Corpus Linguistics: Final Assignment (advanced). Analyzing hyperpartisan documents.

Note: this assignment corresponds to the %35 of the final grade.

The purpose of this assignment is to analyze the usage of words in documents which are considered to be *hyperpartisan*[1], or not.
You will be given two main documents:

- articles-validation-bypublisher-20181122.xml.zip (332 MB): an XML document 150k News articles.
- ground-truth-validation-bypublisher-20181122.xml.zip (5.2MB): a document that states whether the News articles contain hyperartisan arguments or not.

The dataset were used in the semeval-2019 task 4 on hyperpartisan News detection. Whereas the task on semeval was to design a system to automatically detect hyperpartisan news, in this exercise we are going to exploit both corpora (hyperpartisan and not hyperpartisan news), and analyze which terms are the most relevant in each of the sets. For this, analysis we will use the so-called **log odd ratio**.

1 Log odd ratio

The **log odd ratio** is a measure of words compared on two sets of documents (i and j), which in our case corresponds to hyperpartisan and non-hyperpartisan documents, respectively. Each word then can be associated

[1]Hyperpartisan arguments are those that "exhibit blind, prejudiced, or unreasoning allegiance to one party, faction, cause, or person".

with its log-odd ratio r_w, which is a number that can be positive or negative: positive numbers are associated with set i, and negative numbers with set j.

The log-odd ratio r_w is defined as:

$$p_w^{(i)} = \frac{f_w^{(i)}}{N^{(i)}};\ p_w^{(j)} = \frac{f_w^{(j)}}{N^{(j)}}$$
$$o_w^{(i)} = \frac{p_w^{(i)}}{1 - p_w^{(i)}};\ o_w^{(j)} = \frac{p_w^{(j)}}{1 - p_w^{(j)}}$$
$$r_w = \log o_w^{(i)} - \log o_w^{(j)}$$

where $f_w^{(i)}$ is the frequency of word w in group i (hyperpartisan or non-hyperpartisan), and $N^{(i)}$ is the number of words in group i.

For example, suppose that the word *gold* appears $2,500$ times on hyperpartisan documents ($f_{\text{gold}}^{i} = 2500$), and 760 times on non-hyperpartisan documents ($f_{\text{gold}}^{j} = 760$). Furthermore, suppose that there are $25,000$ words in hyperpartisan documents ($N^i = 25000$), and 17500 on non-hyperpartisan documents ($N^j = 17000$). Then:

$$p_{\text{gold}}^{(i)} = \frac{2500}{25000} = 0.1;\ p_{\text{gold}}^{(j)} = \frac{760}{17000} = 0.045$$
$$o_{\text{gold}}^{(i)} = \frac{0.1}{1 - 0.1} = 0.11;\ o_{\text{gold}}^{(j)} = \frac{0.045}{0.955} = 0.047$$
$$r_{\text{gold}} = \log o_{\text{gold}}^{(i)} - \log o_{\text{gold}}^{(j)} = \log 0.11 - \log 0.047 = 0.369$$

and therefore the log odd ratio of *gold* is 0.369.

2 Steps

I recommend you to follow a two-step approach to obtain the log-odd ratios of the words. In the first step, you should generate text files from the XML documents. Then, in a separated step, you should create a file with the words along with its log-odd ratio.

2.1 Generate text files

The first step is to generate two text files for hyperpartisan and non-hyperpartisan news articles, respectively. You have to divide the News articles contained

in articles-validation-bypublisher-20181122.xml.zip into two text files (*hyperpartisan.txt* and `non-hyperpartisan.txt`), according to their ground truth value in ground-truth-validation-bypublisher-20181122.xml.zip. You will need the *lxml* library in python to analyze the XML documents and extract the necessary information.

2.1.1 Tokenize the text

When dividing the articles, it is highly recommended that you tokenize the text using a proper tokenizator. There are several tokenizers available for English in Python, chose the one that fits you best.

2.2 Extract log-odd ratios

Having the two text files, write a script that extracts the log-odd ratios of each word, by applying the equations of Section 2.2. Because the log-odd ratio is sensitive to infrequent words, **discard words that appear less than 20 times in the corpus**.

As a bonus, *you can also extract the log-odd ratios of the bigrams in the corpus.*

2.3 Analyze the results

Having the log-odd ratios, extract the most relevant 50 words in hyperpartisan and non-hyperpartisan documents. **Analyze these words, and write a small document with your findings**. Is there any interesting word on some set? Can you draw some conclusions regarding hyperpartisan text with respect to non-hyperpartisan ones?

If you also computed the log-odd ratios for bigrams, repeat the analysis using bigrams.

2.4 What to present

I've created a "Final Assignment_(advanced)" resource in eGela, and you can use it to upload the required documents (as a zip or tar.gz file). Alternatively, you can send me the files by email.

You have to present the following:

1. The scripts that you have used to calculate the log-odd ratios of the words.

2. The 50 most relevant words for hyperpartisan and non-hyperpartisan documents, according to their log-odd ratios.

3. Optionally, the 50 most relevant bigrams for hyperpartisan and non-hyperpartisan documents, according to their log-odd ratios.

4. A document that describes the steps performed to extract the words (bigrams), as well as your analysis (and thoughs) regarding Section 2.3.

5. Optionally, the document can also include any comment/insight you feel relevant about the course in general.