SOMOS bootcamp

Challenge 1: Pareto

The principle of pareto is very useful in life and in business. Stated simply it says that most of the effects can be obtained by a small subset of the causes. In the business world it is often referred to as the 80-20 rule, 80% of the effects come from 20% of the causes/members of a population. For example, in the business of SOMOS not every site contributes the same, there are few sites that account for most of the revenue of the business and many more sources that do not account as significantly.

When learning a new natural language or programming language the same principle applies. Only a small subset of the words account for most of the corpus of a book. If you want to learn English, would it make sense to study a dictionary? Are all the words equally important? If not, in which order would you learn the words?

1. Create a program that will receive a text as input and that would create a distribution of the word's usage. The expected output would tell us how many times a given word appears. For simplicity It could load a txt file where you could copy/paste a long text.

2. How many different words does the book The Adventures of Sherlock Holmes include?

3. how many different words would a person need to understand to read the 80% of the book?