criteria

Harito ID

2025-10-14

Tiêu chí đánh giá Lab 4: Word Embeddings

Phần 1: Triển khai (50%)

- ☐ **Task 1: Tải và sử dụng model có sẵn (Gensim)**
  - ☐ Tải thành công pre-trained model (ví dụ: `glove-wiki-gigaword-50`).
  - ☐ Lấy được vector của một từ.
  - ☐ Tính được độ tương đồng (similarity) giữa hai từ.
  - ☐ Tìm được các từ đồng nghĩa (most similar).
- ☐ **Task 2: Nhúng câu/văn bản**
  - ☐ Triển khai được hàm nhúng một văn bản bằng cách lấy trung bình vector các từ.
- ☐ **Task 3: Huấn luyện model trên tập dữ liệu nhỏ (Gensim)**
  - ☐ Huấn luyện thành công model Word2Vec từ dữ liệu thô (ví dụ: `en_ewt-ud-train.txt`).
  - ☐ Lưu và tải lại model đã huấn luyện.
- ☐ **Task 4: Huấn luyện model trên tập dữ liệu lớn (Spark)**
  - ☐ Cài đặt và cấu hình `pyspark`.
  - ☐ Đọc và tiền xử lý dữ liệu lớn (ví dụ: `c4-train...json`) bằng Spark.
  - ☐ Huấn luyện thành công model Word2Vec bằng `Spark MLlib`.
- ☐ **Task 5: Trực quan hóa Embedding**
  - ☐ Sử dụng PCA hoặc t-SNE để giảm chiều các word vector xuống 2D.
  - ☐ Vẽ biểu đồ scatter plot để trực quan hóa và quan sát các cụm từ.

Ví dụ: 0 / 12 mục được hoàn thành.

Phần 2: Báo cáo và Phân tích (50%)

- ☐ **Giải thích các bước thực hiện:** Trình bày rõ ràng, mạch lạc các bước đã làm.
- ☐ **Hướng dẫn chạy code:** Nêu cách để thực thi lại notebook và xem kết quả.
- ☐ **Phân tích kết quả (Quan trọng):**
  - ☐ Nhận xét về độ tương đồng và các từ đồng nghĩa tìm được từ model pre-trained.
  - ☐ Phân tích biểu đồ trực quan hóa: Các từ có gần nhau như kỳ vọng không? Có cụm từ nào thú vị không? Giải thích tại sao.
  - ☐ So sánh (nếu có) giữa model pre-trained và model tự huấn luyện.
- ☐ **Nêu khó khăn và giải pháp:** Ghi lại những vấn đề gặp phải và cách bạn đã giải quyết.
- ☐ **Trích dẫn tài liệu:** Ghi rõ các nguồn tham khảo bên ngoài (nếu có).

Ví dụ: 0 / 6 mục được hoàn thành.

Công thức tính điểm

Điểm số cuối cùng được tính như sau:

$$Score = \left(0.5 \times \frac{\text{CodeTasks}}{12} + 0.5 \times \frac{\text{ReportTasks}}{6}\right) \times 10$$

Trong đó: - `CodeTasks` là số lượng mục đã hoàn thành trong Phần 1. - `ReportTasks` là số lượng mục đã hoàn thành trong Phần 2.

Điểm sẽ được làm tròn đến 0.25.

Lưu ý: - Các tiêu chí trên là bắt buộc. - Nộp bài trễ sẽ bị trừ điểm. - Phải trích dẫn bất kỳ nguồn bên ngoài, model, thư viện hoặc công cụ đã sử dụng.