**Lab assignment 2**

**CIS 492**

**Fabio Hinojosa Jimenez**

**2861151**

**No extra credit**

Table of Contents

[Lab 3 3](#_Toc162809083)

[Processing to Build TF-IDF Based Documents Vectors for Content Analysis 3](#_Toc162809084)

[Preprocessing 3](#_Toc162809085)

[Term Frequency (TF) Calculation 3](#_Toc162809086)

[Document Frequency (DF) Calculation 4](#_Toc162809087)

[TF-IDF Calculation 4](#_Toc162809088)

[Document Vectors 4](#_Toc162809089)

[Similarity Calculation 4](#_Toc162809090)

[Find Best Match 5](#_Toc162809091)

[Results 5](#_Toc162809092)

# Lab 3

## Processing to Build TF-IDF Based Documents Vectors for Content Analysis

To start we import the data using ‘csv.reader’ we append only the last row because that is where the text is.

A screenshot of a computer program

Description automatically generated

### Preprocessing

This function preprocess a given text by removing special characters, converting it to lowercase, removing stop words, and performing stemming using Porter Stemmer algorithm.

Later on the tokenize function the text into individual words.

A computer code with text

Description automatically generated

### Term Frequency (TF) Calculation

This function calculates the term frequency (TF) for each term in each document.

A screen shot of a computer code

Description automatically generated

### Document Frequency (DF) Calculation

This function calculates the document frequency (DF) for each term in the given list of topics across all documents.

A screen shot of a computer code

Description automatically generated

### TF-IDF Calculation

This function calculates the TF-IDF (Term Frequency-Inverse Document Frequency) score for each term based on its TF, DF, and the total number of documents.

A computer screen shot of a code

Description automatically generated

### Document Vectors

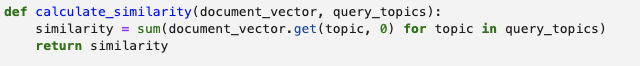
This function constructs document vectors using Lexicon-based TF-IDF scores for the given user query topics. It calculates TF-IDF scores for each term in each document and constructs a vector representation of each document based on the TF-IDF scores.

A screenshot of a computer code

Description automatically generated

### Similarity Calculation

This function calculates the similarity between a document vector and a set of query topics. It measures the relevance of the document to the query topics based on the TF-IDF scores of the terms in the document vector.



### Find Best Match

This function finds the State of the Union Address that best addresses the given query topics. It constructs document vectors for all documents, calculates their similarities to the query topics, and returns the index of the document with the highest similarity.

A screenshot of a computer code

Description automatically generated

## Results

Getting the results using ‘find\_best\_match’ function.

A screenshot of a computer code

Description automatically generated