|  |  |  |
| --- | --- | --- |
|  | **TRƯỜNG ĐẠI HỌC BÁCH KHOA**  **KHOA CÔNG NGHỆ THÔNG TIN** |  |

TIỂU LUẬN CUỐI KỲ

HỌC PHẦN: KHOA HỌC DỮ LIỆU

**NHẬN DẠNG TIÊU ĐỀ BÀI BÁO**

|  |  |  |
| --- | --- | --- |
| HỌ VÀ TÊN SINH VIÊN | LỚP HỌC PHẦN | ĐIỂM BẢO VỆ |
| Lê Hữu Hưng | 20TCLC-KHDL |  |
| Hoàng Nguyên Bách | 20TCLC-KHDL |  |
| Đinh Huy Hoàng | 20TCLC-KHDL |  |

ĐÀ NẴNG, 05/2023**TÓM TẮT**

Trong tiểu luận này, chúng tôi tập trung vào giải quyết vấn đề nhận dạng tiêu đề bài báo từ văn bản. Vấn đề này là một tong những thách thức quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên, do tiêu đề bài báo thường chứa thông tin quan trọng và đóng vai trò quan trọng trong việc thu hút sự chú ý của người đọc.

Phương pháp giải quyết của chúng tôi là sử dụng PhoBert, một mô hình nhúng từ vựng tiếng Việt dựa trên kiến trúc Transformer. PhoBert cho phép chúng tôi biểu diễn các từ trong tiêu đề bài báo dưới dạng các vectơ số thực, giúp mô hình có khả năng hiểu được ngữ nghĩa và tương quan giữa các từ trong tiêu đề.

Kết quả đạt được khi xây dựng mô hình nhận dạng tiêu đề bài báo là khả năng nhận dạng và phân loại chính xác các tiêu đề bài báo từ văn bản. Mô hình của chúng tôi đã được huấn luyện trên một tập dữ liệu đa dạng và đạt được độ chính xác cao trong việc nhận dạng và phân loại các tiêu đề bài báo. Điều này có thể hỗ trợ trong việc tổ chức và phân loại thông tin từ các bài báo, giúp người dùng tiết kiệm thời gian và nỗ lực trong việc tìm hiểu và tìm kiếm thông tin cần thiết từ các tiêu đề bài báo.

**BẢNG PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| Sinh viên thực hiện | Các nhiệm vụ | Tự đánh giá theo 3 mức  (Đã hoàn thành/Chưa hoàn thành/Không triển khai) |
| Lê Hữu Hưng |  | -… |
| Hoàng Nguyên Bách | * Crawl data | Đã hoàn thành |
| Đinh Huy Hoàng |  |  |

**MỤC LỤC**

**CÁC YÊU CẦU VIẾT BÁO CÁO**

**VỀ HÌNH THỨC:**

* Định dạng: căn chỉnh lề kiểu Justify; font chữ 13, times new roman; giãn dòng 1.5, cách trước 6pt; hình ảnh phải đánh số, có chú thích bên dưới hình; bảng phải đánh số, có chú thích bên trên bảng.
* Báo cáo có độ dài từ 15 đến 20 trang và không đóng bìa gương (chỉ cần bìa mềm có màu) gồm: trang bìa (trang 1 của file hướng dẫn này), tóm tắt, bảng phân công nhiệm vụ,mục lục, và các trang nội dung.

**VỀ NỘI DUNG:**

Báo cáo gồm 06 phần sau.

# 1. Giới thiệu

Trong phần này, sinh viên giới thiệu vấn đề cần giải quyết và đề xuất giải pháp tổng quan cần thực hiện.

# 2. Thu thập và mô tả dữ liệu

**2.1. Thu thập dữ liệu**

SV mô tả mô tả giải pháp thu thập dữ liệu gồm nguồn dữ liệu, công cụ thu thập, cách thức sử dụng công cụ, đầu vào và đầu ra của quá trình thu thập, và cho ví dụ minh hoạ.

Yêu cầu:

* SV phải TỰ thu thập (crawl hoặc record) tập dữ liệu.
* Số lượng: >1000 mẫu.

**2.2. Mô tả dữ liệu**

SV mô tả các thống kê tổng quan về tập dữ liệu (số mẫu, số đặc trưng của 1 mẫu, kiểu dữ liệu của mỗi đặc trưng, số mẫu dữ liệu trống của mỗi đặc trưng,...) và xuất ra các thống kê mô tả trực quan về các đặc trưng (ví dụ: dùng boxplot, histogram, scatter plot về độ tương quan,…).

# 3. Trích xuất đặc trưng

SV trình bày việc lựa chọn đặc trưng, làm sạch và chuẩn hóa dữ liệu, giảm chiều dữ liệu, …; và trực quan hoá kết quả/tác dụng của các quá trình trên.

# 4. Mô hình hóa dữ liệu

SV trình bày quá trình xây dựng, xác thực và kiểm thử mô hình gồm:

- Chọn các mô hình phù hợp (ít nhất 2 mô hình) và mô tả ngắn gọn cơ sở lý thuyết của các mô hình đã lựa chọn, mô tả bộ tham số của mỗi mô hình (ví dụ: số lớp của mạng nơ-ron, số node của mỗi lớp, hàm activation được sử dụng,…) thông qua API sử dụng để xây dựng mô hình.

- Chia dữ liệu thành các tập Huấn luyện/Xác thực/Kiểm thử theo tỉ lệ phù hợp.

- Trình bày các tham số của quá trình huấn luyện mô hình (ví dụ: loss function, learning rate, number of epochs,…).

- Trình bày các đồ thị thể hiện hiệu suất (đánh giá bằng Loss hoặc Accuracy) của các mô hình trên các tập Huấn luyện/Xác thực/Kiểm thử.

- Đưa ra các độ đo (metrics) dùng để đánh giá các mô hình (ví dụ: RMSE hoặc MAPE), và so sánh hiệu quả của các mô hình bằng bảng số liệu (hoặc đồ thị) dựa trên các độ đo.

# 5. Kết luận

Sinh viên tổng kết các việc đã làm và các kết quả đạt được. Nêu hướng phát triển để cải thiện kết quả hiện có.

# 6. Tài liệu tham khảo

SV liệt kê các TLTK đã trích dẫn (cite) trong báo cáo tại đây.