MAC - Labo 2

Authors :

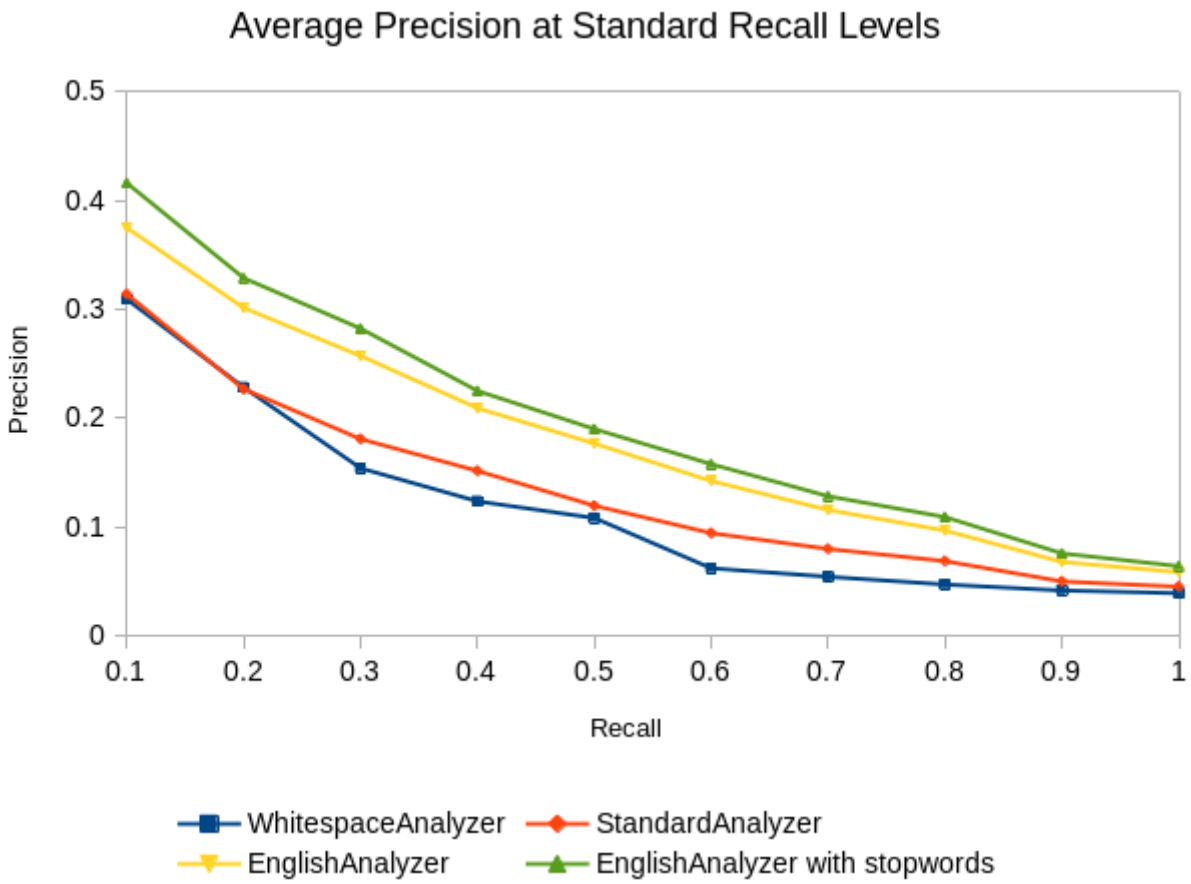
- Gildas HOULMANN
- Thibaud FRANCHETTI

Analyzers comparison

Results

The following graph shows the average precision (AP) at the 11 standard recall levels for the 4 analyzers:

- WhitespaceAnalyzer
- StandardAnalyzer
- EnglishAnalyzer
- EnglishAnalyzer using the `common_words.txt` file as a stopwords list



Discussion

We know that the close our results are to the top-right, the better they are.
We can then see here that the EnglishAnalyser beats all the others. This is due to the face that it is specialized on the language so they can perform better stemming, and globally a better analysis.
The EnglishAnalyser is a little more powerful when using the `common_words.txt` file as a stopwords list.
Below, we can see the StandardAnalyzer and the WhitespaceAnalyzer, pretty close to each other. The

StandardAnalyzer wins by a few because it uses a lowercase filter and a more advanced tokenization. The WhitespaceAnalyzer only uses whitespaces separations to generate tokens, and does nothing more.

Conclusion

In conclusion, when possible, it is better to use an analyzer that is specialized on the language of the documents. Furthermore, stop words have a non-negligible impact on the results.