CSE 6331   Cloud Computing
Summer 2021, © DL, UTA, 2021

Programming Assignment 5
Amazon AWS

Cloud Assignment 5 – dealing with "large" textual information

Description:
Much of the available, digitized information is in the form of text.
(We will consider web, and similar, as a variation of text.)
We will simplify handling to only text, but the concepts could be extended.

More:

Given several (say a few hundred or thousand) text files, which we will generically call documents, we want to find those documents relevant to a user's needs (requests.)
(This is, in general, what "search engines" do.)

For this assignment we will simplify many of the otherwise interesting details, but try to emphasize many of the issues and interesting approaches to searching text.
Almost all original, source documents need to be "cleaned" (may include removing pictures, font details, pagination, and similar.)
For this assignment ALL non ASCII information will be removed.
(Note, we will assume that docs are in English, but Spanish, French and similar languages are not difficult to extend processing, Chinese is more difficult.)

Looking for relevant documents, by doing a simple word scan is simple, very time consuming, and usually gives poor results.
Preprocessing the original docs is important.

The following are some simple methods to cleaning/processing original documents.
- Dealing with upper or lower case letters, usually changing to lower case.
- Remove punctuation (or most)
- Remove very common words ("stop words" such as the, or, and.)
- Additionally (optionally):
    Word stemming (cats -> cat)

(there are many others)

Then the words in the documents are extracted and indexed, with "pointers" back to the individual relevant documents, and where found.

Then, through some sort of interface, a "search" of relevant documents can be done.
The simplest, and usually poorest result is a single word, such as "cloud".
Combinations of words, in close proximity, usually do better, such as "cloud computing".
(You may see terms such as bi-grams, and tri-grams, and other terms or extensions.)

One, free, source of thousands of texts is: https://www.gutenberg.org/

Notice that small optimizations will often give greatly improved results:
For example “red hot” may be the same as “hot red” (interchange word order)
Letters in search may be transposed (or missing): “teh” instead of the,
questionble instead of questionable
There are word lists that may help: (MIT site, github, many others,
depending on what you want.)

Users of this service will interact with your service through web page interfaces, all processing and web service hosting is (of course) cloud based.

Additional Details, functionality:

A user should be able to do searches based on words or word combinations to find relevant documents you should identify a document by name (and any other meta information you have such as author name, publication date) and where in that document that is found (such as line or offset.)
(You should show lines or paragraphs that match, for context.)

Please, submit through Canvas.
All work must be your own,
(Same as previous assignments)