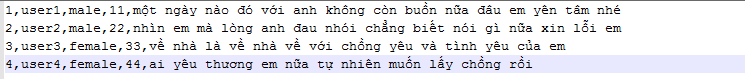
Danh sách 4 status ban đầu:

*1,user1,male,11,một ngày nào đó với anh không còn buồn nữa đâu em yên tâm nhé*

*2,user2,male,22,nhìn em mà lòng anh đau nhói chẳng biết nói gì nữa xin lỗi em*

*3,user3,female,33,về nhà là về nhà về với chồng yêu và tình yêu của em*

*4,user4,female,44,ai yêu thương em nữa tự nhiên muốn lấy chồng rồi*



File ban đầu của em gồm 4 status như trên. Trong đó có 2 male, 2 female

Bước 1: Sau khi chạy vnTokenizer thì kết quả thu được là:

*1,user1,male,11,một ngày nào\_đó với anh không còn buồn nữa đâu em yên\_tâm nhé*

*2,user2,male,22,nhìn em mà lòng anh đau\_nhói chẳng biết nói\_gì nữa xin\_lỗi em*

*3,user3,female,33,về nhà là về nhà về với chồng yêu và tình\_yêu của em*

*4,user4,female,44,ai yêu\_thương em nữa tự\_nhiên muốn lấy chồng rồi*

Bước 2: Sau đó em đếm các từ khác nhau: có 35 từ khác nhau trong tổng số 47 từ

|  |  |  |
| --- | --- | --- |
| **STT** | **Từ** | **Số lần** |
| 1 | một | 1 |
| 2 | ngày | 1 |
| 3 | nào\_đó | 1 |
| 4 | với | 2 |
| 5 | anh | 2 |
| 6 | không | 1 |
| 7 | còn | 1 |
| 8 | buồn | 1 |
| 9 | nữa | 3 |
| 10 | đâu | 1 |
| 11 | em | 5 |
| 12 | yên\_tâm | 1 |
| 13 | nhé | 1 |
| 14 | nhìn | 1 |
| 15 | mà | 1 |
| 16 | lòng | 1 |
| 17 | đau\_nhói | 1 |
| 18 | chẳng | 1 |
| 19 | biết | 1 |
| 20 | nói\_gì | 1 |
| 21 | xin\_lỗi | 1 |
| 22 | về | 3 |
| 23 | nhà | 2 |
| 24 | là | 1 |
| 25 | chồng | 2 |
| 26 | yêu | 1 |
| 27 | và | 1 |
| 28 | tình\_yêu | 1 |
| 29 | của | 1 |
| 30 | ai | 1 |
| 31 | yêu\_thương | 1 |
| 32 | tự\_nhiên | 1 |
| 33 | muốn | 1 |
| 34 | lấy | 1 |
| 35 | rồi | 1 |

Bước 4: Sau đó em tạo 4 file định dạng libsvm:

Trong đó nhãn +1 là male, -1 là female

Facebook\_data\_fake.libsvm: File này đếm số lượng từ trong status đó

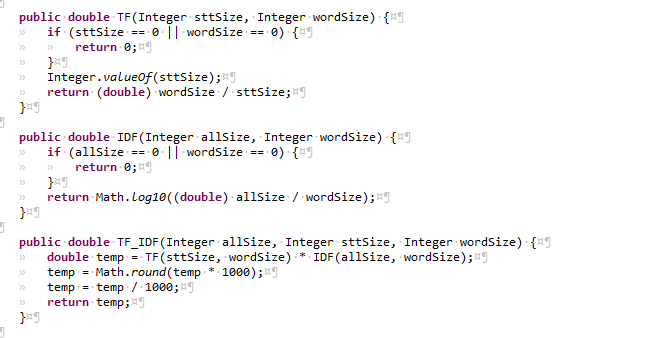
+1 1:1 2:1 3:1 4:1 5:1 6:1 7:1 8:1 9:1 10:1 11:1 12:1 13:1

+1 5:1 9:1 11:2 14:1 15:1 16:1 17:1 18:1 19:1 20:1 21:1

-1 4:1 11:1 22:3 23:2 24:1 25:1 26:1 27:1 28:1 29:1

-1 9:1 11:1 25:1 30:1 31:1 32:1 33:1 34:1 35:1

Facebook\_data\_fake\_tfidf.libsvm: Chỉ số TF-IDF



wordSize: là số lượng xuất hiện của word trong status đang xét

sttSize: là tổng số lượng các word trong status đang xét

allSize: là tổng số status (=4)

+1 1:0.046 2:0.046 3:0.046 4:0.046 5:0.046 6:0.046 7:0.046 8:0.046 9:0.046 10:0.046 11:0.046 12:0.046 13:0.046

+1 5:0.05 9:0.05 11:0.05 14:0.05 15:0.05 16:0.05 17:0.05 18:0.05 19:0.05 20:0.05 21:0.05

-1 4:0.046 11:0.046 22:0.029 23:0.046 24:0.046 25:0.046 26:0.046 27:0.046 28:0.046 29:0.046

-1 9:0.067 11:0.067 25:0.067 30:0.067 31:0.067 32:0.067 33:0.067 34:0.067 35:0.067

Facebook\_data\_fake\_binary.libsvm: Xuất hiện là 1, không xuất hiện là 0

+1 1:1 2:1 3:1 4:1 5:1 6:1 7:1 8:1 9:1 10:1 11:1 12:1 13:1

+1 5:1 9:1 11:1 14:1 15:1 16:1 17:1 18:1 19:1 20:1 21:1

-1 4:1 11:1 22:1 23:1 24:1 25:1 26:1 27:1 28:1 29:1

-1 9:1 11:1 25:1 30:1 31:1 32:1 33:1 34:1 35:1

Facebook\_data\_fake\_ngram.libsvm: 1-gram là 1, 2-gram là 2….

+1 1:1 2:1 3:2 4:1 5:1 6:1 7:1 8:1 9:1 10:1 11:1 12:2 13:1

+1 5:1 9:1 11:1 14:1 15:1 16:1 17:2 18:1 19:1 20:2 21:2

-1 4:1 11:1 22:1 23:1 24:1 25:1 26:1 27:1 28:2 29:1

-1 9:1 11:1 25:1 30:1 31:2 32:2 33:1 34:1 35:1