**BÁO CÁO KẾT QUẢ THỰC HIỆN BÀI TẬP CUỐI KỲ - XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

**Tên Đề tài: HỆ THỐNG SỬA LỖI CHÍNH TẢ TIẾNG VIỆT CHO VĂN BẢN SỬ DỤNG TRÍ TUỆ NHÂN TẠO**

Họ và tên TV1: Lê Văn Phước Lớp: 20AD Chữ ký:

Họ và tên TV2: Hồ Luận Lớp: 20AD Chữ ký:

1. **Thuật toán, mô hình, thư viện, mã nguồn sử dụng trong đề tài:**

* **Thư viện:** rarfile, tqdm, pickle, re, unidecode, numpy, nltk, keras, tensorflow, collections.Counter, editdistance, flask.
* **Thuật toán:** 4-gram, 5-gram, one-hot encoding
* **Mô hình:** blstm
* **Mã nguồn mở:** github.com/VietHoang1512/Vietnamese-Spell-Correction
* **Mã nguồn:** [drive.google.com/drive/folders/19a\_fyp6nK0ES7GvnZpuX4oGEUMpFS5WH](https://drive.google.com/drive/folders/19a_fyp6nK0ES7GvnZpuX4oGEUMpFS5WH)

1. **Mô tả Dữ liệu (ngắn gọn):**

* Dữ liệu huấn luyện từ tập VNTC, tổng hợp từ các trang báo như vnexpress.net, tuoitre.vn, thanhnien.vn, nld.com.vn. Các bài báo thuộc nhiều lĩnh vực, chia thành tập huấn luyện và kiểm thử.
* Bộ dữ liệu sau khi cắt giảm gồm 10 lĩnh vực, mỗi lĩnh vực có 1499 văn bản.
* Link dữ liệu: https://github.com/duyvuleo/VNTC/tree/master/Data/ 10Topics/Ver1.1

1. **Liệt kê các chức năng đã hoàn thành**
2. Tiền xử lý

+ Chia dữ liệu thành từng câu một, loại bỏ đi các câu chứa kí tự đặc biệt và kí tự phi Latin

+ Tách các câu thành các các cụm từ dài nhất có thể

+ Từ các cụm từ đó ta tách thành 4-gram(phương pháp 1); 5-gram(phương pháp 2)

+ Sử dụng cấp độ character-level tokenization và phương pháp one-hot encoding để vector hóa từ và văn bản.

1. Xây dựng hàm tạo lỗi chính tả cho các cụm từ
2. Xây dựng, trainning mô hình học sâu sửa lỗi chính tả cho cụm 4 từ(phương pháp 1).
3. Xây dựng, trainning mô hình học sâu sửa lỗi chính tả cho cụm 5 từ(phương pháp 2).
4. Dùng mô hình và các phương pháp xử lý ngôn ngữ tự nhiên để sửa lỗi chính tả cho văn bản
5. Đánh giá hai phương pháp sửa lỗi chính tả cho văn bản (phương pháp 1, phương pháp 2)

Với thuật toán editdistance ta có thông số lỗi: Character Error Rate, Word Error Rate, Sequence Error Rate.

Qua các thông số trên ta có thể so sánh được phương pháp nào tốt hơn

1. Website sửa lỗi chính tả cho văn bản
2. **Phân chia công việc trong nhóm**

|  |  |
| --- | --- |
| Thành viên 1: Lê Văn Phước;  + Dùng mô hình và các phương pháp xử lý ngôn ngữ tự nhiên để sửa lỗi chính tả cho văn bản  + Phát triển thêm phương pháp 4-gram  + Huấn luyện các mô hình  + Xây dựng phương pháp đánh giá hai phương pháp 4-gram và 5-gram  + Làm website  + Làm slide, báo cáo | Thành viên 2: Hồ Luận  + Dịch, tìm hiểu các phương pháp của các bài báo nghiên cứu khoa học  + Làm slide |