BÁO CÁO TIẾN ĐỘ LẦN 2

I. Cấu trúc thư mục

Link github của đồ án: https://github.com/TruongNgocTai/DoAnThucHanh/

Các thư mục:

- ThuThapDuLieuVaTienXuli_Amazon.ipynb : file source code của đồ án
- DuLieuTho.json : file dữ liệu vừa crawl về chưa xử lí, lưu dưới dạng chuỗi.
- DuLieuDaQuaTienXuLi.json : dữ liệu đã được tiền xử lí.
- Pos_reviews.txt, neg_reviews.txt, neur_reviews.txt : các file chứa các reviews được phân loại positive – negative – neutral tương ứng.
- Folder productPicture : chứa ảnh các sản phẩm đại diện. (vì khi download hơn 4000 reviews nhưng số lượng review thuộc neg và neutr vẫn không đủ nên phải crawl thêm, và vì nhiều nên chỉ để vài ảnh đại diện).

Đặc biệt: dữ liệu sau khi qua làm sạch tự động thì vẫn chưa sạch hoàn toàn, và khi chia tách dữ liệu được chia dựa theo số sao, dẫn đến khi dữ liệu được chia thì sẽ được duyệt lại 1 lần nữa bằng tay để đảm bảo review đó đúng lớp.

II. Tiến độ

Đã crawl đủ dữ liệu
Đã làm sạch dữ liệu
Đã chuyển đổi thành vector số: sử dụng tf-idf
Đã đưa vào model huấn luyện: ử dụng model phân lớp SVC – support vector classificatioin

III. Hướng chọn mô hình huấn luyện

Sử dụng mô hình GridSearch CV() và model phân lớp Support Vector Classification - SVC()
Sử dụng phương pháp kiểm tra chéo(cross-validation) k-fold.

Ở trong source code: mô hình GridSearchCV() thực hiện cross-validation với 5-fold cho tất cả các dữ liệu(vector) đi qua model SVC(3 lớp) và chọn tham số tốt nhất.