Programming Exercises (Files)

1. A hapax legomenon (often abbreviated to hapax) is a word which occurs only once in either the written record of a language, the works of an author, or in a single text. Define a function that given the file name of a text will return all its hapaxes. Make sure your program ignores capitalization. [open http://www.gutenberg.org/ and download an e-book as plain text, use the file for texting your program]
2. Write a program that given a text file will create a new text file in which all the lines from the original file are numbered from 1 to n (where n is the number of lines in the file).
3. Write a program that will calculate the average word length of a text stored in a file (i.e the sum of all the lengths of the word tokens in the text, divided by the number of word tokens). [open http://www.gutenberg.org/ and download an e-book as plain text, use the file for texting your program]
4. A sentence splitter is a program capable of splitting a text into sentences. The standard set of heuristics for sentence splitting includes (but isn't limited to) the following rules:

   Sentence boundaries occur at one of "." (periods), "?" or "!", except that .

   Periods followed by whitespace followed by a lower case letter are not sentence boundaries.

   a) Periods followed by a digit with no intervening whitespace are not sentence boundaries.
   b) Periods followed by whitespace and then an upper case letter, but preceded by any of short list of titles are not sentence boundaries.
   c) Sample titles include Mr., Mrs., Dr., and so on.
   d) Periods internal to a sequence of letters with no adjacent whitespace are not sentence boundaries
   e) (for example, www.aptex.com, or e.g).
   f) Periods followed by certain kinds of punctuation (notably comma and more periods) are probably not sentence boundaries.

   Your task here is to write a program that given the name of a text file can write its content with each sentence on a separate line. Test your program with the following short text:

   Mr. Miyagi bought cheapsite.com for 1.5 million dollars, i.e. he paid a lot for it. Did he mind? Adam Jones Jr. thinks he didn't. In any case, this isn't true... Well, with a probability of .9 it isn't.

   The result should be:

   Mr. Miyagi bought cheapsite.com for 1.5 million dollars, i.e. he paid a lot for it.
   Did he mind?
   Adam Jones Jr. thinks he didn't.
   In any case, this isn't true...
   Well, with a probability of .9 it isn't.

5. Extra Challenge (OPTIONAL)

   A certain children's game involves starting with a word in a particular category. Each participant in turn says a word, but that word must begin with the final letter of the previous word. Once a word has been given, it cannot be repeated. If an opponent cannot give a word in the category, they fall out of the game. For example, with "animals" as the category,

   Child 1: dog
   Child 2: goldfish
   Child 1: hippopotamus
   Child 2: snake
   ...

   Your task in this exercise is as follows: Take the following selection of 70 English Pokemon names and generate the/a sequence with the highest possible number of Pokemon names where the subsequent name starts with the final letter of the preceding name. No Pokemon name is to be repeated.

   audino bagon baltoy banette bidoof braviary bronzor carracosta charmeleon cresselia croagunk darmanitan deino emboar emolga exeggcute gabite girafarig gulpin haxorus heatmor heatran ivysaur jellicent jumpluff kangaskhan kricketune landorus ledyba loudred lumineon lunatone machamp magnezone mamoswine nosepass petilil pidgeotto pikachu pinsir poliwrath poochyena porygon2 porygonz registeel relicanth remoraid rufflet sableye scolipede scrafty seaking sealeo silcoon simisear snivy snorlax spoink starly tirtouga trapinch treecko tyrogue vigoroth vulpix wailord wartortle whismur wingull yamask