LSA (Latent Semantic Analysis)

Latent Semantic Analysis is an efficient way of analyzing the text and finding the hidden topics by understanding the context of the text. The main concept and work of LSA are to group together all the words that have a similar meaning.

Latent Semantic Analysis also known as Latent Semantic Indexing(LSI) is a dimension reduction technique as well.

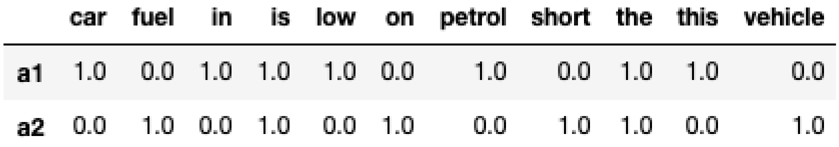
Why LSA ?

Most simple way of finding similar documents is by using vector representation of text and cosine similarity. Vector representation represents each document in the form of vector. This vector is known as document-term matrix.

For example:

a1 = "the petrol in this car is low"
a2 = "the vehicle is short on fuel"

The document term matrix is



	car	fuel	in	is	low	on	petrol	short	the	this	vehicle
a1	1.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0
a2	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0

Similarity between documents if found out using cosine similarity between the documents matrix. The similarity between documents a1 and a2 is 0.3086067 which is too low and should have been high since the documents are mostly similar in context. This is the disadvantage of document-term matrix or the vector representation technique. Another disadvantage is the vocabulary size as the language has the huge vocabulary causing the matrix to be bigger and computationally expensive. This disadvantages of the vector representation have led to the requirement of new technique for finding the similarity among the documents and finding the hidden topics.

LSA involves SVD, which is computationally inexpensive as compared to document-term matrix or the vector representation technique.

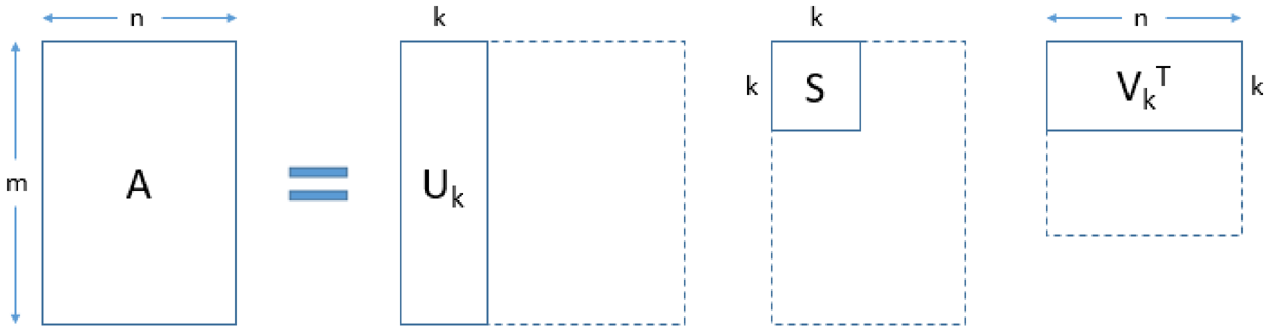
How it works

Latent Semantic Analysis is one of the foundational techniques in topic modeling. The core idea is to take a matrix of what we have documents and terms and decompose it into a separate document-topic matrix and a topic-term matrix.

LSA take in the document-term matrix with a tf-idf score or it take in a bag of word corpus as well and performs dimensionality reduction on A.

This dimensionality reduction can be performed using truncated SVD, SVD or singular value decomposition is a technique in linear algebra that factorizes any matrix A into the product of 3 separate matrices i.e. A = U*S*V where S is a diagonal matrix of the singular values of A. Critically truncated SVD reduces dimensionality by selecting only the t largest singular values and only keeping the first t columns of U and V. In topic modelling case, t is a hyperparameter we can select and refers to the number of topics we want to find.

$$A = USV^T$$



U (m × t) emerges as our document-topic matrix. Each row of
the matrix Uk (document-term matrix) is the vector representation of the corresponding document i.e. rows represent document vectors expressed in terms of topics.

V (n × t) becomes our term-topic matrix. Each row of the matrix Vk (term-topic matrix) is the vector representation of the corresponding terms i.e. rows represent term vectors expressed in terms of topics.

In both U and V, the columns correspond to one of our t topics. SVD gives us vectors for every document and term in our data. With these document vectors and term vectors, we can now easily apply measures such as cosine similarity to evaluate the similarity of different documents, the similarity of different words and the similarity of terms and documents.