# STA 3180 Statistical Modelling: Text Mining

# STA 3180 Statistical Modelling - Text Mining Lecture Notes

Text mining is a process of extracting meaningful information from unstructured text data. It involves the use of natural language processing (NLP) and machine learning algorithms to identify patterns, trends, and relationships in large amounts of text. Text mining can be used to gain insights into customer sentiment, product reviews, and other forms of text-based data.

## Key Concepts

**Natural Language Processing (NLP):** NLP is a branch of artificial intelligence that deals with understanding and manipulating human language. It is used to analyze text data and extract meaningful information from it.

**Text Mining:** Text mining is the process of extracting meaningful information from unstructured text data. It involves the use of NLP and machine learning algorithms to identify patterns, trends, and relationships in large amounts of text.

**Text Preprocessing:** Text preprocessing is the process of cleaning and preparing text data for further analysis. This includes tasks such as tokenization, stop word removal, stemming, and lemmatization.

**Tokenization:** Tokenization is the process of breaking down a text into individual words or phrases.

**Stop Word Removal:** Stop word removal is the process of removing words that are commonly used but have little meaning, such as “the”, “a”, and “of”.

**Stemming:** Stemming is the process of reducing words to their root form. For example, the words “running”, “ran”, and “runs” would all be reduced to the root form “run”.

**Lemmatization:** Lemmatization is the process of reducing words to their base form. For example, the words “running”, “ran”, and “runs” would all be reduced to the base form “run”.

## Practice Multiple Choice Questions

Q1. What is text mining?

A. The process of extracting meaningful information from unstructured text data

B. The process of cleaning and preparing text data for further analysis

C. The process of breaking down a text into individual words or phrases

D. The process of reducing words to their root form

Answer: A. The process of extracting meaningful information from unstructured text data

Explanation: Text mining is the process of extracting meaningful information from unstructured text data. It involves the use of natural language processing (NLP) and machine learning algorithms to identify patterns, trends, and relationships in large amounts of text.