Basics Zipf's Law Exercise

From WP:

Zipf's law /ˈzɪf/, an empirical law formulated using mathematical statistics, refers to the fact that many types of data studied in the physical and social sciences can be approximated with a Zipfian distribution, one of a family of related discrete power law probability distributions.

Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table.

1. Use the text file corpus/en.txt and corpus/es.txt
2. Write a program to read the corpus. Tokenize it using whatever tokenizer from NLTK or write your own tokenizer.
3. Write a program to check Zipf's first law (f = K/r) on this real corpus: Count word frecuencies, sort them by rank, and plot the curve.
4. Compute the proportionality constant (K) between rank and frequency for each word. Compute its average and deviation . Discuss the results . Are they consistent with Zipf's Law ?
5. Perhaps you have found problems with the tokenization (Word case, punctuation marks, numbers, etc. Try to fix them and repeat the ítems 3 and 4.
6. Now move to the char level. Repeat the ítems 3 and 4 using now as units not words but chars (letters and punctuation marks).
7. If your program is in python you can use access functions to the text files in auxiliar.py. For plotting there are several python libraries, one of them is matplotlib.