Từ nguồn dữ liệu đã cho, sinh viên thực hiện:

(Code được làm trên file notebook)

1. **Tiền xử lý dữ liệu: (STT_Cau1.ipynb)**
   - Chuyển văn bản thành chữ thường;
   - Loại bỏ URL;
   - Xóa dấu câu;
   - Xóa chữ số;
   - Tách câu;
   - Chuyển biểu tượng cảm xúc thành văn bản (Emojis and Emoticons);
   - Xóa các từ không có nghĩa (Stop Words);
   - Chuẩn hóa văn bản (Standardizing Text);
   - Sửa lỗi chính tả (Correcting Spelling);
   - Tách từ (Tokenizing);

   Lưu kết quả vào từng file riêng, đặt vào thư mục: **\Data_Preprocessing**

2. **Thực hiện khám phá dữ liệu: (STT_Cau2.ipynb)**

   a. Thống kê tần suất từ xuất hiện trong văn bản;

   b. Trực quan hóa bằng đồ thị và word cloud;

3. **Xây dựng vector đặc trưng văn bản:**

   a. Bag-of-Words (BOW); **(STT_Cau3a.ipynb)**

   b. N-grams (Bag-of-N-grams); **(STT_Cau3b.ipynb)**

   c. TF-IDF (Term Frequency-Inverse Document Frequency); **(STT_Cau3c.ipynb)**