Strategy used:

To recap, the database in csv format had the following structure:

000001,3,Fears for T N pension after talks,Unions representing workers at Turner Newall say they are 'disappointed' after talks with stricken parent firm Federal Mogul.

The items separated by commas are as follows: id, subject (1, 2, 3 or 4), title, news
To facilitate the manipulation of each item in this database, the concept of regular expression (Regex) was used in order to calculate the TFIDF of each word in the file containing 7600 news items.

First MapReduce:

In this initial stage, special characters were cleaned and the database was completely standardized to lowercase. The key was defined as the combination of the news id and the word, which are separated by underscore. And the value was set to 1 to be added in the next step, Reduce. The output from Map to Reduce was:

000002_the <1, 1>

Upon reaching this key-value in Reduce, he simply added the values to obtain the value of the number of times the word in question appeared in a given document, this being the step to discover the "*tfd*" variable of the formula for the calculation of the TFIDF. Finally, the output from MapReduce was:

000002_the 2

Second MapReduce:

This step aims to calculate the number of documents in which a word appeared, using the previous MapReduce output as a database. A new regular expression was defined to facilitate the use and manipulation of data, in this step the key was defined as the word and its values as the ids concatenated with the number of times the word appeared in these news. The map output was as follows:

distinguished <001437_1, 001357_1, 001407_1>

Upon receiving this information, two repetition loops were used: the first to count how many values there were in the values received from the key-value, thus being possible to define the number of documents in which a word appeared, in addition these values were copied to a list with the purpose of being used in the next loop, because after iterating once over the values of the key-value, it would be emptied, making its use unfeasible. In the second loop, the list of copied values was used to define the key as an id concatenated with the word and variable obtained in the first MapReduce and the value was defined as the count of news in which a certain word appeared, this being the last variable "*dft*" that was left to be used in the calculation of the TFIDF. The output of this MapReduce was:

001437_distinguished_1 3
001357_distinguished_1 3
001407_distinguished_1 3

Last Map:

In this last phase, the concept of regular expression was also used to facilitate the extraction of variables, to calculate the TFIDF of each word, from the database which would be output from the previous MapReduce. The key was defined as the concatenation of the news id with the word and the value as TFIDF, thus, finally, the output of this last process:

001357_distinguished 3.278753600952829
001407_distinguished 3.278753600952829
001437_distinguished 3.278753600952829