VNUHCM - UNIVERSITY OF SCIENCE
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF KNOWLEDGE ENGINEERING

INTRODUCTION TO NATURAL LANGUAGE PROCESSING

PROJECT 01:

PRE-PROCESSING

GVHD: Dr. Nguyen Hong Buu Long
GVTH: Dr. Le Thanh Tung
Dr. Luong An Vinh

HO CHI MINH CITY
OCTOBER/2023

Contents

1 Introduction

1.1 Natural Language Processing

Natural Language Processing (NLP) is a fascinating and rapidly evolving field at the intersection of computer science, artificial intelligence, and linguistics. It focuses on the interaction between computers and human language, aiming to enable machines to understand, interpret, and generate human language in a way that is both meaningful and useful.

NLP seeks to bridge the gap between the complex, nuanced nature of human communication and the computational abilities of machines. It encompasses a wide range of tasks, including text analysis, language translation, sentiment analysis, speech recognition, and chatbot development, among others.

The core challenge in NLP is to teach computers to process and comprehend natural language, which is inherently ambiguous and context-dependent. Researchers and practitioners in this field employ various techniques, such as machine learning, deep learning, and linguistic analysis, to build models and algorithms capable of handling the intricacies of language.

NLP has found applications in numerous domains, from virtual assistants like Siri and Alexa to content recommendation systems, language translation services, and sentiment analysis tools used in social media monitoring. As NLP continues to advance, it holds the promise of revolutionizing the way we interact with technology, making human-machine communication more seamless and accessible than ever before.

1.2 Introduction to Python

Python is one of the most popular and versatile programming languages used extensively in the field of Natural Language Processing (NLP). Its simplicity, readability, and a rich ecosystem of libraries and tools make it an ideal choice for NLP practitioners and researchers.

Why Python?

- Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc).
- Python has a simple syntax similar to the English language.
- Python has syntax that allows developers to write programs with fewer lines than some other programming languages.
- Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.
- Python can be treated in a procedural way, an object-oriented way or a functional way.

You are required to use Python (version 3.0) on the Google Colab platform to carry out this project. You can learn Python through the following websites:

- https://www.w3schools.com/python/python_intro.asp
- https://www.programiz.com/python-programming

1.3 Introduction to dataset

The IMDb (Internet Movie Database) movie dataset is a valuable resource for Natural Language Processing (NLP) tasks and sentiment analysis. It contains a wealth of information about movies, including user reviews and ratings, making it an ideal dataset for exploring how people perceive and review films. In this introduction, we will discuss how to work with the IMDb movie dataset for NLP purposes in Python, including steps to show sample data and perform basic analysis.

2 Project Requirement

2.1 Theory Part

The students are required to research and answer the following questions, accompanied by corresponding examples in English. Try to provide clear and concise responses.

1. Define the term "stopwords" and provide examples in English.
2. Explain the significance of Term Frequency (TF) and Inverse Document Frequency (IDF) in text analysis. Why is it essential to use both TF and IDF together?
3. Evaluate the advantages and disadvantages of using TF-IDF in text processing.
4. In the scikit-learn library, investigate how IDF is implemented. Compare it to the standard IDF formula and explain any differences if present.

2.2 Programming

You are required to perform the following tasks using Python on the Colab platform. The tasks include:

1. Write a program to download the IMDb movie review dataset.
2. Write a program to calculate the number of data samples, the maximum, minimum, and average length of reviews, and create a chart showing the ratio of positive and negative classes in the dataset.
3. Write a program to tokenize the words in the reviews, lovercase all words, remove special characters and remove the stopwords.
4. Write a program to calculate the term frequency (TF) of each word in the reviews.
5. Write a program to calculate the inverse document frequency (IDF) of words in the reviews, considering each review as a separate document.
6. Write a function that takes a review sentence as input and returns its TF-IDF representation using the information obtained in the previous steps.

3 Submission

The submission file must be in the following format: [**Student_ID.zip**] or [**Student_ID.zip**], is the compression of the [**Student_ID**] folder. This folder contains:

- The [**Student_ID.ipynb**] downloaded from Google Colab. Before submitting, please choose **"Run All"** and make sure that your file can run normally
- In the Python Notebook, please add a text box and insert the information including your name and your student ID.
- The theory part should be **structured, logical, clear** and **coherent**.
- All links and books related to your submission must be mentioned.