# Assessment 3 Information

| | | | |
|---|---|---|---|
| **Subject Code:** | TECH3300 | | |
| **Subject Name:** | Machine Learning Applications | | |
| **Assessment Title:** | Design a Text Retrieval System | | |
| **Assessment Type:** | Coding and Presentation | | |
| **Assessment Length:** | 2000 | Words | (+/-10%) |
| **Weighting:** | 40 % | | |
| **Total Marks:** | 40 | | |
| **Submission:** | MyKBS | | |
| **Due Date:** | Week 13 | | |

## Your Task

Design a text retrieval system to find similar movies/shows based on the descriptions.

## Assessment Description

We humans communicate using different languages, either by speaking or writing. Text data is abundant in the real world. It's a challenging task to work with natural languages. Your team lead has assigned you one such task of recommending movies based on the movie description.

### Data

A movies/shows dataset with description is curated by pre-processing the Kaggle IMDb Movies/Shows with Descriptions dataset (https://www.kaggle.com/datasets/ishikajohari/imdb-data-with-descriptions) and is provided to you in MyKBS. You are encouraged to explore the original source.

The original dataset is pre-processed and is provided in 2 files - train.csv and test.csv. MyKBS provides you these files each containing following columns:

- title: Title of the movie/show.
- description: Description of the movie/show.

You are required to train a text retrieval system using the train.csv file. And test the system using the test.csv file.

## Problem Statement

As an individual, you are required to download the data sets, i.e., train.csv and test.csv files from MyKBS. You must build a text retrieval system to find similar movies/shows based on the descriptions. You should systematically approach the problem by addressing the below tasks:

- Load the data sets and pre-process them to fit your requirements. You must use at least two pre-processing techniques. (5 marks)
- Design a text retrieval system using TF-IDF (with inverted file) algorithm. (10 marks)
- Find the top 3 movies/shows matches in the train.csv based on the descriptions provided in the test.csv. (5 marks)
- You are to record a 5-minute video accompanying PowerPoint slides to elaborate the approach and performance of the system using relevant metric(s). In recording this video, you will need to prepare accompanying PowerPoint slides thar are clear, concise, of the required quality and references in accordance with the Kaplan Harvard Referencing style. (20 marks)

## Learning Objectives

This assessment aims to achieve the following subject learning outcomes:

| | |
|---|---|
| **LO1:** | Explore programming functions to source, store and prepare data for machine learning applications. |
| **LO3:** | Design algorithmic models for the application of machine learning in information technology. |
| **LO4:** | Create advanced insights of strategic organisational value with the aid of machine learning. |

## Assessment Guidelines

You are required to follow the below guidelines:

- You should write your Text Retrieval System code using Python 3 programming language.

- The use of any Python third-party package(s) is restricted to the following tasks:

  - Loading the datasets. E.g., Pandas.
  - Any necessary text pre-processing steps. E.g., Natural Language Toolkit, etc.
  - Performing necessary calculations during the building of the system. E.g., NumPy.
  - Calculating the performance of the system. E.g., Scikit Learn, Matplotlib, Plotly, etc.

- You should **NOT** use any third-party package for calculating TF-IDF (with inverted file).

- You should **ONLY** use the provided files, i.e., train.csv and test.csv for training/testing your system.

- Finally, submit your Python code and the recorded presentation via MyKBS.

## Important Study Information

## Academic Integrity Policy

KBS values academic integrity. All students must understand the meaning and consequences of cheating, plagiarism and other academic offences under the Academic Integrity and Conduct Policy.

- What is academic integrity and misconduct?
- What are the penalties for academic misconduct?
- What are the late penalties?
- How can I appeal my grade?

The answers to these questions can be accessed at https://www.kbs.edu.au/about-us/school-policies.

## Length Limits for Assessments

Penalties may be applied for assessment submissions that exceed prescribed limits.

## Study Assistance

Students may seek study assistance from their local Academic Learning Advisor or refer to the resources on the MyKBS Academic Success Centre page. Further details can be accessed at https://elearning.kbs.edu.au/course/view.php?id=1481

## Assessment Marking Guide

| **Marking Criteria**<br>**____ / 40** | **F (Fail)**<br>**0 – 49 %** | **P (Pass)**<br>**50 – 64 %** | **C (Credit)**<br>**65 – 74 %** | **D (Distinction)**<br>**75 – 84 %** | **HD (High Distinction)**<br>**85 – 100 %** |
|---|---|---|---|---|---|
| Load and pre-process the data.<br><br>____ / 5 | Inaccurate way(s) to load/pre-process the data or no attempt made. | Partially able to load/pre-process the data with some errors in both the flows. | Able to load/pre-process the data with some errors in one of the flows. Only implements 1 pre-processing technique. | Correctly load/pre-process the data with no errors. Prepares data for the system using 2 pre-processing techniques. | Demonstrates creative ways to pre-process to get better results in the retrieval process. |
| TF-IDF (with inverted file) implementation.<br><br>____ / 10 | Python code does not work correctly. | Python code works with errors in the TF-IDF and inverted file algorithm part. | Python code works with errors in TF-IDF or inverted file algorithm part. | Python code works perfectly for TF-IDF with inverted file algorithm. Code is documented with some comments. | TF-IDF with inverted file algorithm works flawlessly and is optimised for the speed. Python code is documented with exemplary comments to explain the processing of the code. |
| Retrieve Top 3 similar movies/shows.<br><br>____ / 5 | Top 3 similar movies/shows were not shown. | Flaws in retrieving top 3 similar movies/shows. | Some of the top 3 similar movies/shows are not correct. | The top 3 similar movies/shows are reported. | The top 3 similar movies/shows are reported in a legible manner. With proper comments to explain the process. |

| | | | | | |
|---|---|---|---|---|---|
| Recorded Presentation.<br><br>____ / 20 | You prepared and used generic PowerPoint slides that aid the recording and delivery of your video in a limited way.<br><br>Neither in-text referencing nor reference list adhere to Kaplan Harvard Referencing Style. | You prepared and used detailed PowerPoint slides that adequately aid the recording and delivery of your video.<br><br>Both in-text referencing and the resultant reference list adhere to Kaplan Harvard Referencing Style, with many errors. | You prepared and used detailed PowerPoint slides that sufficiently aid the recording and delivery of your video.<br><br>Both in-text referencing and the resultant reference list adhere to Kaplan Harvard Referencing Style, with some errors. | You prepared and used detailed PowerPoint slides that comprehensively aid the recording and delivery of your video.<br><br>Both in-text referencing and the resultant reference list adhere to Kaplan Harvard Referencing Style, with only occasional minor errors. | You prepared and used excellent PowerPoint slides that enhance the quality of the recording and delivery of your video.<br><br>Both in-text referencing and the resultant reference list adhere strictly to Kaplan Harvard Referencing Style, with no errors. |