TECHNICAL ASSIGNMENT

Author: Arturo Sánchez Palacio

Date: 10/08/2020

In this assignment two main tasks are commanded: text classification and topic modeling. In order to approach these two tasks we start with a brief exploratory analysis over our data. The dataset[1] is formed by texts from four categories: Wikipedia articles, extracts from books, reviews and conference papers in three different languages: English, French and Spanish. The following diagram shows how texts are distributed depending on category and language:

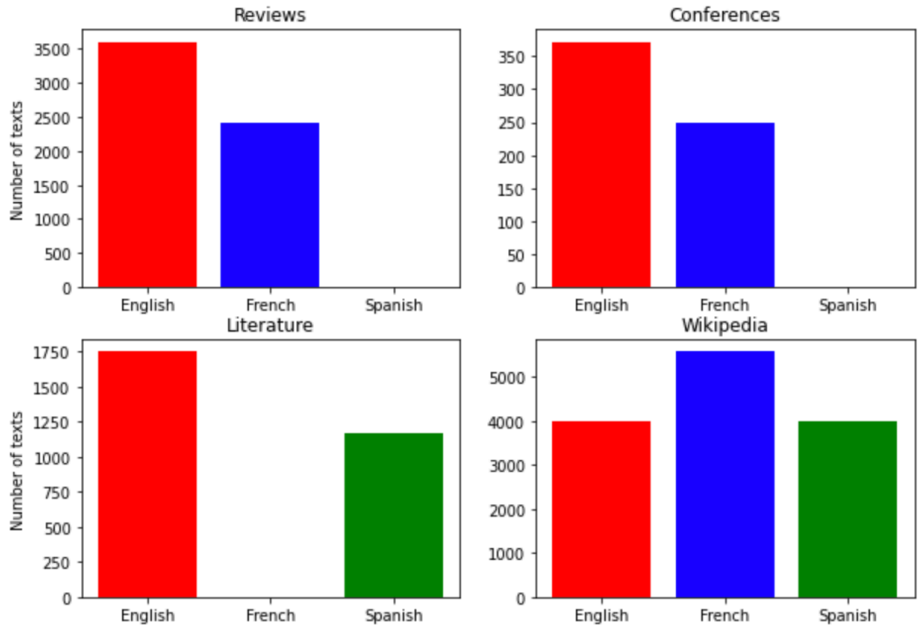

Diagram 1. Distribution of the texts depending on category and language.

In order to perform classification a decision must be made. Should texts be translated to a common language and then perform 'usual' classification or should a multilingual model be used? I rejected the first idea because of two main reasons:

- Using two models (Machine Translation and Text Classification) would multiply the possibilities of missing relevant information. Machine Translation sometimes alters the structure of texts which seems quite relevant in this case. In addition every model has its own bias so using two models implies two sources of bias instead of one.
- Multilingual models have proved impressive results in the last years generating embeddings that are pre-tained and can be used with a small fine-tuning. In my experience using these models have always shown better results than the combination of two different models.

I considered mainly two embeddings for this task: BERT and MUSE. I finally chose BERT because its implementations are usually better optimized and more debugged. BERT model allows us to use its embedding by only adding a few layers in order to adapt it to our classification task. If you want to know more about BERT don't hesitate to read my article about it on Medium in the following link. I used a module called SimpleTransformer that really lowers the difficulty of using BERT, specially because it automates the specific preprocessing BERT requires. I had

[1]Ferrero, Jérémy & Agnès, Frédéric & Besacier, Laurent & Schwab, Didier. (2016). A Multilingual, Multi-Style and Multi-Granularity Dataset for Cross-Language Textual Similarity Detection.

worked with this module before and it has always shown very interesting performance results. BERT's main disadvantage is that it takes a great amount of time to train (even when only fine tuning for specific tasks). I was quite pressed with the timing so in order to simplify the training and speed up the obtention of results I decided to cut the texts and work only with the first 35.000 characters. This choice seemed quite bold but yielded surprisingly good results.

Training the model took about 14 hours on my computer and led to outstanding results by only misclassifying 3 texts (which means an accuracy of 99.87%). In addition the tagging of two of these texts seems a bit inconsistent (check the notebook: classification.ipynb for more detail). This result seems quite surprising but when we explore the texts we can see that categories are actually quite disjoint so a well trained model should be able to achieve a really high accuracy.

In order to perform topic analysis I designed a process in which we extract topics depending on the language (using custom models for preprocessing depending on the language and then performing LDA). I designed a small function in order to erase special characters, stop words and capital letters and then lemmatize each word. Once this is done I used coherence plots in order to decide which is the optimal number of topics for our analysis. Let's check an example:

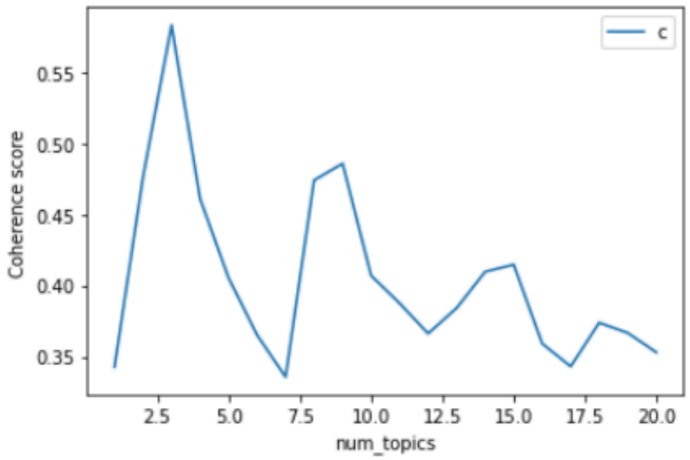

Diagram 2. Coherence plot for LDA topic modeling in English.

On the previous graph we see that the optimal number of topics according to coherence is 3 so we would choose the model with three topics.

Finally we used the module pyLDAvis which is very famous because of its interactive plots in order to visualize LDA results. See an example:

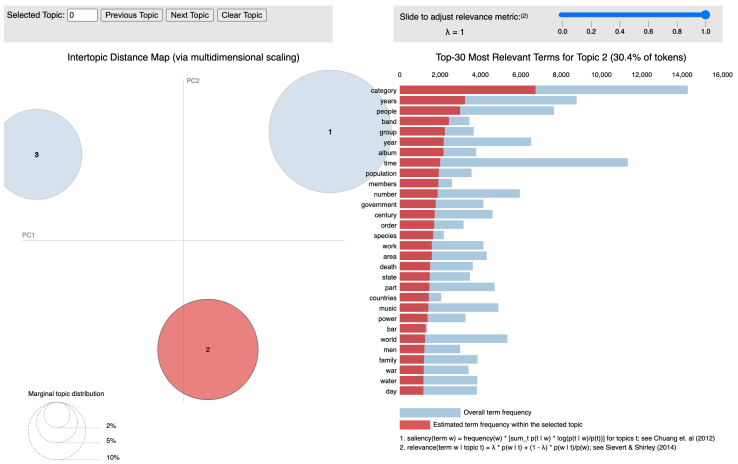

Diagram 3. English LDA

In the plot we can see the model built with three topics and precisely the information of the second topic which seems music related since it has words like band, group, album or members.

After performing a joint analysis of the results yielded on each model these are some of the topics that are clearly present and detected by our models[2]:

- **Sports** shown by words like tennis, football, player, team, club, season…
- **Cinema** shown by words like film, director, actor, actress, arts…
- **Music** shown by music, album, group, song, lyrics, cd, success, jazz, metal…
- **History** shown by words like city, art, war, king, queen, empire, battle, troop…

These were the topics that were clearer to me when analyzing the results. Two more topics which are Science and Computer Science could also be considered though I did not find enough evidence to support them, only some some words like software, system, services, function, numbers… However none of the models captured these topics as separated entities so I decided to leave them only as a side note.

[2] In order to get more precise information do not hesitate to explore the html files where you will be able to interact with the results of this analysis.