

# Problem Description

Search the most recent tweets containing a specific set of words and analyze what other words appear frequently. This gives us other words, ideas, issues etc., related to the given set of words, in the context of social media discussions.

- R scripts are used to extract the most recent 3000 tweets containing the words "america" and "france".
- After cleaning up the text containing the tweets using various functions (in packages written for R), the word cloud is produced.
- The word cloud prints the related words to the given set, in a font-size that is proportional to their frequencies.
- The example set of words chosen for illustration purposes are {america, france} . Some of the most frequently used words along with the given set are {syrians, war, flee, smugglers, etc. }

The following is an idea that has not been implemented in the code:

- Given a context and a set of keywords, by analyzing the frequencies of other words appearing along with the given set, one could form clusters of words with similar frequencies and then put a crude metric that measures, more or less, how closely these clusters are related to the given set. For example, in the context of current events and social media and  $\{\text{america, france}\}$  as the the given set  $S$ , the cluster containing the word syrians is closer to  $S$  than the cluster containing tegucigalpa.
- This process could potentially be refined so that the metric that relates the clusters to the given set, can be used in the context of talent acquisition and the given set being keywords in a particular discipline, such as, C++, Unix, Ruby etc.