**Clustering**

Process

* Construct a document matrix for all cleaned data by using TF-IDF
* Try clustering with mixture components from 2 to 5 by using Gaussian Mixture Models
* Characterize the clusters with some words likely happen in disaster/non-disaster

Pick “wildfire”, “earthquake”, “thunderstorm”, and “crash” as disaster-relevant words

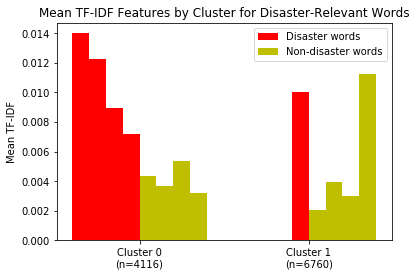
Pick “love”, “Obama”, “summer”, “movie” as non-disaster-relevant words

* Visualize the mean TF-IDF graph

Result:

* Mixture component = 2

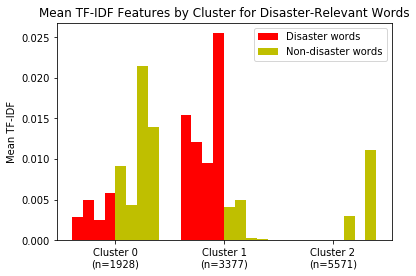
The number of documents assigned to each cluster is 4116 and 6760.



According to the plot above, cluster 0 seems to correspond to disaster words while cluster 1 corresponds to non-disaster words.

* Mixture component = 3

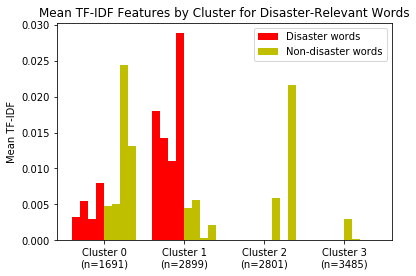
The number of documents assigned to each cluster is 1928, 3377 and 5571.



According to the plot above, cluster 1 seems to correspond to disaster words while cluster 0 and 2 corresponds to non-disaster words.

* Mixture component = 4

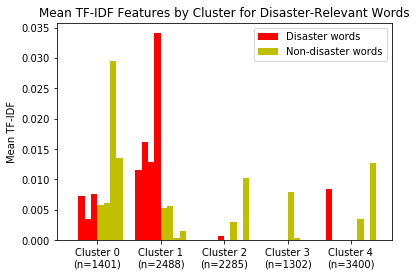
The number of documents assigned to each cluster is 1691, 2899, 2801 and 3405.



According to the plot above, cluster 1 seems to correspond to disaster words while cluster 0, 2 and 3 corresponds to non-disaster words.

* Mixture component = 5

The number of documents assigned to each cluster is 1401, 2488, 2285, 1302 and 3400.



According to the plot above, cluster 1 seems to correspond to disaster words while cluster 0, 2, 3 and 4 corresponds to non-disaster words.

Comparing the four plots, it seems that the best choice is n\_clusters = 3, which works well in many cases.