## **Part 2 Original data Analysis -Part 2**

## **1 Task Definition**

1 Find top 300 negative words in the negative comments in decreasing order

2 Find top 300 positive words in the positive comments in decreasing order

3 Choose 5-10 from top negative words, find their top 100 co-occurrence words in negative comments.

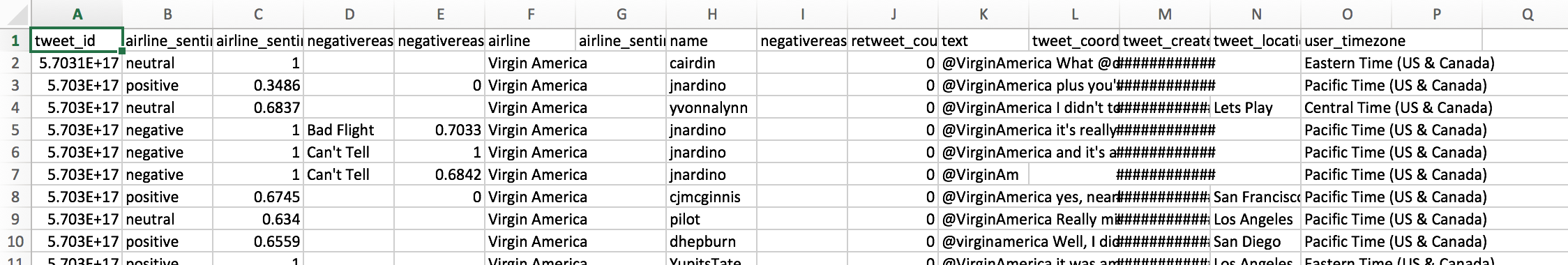
4 Choose 5-10 from top positive words, find their top 100 co-occurrence words in the positive words

## **2 Tools and Input format**

Tool: Mapreduce in Hadoop

Input: a csv file with detailed information about tweets

Screen cut:



## **3 Algorithm details about task 1-4**

Task 1 and 2 key point:

1) remove the stop words, accelerate the process by using hash set, which can determine weather the word is stop word in O(1) time

2) in the map part, analysis each line of tweets, in Task 1, we only analysis and emit the negative comments, correspondingly, in Task 2, we only analysis and emit the positive part.

3) in the reduce process, define a sorter in the combine part which can sort result by our defined method, specially value here.

Task 2 and 3 key point:

1) We define our own Writable Map which can carry the co-occurance words and their number

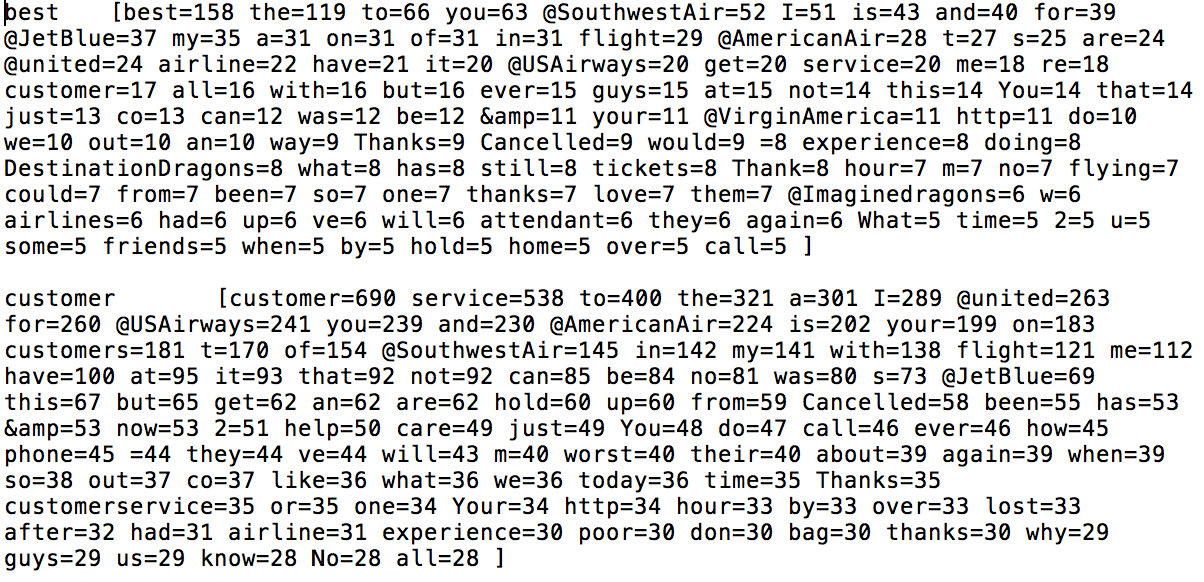
2) We only pick up 5-15 top words from either positive or negative top words which make sense.

3) Because we need to print out the co-occurrence in the decreasing order about top 100 words, the key point in this project is the toString method() of our self defined Writable map, here we not only define the format to list the word, but also limit the number to 100, more importantly, we convert to writable map to a arraylist, in order to sort it by self defined sorter, then print out the sorted arralylist before index 100

# ***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

# **4 Output**

raw output screen



We process the raw out put to tables:

Negative:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Top 1 | | Top 2 | | Top 3 | | Top 4 | | Top 5 | |
| customer | service | 538 | United | 289 | For | 260 | USAirways | 241 | AmericanAir | 224 |
| delayed | Flight | 258 | United | 163 | My | 135 | USAirways | 123 | JetBlue | 75 |
| flight | United | 886 | Cancelled | 747 | USAirways | 745 | AmericanAir | 690 | SouthwestAir | 589 |
| help | Air | 230 | USAirway | 214 | United | 207 | Flight | 202 | Helpful | 202 |
| hold | Airways | 275 | Flight | 170 | Been | 179 | Cancelled | 120 | Hour | 109 |
| service | Customer | 538 | United | 307 | USAirways | 259 | You | 229 | Flight | 156 |
| time | United | 259 | Flight | 232 | USAirway | 218 | My | 217 | AmericanAir | 168 |

Positive

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Top 1 | | Top 2 | | Top 3 | | Top 4 | | Top 5 | |
| best | Southwestair | 52 | Jetblue | 37 | My | 37 | Flight | 29 | Airline | 28 |
| customer | Service | 400 | United | 263 | For | 260 | You | 239 | Your | 199 |
| great | Flight | 61 | Jetblue | 55 | Southwestair | 52 | Airways | 47 | United | 46 |
| love | Southwest | 83 | Jetblue | 56 | Would | 38 | Me | 35 | Airline | 35 |
| thank | Thanks | 553 | United | 213 | Airline | 216 | JetBlue | 153 | USAirways | 141 |
| thanks | United | 134 | Air | 117 | JetBlue | 114 | Service | 34 | Time | 26 |
| service | Customer | 538 | United | 307 | Usairways | 259 | You | 229 | Americanairline | 225 |

# **5 Analysis and Conclusion**

1) We find the jetblue and Southwest air has happened to appear quite frequently with positive top words, specifically for southwestair, it has come to top 5 co-occurrence with best, great, for 52, 52 time. Jetblue is even better, is has come to the top 5 of best, great, love for 27, 55, 46 times. So we know jetblue and southwest aire are quite popular with good words, we have a sense of positive feeling of it.

2) if I am a customer, I would like to choose the southwest and jetblue at the same price and equally time schedule. If I am the worker of air line company, maybe I need to think about what is the difference between us and southwest air company, do we need to consider jetblue?

3) I suggest do not draw any conclusion further than that, just based independently on this part, this part is serviced at a clue or hint, but not strong evidence, we have machine learning part in the following.