5. *Presidential logorrhea*

In this problem, you will analyze the lengths of the State of the Union addresses.

(a) Write Python code that parses each SOU address, finding end-of-sentence markers. Don't worry about being too precise about sentence boundaries—as a first approximation, you could find words ending in a period. (But what about "Mr."?)

(b) For each year, compute the number of sentences in the address, and the mean sentence length in words for that year. Plot these data and two linear regressions, one plot for the number of sentences by year, another for the average sentence length by year. Note that the definition of "word" and "sentence" is imprecise. You can experiment with different parsing rules, and see if the results change qualitatively. Describe the trends that you see, and give some explanation for them. You should compute the linear regressions directly—for example, you may use the linear algebra routine numpy.linalg.solve but do not use a package that computes the regression.

(c) Now, compute two regressions of the total number of words in a SOU versus year–one for the years 1790 to 1912, another for the years 1913 to the present. What trends do you see? Lookup the history of the State of the Union addresses (for example on Wikipedia) to explain the regressions.

(d) Which President has the longest sentences on average? Which has the shortest sentences? Compute the median, 25% and 75% quantiles across all Presidents. What was the longest and shortest sentence ever spoken (or written) in a SOU?