BÁO CÁO DỰ ÁN: CÔNG CỤ PHÂN TÍCH VĂN BẢN (TEXT ANALYSIS TOOL)

🎯 1. TÓM TẮT DỰ ÁN

- **Tên dự án**: Text Analysis Tool
- **Mục tiêu**: Xây dựng một ứng dụng web có khả năng phân tích văn bản, tận dụng sức mạnh của các thư viện Xử lý Ngôn ngữ Tự nhiên (NLP).
- **Công nghệ chính**: NLTK, SpaCy, Streamlit.

🚀 2. CÁC TÍNH NĂNG CHÍNH

2.1. Tokenization (Phân đoạn từ)

- Tách văn bản đầu vào thành các đơn vị nhỏ nhất (tokens) bằng cả hai thư viện NLTK và SpaCy.
- Hiển thị tổng số lượng tokens và danh sách chi tiết các tokens đã được phân tách.

2.2. Part-of-Speech Tagging (Gắn nhãn từ loại)

- Xác định và gắn nhãn từ loại (danh từ, động từ, tính từ, v.v.) cho mỗi token trong văn bản.
- Cung cấp bảng giải thích chi tiết ý nghĩa của các nhãn từ loại (POS tags) để người dùng dễ dàng tham khảo.
- Trình bày kết quả dưới dạng bảng dữ liệu có thể sắp xếp và lọc.

2.3. Named Entity Recognition (Nhận dạng thực thể có tên)

- Tự động nhận dạng và phân loại các thực thể được đặt tên trong văn bản như: Tên người (Person), Tổ chức (Organization), Địa điểm (Location), Ngày tháng (Date), Tiền tệ (Money), v.v.
- Làm nổi bật (highlight) các thực thể đã nhận dạng ngay trên văn bản gốc để dễ dàng quan sát.
- Tạo biểu đồ trực quan hóa sự phân bố của các loại thực thể khác nhau trong văn bản.

2.4. Giao diện web thân thiện

- Sử dụng Streamlit để xây dựng một giao diện người dùng trực quan, đơn giản và dễ tương tác.
- Thiết kế layout hai cột khoa học: một cột cho việc nhập liệu và tùy chỉnh, cột còn lại để hiển thị kết quả phân tích.
- Tích hợp sidebar để người dùng nhập văn bản và lựa chọn các chức năng phân tích.

📊 3. KẾT QUẢ PHÂN TÍCH VÍ DỤ

Văn bản mẫu:

Apple Inc. was founded by Steve Jobs in California on April 1, 1976. The company is now worth over $3 trillion.

Kết quả phân tích:

- **Tokens**: ["Apple", "Inc.", "was", "founded", "by", "Steve", "Jobs", ...]

- **POS Tags (Từ loại)**: [("Apple", "NNP"), ("Inc.", "NNP"), ("was", "VBD"), ...]
- **Entities (Thực thể)**:
  - Apple Inc. → **ORG** (Tổ chức)
  - Steve Jobs → **PERSON** (Người)
  - California → **GPE** (Địa điểm)
  - April 1, 1976 → **DATE** (Ngày tháng)
  - $3 trillion → **MONEY** (Tiền tệ)

🛠 4. HƯỚNG DẪN CÀI ĐẶT VÀ SỬ DỤNG

Yêu cầu hệ thống:

- Đã cài đặt Conda hoặc Miniconda.
- Python phiên bản 3.9 trở lên.

Các bước cài đặt:

1. **Tạo môi trường Conda:**
   conda env create -f environment.yml

2. **Kích hoạt môi trường:**
   conda activate text-analysis-tool

3. **Tải về mô hình ngôn ngữ của SpaCy:**
   python -m spacy download en_core_web_sm

4. **Chạy ứng dụng:**
   streamlit run app.py

📊 5. THỐNG KÊ VĂN BẢN

Ứng dụng cung cấp các số liệu thống kê cơ bản và hữu ích về văn bản đầu vào, bao gồm:

- Tổng số ký tự
- Tổng số từ
- Tổng số câu
- Tổng số thực thể (entities) được nhận dạng

🎨 6. GIAO DIỆN NGƯỜI DÙNG

- **Sidebar**: Khu vực để người dùng nhập văn bản cần phân tích và lựa chọn các tùy chọn.
- **Layout 2 cột**: Bố cục rõ ràng giúp hiển thị song song phần nhập liệu và kết quả.
- **Bảng dữ liệu tương tác**: Kết quả được trình bày trong bảng có thể sắp xếp, tiện lợi cho việc tra cứu.
- **Biểu đồ trực quan**: Sử dụng thư viện Plotly để vẽ biểu đồ, giúp việc phân tích trở nên sinh động và dễ hiểu hơn.
- **Highlight Entities**: Các thực thể được làm nổi bật bằng màu sắc khác nhau trực tiếp trên văn bản gốc.

🔧 7. CẤU TRÚC DỰ ÁN

Cấu trúc thư mục của dự án được tổ chức như sau:

```
Text-Analysis-Tool/
├── app.py              # File mã nguồn chính của ứng dụng Streamlit
├── environment.yml     # File định nghĩa môi trường Conda
├── README.md           # File hướng dẫn tổng quan về dự án
├── NLP_ClassDay2.pdf   # Tài liệu tham khảo
├── task1.png           # Hình ảnh minh họa task 1
└── task2.png           # Hình ảnh minh họa task 2
```

📚 8. CÁC THƯ VIỆN SỬ DỤNG

- **Streamlit**: Framework chính để xây dựng giao diện ứng dụng web.
- **SpaCy**: Thư viện NLP hiện đại, hiệu suất cao, dùng cho các tác vụ nhận dạng thực thể và gắn nhãn từ loại.
- **NLTK (Natural Language Toolkit)**: Một thư viện NLP toàn diện, được sử dụng cho tác vụ phân đoạn từ.
- **Pandas**: Dùng để xử lý và hiển thị dữ liệu dưới dạng bảng (DataFrame).
- **Plotly**: Thư viện dùng để tạo các biểu đồ tương tác và trực quan hóa dữ liệu.

🎯 9. KẾT LUẬN

Dự án **Text Analysis Tool** đã hoàn thành mục tiêu đề ra, xây dựng thành công một ứng dụng web hoàn chỉnh cho việc phân tích văn bản với các tính năng NLP cốt lõi. Ứng dụng sở hữu giao diện thân thiện, dễ sử dụng, đồng thời cung cấp các kết quả phân tích chi tiết, trực quan và có giá trị cho người dùng.