Topic Modelling Techniques

using

Linear Algebra

May 2020

**1. Abstract**

In machine learning and natural language processing1, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modelling2 is a frequently used unsupervised text-mining tool for discovery of hidden semantic structures in a text body. Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: "dog" and "bone" will appear more often in documents about dogs, "cat" and "meow" will appear in documents about cats, and "the" and "is" will appear equally in both. A document typically concerns multiple topics in different proportions; thus, in a document that is 10% about cats and 90% about dogs, there would probably be about 9 times more dog words than cat words.

The "topics" produced by topic modelling techniques are clusters of similar words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document's balance of topics is. In the age of information, the amount of the written material we encounter each day is simply beyond our processing capacity. Topic models can help to organize and offer insights for us to understand large collections of unstructured text bodies. Originally developed as a text-mining tool, topic models have been used to detect instructive structures in data such as genetic information, images, and networks. They also have applications in other fields such as bioinformatics.

**2. Keywords**

Machine learning, Natural Language Processing, Topic Modelling, Text Mining, Unsupervised Learning, Statistical Model, Semantic Structure, Dimensionality Reduction, Feature Extraction, Information Retrieval, Eigenvectors, Eigenvalues, Orthogonality, LSA, SVD, NMF

**3. Introduction**

We have used the 20 News Group Dataset3, originally made by Ken Lang for his paper on filtering news from the net. It is an extremely popular dataset using for a variety of natural language processing tasks from text classification to text clustering and even topic modelling. The dataset, which has also been published on Kaggle as well as other renowned machine learning repositories such as the UCI Machine Learning Repository and is even a part of scikit-learn, contains almost 20,000 news documents for each of its 20 news categories. Some of the topics include hardware, motorcycles, electronics, space, sales, politics, and sports.

Our goal is to find the underlying topics within this dataset and try and obtain all the words corresponding to each topic in each news group, which will then be compared to the group’s topic for accuracy measures. To perform topic modelling, we need to extract appropriate features so we can pick the words that offer most relevance to each topic and these words will indicate the underlying topics in each set of documents. Our goal is to use the two most powerful linear algebraic techniques in topic modelling namely, Latent Semantic Analysis (LSA) and Non-Negative Matrix Factorization (NMF) and compare the results of these two approaches, view the topics identified by each algorithm and perform a simple cluster analysis. Both algorithms work on the principle of matrix factorization. LSA is a powerful statistical technique that performs matrix factorization on a document-term or term-document matrix using truncated Singular Value Decomposition (SVD) to reduce its dimension and obtain the features which provide us with the most singular values. NMF factorizes a matrix into three non-negative matrices where each of these matrices hold the relevance of words to topics and features.

**3. Latent Semantic Analysis – LSA**

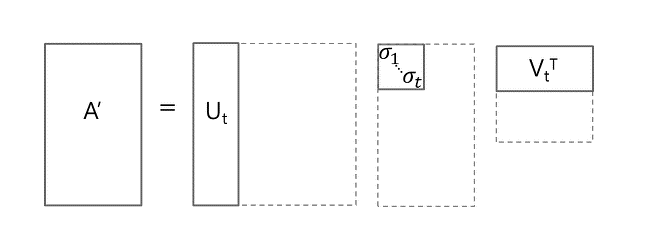
**3.1 Linear Algebra Background**

**Theorem 1.** *The singular value decomposition of an* ***m x n*** *real or complex matrix* **M** *is a factorization of the form* **USVT***, where* **U** *is an* ***m x m*** *real or complex unitary matrix,* **S** *is an* ***m x n*** *rectangular diagonal matrix with non-negative real numbers on the diagonal, and* **V** *is and* ***n x n*** *real or complex unitary matrix. If* **M** *is real,* **U** *and* **VT***are real orthonormal matrices.*

**M = U S VT** (1)

**Theorem 2.** *The diagonal entries* **S** *are known as the singular values of* **M***.*

**Theorem 3.** *The truncated singular value decomposition of a matrix* **M** *is performed by taking only the* ***t*** *column vectors of* **U** *and* ***t*** *row vectors of***VT** *corresponding to the* ***t*** *largest singular values in* **S** *and discarding the rest of the matrices.*



**3.2 Proof**

**3.3 Implementation**

**References**

**1** https://en.wikipedia.org/wiki/Natural\_language\_processing

**2** https://en.wikipedia.org/wiki/Topic\_model