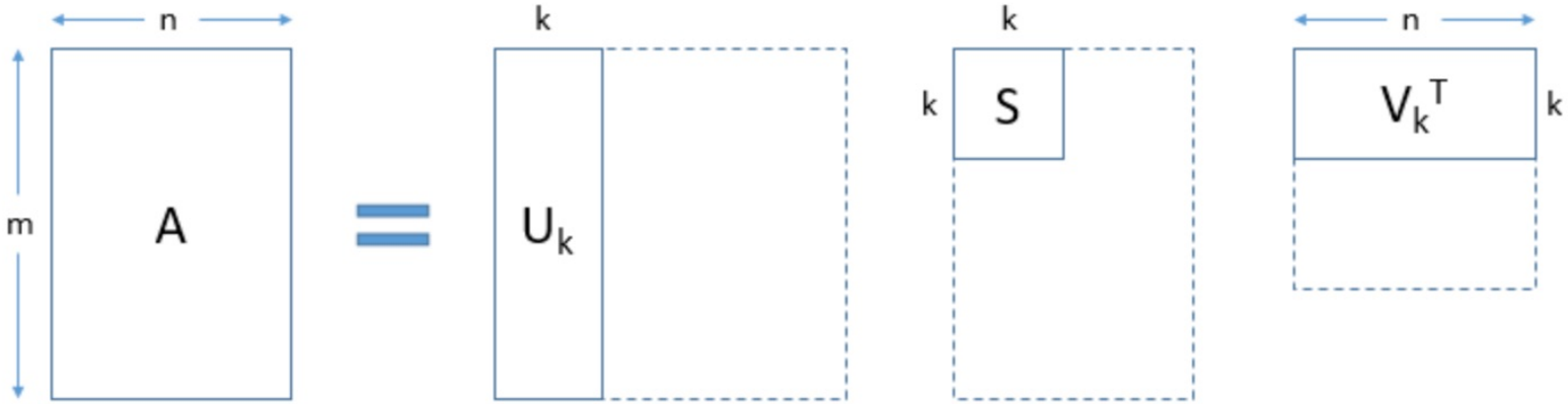
Terms

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 LatentDirichletAllocat

>>> LDA = LatentDirichletAllocation(n_components=7,random_st

>>> 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.