–TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÀI TẬP GIỮA KỲ**

**HỌC MÁY**

*Người thực hiện*: **Nguyễn Minh Nhựt – 518H0545**

Lớp **: 18H50203**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÀI TẬP GIỮA KỲ**

**HỌC MÁY**

*Người thực hiện*: **Nguyễn Minh Nhựt – 518H0545**

Lớp **: 18H50203**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

# MỤC LỤC

[MỤC LỤC 1](#_Toc65793545)

**Text Classification**

1.1 Mô tả cách xây dựng dữ liệu và bộ dữ liệu

1.2 Sơ đồ tổng quát

1.3 Đặc trưng của mô hình

1.3.1. Tiền xử lý

1.3.2 Feature

1.3.3 Giải thuật

1.3.4 Cách đánh giá

1.4 Thực nghiệm

1.4.1 Công nghệ sử dụng

1.4.2 Các tham số của mô hình

1.4.3 Kết quả đạt được

1.4.4 Thảo luận

**Tự đánh giá**

# TEXT CLASSIFICATION

## 1.1 Mô tả cách xây dựng dữ liệu và bộ dữ liệu

### 1.1.1 Trình bày chi tiết đã xây dựng dữ liệu như thế nào?

Đầu tiên, em xin giới thiệu những thư viện đã dùng để xây dựng bộ dữ liệu (dataset) như sau:

* Selenium – thư viện chính cho việc scrape data từ trang VnExpress
* Xlwt – thư viện dùng để tương tác chính với file xls
* Time – thư viện dùng để set sleep time giữa những lần truy xuất url từ các trang báo

### 1.1.2 Dữ liệu dung lượng bao nhiêu, bao nhiêu tập tin, bao nhiêu lớp

Hướng xây dựng bộ dữ liệu đầu tiên là

### 1.1.3 Phân tích dữ liệu (văn bản dài nhất, ngắn nhất, số từ, …)

### 1.1.4 Cách phân chia dữ liệu: Train, Test, Validation

### 1.1.5 Phân tích các bộ dữ liệu trên

**1.2 Sơ đồ tổng quát**

Vẽ sơ đồ tổng quát giải quyết bài toán

**1.3 Đặc trưng của mô hình**

1.3.1 Tiền xử lý

Trình bày cụ thể về tiền xử lý dữ liệu

1.3.2 Feature

Trình bày chi tiết về feature của bài toán

1.3.3 Giải thuật

Trình bày chi tiết về giải thuật sử dụng trong bài toán

1.3.4 Cách đánh giá

Trình bày về các độ đo đánh giá bài toán

**1.4 Thực nghiệm**

1.4.1 Công nghệ sử dụng

Trình bày các thư viện sử dụng trong bài toán

(chỉ liệt kê và các version tương ứng)

1.4.2 Các tham số của mô hình

Bài toán có các tham số nào, lý do chọn các tham số đó

1.4.3 Kết quả đạt được

Các modul code chính

Kết quả đạt được theo các độ đo ở trên như thế nào

1.4.4 Thảo luận

Phân tích kết quả đạt được và thảo luận

**TỰ ĐÁNH GIÁ**

|  |  |  |
| --- | --- | --- |
| **Nội dung** | **Điểm** | **Ghi chú** |
| 1.1 Mô tả cách xây dựng dữ liệu và bộ dữ liệu **(2 điểm)** |  |  |
| 1.2 Sơ đồ tổng quát **(2 điểm)** |  |  |
| 1.3 Đặc trưng của mô hình **(3 điểm)** |  |  |
| 1.3.1. Tiền xử lý **(0.5 điểm)** |  |  |
| 1.3.2 Feature **(1 điểm)** |  |  |
| 1.3.3 Giải thuật **(1 điểm)** |  |  |
| 1.3.4 Cách đánh giá **(0.5 điểm)** |  |  |
| 1.4 Thực nghiệm **(3 điểm)** |  |  |
| 1.4.1 Công nghệ sử dụng **(0.5 điểm)** |  |  |
| 1.4.2 Các tham số của mô hình **(0.5 điểm)** |  |  |
| 1.4.3 Kết quả đạt được **(1.25 điểm)** |  |  |
| 1.4.4 Thảo luận **(0.75 điểm)** |  |  |
| **Tổng điểm** |  |  |

**TÀI LIỆU THAM KHẢO**

1. Natural Language Processing Recipes – Unlocking Text Data with Machine Learning and Deep Learning using Python – Akshay Kulkarni, Adrsha Shivananda