# Project 22:

# Develop a classification model to group articles based on the placement of texts within the articles

**Project Supervisor:** Prabha Rajagopal

## Abstract

*Finding information is essential for many. The internet especially has become a go to place for many different search queries. However, some of these articles appearing in the search may or may not be relevant. And even if they are relevant to the search query, it may not necessarily be easy to find the information in the articles or links. Hence, it may be beneficial to the users if the articles that are easier to find information are prioritized in the displayed search results. Therefore, this project attempts to develop a classification model to group articles based on the placement of texts within the articles.*

## 1. Project Overview

Users tend to prefer articles which are easier to find the information they are looking for. However, not all articles display their information in an easy to locate places. Some articles place their information in strategic places such as abstract, keywords, or introduction. This project intends to develop a model that can classify articles based on the appearances of words in the text such as abstract, keyword, introduction or paragraphs.

## 2. Objectives

To classify articles based on the placement of texts within the articles.

To develop a classification model to predict new articles based on the location of text within the articles.

## 3. Background

Information retrieval (IR) systems such as search engines retrieve large amount of documents or links related to a specific query. Of these retrieved documents, some may be relevant while others irrelevant. Depending on the retrieved relevant documents, the effectiveness of IR systems can be measured using specific metrics. Then the performance of the IR systems can be compared with other IR systems to determine which of the systems are performing well in retrieving relevant documents.

The information retrieval evaluation consists of two different types; the user- based and system-oriented evaluation. The system-oriented evaluation has always prioritized relevance of the document to a user’s query as a measure for user satisfaction. However, studies have shown that the outcome of system-based evaluation does not agree with real user satisfaction (Hersh et al., 2000; Turpin & Hersh, 2001) due to the factors such as effort, system and user effectiveness, and user characteristics (Al-Maskari &

Sanderson, 2010) which influence user satisfaction. Other recent studies also have highlighted that effort in retrieving relevant documents is equally important for user satisfaction (Verma, Yilmaz, & Craswell, 2016; Yilmaz, Verma, Craswell, Radlinski, & Bailey, 2014) and have shown differences in IR system evaluation when effort were incorporated (Rajagopal, Ravana, Koh, & Balakrishnan, 2019). The effort in this context is referring to the amount of work needed by the user to find, identify and understand the relevant content in the document.

Findability is the ease with which information contained on a website can be found, both from outside the website (using search engines and the like) and by users already on the website. Although findability has relevance outside the World Wide Web, the term is usually used in that context (https://en.wikipedia.org/wiki/Findability).

## 4. Proposed Methodology

1. Import/clean/tidy/transform data.

2. Determine the appearances of title words in each article (where does the title words appear?).

3. Produce a classification model.

4. Measure the performance of the model.

5. Analyze the characteristics of each group of articles. (eg: easy findability articles have title words that appear in the abstract; difficult fundability articles have title words that appear in first paragraphs).

6. Include suitable visualization to showcase your project outcome.

## 5. Scope of Technology

1. R/python programming

2. Text analysis

## 6. References (if any)

Suggested dataset: https://plos.org/text-and-data-mining/

1. Al-Maskari, A., & Sanderson, M. (2010). A review of factors influencing user satisfaction in information retrieval. Journal of the American Society for Information Science and Technology, 61(5),

## 7. Additional Documents (if any)