Assignment #2: Files and String Processing

The Problem

The following index (reference: Rudolf Flesch, *How to Write Plain English*, Barnes & Noble Books, 1979) was invented by Rudolf Flesch as a simple tool to gauge the legibility of a document without using linguistic analysis.

1. Count all **words** in the file. A **word** is any sequence of characters delimited by white space or the end of a sentence, whether or not it is an *actual* English word.
2. Count all **syllables** in each word. To make this simple, use the following rules:
   - Each group of adjacent **vowels** (a, e, i, o, u, y) counts as one syllable (for example, the "ea" in "real" counts as one syllable, but the "e..a" in "regal" count as two syllables). However, an "e" at the end of a word does not count as a syllable. Each word has at least one syllable even if the previous rules give a count of zero.
3. Count all **sentences**. A **sentence** is a group of words terminated by a period, colon, semicolon, question mark, or exclamation mark. Multiples of each of these characters should be treated as the end of a single sentence. For example, "Fred says so!!!" is one sentence.
4. The **index** is computed by:
   index = 206.835 – 84.6 * (#**syllables** / #**words**) – 1.015 * (#**words** / #**sentences**)
   rounded to the nearest integer (use the round function rather than ceiling or floor).

The index is a number, usually between 0 and 100, indicating how difficult the text is to read. Some examples for random material for various publications are:

Comic book	95
Consumer ads	82
Sports Illustrated	65
Time magazine	57
New York Times	39
Auto insurance policy	10
Internal Revenue Code in the U.S.	-6

The purpose of the index is to force authors to rewrite their text until the index is appropriately high enough. This is achieved by reducing the length of sentences and by removing long words. For example, the sentence:

The following index was invented by Flesch as a simple tool to estimate the legibility of a document without linguistic analysis.

can be rewritten as:

Flesch invented an index to check whether a document is easy to read. To compute the index, you need not look at the meaning of the words.

Input
Your program will read the text to be analyzed from a file. The filename is to be given as a command line parameter to the program. You will name the program fleschIndex.c and will execute the code on a file by doing the following:

./fleschIndex <filename>

For example, if you have a file with an essay and the file was named *myEssay.txt* then you would do the following to find the Flesch index:

./fleschIndex myEssay.txt

Output
The output (to **stdout**) from your program will be the following:

1. The Flesch/legibility index that you have computed
2. The number of syllables in the input
3. The number of words in the input
4. The number of sentences in the input

It will have the following format (and must match exactly):

Flesch Index = 87
Syllable Count = 10238
Word Count = 2032
Sentence Count = 193

Project Guidelines
Sample testing files will be posted to **CourseLink** a week before the due date.

You must submit a zip file containing the following:

- Your source code (.c)
- A functioning makefile

Your makefile MUST successfully compile your program on the school server and produce an executable named **fleschIndex**. You may include other flags (such as the math library) if needed.

Your program should read the entire file into memory and then do the counting and computation of the index.

You must do the following error checking / make the following assumptions:

- Your program will terminate with a meaningful error message if you cannot open or read the file given on the command line.
- Numbers (in the form of 2, 45, 865, *etc*.) will be counted as one word and one syllable.

- There will not be currency in the text (nothing of the form $12.50).
- The only valid punctuation characters are the sentence termination characters: period, colon, semicolon, question mark, and exclamation mark. All other punctuation such as commas, #, @, *etc*. are to be **ignored**.