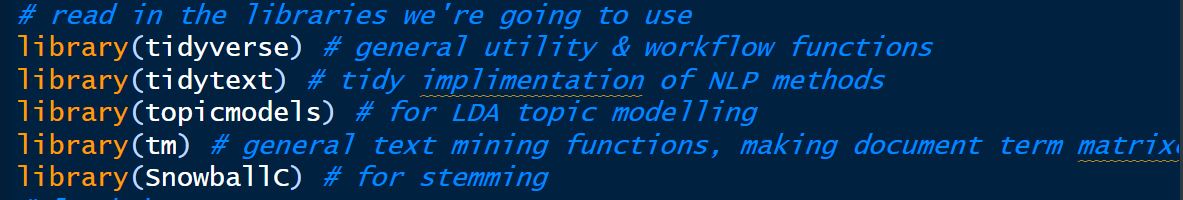
**Topic Modeling for Twitter Dataset**

**Topic modeling**: The NLP task of identifying automatically identifying major themes in a text, usually by identifying informative words.

Before we can get started on topic modelling, we need to get our environment set up. First, let's load in the packages we're going to use. Here, I've added a comment letting you know why we need each package.



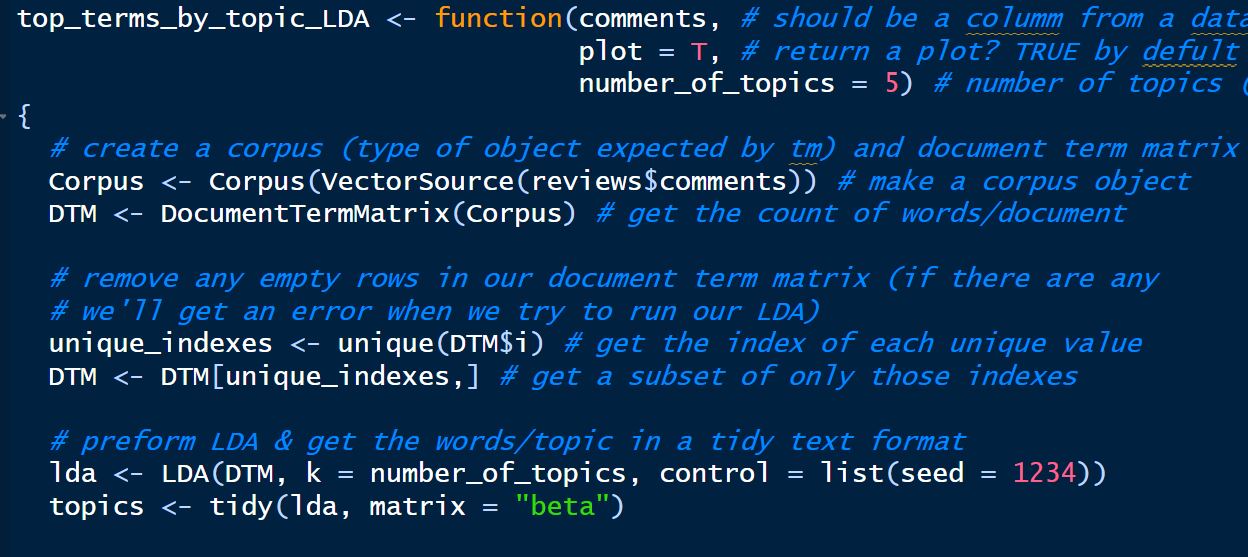
And now load the dataset of twitterdata.csv



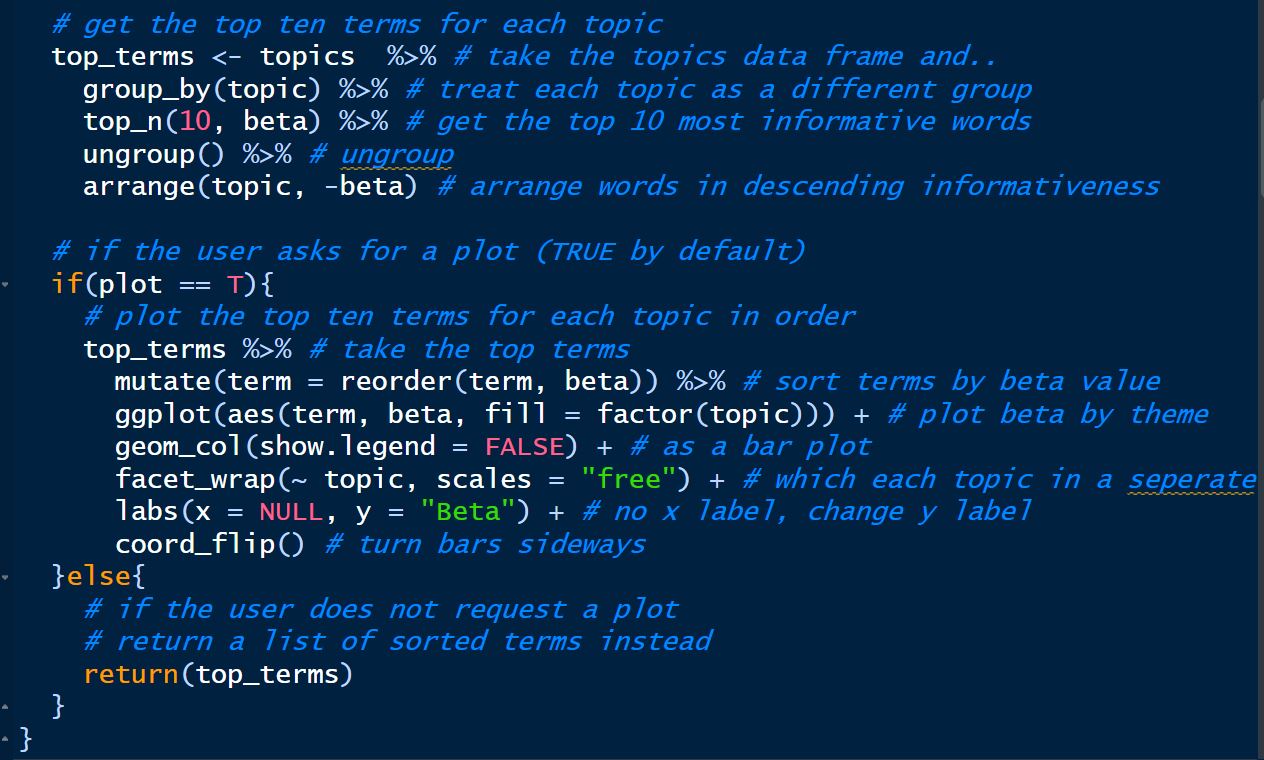
Now , using the LDA (Latent Dirichlet Allocation) Model, we can make a function to get and plot the most informative term by a specified number of topics.

**Unsupervised topic modeling with LDA** :

So , the unsupervised topic modeling using LDA will be:



Setting up the corpus , we can store all the user comments that have been on the dataset.

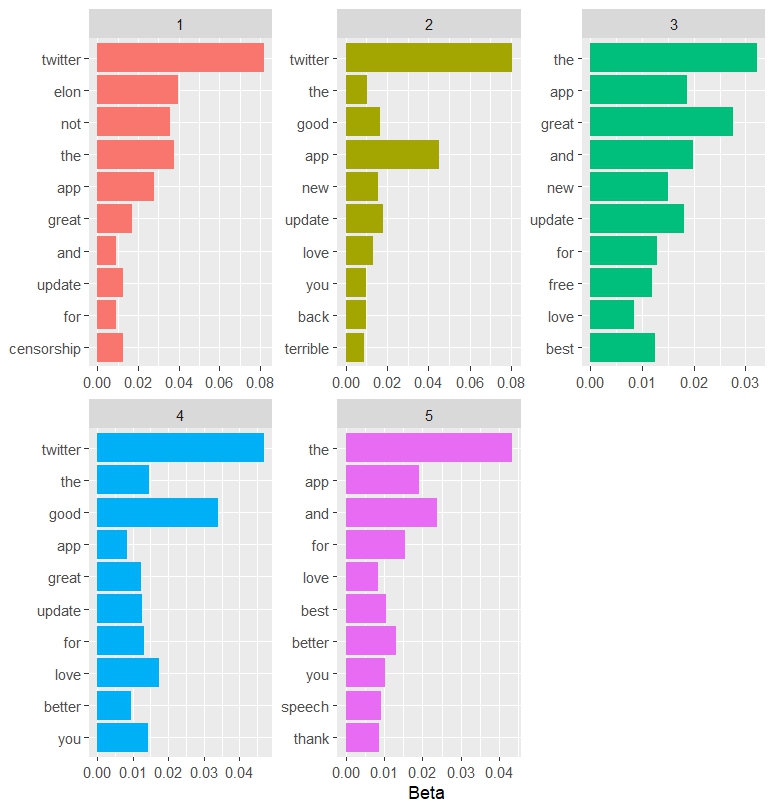


And finally building up the unsupervised LDA model to plot the specified word for the twitter data set.

Now after building the model, we plot the most 10 informative words for the topic modeling.

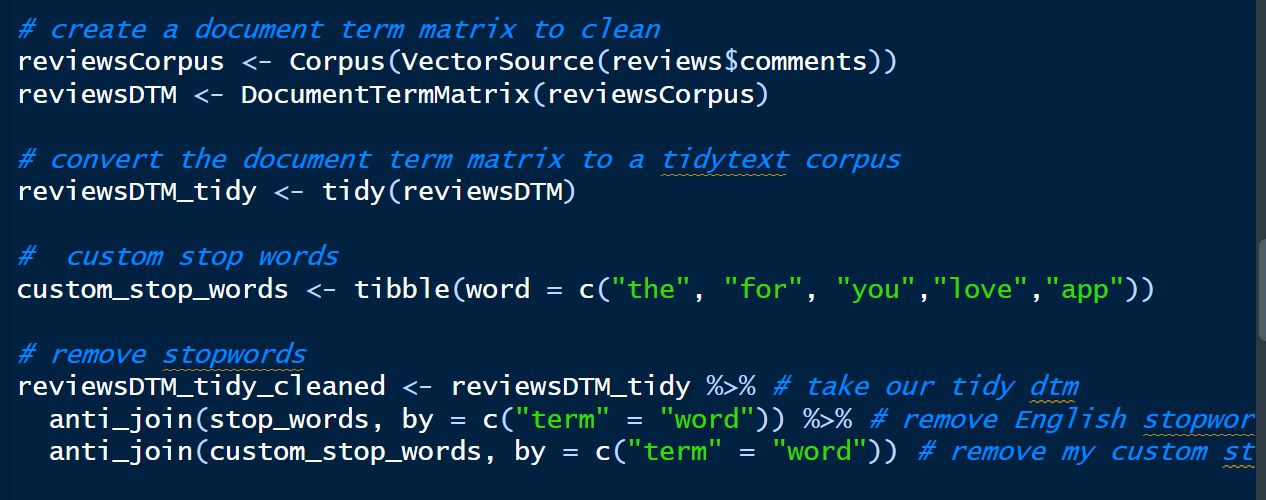


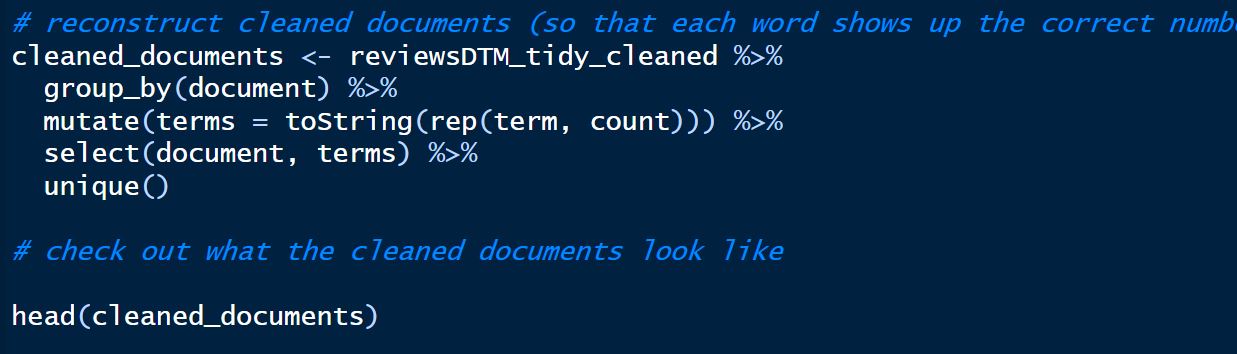
And the output will be :



From the unsupervised LDA model it predicts some user comment that can be aliened with the user correlation between security and privacy updates and user satisfaction.

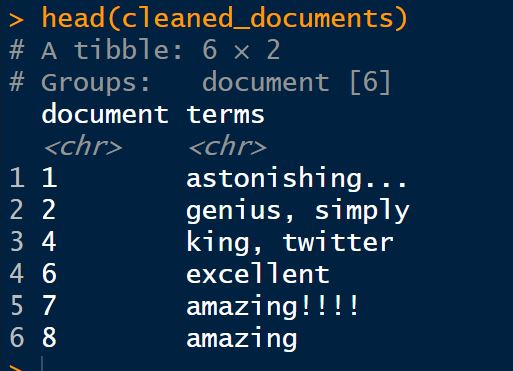
So, we can see that many unnecessary words (‘for’, ‘you’, ‘and’) are frequently occurred. So, we need to clean our corpus data for get better results.

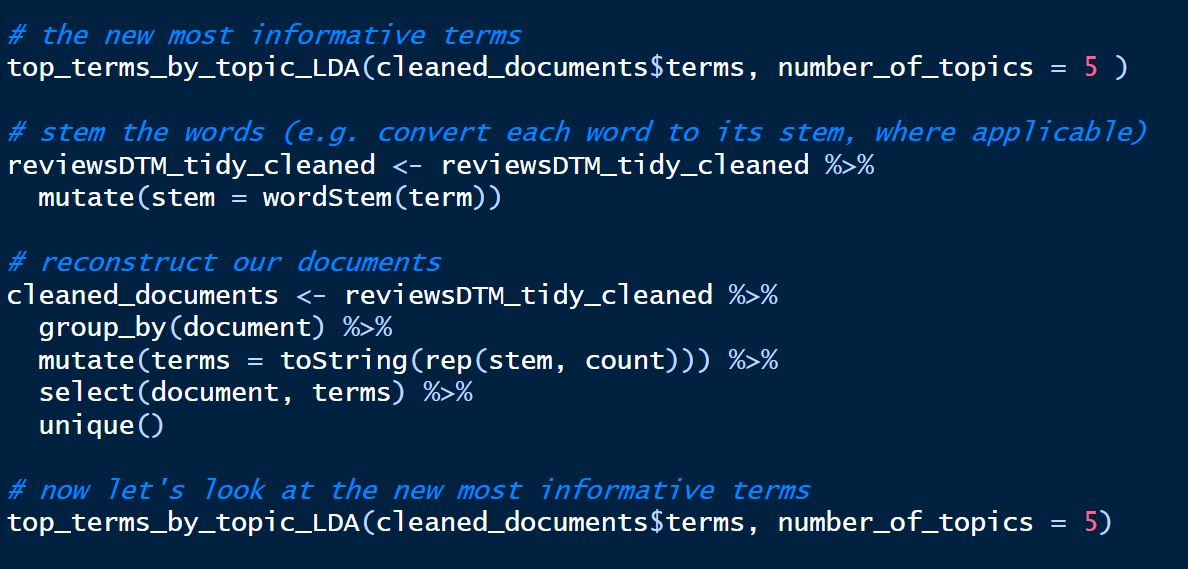




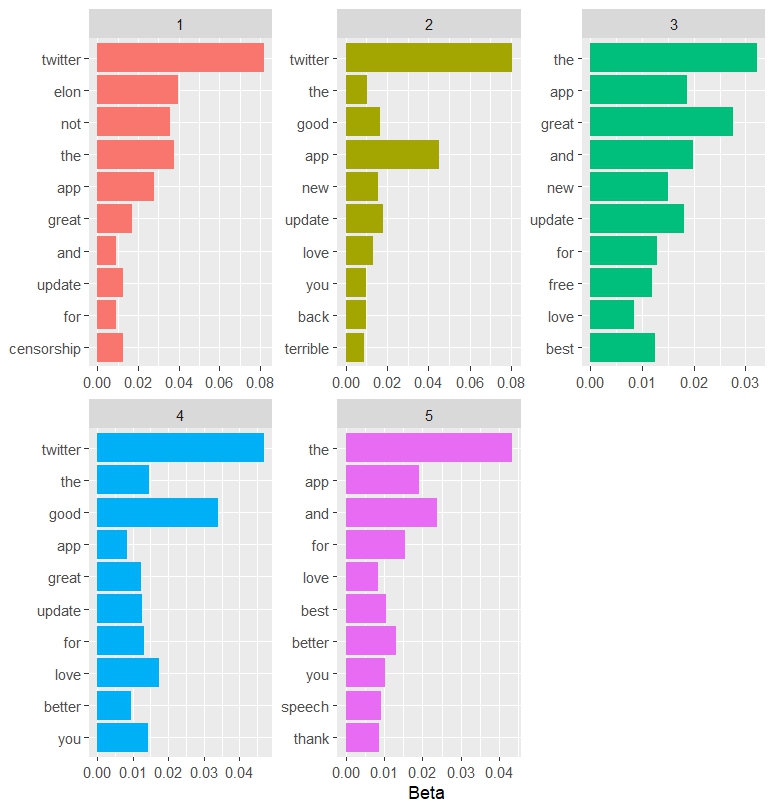
So, we create a document term matrix (dtm) to clean up all the unnecessary words which not relevant to the analysis. And also create a cleaned documents that can give us the topic for modeling.

Some out put of cleaned corpus is:





After cleaning the corpus, we have created a new informative term of the dataset.



From the new output we can see that it still generated some unnecessary words that might not be helpful for our prediction of user satisfaction and security and privacy update.

So, for better understanding we can used Supervised topic modeling with TF-IDF.

**Supervised topic modeling with TF-IDF :**

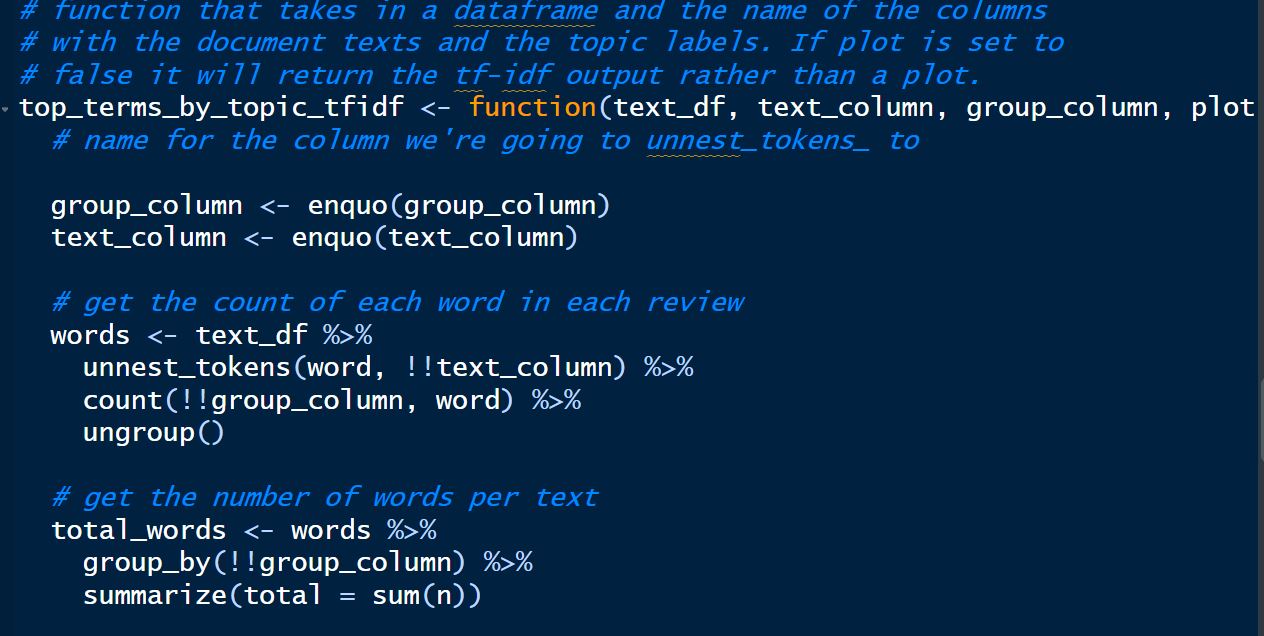
Now that we've given unsupervised topic modelling a go, let's try supervised topic modeling. For this, we're going to use something call TF-IDF, which stands for "term frequency-inverse document frequency".

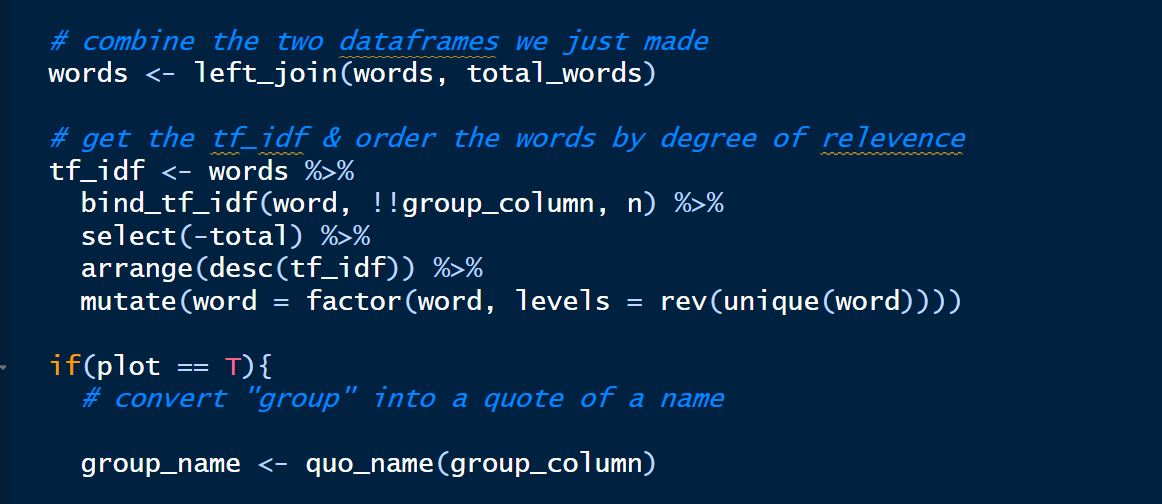
The general idea behind how TF-IDF works is this:

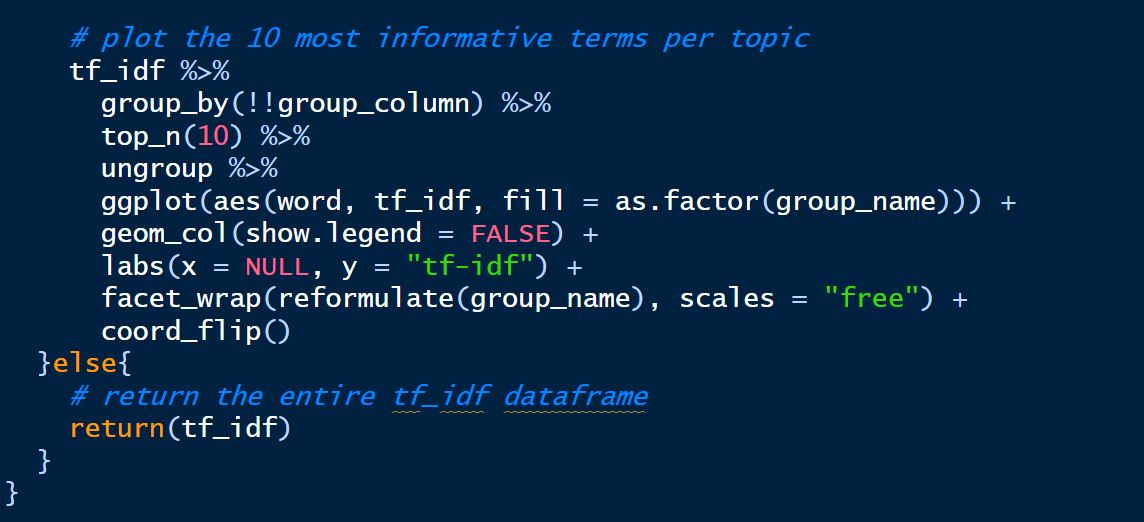
* Words that are very common in a specific document are probably important to the topic of that document
* Words that are very common in all documents probably aren't important to the topics of any of them

So, a term will receive a high weight if it's common in a specific document and also uncommon across all documents.

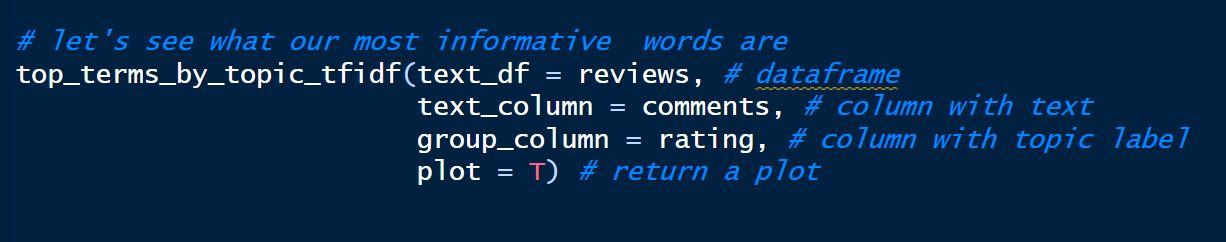
In order to streamline our analysis, we are writing a function that takes in a data frame, the name of the column that has the texts in it and the name of the column that has the topic labels in it.



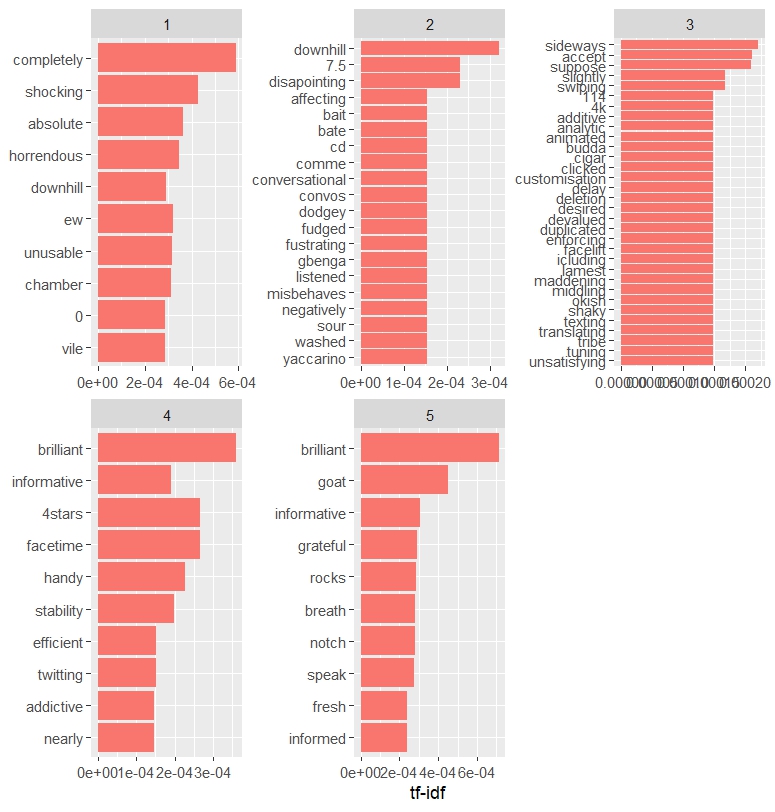




So, we have built the function of supervised model. Now let’s check the most informative words that generated with TF-IDF

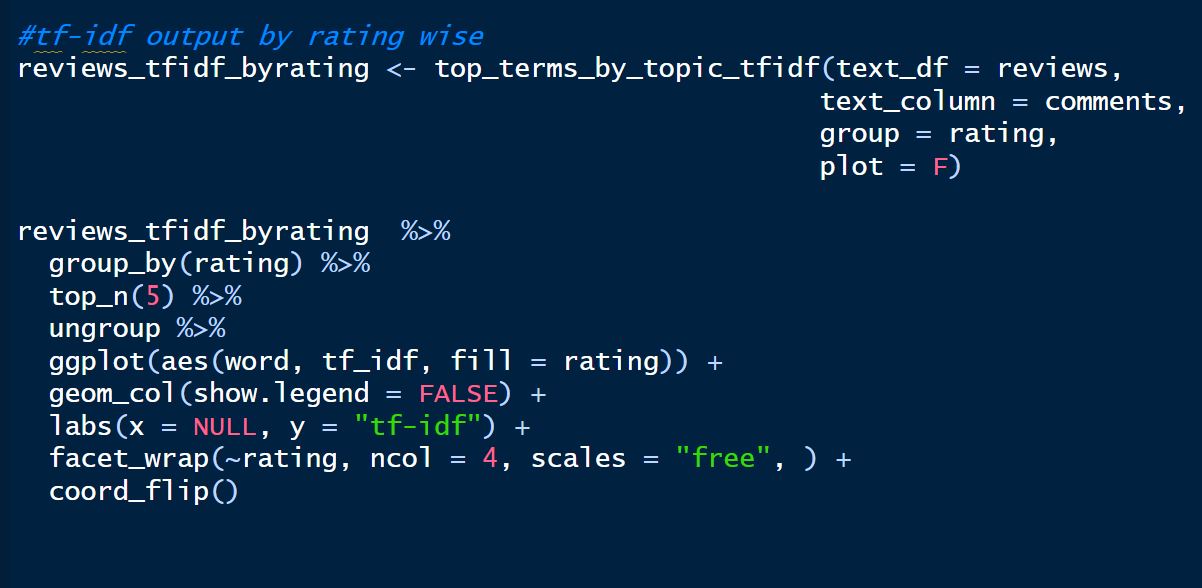


The output result of the supervised TF-IDF model is:



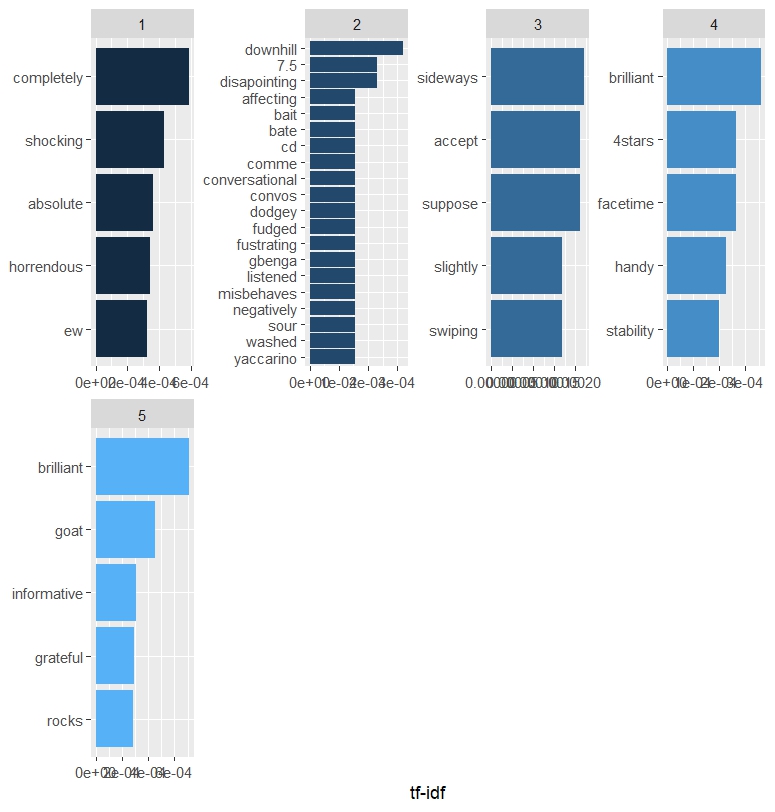
From the out we can see that we have unique words that can determine the user behavior pattern.

From this, we can see that negative reviews include words like "unusable", "disappointing", "unsatisfying" and "delay", while positive reviews also come up (both "brilliant" and "informative" show up in the top ten words).



let's check out the comments for each rating. We are going to use the function above to return just the TF-IDF output and then plotting it by rating.

And the final output of using the reviews by ratings is:



From this, we can see that there are fewer negative reviews words while positive reviews words are showing positive user experience(both "brilliant" ,"informative" and ‘stability’ show up in the top five words).

**Conclusion:**

From overall analysis we can conclude that , the twitter datasets provide us some insides about user satisfaction level but not corelated with the user security and privacy updates.