Programming Assignment #5

Word sequences from a file

CS 3358.501, Summer I 2012
Instructor: Jill Seaman

Due: Tuesday, 6/26/2012 (upload electronic copy by 4:30pm)

Problem:

You will be writing **part** of a tool that can be used to catch plagiarists. The goal of the tool is to determine similarities between documents in a large set to find out if plagiarism is going on within the group.

For this assignment you will write a program that will be able to produce a list of all the n-word sequences in a document. A word is a sequence of characters containing no whitespace (spaces or newlines). An n-word sequence is just a list of n words that occur in that order in the document.

For example, if the document starts this way:

```
Death of a Salesman:
In the play, Arthur Miller's Death of a Salesman: Willy Loman, a sympathetic
salesman and despicable father who's "life is a casting off" has some traits that
match Aristotle's views of a tragic hero. Willy's series of "ups and downs" is identical
to Aristole's views of proper tragic figure; a king with flaws. His faulty personality, the
financial struggles, and his inabiltity are three substantital flaws that contribute to his
failure and tragic end.
```

The list of 6-word sequences would start out this way:

```
Death of a Salesman: In the
of a Salesman: In the play,
a Salesman: In the play, Arthur
Salesman: In the play, Arthur Miller's
In the play, Arthur Miller's Death
the play, Arthur Miller's Death of
play, Arthur Miller's Death of a
Arthur Miller's Death of a Salesman:
Miller's Death of a Salesman: Willy
Death of a Salesman: Willy Loman
...
```

Your program should take the name of a file in the working directory, and the value for n (the number of words in a sequence) as input. These can be either input on the command line, or your program can prompt the user to enter them. Then your program should output the list of all the n-word sequences to the screen (or a separate text file).

NOTES:

- Just one *.cpp file.
- I (strongly) recommend using a queue to assist with producing the sequences from the file. You will need a queue that will allow you to enqueue, dequeue, and then see the contents of the entire queue. There is a data structure in the STL that allows you to push_back and pop_front, and you can access each element using the square bracket notation (queue[10]). It is called a deque ("deck").
- The next assignment (PA#6) will use the solution from this assignment to add more functionality to the plagiarism detection tool.

Style:

See the Style Guidelines document on the course website.

Logistics:

Please submit you solution in a single file. You can call it process_files_xxxxxx.cpp.

The xxxxx is your TX State NetID (your txstate.edu email id).

Submit: an electronic copy only, using the Assignments tool on the TRACS website for this class.