**THỰC HÀNH BUỔI 1**

**TÌM HIỂU, CÀI ĐẶT THƯ VIỆN, NGUỒN DỮ LIỆU CÔNG KHAI**

|  |
| --- |
| Họ tên và MSSV:  Lớp: Ngày học: |

# **Các thư viện thường được sử dụng trong Python**

- Numpy - Keras - Pandas - Pytorch

- Tensorflow - Matplotlib - OpenCV - Seabon

- Scipy - Theano - Simpleitk - Pillow

- Selenium - Scrapy - Requests - Beautiful Soup

- Pytest - Mahotas

- NTLK - Scikit-Learn - Gensim - TextBlob

- Spacy - Pattern - Transformers

* **BÀI TẬP 1**

Sinh viên tìm hiểu các thư viện nêu trên (đặc điểm, chức năng, điểm nổi bật, cách thức và lệnh cài đặt) và cho biết thư viện nào hỗ trợ cho NLP.

Sinh viên nhập câu trả lời tại đây

* **BÀI TẬP 2**

Sinh viên chạy thực nghiệm các đoạn code sau và trả lời các câu hỏi:

**Code 1**

from textblob import TextBlob

text = "I love NLP. It's fantastic!"

blob = TextBlob(text)

sentiment = blob.sentiment.polarity

print(sentiment)

**Code 2**

import gensim

from gensim import corpora

documents = ["Apple and Microsoft are leading technology companies.",

"Python and Java are popular programming languages.",

"Football and basketball are exciting sports."]

stopwords = {"and", "are"}

texts = [[word for word in doc.split() if word.lower() not in stopwords] for doc in documents]

dictionary = corpora.Dictionary(texts)

corpus = [dictionary.doc2bow(text) for text in texts]

lda = gensim.models.LdaModel(corpus, num\_topics=3, id2word=dictionary, passes=15)

print(lda.print\_topics(num\_topics=3, num\_words=3))

**Code 3**

import spacy

nlp = spacy.load("en\_core\_web\_sm")

sentence = "The cat chased the mouse"

doc = nlp(sentence)

for token in doc:

print(token.text, "-->", token.dep\_)

1. Thư viện sử dụng trong đoạn code có chức năng gì?
2. Kết quả chạy thực nghiệm
3. Nhận xét kết quả

Sinh viên nhập câu trả lời tại đây

# **Các nguồn Dataset**

- **Kaggle**: https://www.kaggle.com/

- **Papers with Code**: https://paperswithcode.com/datasets

- **UCI Machine Learning Repository**: https://archive.ics.uci.edu/ml/index.php

- **Registry of Open Data on AWS**: https://registry.opendata.aws

- **Google Dataset Search**: https://datasetsearch.research.google.com/

- **Microsoft Datasets**: https://msropendata.com/

- **Reddit datasets**: https://www.reddit.com/r/datasets/top/?sort=top&t=all

- **CMU Libraries**: <https://guides.library.cmu.edu/az.php>

- **Public Datasets trên Github**: <https://github.com/awesomedata/awesome-public-datasets#machinelearning>

- **YouTube Dataset**: <https://research.google.com/youtube8m/>

- **OpenML**: <https://www.openml.org/search?type=data&sort=runs&status=active>

# BÀI TẬP 3

Sinh viên tìm hiểu các nguồn dataset nêu trên và cho biết:

- Đặc điểm của từng nguồn.

- Bộ dữ liệu công khai trên các nguồn thuộc lĩnh vực gì (nêu rõ từng nguồn).

- Dạng dữ liệu công khai.

- Ưu điểm, hạn chế (nếu có) của từng nguồn.

- Nguồn nào cung cấp dữ liệu cho lĩnh vực NLP.

- Chọn 1 bộ dữ liệu để tìm hiểu và cho biết các thông tin cơ bản về bộ dữ liệu (vd: lĩnh vực, số mẫu, kích thước, thuộc tính, ...)

- Nêu 01 nguồn cung cấp dữ liệu NLP uy tín (khác với các nguồn nêu trên).

Sinh viên nhập câu trả lời tại đây

|  |
| --- |
| **FILE PHẢI NỘP CHO BÀI THỰC HÀNH NÀY** |
| **Một file Word** chứa đáp án các yêu cầu |

*--- Hết ---*