**Document Text analysis with Latent Semantic Analysis**

Abstract

Contents

# Chapter 1 Introduction

The project is to develop an application to “model systematic communication by statistically analysing the difference between law and science, science and non-science and psychology” analyse and derive the text corpus of different types of documents semantically. Like documents with scientific research papers, non-scientific papers, psychology papers, etc., by extracting the statistical insights from the data using Latent Semantic Analysis (LSA). LSA is a natural language processing technique based on linear algebra that tries to capture and code the semantics of words and documents at different levels of document like paragraph level, document level, etc. (Pilato and Vassallo, 2015).

My interest in machine learning and the depth of knowledge I can acquire from the subject, and the research possibilities, and the challenging level of this project made me choose this project.

A lot of research has already been produced in this area. From(Mohamed and Watada, 2010), an LSA approach to document classification for knowledge application acquisition has been proposed, but the classification of documents is done by generating a set of words with high weights from the document and matching them to a predefined dictionary of categorised words obtained from training the documents. The scope of using a predefined dictionary to the model might not be the ultimate or the best solution to categorise the documents effectively. To avoid this limitation, a statistical method needs to be developed to derive the semantic distance between the weighted words. Knowing that there are many algorithms to analyse the document semantically, knowing the best algorithm to produce weighted words to analyse the document is a problem. There are many statistical methods to test the results, like the chi-square test, t-test, Mantel test. Choosing one that suites to the project results is a problem. So, the research question raises here. What machine learning algorithm is effective to analyse the document corpus? And What test best fits to test the results statistically?

# Chapter 2 Literature Review

# Chapter 3 Implementation

# Chapter 4 Experiment Design And Results