Feature extraction for text data

Nicolas Vo

Table of contents

1. CORPUS

1.1. Presentation

AG is a collection of more than 1 million news articles gathered from more than 2000 sources and is provided by the academic community for research purposes. The AG's news topic classification dataset is constructed by choosing 4 largest classes from the original corpus. Each class contains 30,000 training samples and 1,900 testing samples. The total number of training samples is 120,000 and testing 7,600. It is used as a text classification benchmark for many papers.

Source: https://github.com/mhjabreel/CharCNN/tree/master/data/ag_news_csv

The **Brown Corpus** was compiled as a general corpus (text collection) in the field of corpus linguistics. It contains 500 samples of English-language text, totaling roughly one million words, compiled from works published in the United States in **1961**.

Source: https://en.wikipedia.org/wiki/Brown_Corpus

The **20 Newsgroups** data set is a collection of approximately 18,846 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. The data is organized into 20 different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hardware / comp.sys.mac.hardware), while others are highly unrelated (e.g misc.forsale / soc.religion.christian). The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.

Source: http://qwone.com/~jason/20Newsgroups/

1.2. Descriptive analysis

AG is perfectly distributed with 30,000 samples in each category. On the other hand, the **Brown Corpus** isn't with *learned* and *belles_lettres* accounting for almost a 30 percent of all entries whereas *reviews* and *religion* have less than 20 entries and *humor* and *science_fiction* less than 10. **20 Newsgroup** is overall well distributed.

	category	text	percentage
0	learned	80	16.0
1	belles_lettres	75	15.0
2	lore	48	9.6
3	news	44	8.8
4	hobbies	36	7.2
5	government	30	6.0
6	adventure	29	5.8
7	fiction	29	5.8
8	romance	29	5.8
9	editorial	27	5.4
10	mystery	24	4.8
11	religion	17	3.4
12	reviews	17	3.4
13	humor	9	1.8
14	science_fiction	6	1.2

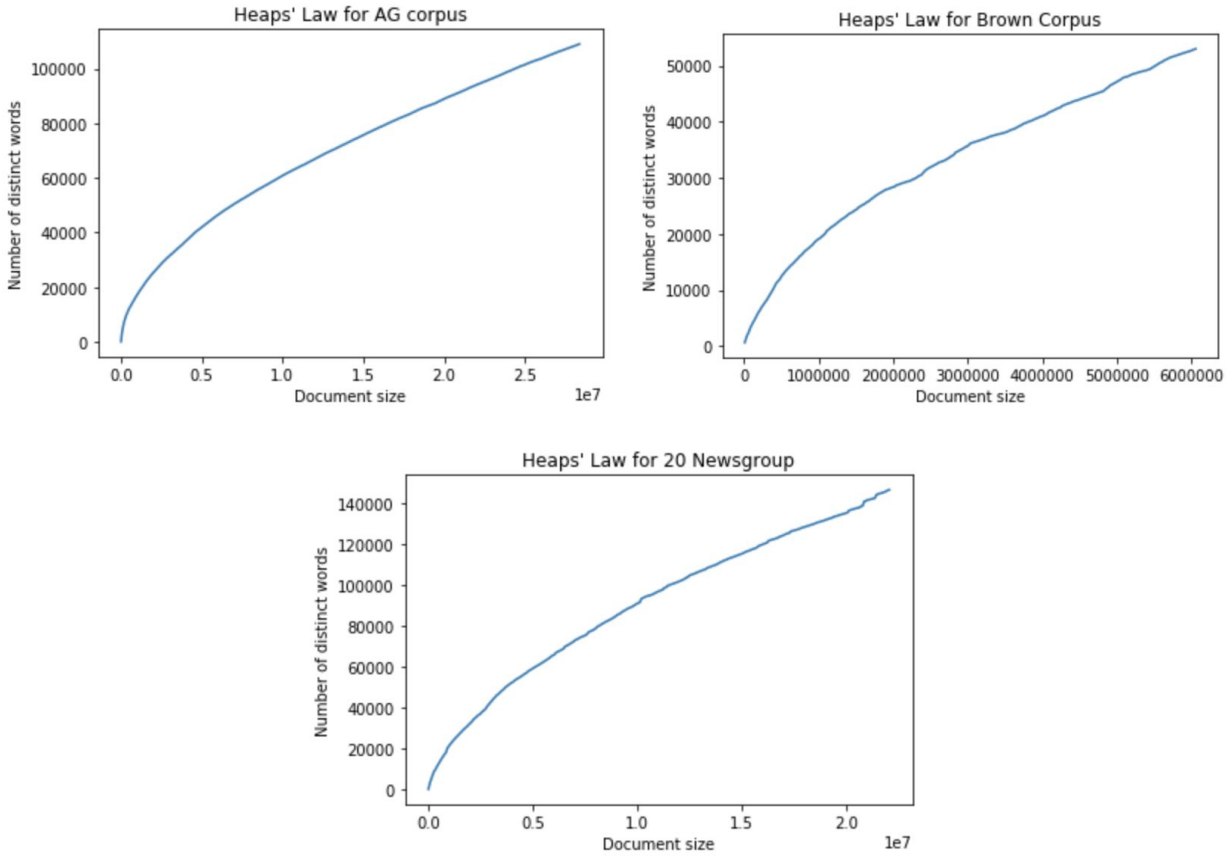
	category	text	percentage
0	rec.sport.hockey	600	5.303164
1	soc.religion.christian	599	5.294326
2	rec.motorcycles	598	5.285487
3	rec.sport.baseball	597	5.276648
4	sci.crypt	595	5.258971
5	rec.autos	594	5.250133
6	sci.med	594	5.250133
7	comp.windows.x	593	5.241294
8	sci.space	593	5.241294
9	sci.electronics	591	5.223617
10	comp.os.ms-windows.misc	591	5.223617
11	comp.sys.ibm.pc.hardware	590	5.214778
12	misc.forsale	585	5.170585
13	comp.graphics	584	5.161747
14	comp.sys.mac.hardware	578	5.108715
15	talk.politics.mideast	564	4.984974
16	talk.politics.guns	546	4.825879
17	alt.atheism	480	4.242531
18	talk.politics.misc	465	4.109952
19	talk.religion.misc	377	3.332155

On the left, the distribution in the Brown Corpus,
on the right, the distribution of the 20 Newsgroup data set.

The three corpuses have different mean text lengths: it is around 236 for AG, 12,110 for the Brown Corpus and 1949 for 20 Newsgroup. Note that the title and content were concatenated in the AG data set. Therefore, the data is heterogeneous and going from one corpus to another requires manipulations, decisions and critical thinking.

1.3. Heaps' and Zipf' laws

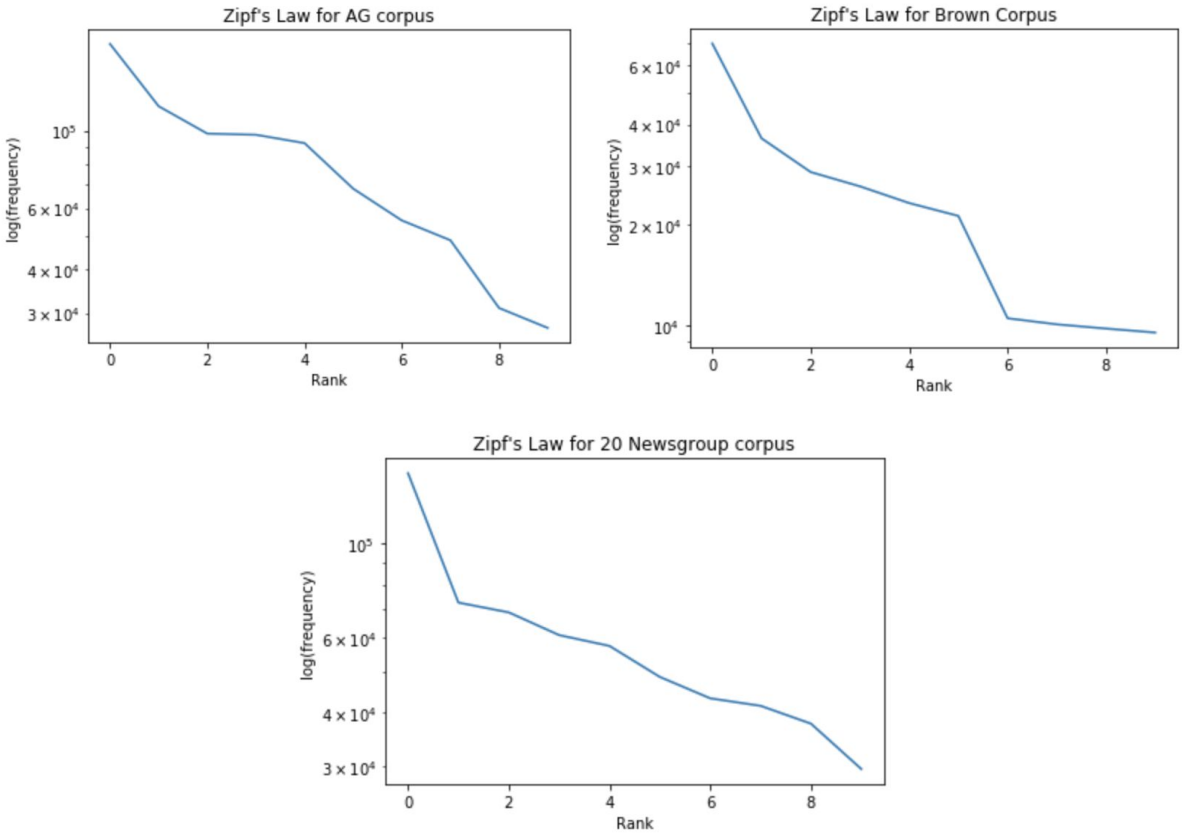
These are two empirical laws often used in the field of linguistics. Heaps' law plots the number of distinct words in a vocabulary with regard to the text size and shows that they grow together, albeit at a reasonable rate, therefore large collections increase the difficulty in covering a vocabulary in its entirety. We plot Heaps' law for all three corpuses.



Heaps' law for the AG, Brown and 20 Newsgroup corpuses

Zipf's law states that given a large sample of words used, the frequency of any word is inversely proportional to its rank in the frequency table. So word number n has a frequency proportional to 1/n. Thus the most frequent word will occur about twice as often as the second most frequent word, three times as often as the third most frequent word, etc. We plot Zipf's law with a logarithmic scale for the word frequency and observe the predominance of first rank words over the rest. The first ten ranks are usually *the, to, of, a, and, in, is, i, that, it,* with some small variations in the ordering.

Source: https://simple.wikipedia.org/wiki/Zipf%27s_law



Zipf's law for the AG, Brown and Newsgroup corpuses

2. PREPROCESSING

Preprocessing involves traditional methods of lowercasing text, tokenization into word tokens, removing punctuation, stopwords. Advanced methods involve **stemming** done with both pystemmer and NLTK's SnowballStemmer and spaCy **lemmatization** was also considered, but found to be extremely slow and therefore, it is imported but left unused. We note that lemmatization would have helped with a syntactic understanding of the data, which is not the main goal of this study.

Further processing consists in vectorizing the document tokens into a **document-term matrix** which has document rows, vocabulary columns and word count values. Then, rare words (hapax) are filtered by setting a minimum of word occurrence (3 by default). Finally, **TF-IDF** transformation is applied in order to give words a weight which underlines their high frequency in a document and low frequency in the corpus vocabulary. Documents are deemed similar using pairwise metrics, i.e. angle between document vectors between 0 and 1.

The use of **n-grams** consists in looking at tokens as couples of two or sometimes three words. The NLTK implementation was used.

model	execution_time	error	score	matrix_shape
basic	99	12079	0.8922	26938
pystemmer	100	12272	0.8899	17955
SnowballStemmer	123	12272	0.8899	17956
tfidf	92	12609	0.8870	26938
pystemmer_tfidf	117	12732	0.8850	17955
SnowballStemmer_tfidf	131	12732	0.8850	17956
ngrams	103	14944	0.8400	83324

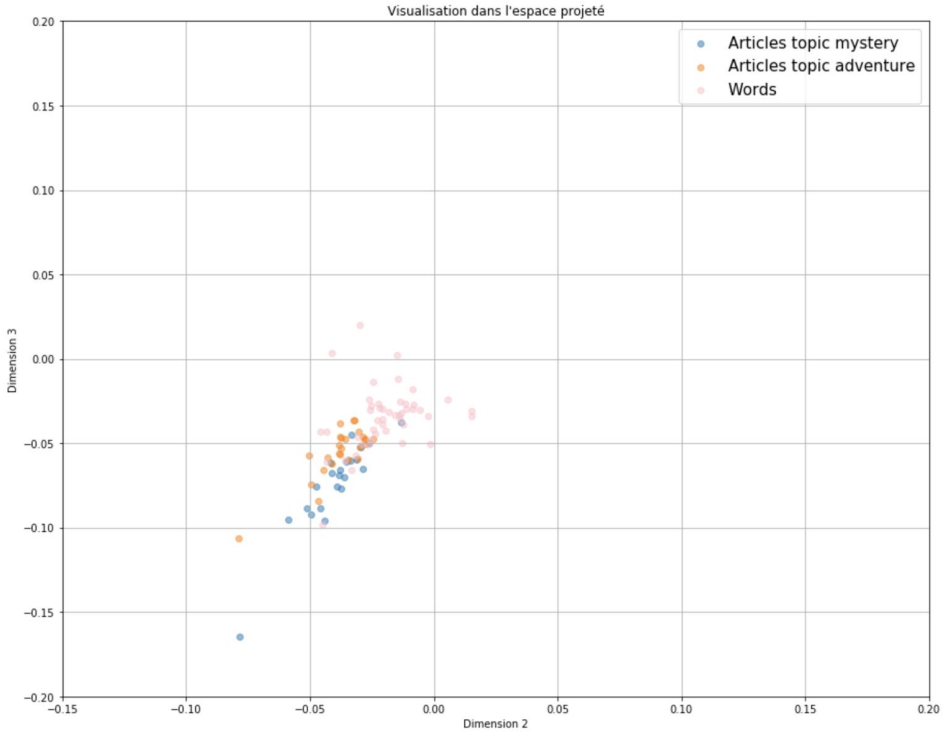
Table of preprocessing performances with execution time in seconds

All in all, the performances slightly degrade when introducing stemming and TF-IDF transformation to the processing. But one the one hand, stemming greatly reduces the corpus size and TF-IDF brings a lot with term weighting instead of simple occurrence counting. N-grams on the other hand increase both error rate and corpus size and are deemed non optimal. The pystemmer implement does outperform the NLTK Porter stemmer as expected.

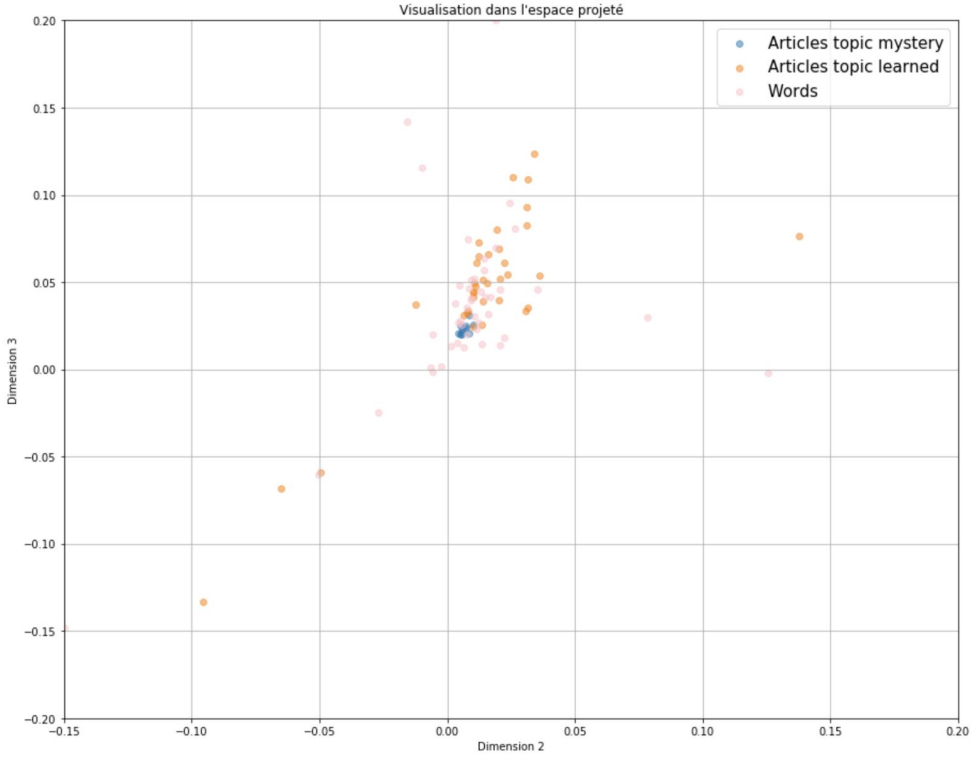
The whole project relies on the Naive Bayes classifier which is a probabilistic model which assumes all variables are independent. It is computationally cheap and has great performances in the field of Natural Language Processing.

3. APPLICATION TO THE BROWN CORPUS

The goal is to train a model on the AG data set and predict AG labels out of the Brown Corpus. The issues are the differences in labels, number of documents and document sizes. We deal with the labels by trying to match together those we think are suitable. We remark that the Brown Corpus was collected in the 1960's, therefore the technology in *sci/tech* was not as developed as it is today. We note a lack of *business* related documents and the AG corpus does not have labels similar to *religion* or *romance*, for example. The Latent Semantic Analysis helps finding hidden meanings behind words. It uses the Singular Value Decomposition in order to reduce the number of rows while preserving the similarity structure along columns, which mathematically means reducing a matrix of rank N into K linearly combined, orthogonal and unique vectors. We use it in order to find similar labels as seen below.

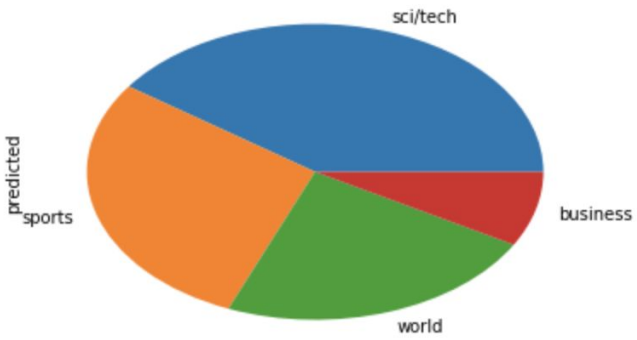


LSA of Brown Corpus for mystery and adventure labels



LSA of Brown Corpus for mystery and learned labels

Although close to each other, the *mystery* and *learned* labels do not overlap whereas *mystery* and *adventure* ones do. Therefore, we can consider attributing a same label to *mystery* and *adventure*. We also take a look at the predicted labels the AG-trained model outputs.



Pie chart of the predicted labels for the Brown Corpus

The results are skewed towards *sci/tech* and *sports* whereas less than 10 percent of documents are labeled as *business*: the *learned* documents are indeed academic texts containing scientific terms and there not many documents treating finances or economy.

The first pipeline is trained with Brown data for both vectorization and model: it classifies *business* and *sports* wrong, *world* correctly half of the time and *sci/tech* perfectly (see detailed results below). The second pipeline has the AG vocabulary but its classifier is fitted with Brown data: it only outputs *world* and does poorly at it with 47 percent of precision. The mediocre results can be explained by the small number of documents and highly heterogeneous distribution of the Brown Corpus, as well as the large text sizes which the AG-trained vectorizer is not covering at all. Therefore, the second pipeline has overall bad results whereas the first classifies using the Brown Corpus vocabulary for vectorization. On top of being small, the Brown Corpus also needed to be further split in train and test sets which did not help with the number of documents used for training.

```
Nombre d'erreurs : 62
Error percentage : 49.6
              precision    recall  f1-score   support

    business       0.00      0.00      0.00        12
    sci/tech       1.00      0.17      0.29        24
      sports       0.00      0.00      0.00        30
       world       0.49      1.00      0.66        59

   micro avg       0.50      0.50      0.50       125
   macro avg       0.37      0.29      0.24       125
weighted avg       0.42      0.50      0.36       125
```

Classification and evaluation of first pipeline

```
Nombre d'erreurs : 66
Error percentage : 52.800000000000004
              precision    recall  f1-score   support

    business       0.00      0.00      0.00        12
    sci/tech       0.00      0.00      0.00        24
      sports       0.00      0.00      0.00        30
       world       0.47      1.00      0.64        59

   micro avg       0.47      0.47      0.47       125
   macro avg       0.12      0.25      0.16       125
weighted avg       0.22      0.47      0.30       125
```

Classification and evaluation of second pipeline

4. APPLICATION TO THE 20 NEWSGROUP CORPUS

4.1. AG corpus applied to 20 Newsgroup corpus

The 20 Newsgroup labels are similar to both those of the AG and Brown corpuses. The documents categorized *computer* and *science* fit the AG *sci/tech* label well. The custom labels for NG result in the following distribution: 5300 *sci/tech*, 3000 *world*, 2000 sports and 600 business. It is unbalanced even though the original data set very evenly proportioned. The model transforms NG test data and predicts it based on the AG model and scores good with 34 percent for misclassification. The *world* and *sports* classes are especially well predicted.

```
Nombre d'erreurs : 2614
Error percentage : 34.7052575677111
              precision    recall  f1-score   support

    business       0.16      0.08      0.10       390
    sci/tech       0.61      0.96      0.75      3534
      sports       0.77      0.53      0.62      1590
       world       0.96      0.33      0.49      2018

   micro avg       0.65      0.65      0.65      7532
   macro avg       0.62      0.47      0.49      7532
weighted avg       0.71      0.65      0.62      7532
```

Classification and evaluation of AG on NG

4.2. Brown Corpus applied to 20 Newsgroup corpus

The Brown Corpus has labels which fit NG a little more. For instance, the NG documents related to *religion* can simply be labeled *religion*, *computer* and *science* can be labeled *learned*, *politics* can be *government*. The pipeline is trained with Brown and predicts the NG test set with Brown labels. At first, the score seems mediocre with 84 percent of errors. However, *learned* reaches 84 percent in precision and *religion* achieves 100 percent in correct classifications.

```
Nombre d'erreurs : 6319
Error percentage : 83.89537971322358
                precision    recall  f1-score   support

     adventure       0.00      0.00      0.00         0
belles_lettres       0.00      0.00      0.00         0
     editorial       0.00      0.00      0.00         0
       fiction       0.00      0.00      0.00         0
    government       0.17      0.00      0.00      1050
       hobbies       0.05      0.00      0.00       797
       learned       0.84      0.25      0.38      3930
          lore       0.00      0.00      0.00         0
       mystery       0.00      0.00      0.00         0
          news       0.37      0.30      0.33       787
      religion       1.00      0.00      0.00       968
       romance       0.00      0.00      0.00         0

     micro avg       0.16      0.16      0.16      7532
     macro avg       0.20      0.05      0.06      7532
  weighted avg       0.63      0.16      0.24      7532
```

Classification and evaluation of Brown on NG

It can certainly be argued that the labels we applied were generous since we only used 5 of them. Nonetheless, the *religion*-related labels were pretty straightforward and could be considered as *religion* with confidence.

4.3. 20 Newsgroup corpus applied to itself

NG has enough data to classify its 20 classes. The model trained on NG's training data and evaluated on its test data shows good out of the box performances.

```
Nombre d'erreurs : 1699
Error percentage : 22.5570897503983
                          precision    recall  f1-score   support

              alt.atheism       0.73      0.53      0.61       319
            comp.graphics       0.70      0.67      0.68       389
  comp.os.ms-windows.misc       0.68      0.74      0.71       394
 comp.sys.ibm.pc.hardware       0.64      0.74      0.68       392
    comp.sys.mac.hardware       0.78      0.81      0.79       385
           comp.windows.x       0.84      0.70      0.76       395
             misc.forsale       0.87      0.69      0.77       390
                rec.autos       0.86      0.88      0.87       396
          rec.motorcycles       0.89      0.87      0.88       398
       rec.sport.baseball       0.92      0.92      0.92       397
         rec.sport.hockey       0.94      0.95      0.94       399
                sci.crypt       0.74      0.94      0.83       396
          sci.electronics       0.75      0.61      0.67       393
                  sci.med       0.92      0.80      0.85       396
                sci.space       0.89      0.88      0.88       394
   soc.religion.christian       0.54      0.95      0.69       398
       talk.politics.guns       0.61      0.94      0.74       364
    talk.politics.mideast       0.93      0.87      0.90       376
       talk.politics.misc       0.84      0.51      0.63       310
       talk.religion.misc       0.86      0.20      0.33       251

                micro avg       0.77      0.77      0.77      7532
                macro avg       0.80      0.76      0.76      7532
             weighted avg       0.80      0.77      0.77      7532
```

Classification and evaluation of NG on NG

We push it a step further and focus on the *computer* documents. A final model is trained on *computer* labeled training documents and evaluated on computer labeled test documents. The test set does not have *comp.graphics* documents which is disappointing. The model does predict the four other classes with a high precision, which shows how domain-specific classification can lead to better results.

```
Nombre d'erreurs : 339
Error percentage : 21.64750957854406
                          precision    recall  f1-score   support

            comp.graphics       0.00      0.00      0.00         0
  comp.os.ms-windows.misc       0.76      0.79      0.78       394
 comp.sys.ibm.pc.hardware       0.80      0.79      0.79       392
    comp.sys.mac.hardware       0.85      0.85      0.85       385
           comp.windows.x       0.95      0.70      0.81       395

                micro avg       0.78      0.78      0.78      1566
                macro avg       0.67      0.63      0.65      1566
             weighted avg       0.84      0.78      0.81      1566
```

5. CONCLUSION

Text mining is often more a matter of art than science. The preprocessing steps are standard, yet the parts concerning the clustering and decomposition into topics are difficult to correctly tune. This study shows how complex it is to apprehend new corpuses with known ones. There is a constant process of adaptation with respect to evolving labels and topics, and depending on the context and desired granularity. It is essential to have data sets which balance number of documents and vocabulary size.