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

Tên Đề tài: TÓM TẮT VĂN BẢN TIẾNG VIỆT

Họ và tên TV1: Hoàng Hữu Tiến Đạt Lớp: 21AD Chữ ký:

Họ và tên TV2: Dương Tấn Huy Lớp: 21AD Chữ ký:

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

- KMeans Clustering: Thuật toán nhóm các câu thành cụm dựa trên đặc trưng nội dung, sau đó chọn câu đại diện từ mỗi cụm để tóm tắt văn bản.
- TextRank: Thuật toán xây dựng đồ thị dựa trên độ tương đồng giữa các câu, áp dụng PageRank để xếp hạng và xác định các câu quan trọng nhằm tóm tắt nội dung.
- LSA (Latent Semantic Analysis): Sử dụng phân tích giá trị suy biến (SVD) để xác định các câu mang ý nghĩa chính trong văn bản, hỗ trợ tóm tắt ngữ nghĩa.
- Mô hình Word2Vec pretrained (cc.vi.300.bin): Mô hình nhúng từ tiếng Việt được tải từ fasttext, sau đó giảm chiều từ 300 xuống 100 (cc.vi.100.vec) trên Kaggle, dùng để biểu diễn từ hoặc câu dưới dạng vector cho các tác vụ tóm tắt.
- Thư viện mã nguồn mở: Dự án sử dụng các thư viện như numpy, pandas, matplotlib, seaborn, pyvi, nltk, scikit-learn, gensim, networkx, scrapy, re và streamlit.

II. Mô tả Dữ liệu (ngắn gọn):

- Nguồn dữ liệu: Dữ liệu văn bản tiếng Việt được thu thập từ các trang báo điện tử (bao gồm 24h.com.vn và soha.vn thông qua Scrapy).
- Thông số, độ lớn dữ liệu: Thư mục data/test chứa 1.000 tệp văn bản, mỗi tệp dài khoảng 300-600 từ, tổng cộng ít nhất khoảng 300.000-600.000 từ.

III. Liệt kê các chức năng đã hoàn thành:

1. Thu thập dữ liệu: Dùng Scrapy thu thập 1.000 bài báo (300–600 từ mỗi bài) từ các trang tin (24h.com.vn, soha.vn), lưu tại data/test.
2. Tiền xử lý: Tách từ, chuẩn hóa văn bản với PyVi và NLTK.
3. Giảm chiều: Chuyển mô hình Word2Vec từ cc.vi.300.bin sang cc.vi.100.vec.
4. Vector hóa: Dùng Word2Vec 100 chiều để biểu diễn câu thành vector ngữ nghĩa.
5. Tóm tắt văn bản: Áp dụng 3 phương pháp (KMeans Clustering, TextRank, LSA).
6. Giao diện: Dùng Streamlit cho phép nhập văn bản, chọn phương pháp, hiển thị kết quả.
7. Đánh giá & trực quan hóa: Dùng cosine similarity và SVD để đánh giá nội dung/ngữ nghĩa (evaluate.py). Vẽ biểu đồ so sánh hiệu suất các phương pháp.

IV. Phân chia công việc trong nhóm:

Công việc	Hoàng Hữu Tiến Đạt	Dương Tấn Huy
Thu thập dữ liệu	x	x
Tiền xử lý		x
Giảm chiều	x	
Vector hóa	x	
Tóm tắt văn bản	x	x
Giao diện	x	x
Đánh giá & trực quan hóa		x