22nd-12-2020

Chapter-5
Topic modelling

Objectives:

1) Describe topic modelling and its use cases
- Describe topic modelling algorithms
- Describe the working of LSA and LDA
- Describe topic fingerprinting
- Implement topic modelling using LSA and LDA.

Topic modelling:

- A simple way to capture meaning from a collection of documents
- Unsupervised learning algo.

Why topic modelling:

- **Topic discovery:**
Find a set of topics that can be used to classify a set of documents

- **Discovering themes :**
  To get a general idea of what themes or topics are present in document.

- **EDA :**
  Topic modelling can be used to check whether data is balanced or ~~not~~ skewed, which then can be used to select the ML algo.

- **Document clustering :**
  Topic modelling allows soft clustering

- Dimensionality reduction
- Historical analysis.

⇒ Topic modelling Algorithm :

Assumptions :

- Topics contain a set of words
- Documents contain a set of topics.

Algo's → LSA (latent Semantic Analysis)
LDA (Latent Dirichlet Allocation)

⇒ Better Defination
(from Analytics Vidhya video)

Topics → A repeated group of statistically significant tokens or words in a corpus.

Topic modelling - Process to find topics from documents in an unsupervised manner.

(Out of Book content)

⇒ Overview of SVD (Singular value decomposition)
(Since it will be used in LSA)

Used for :
- Data reduction
- Data driven generalization of Fourier transform

Let X be a matrix.

$$X = \begin{bmatrix} | & | & | & & | \\ x_1 & x_2 & x_3 & \cdots & x_m \\ | & | & | & & | \end{bmatrix}$$ where $x_k \in \mathbb{R}^n$

$\downarrow$ $(n, 1)$

Shape of $x \rightarrow (n, m)$

Using SVD we can represent this X as a product of 3 other matrices.

$$X = U \cdot \Sigma \cdot V^T$$

U: unitary, orthogonal; Σ: diagonal matrix

$$UU^T = U^TU = I_{n \times n}$$

$$VV^T = V^TV = I_{m \times m}$$

Σ diagonal $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$.

$$\begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ x_1 & x_2 & \cdots & x_m \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \cdots & u_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \\ & 0 & \end{bmatrix} \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_m \\ | & | & & | \end{bmatrix}^T$$

(n, m) = (n, n) × (n, m) × (m, m)

ordered by importance

$$X = U \Sigma V^T$$

eigen vectors; importance; mixture of U's that makes X's

U, Σ, V^T → of X, they are guaranteed to exist and are unique.

⇒ SVD matrix approximation

$$X = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ x_1 & x_2 & \cdots & x_m \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} = U \Sigma V^T$$

$$= \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \cdots & u_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_m \\ & 0 & & \end{bmatrix} \begin{bmatrix} - & v_1^T & - \\ - & v_2^T & - \\ - & v_3^T & - \\ - & v_m^T & - \end{bmatrix}$$

$$= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_m u_m v_m^T + 0$$

- Even V is a large matrix (n, n) only the first m cols are significant for us. $(n >> m)$

- After the first m cols, they are scaled by 0, (from diagonal matrix) $\sigma_i = 0 \ \forall i > m$.

- ∴ We need to use only the first m cols of U to get X.

$$X = \underbrace{\hat{U} \hat{\Sigma} V^T}$$

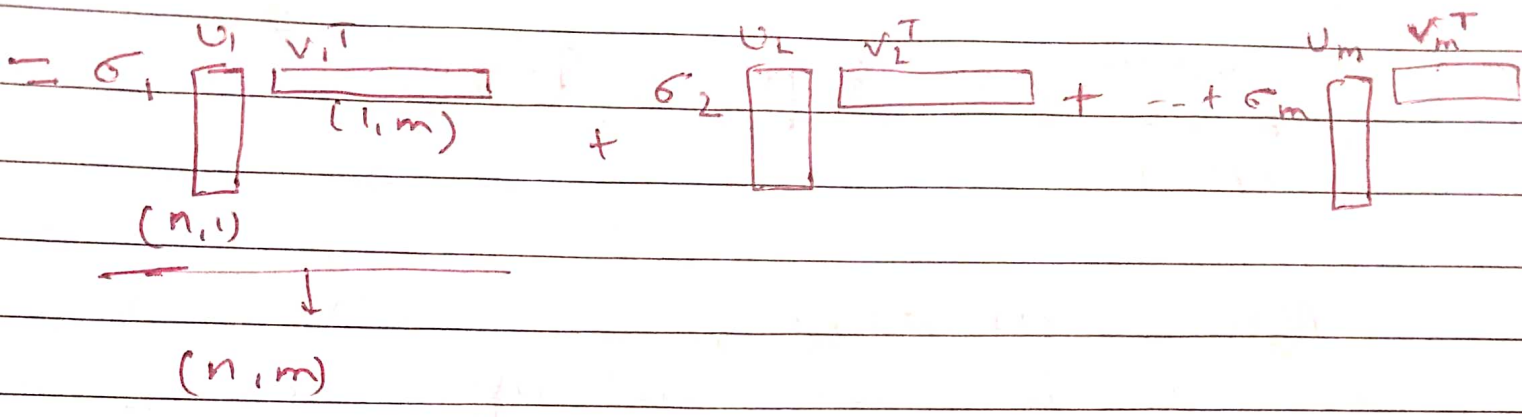
↓

Economy SVD

We calculate only first m cols of x.

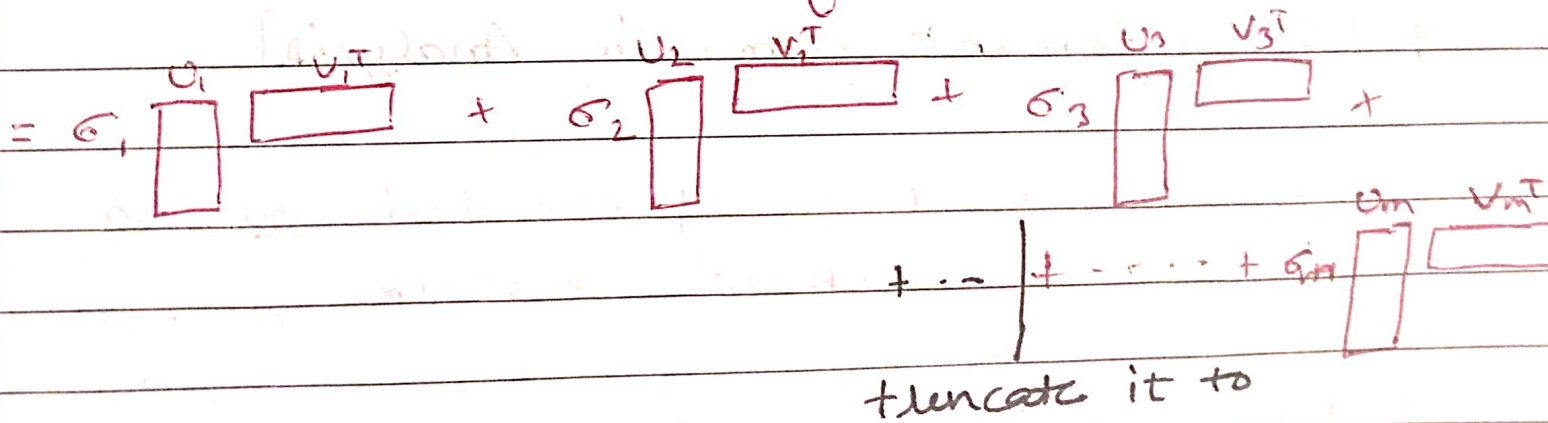
$= \sigma_1 U_1 V_1^T + \sigma_2 U_2 V_2^T + \cdots + \sigma_m U_m V_m^T$

Representing them as a sum of rank 1 matrices.



The above expression contains m terms.

But we will truncate it to r terms which will give us the best estimate of X.



Now,

$$X \approx \tilde{U} \tilde{\Sigma} \tilde{V}^T = \tilde{X}$$

How to select r,

According to Eckard - Young Theorem [1936],

$$\underset{\tilde{X}, \text{ s.t. } \text{rank}(\tilde{X}) = r}{\operatorname{argmin}} \|X - \tilde{X}\|_F = \tilde{U}\tilde{\Sigma}\tilde{V}$$

After truncating,
U is not unitary.

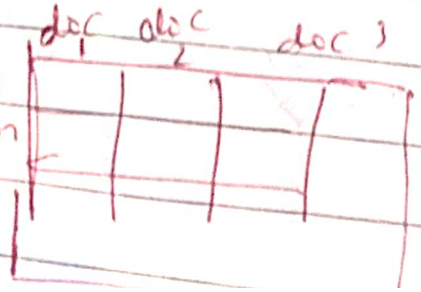
⇒ Interpretation of SVD in terms of correlation

[Continuing the book]

⇒ LSA [Lemant Semantic Analysis]

- Our corpus is represented as a term-to-document matrix.

Rows — terms (words)
columns — documents.



DATE
PAGE No.

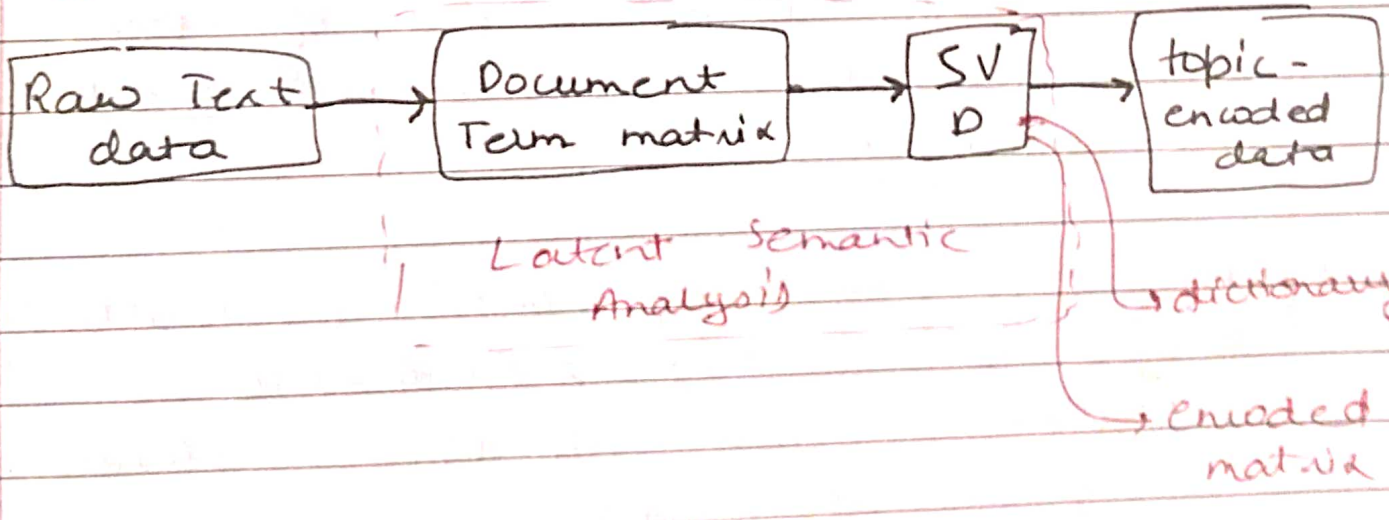
Using SVD, this matrix can be broken down into product of 3 matrices.

Term-to-document	=	Term to topics	x	topic importance	x	topic to documents
$m \times m$		$m \times N$		$N \times N$		$N \times m$

⇒ From Youtube

Latent features

- Features that are hidden in the data which cannot be directly measured.
- These features are essential to the data, but are not original features of the dataset.



18th-12-2020

- In LSA, we have to choose number of topics beforehand.

To select the optimal number of topics, we create LSA model for different topics in range and check their coherence value

⇒ Latent Dirichlet Allocation

- LDA is more often used for topic modelling
- LDA is a generative statistical model that allows a set of items to be sorted into unobserved groups by similarity.
- LDA is reasonably good particularly useful for finding reasonably accurate mixtures of topics within a given document.

⇒ From Youtube :
Luis Serrano

The problem :

- Let's suppose that we have some documents, each document can belong to a topic or combination of other topics.

LDA approach :

- We get two dirichlet distribution one associates documents to topics and other associates topics-to-words.

- With the help of these dirichlet dist we generate a multinomial dist.

- First we generate topics and then we generate words.

⇒ From ~~From~~ From ppt
(mphil in advance computer science)

⇒ Introduction to probalistic topic models :-

- We want to find themes (or topics) in document.
  - useful for search or browsing
- We don't want to do supervised topic classification:-
  rather not fix topics in advance. nor do manual annotation.
- Need an approach which automatically takes out the topics.
- This is essentially a clustering problem - can think of both words and documents are being clustered.

⇒ Key assumptions behind LDA

- Documents exhibit multiple topics
- LDA is a probabilistic ~~approach~~ model with a corresponding generative process.
  - each doc. is assumed to be generated by this (simple process)
- A topic is a distribution over a fixed vocabulary
  - these topics are assumed to be generated first, before the documents.
- Only the number of topics is specified in advance.

⇒ The Generative process

To generate a document:

1. Randomly choose a distribution over topics
2. For each word in the document:
   a. Randomly choose a topic from the dist over topics
   b. Randomly choose a word from the corresponding topic (dist over the vocabulary).

⇒ Topic Fingerprinting :

Document fingerprinting :

A set of numbers that summarizes a document's content and allows you to perform simple math functions.

- We can use topic modelling to figure out what topics are present in a document and its relevance.

- Suppose we use topic modelling with num of topics = 50, then we can represent and visualize each document of ours as a vector of len = 50.

We can use this vectors to find similar documents and use them for any other math operations.