Các bước xử lý status đã lấy được trên mạng xã hội:

**Bước 1:**

Sau khi lấy được các status trên mạng facebook em sẽ lưu vào file với định dạng như sau:

Id người dung, tên người dung, giới tính, id status, status

Tên file: full\_status.csv



Status hiện tại chưa được xử lý, có rất nhiều status không phải tiếng việt, có nhiều ký tự, hình vẽ, quá ngắn, quá dài, cần phải loại bỏ những status như vậy:



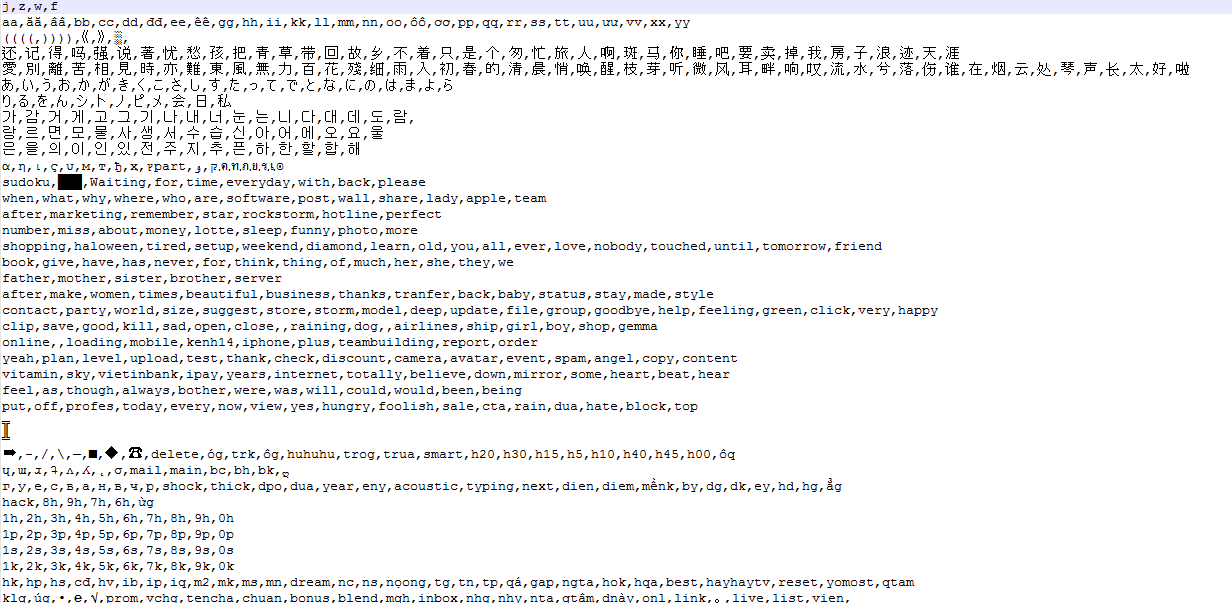




**…**

Em định nghĩ ra 1 file là **filter.data.** Nếu status nào có những từ ký tự nào trùng với danh sách trong file filter.data thì em sẽ loại bỏ đi.





Em cũng loại bỏ status có từ ký tự nhỏ hơn 6 và lớn hơn 255, status có số ký tự lớn gấp đồi từ cũng bị loại bỏ.

Những status là tiếng việt nhưng viết không dấu cũng bị loại bỏ (Nhưng không lọc được sạch hết).

Sau đó em được 1 file **full\_status\_filter.csv**.

**Bước 2:**

Khi đã được **full\_status\_filter.csv** file em sẽ cho file chạy qua vntokenizer tool. Loại bỏ những status chạy lâu đi. Cuối cùng em sẽ được **vn\_tokenizer\_status.csv** file



**Bước 3:**

Sau đó em đếm số word xuất hiện trong toàn bộ file. Nếu word nào xuất hiện ít hơn 5 lần em sẽ bỏ qua, nếu word nào length nhỏ hơn bằng 1 cũng bị loại rồi tạo ra 4 file định dạng LibSVM

**facebook\_status\_binary.libsvm**: Xuất hiện là 1, không xuất hiện là 0

**facebook\_status\_default.libsvm**: Đếm số word xuất hiện trong status đó

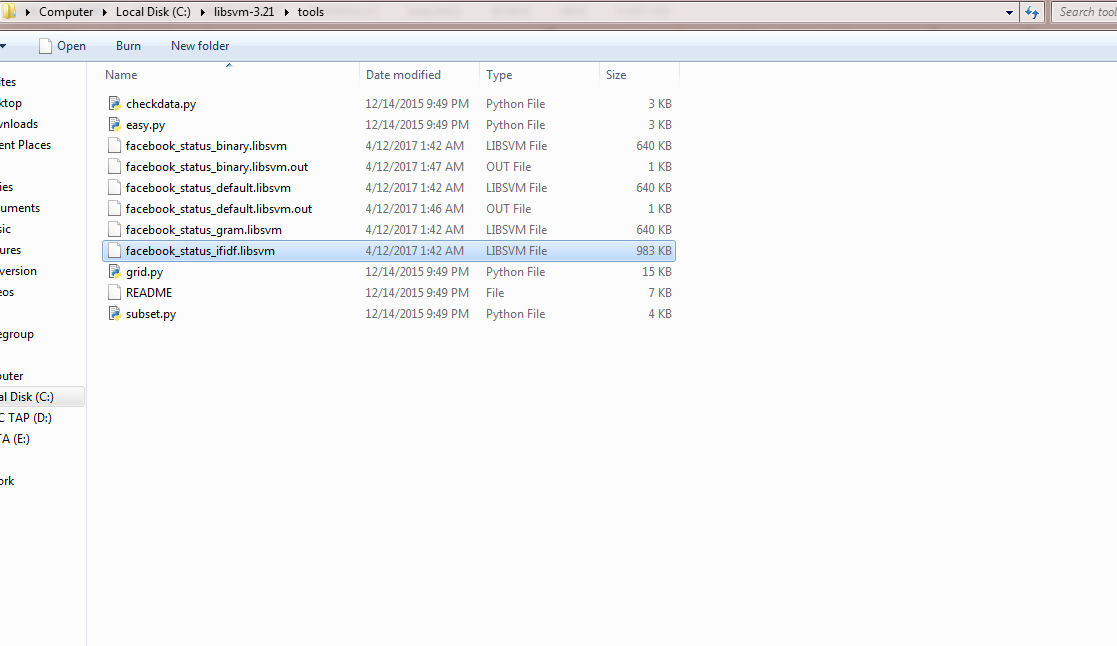
**facebook\_status\_ngram.libsvm**: Đếm số ngram của các word trong status đó. 1-gram là 1, 2-gram là 2…

**facebook\_status\_ifidf.libsvm**: Tính chỉ số if-idf của word trong status đó với toàn bộ văn bản.

**Bước 4:**

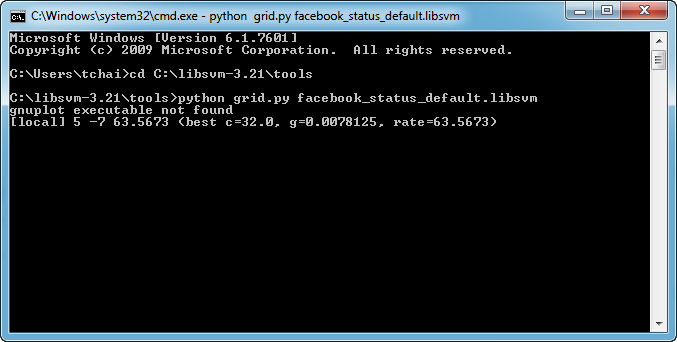
Khi đã có 4 file em lần lượt cho vào training

Trỏ tới đường dẫn có chứa libsvm: cd c:\libsvm-3.21\tools

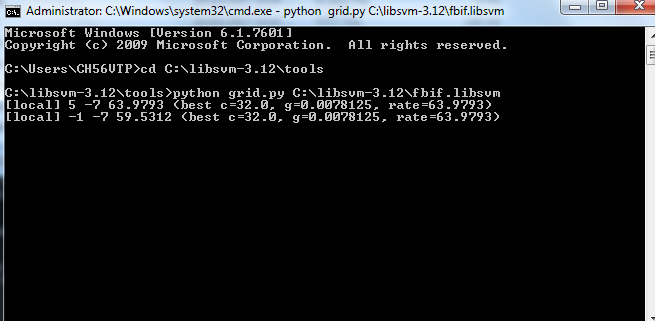


Dùng cmd chạy 4 câu lệnh sau:

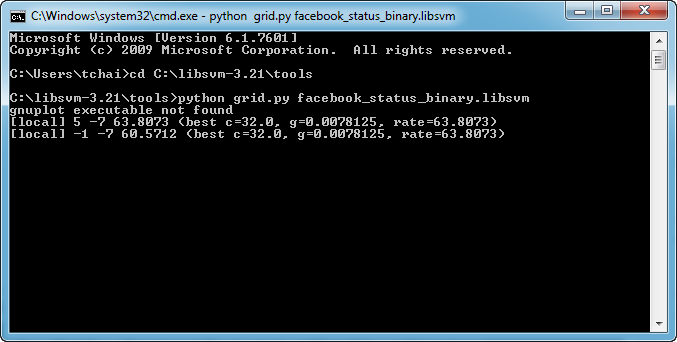
**python grid.py facebook\_status\_default.libsvm**



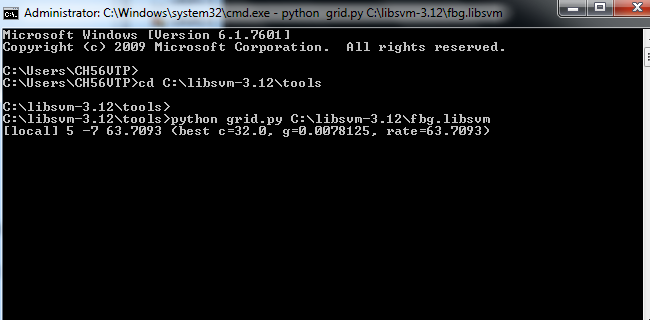
**python grid.py facebook\_status\_ifidf.libsvm**



**python grid.py facebook\_status\_binary.libsvm**



**python grid.py facebook\_status\_gram.libsvm**



END!