Topic Modelling

A Topic Model can be defined as an unsupervised technique to discover topics across various text documents. Since topic modeling doesn't require training, it's a quick and easy way to start analyzing your data.

The following are several scenarios when topic modeling can prove useful.

Text classification - Topic modeling can improve classification by grouping similar words together in topics rather than using each word as a feature

Uncovering Themes in Texts - Useful for detecting trends in online publications for example

Recommender Systems - Using a similarity measure we can build recommender systems. If our system would recommend articles for readers, it will recommend articles with a topic structure similar to the articles the user has already read

Topic Modeling Algorithms

There are several algorithms for doing topic modeling. The most popular ones include

LDA – Latent Dirichlet Allocation – The one we'll be focusing in this tutorial. Its foundations are Probabilistic Graphical Models

LSA or LSI – Latent Semantic Analysis or Latent Semantic Indexing – Uses Singular Value Decomposition (SVD) on the Document-Term Matrix. Based on Linear Algebra

NMF – Non-Negative Matrix Factorization – Based on Linear Algebra

Here are some things all these algorithms have in common:

The number of topics as a parameter.

All of the algorithms take the Document Word Matrix or Document Term Matrix as input

All of them gives out Document Topic Matrix and Topic Term Matrix as output matrices.