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 3: Sau đó em tạo 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 và đây cũng chính là 1-gram

+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\_2gram.libsvm**: Đây là file 2-gram

Danh sách các từ khác nhau là 42 trong tổng số 43 cụ từ 2gram

|  |  |  |
| --- | --- | --- |
| 1 | một ngày | 1 |
| 2 | ngày nào\_đó | 1 |
| 3 | nào\_đó với | 1 |
| 4 | với anh | 1 |
| 5 | anh không | 1 |
| 6 | không còn | 1 |
| 7 | còn buồn | 1 |
| 8 | buồn nữa | 1 |
| 9 | nữa đâu | 1 |
| 10 | đâu em | 1 |
| 11 | em yên\_tâm | 1 |
| 12 | yên\_tâm nhé | 1 |
| 13 | nhìn em | 1 |
| 14 | em mà | 1 |
| 15 | mà lòng | 1 |
| 16 | lòng anh | 1 |
| 17 | anh đau\_nhói | 1 |
| 18 | đau\_nhói chẳng | 1 |
| 19 | chẳng biết | 1 |
| 20 | biết nói\_gì | 1 |
| 21 | nói\_gì nữa | 1 |
| 22 | nữa xin\_lỗi | 1 |
| 23 | xin\_lỗi em | 1 |
| 24 | về nhà | 2 |
| 25 | nhà là | 1 |
| 26 | là về | 1 |
| 27 | nhà về | 1 |
| 28 | về với | 1 |
| 29 | với chồng | 1 |
| 30 | chồng yêu | 1 |
| 31 | yêu và | 1 |
| 32 | và tình\_yêu | 1 |
| 33 | tình\_yêu của | 1 |
| 34 | của em | 1 |
| 35 | ai yêu\_thương | 1 |
| 36 | yêu\_thương em | 1 |
| 37 | em nữa | 1 |
| 38 | nữa tự\_nhiên | 1 |
| 39 | tự\_nhiên muốn | 1 |
| 40 | muốn lấy | 1 |
| 41 | lấy chồng | 1 |
| 42 | chồng rồi | 1 |

+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

+1 13:1 14:1 15:1 16:1 17:1 18:1 19:1 20:1 21:1 22:1 23:1

-1 24:2 25:1 26:1 27:1 28:1 29:1 30:1 31:1 32:1 33:1 34:1

-1 35:1 36:1 37:1 38:1 39:1 40:1 41:1 42:1

**Facebook\_data\_fake\_3gram.libsvm**: Đây là file 3-gram

Danh sách các từ khác nhau là 42 trong tổng số 39 cụ từ 3gram

|  |  |  |
| --- | --- | --- |
| 1 | một ngày nào\_đó | 1 |
| 2 | ngày nào\_đó với | 1 |
| 3 | nào\_đó với anh | 1 |
| 4 | với anh không | 1 |
| 5 | anh không còn | 1 |
| 6 | không còn buồn | 1 |
| 7 | còn buồn nữa | 1 |
| 8 | buồn nữa đâu | 1 |
| 9 | nữa đâu em | 1 |
| 10 | đâu em yên\_tâm | 1 |
| 11 | em yên\_tâm nhé | 1 |
| 12 | nhìn em mà | 1 |
| 13 | em mà lòng | 1 |
| 14 | mà lòng anh | 1 |
| 15 | lòng anh đau\_nhói | 1 |
| 16 | anh đau\_nhói chẳng | 1 |
| 17 | đau\_nhói chẳng biết | 1 |
| 18 | chẳng biết nói\_gì | 1 |
| 19 | biết nói\_gì nữa | 1 |
| 20 | nói\_gì nữa xin\_lỗi | 1 |
| 21 | nữa xin\_lỗi em | 1 |
| 22 | về nhà là | 1 |
| 23 | nhà là về | 1 |
| 24 | là về nhà | 1 |
| 25 | về nhà về | 1 |
| 26 | nhà về với | 1 |
| 27 | về với chồng | 1 |
| 28 | với chồng yêu | 1 |
| 29 | chồng yêu và | 1 |
| 30 | yêu và tình\_yêu | 1 |
| 31 | và tình\_yêu của | 1 |
| 32 | tình\_yêu của em | 1 |
| 33 | ai yêu\_thương em | 1 |
| 34 | yêu\_thương em nữa | 1 |
| 35 | em nữa tự\_nhiên | 1 |
| 36 | nữa tự\_nhiên muốn | 1 |
| 37 | tự\_nhiên muốn lấy | 1 |
| 38 | muốn lấy chồng | 1 |
| 39 | lấy chồng rồi | 1 |

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

+1 12:1 13:1 14:1 15:1 16:1 17:1 18:1 19:1 20:1 21:1

-1 22:1 23:1 24:1 25:1 26:1 27:1 28:1 29:1 30:1 31:1 32:1

-1 33:1 34:1 35:1 36:1 37:1 38:1 39:1