TFIDF representation

By Alison de Almeida Sales

• Project definition:

Much of the information produced by humanity occurs digitally and a portion considerable is in textual format. Consequently, there is a need to extract information from digital texts in an automatic way allowing the performance of various analyses of the data.

Many Machine Learning techniques are capable of learning information from texts and provide the most different analysis of the set of documents. However, the vast majority of these techniques are not able to handle texts in their raw format, that is, the way what we human beings read. In this way, we need to transform the documents into a numerical structure that is convenient for using Learning Algorithms Machine.

In this project, you will have to pre-process a text base using the Hadoop framework and the Map Reduce approach to calculate the TFIDF of each word.

TFIDF is a weight (value) that we assign to associate a word with a document. form to represent its importance. Therefore, we can calculate the TFIDF of a term (word) t associated with a document d through equation 1.

$$(1) \quad TFIDF_{t,d} = tf_d \times idf_t$$

tfd(term frequency) = number of times the term t appears in document d

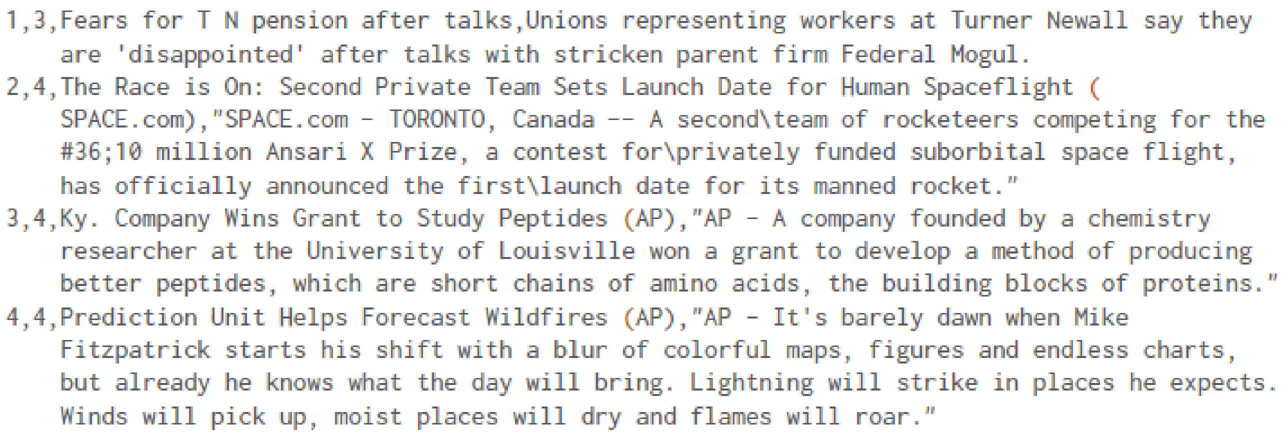
$$(2) \quad idf_t \text{ (inverse document frequency)} = \log_{10}\left(\frac{N}{(1 + df_t)}\right)$$

dft = number of documents in which the term t appears
N = total number of documents

- **Text base:**

You will have to work with a base of news texts. The base consists of a single file where each line contains information from a document different.

Example - excerpt from the text base:



```
1,3,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.
2,4,The Race is On: Second Private Team Sets Launch Date for Human Spaceflight (
    SPACE.com),"SPACE.com - TORONTO, Canada -- A second\team of rocketeers competing for the
    #36;10 million Ansari X Prize, a contest for\privately funded suborbital space flight,
    has officially announced the first\launch date for its manned rocket."
3,4,Ky. Company Wins Grant to Study Peptides (AP),"AP - A company founded by a chemistry
    researcher at the University of Louisville won a grant to develop a method of producing
    better peptides, which are short chains of amino acids, the building blocks of proteins."
4,4,Prediction Unit Helps Forecast Wildfires (AP),"AP - It's barely dawn when Mike
    Fitzpatrick starts his shift with a blur of colorful maps, figures and endless charts,
    but already he knows what the day will bring. Lightning will strike in places he expects.
    Winds will pick up, moist places will dry and flames will roar."
```

The information for each line is separated by a comma, so for each document we have four pieces of information (columns):

1. the first column contains the document id;
2. the second column has the subject of the document represented by a numeric value;
3. the third column has the title of the document;
4. and the last column shows the content of the document.

In all, the text base has 7600 documents organized into 4 different subjects. Not you will need to use the subject to calculate the TFIDF metric, however, this information may be relevant for further data analysis.