Lab 2

We need to extract *tfidf* weights given the collection, and produce in output a file formatted like this:

word1 doc_id weight
word2 doc_id weight
word3 doc_id weight
...

We already have:
1. stemmed words frequency
2. a stoplist
3. the keywords for every document

Stopwords have already been removed from the keywords so we can ignore *stopping*.
Since keywords are already stemmed we can ignore *stemming* too.

The script *weights.py* builds a dictionary with the unique words, then builds a matrix of n_docs x n_words with occurencies using the frequency file:

| doc_id | w1 | w2 ...
+---------+---+------
| doc1 | 5 | 0 ...
| doc2 | 1 | 2 ...
| doc3 | 0 | 3 ...
+---------+---+------

with this matrix we use the *TfidfTransformer* of the scikit-learn module to get a matrix with the tfidf weights. Finally we save them to the file.