Topic Modelling and Named Entity Recognition

Avinash Navlani

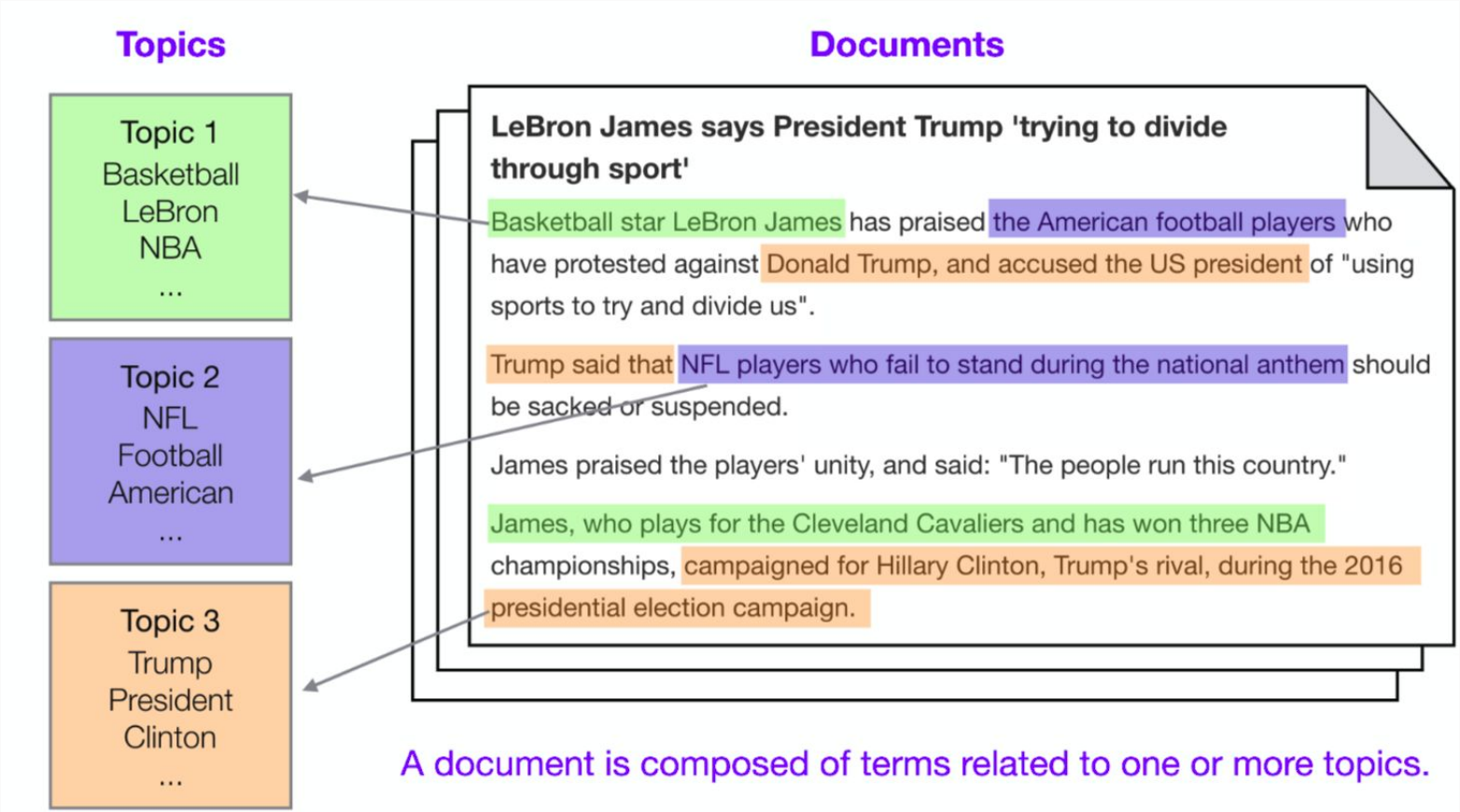
Session Highlights

1. What is Topic Modelling.
2. Text Classification
3. What Algorithms are available for topic modelling.
4. Use Cases of Topic Modelling.
5. What is LSA?
6. How LSA works?
7. Implement LSA or Use LSA for finding topics of document.
8. Advantages and Disadvantages of LSA
9. What is LDA?
10. How LDA Works?
11. What is NMF? How NMF works?
12. What is the best way to determine k (number of topics) in topic modeling?
13. Named Entity Recognition

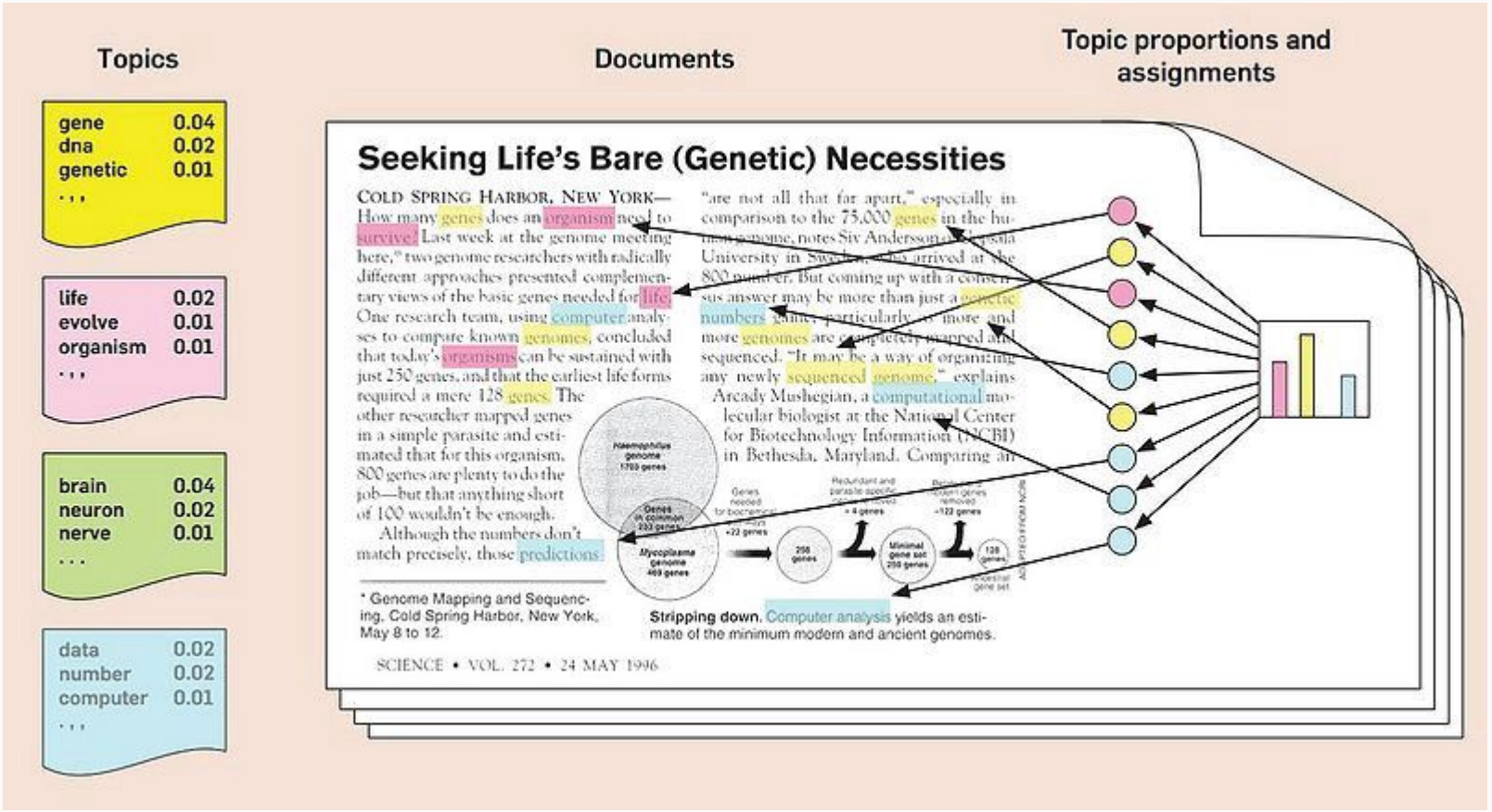
Topic Modelling

- Topic Modelling aims to automatically discover the hidden thematic structure in a large corpus of text documents.
- Topic modelling is an unsupervised text mining approach.
- A single document can potentially be associated with multiple topics.

Topic Modelling (Conti...)

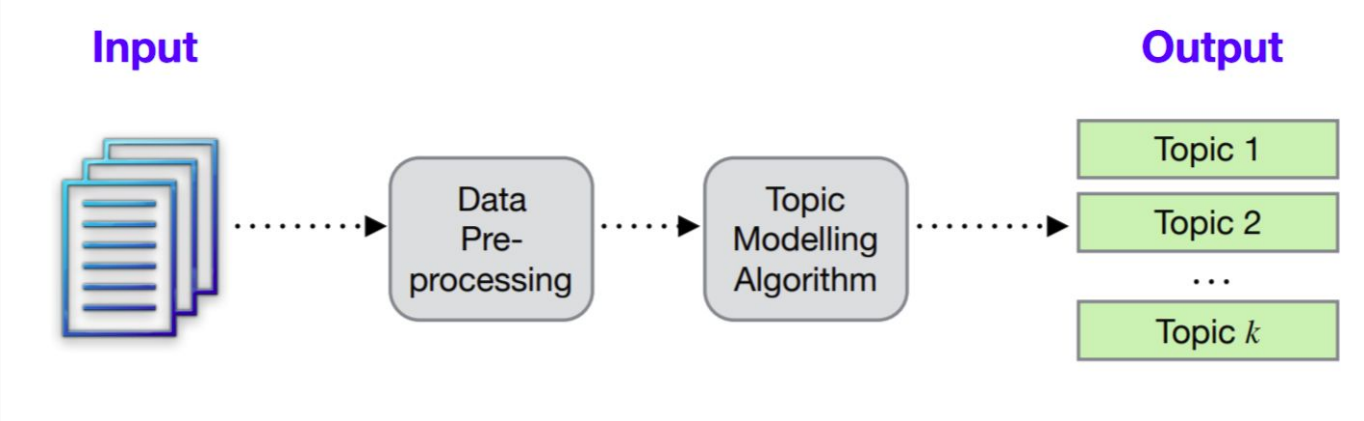


Topic Modelling (Conti...)



Topic Modelling-Definition

Topics can be defined as “a repeating pattern of co-occurring terms in a corpus”. A good topic model should result in – “health”, “doctor”, “patient”, “hospital” for a topic – Healthcare, and “farm”, “crops”, “wheat” for a topic – “Farming”.



Text Classification

Which medical speciality does this relate to?

TINEA PEDIS, or ATHLETE'S FOOT, is a very common fungal skin infection of the foot. It often first appears between the toes. It can be a one-time occurrence or it can be chronic. The fungus, known as Trichophyton, thrives under warm, damp conditions so people whose feet sweat a great deal are more susceptible. It is easily transmitted in showers and pool walkways. Those people with immunosuppressive conditions, such as diabetes mellitus, are also more susceptible to athlete's foot.

Nephrology

Neurology

Podiatry

Topic Modelling Use Cases

- **Automatically Summarizing Resumes:**
  - we could facilitate evaluation of resumes at a quick glance, thereby simplifying the effort required in shortlisting candidates among a pile of resumes.
- **Optimizing Search Engine Algorithms:**
  - A more efficient approach would be to run an NER model on the articles once and store the entities associated with them permanently.
- **Powering Recommender Systems:**
  - recommender systems which automatically filter relevant content we might be interested in and accordingly guide us to discover related and unvisited relevant contents based on our previous behaviour.
- **Simplifying Customer Support:**
  - recognizing relevant entities in customer complaints and feedback such as Product specifications, department or company branch details, so that the feedback is classified accordingly and forwarded to the appropriate department responsible for the identified product.

Topic Modelling Use Cases (Conti...)

- Healthcare:
  - A quickly growing use of natural language processing (NLP) exists in the healthcare industry. Recent advancements in technology have made it possible to extract useful and very valuable information from unstructured medical records. This can be used to correlate patient information and look for treatment patterns.

Available Algorithms for Topic Modelling

- **Latent Semantic Analysis (LSA)**
- **Latent Dirichlet Allocation (LDA)**
- **Non-negative Matrix Factorisation (NMF)**

- LSI (also known as Latent Semantic Analysis, LSA) learns latent topics by performing a matrix decomposition (SVD) on the term-document matrix.
- LDA is a generative probabilistic model, that assumes a Dirichlet prior over the latent topics.
- LDA is based on probabilistic graphical modeling while NMF relies on linear algebra.

Text Preprocessing

- Bag-of-words model(BoW) is the simplest way of extracting features from the text. BoW converts text into the matrix of occurrence of words within a document. This model concerns about whether given words occurred or not in the document
- In Term Frequency(TF), you just count the number of words occurred in each document. The main issue with this Term Frequency is that it will give more weight to longer documents. Term frequency is basically the output of the BoW model.
- IDF(Inverse Document Frequency) measures the amount of information a given word provides across the document. IDF is the logarithmically scaled inverse ratio of the number of documents that contain the word and the total number of documents.

LSA(Latent Semantic Analysis)

- LSI or LSA is used typically as a dimension reduction or noise reducing technique.
- LSI (also known as Latent Semantic Analysis, LSA) learns latent topics by performing a matrix decomposition (SVD) on the term-document matrix.
- Some of the well known and simple applications in which this technique is used are document clustering in text analysis, recommender systems, building user profiles etc.
- In order to understand the application of this technique, you need to first understand the concept and intuition of dimensionality reduction, eigen vectors, eigen space, Principal Component Analysis, SVD etc.

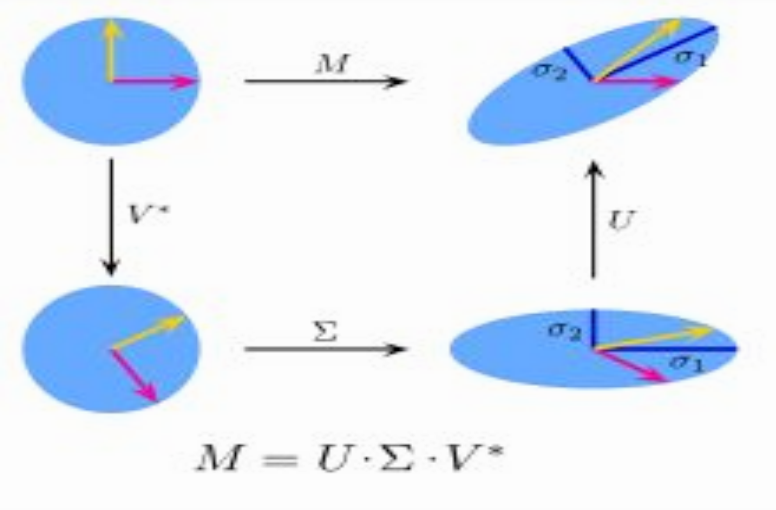
Singular Value Decomposition

- **Matrix Factorization:** It is a representation of a matrix into a product of matrices. There are many different matrix factorization and each used for different class of problems.

$$M=U\Sigma V^*$$

  - U is a left singular matrix
  - Σ is a m×n diagonal matrix with non-negative real numbers.
  - V is a n × n right singular matrix
  - V*transpose of the n ×n.

Singular Value Decomposition (Conti...)

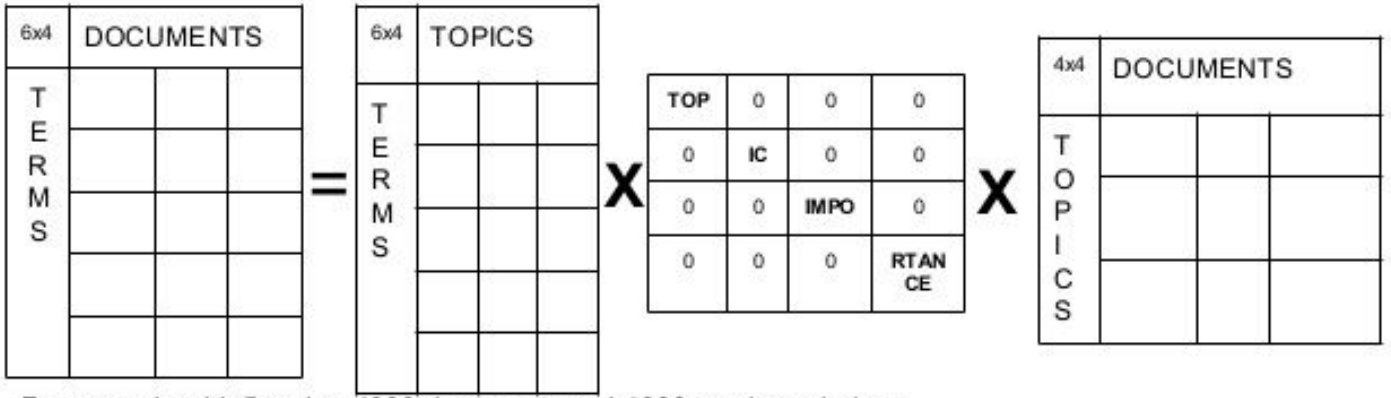


The expression **UΣV*** can be interpreted as a composition of three geometrical transformations: a rotation, a scaling, and another rotation.

LSA

Nothing more than a **singular value decomposition (SVD)** of document-term matrix:

Find three matrices U, Σ and V so that: $X = U\Sigma V^t$



For example with 5 topics, 1000 documents and 1000 word vocabulary:
Original matrix: 1000 x 1000 = 10^6
LSA representation: 5x1000 + 5 + 5x1000 ~ 10^4
-> 100 times less space!

Latent Semantic Analysis (LSA)

LSA is essentially low-rank *approximation* of document term-matrix

3	4	1	0
4	3	0	1
3	4	4	3
0	1	4	3
2	0	3	3
0	1	3	4

=

Word assignment to topics

	IT	cars
linux	-0.33	-0.53
modem	-0.32	-0.54
the	-0.62	-0.10
clutch	-0.38	0.42
steering	-0.36	0.25
petrol	-0.37	0.42

X

Topic Importance

11.4	
	6.27

X

Topic distribution acr documents

	D1	D2	D3	D4
IT	-0.42	-0.48	-0.57	-0.51
cars	-0.56	-0.52	0.45	0.46

Advantages

1) Easy to implement, understand and use.
There are many practical and scalable implementations available. Some of them are mahout (java), gensim (python), scipy (svd python), sklearn(python).

2) Performance: LSA is capable of assuring decent results , much better than plain vector space model. It works well on dataset with diverse topics.

3) Synonymy: LSA can handle Synonymy problems to some extent

4) Runtime : Since it only involves decomposing your term document matrix, it is faster, compared to other dimensionality reduction models

Disadvantages

1) Representation is dense, so hard to index based on individual dimensions.

2) It is a linear model, so not the best solution to handle nonlinear dependencies

3) The latent topic dimension can not be chosen to arbitrary numbers. It depends on the rank of the matrix, so can't go beyond that.

4) The model is not humanly readable. Debug/evaluation is possible through finding similar words for each word in the latent space though. But otherwise not easy to interpret like, say LDA

Implementing LSA

In python, we can implement LSA via two libraries:

-Scikit learn

-Gensim

Latent Dirichlet Analysis

Latent Dirichlet Allocation is the most popular topic modeling technique.

LDA is a matrix factorization technique and probabilistic modelling technique. In vector space, any corpus (collection of documents) can be represented as a document-term matrix.

The following matrix shows a corpus of N documents D1, D2, D3 … Dn and vocabulary size of M words W1,W2 .. Wn. The value of i,j cell gives the frequency count of word Wj in Document Di.

	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

Latent Dirichlet Analysis(Conti...)

LDA converts this Document-Term Matrix into two lower dimensional matrices – M1 and M2.

M1 is a document-topics matrix and M2 is a topic – terms matrix with dimensions (N, K) and (K, M) respectively, where N is the number of documents, K is the number of topics and M is the vocabulary size.

LDA makes use of sampling techniques in order to improve these matrices.

	K1	K2	K3	K
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
Dn	1	0	1	0

	W1	W2	W3	Wm
K1	0	1	1	1
K2	1	1	1	0
K3	1	0	0	1
K	1	1	0	0

Latent Dirichlet Analysis(Conti...)

- It Iterates through each word “w” for each document “d” and tries to adjust the current topic – word assignment with a new assignment. A new topic “k” is assigned to word “w” with a probability P which is a product of two probabilities p1 and p2.
- For every topic, two probabilities p1 and p2 are calculated. P1 – p(topic t / document d) = the proportion of words in document d that are currently assigned to topic t. P2 – p(word w / topic t) = the proportion of assignments to topic t over all documents that come from this word w.
- The current topic – word assignment is updated with a new topic with the probability, product of p1 and p2 . In this step, the model assumes that all the existing word – topic assignments except the current word are correct. This is essentially the probability that topic t generated word w, so it makes sense to adjust the current word’s topic with new probability.
- After a number of iterations, a steady state is achieved where the document topic and topic term distributions are fairly good. This is the convergence point of LDA.

Non-negative Matrix Factorization(NMF)

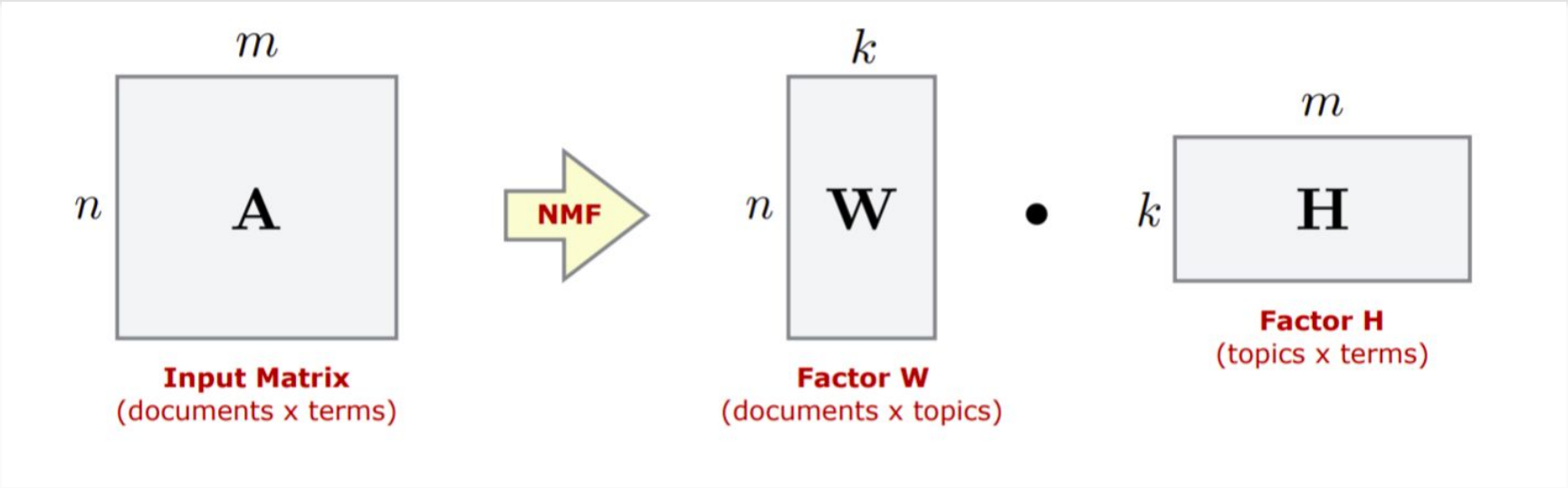
- Non-negative Matrix Factorization is a Linear-algebraic model, that factors high-dimensional vectors into a low-dimensionality representation.
- NMF takes advantage of the fact that the vectors are non-negative. By factoring them into the lower-dimensional form, NMF forces the coefficients to also be non-negative.
- Find two non-negative matrices (W, H) whose product approximates the non- negative matrix X. This factorization can be used for example for dimensionality reduction, source separation or topic extraction.

Non-negative Matrix Factorization(Conti...)

Given the original matrix **A**, we can obtain two matrices **W** and **H**, such that A= WH. NMF has an inherent clustering property, such that W and H represent the following information about A:

- A (Document-word matrix)—input that contains which words appear in which documents.
- W (Basis vectors)—the topics (clusters) discovered from the documents.
- H (Coefficient matrix)—the membership weights for the topics in each document.

Non-negative Matrix Factorization(Conti...)



LDA Vs NMF

- LDA is good in identifying coherent topics where as NMF usually gives incoherent topics.
- However, in the average case NMF and LDA are similar but LDA is more consistent.
- NMF provide better understanding than the topics given by LDA.
- LDA being more semantically interpretable while NMF being faster than LDA.

What is the best way to determine k (number of topics) in topic modeling?

1. Topic coherence is a realistic measure of how good the topics produced by LDA really are.
2. One way to determine the number of topics is to consider each topic as a cluster and then to evaluate the effectiveness of such clusters using known metrics such as Silhouette (clustering) - Wikipedia.

Named Entity Recognition

- Named Entity Recognition is a process where an algorithm takes a string of text (sentence or paragraph) as input and identifies relevant nouns (people, places, and organizations) that are mentioned in that string.

- **Named-entity recognition** (**NER**) (also known as **entity identification**, **entity chunking** and **entity extraction**) is a sub-task of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Named Entity Recognition (Conti...)

News and publishing houses generate large amounts of online content on a daily basis and managing them correctly is very important to get the most use of each article. Named Entity Recognition can automatically scan entire articles and reveal which are the major people, organizations, and places discussed in them. Knowing the relevant tags for each article help in automatically categorizing the articles in defined hierarchies and enable smooth content discovery.

Implementation of NER

- NLTK
- Spacy
- Core NLP

Spacy is very powerful and industrial strength package for almost all natural language processing tasks. That’s why we are using Spacy in our project.

NER using Spacy

- WE can generated two types of entities from given text using spacy:
  - Pretend Entities: Already defined in spacy library and we can easily extract those entity by calling simple function.
  - Custom Entities: It is the one of the powerful approach of spacy. This help us to design new entities. It changes the schema and define **custom entities**.

Summary

- Topic Modelling gives us the thematic or contextual information about the given text/article/blog.
- NER offers us to find important entities(pretend and custom entities) in given text.
- Combination of both will help us summarize document and generate tags about the document.
- Further exploration of text requires sentiment analysis of document, which gives us polarity of text/article/document.


Q&A
SITES
LIST
You have
Questions
We have
Answers

Thank You

video
networking
social media
process
conversations
content
authority
share
seo
instant
like
building
production
collaboration
always
profile
research
forum
webinternet
marketing