Corpus Linguistics: Final Assignment (regular).

Note: this assignment corresponds to the %25 of the final grade.

The purpose of this assignment is to extract some interesting collocations from the Brown Corpus, a collection of 500 documents of English text that contains more than one million words. Specifically, I want you to analyze the collocations involving the words "strong" and "powerful".
These are the steps you need to follow:

1 Extract text files

Convert the brown corpus into a text file using the `brown2txt.py` script. The script takes a document as input, and converts it to a text file. When converting the input, take into account:

- Each output word should consist on a "word/POS" combination. For instance, if the word form is "Fulton", and its POS value is "NP", the you should output the string "Fulton/NP".
- Each sentence must be in a separated line.

1.1 Try the script

Download the brown corpus from the eGela ("Brown corpus in XML format") into your folder. Then, you can try the script like this:

```
% python3 brown2txt.py brown_tei/a01.xml
```

And the first line in the output should look like this:

```
The/AT Fulton/NP County/NN Grand/JJ Jury/NN said/VBD Friday/NR an/AT
investigation/NN of/IN Atlanta's/NPg recent/JJ primary/NN election/NN
produced/VBD ``/pct no/AT evidence/NN ''/pct that/CS any/DTI
irregularities/NNS took/VBD place/NN ./pct
```

1.2 Convert all the corpus

Once the `brown2txt.py` is correct, you must convert the full Brown corpus. For this, you can use the `do_brown_txt.sh` script by running the following:

```
% bash do_brown_tab.sh
```

the script expects the input documents to be on a folder named `brown_tei`, and it will create a folder named `brown_txt` with the output text files.

2 Extract the collocations

Next, you must extract all "Adjective-Noun" collocations from the text corpus created in the last step. The corpus uses the same format we had in the gutenberg corpus, when extracting collocations according to the POS values. Therefore, you can use the `bigrams_pos.py` script to extract the collocations. You may need the POS tag labels for Brown corpus, which are available here.

Once the collocations are extracted, select only those involving the adjectives "strong" and "powerful", ordered by frequency in descending order (note: you can use the Unix commands '`grep`', '`egrep`', '`sort`', etc). Create two result files:

- `strong_JN.txt`
- `powerful_JN.txt`

3 Analyze the results

Perform a little analysis on the files produced in the last step, and write a short document describing when and how the adjectives "strong" and "powerful" are used. Write your thoughts about how those words are used, and the differences among them.

3.1 What to present

I've created a "Final Assignment_(regular)" 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 need to present the following:

- The files `strong_JN.txt` and `powerful_JN.txt` as produced in Section 2
- Your analysis (and thoughs) of Section 3
- Optionally, you can also include any comment/insight you feel relevant about the course in general.

If you have any question, issues, comments, etc, please do not hesitate to contact me.