Assignment - 2 on Suffix Trees - UE15CS311 (AdvAlgo)

# Problem Definition:

Given a set of documents, build a suffix tree to perform the following operations.

1. List all the occurrences of a query-string in the set of documents.
2. List only the first occurrence of a given query-string in every document. If the query-string is not present, find the first occurrence of the longest substring of the query-string.
3. For a given query-string (query of words), list the documents ranked by the relevance.

# Data-Set and Resources:

Data-set 1: Aesop Tales ([AesopTales.txt](https://drive.google.com/open?id=0B-E2SGzqkISrSy1mdm5TUEN5NUk)), where each tale is a document. It's provided as a single text file with all the tales concatenated with two blank lines separating two tales. First line of the tale is the title of the tale.

1. **List all the occurrences of a query-string in the set of documents.** One result per occurrence. For every occurrence, a result to have the title of the tale along with the sentence (sentences if the query-string spans over multiple sentences) having the matched text (text snippet surrounding the matched text).
2. **List only the first occurrence of a given query-string in every document. If the query-string is not present, find the first occurrence of the longest substring of the query-string.** For every occurrence, a result to have the title of the tale along with the sentence (sentences if the query-string spans over multiple sentences) having the matched text (text snippet surrounding the matched text). One result per document (tale) if there is at least one character match.
3. **For a given query-string (query of words), list the documents ranked by the relevance.** Define relevance. The following criteria are some guidelines for defining the relevance but you can have your own reasonable definition.
   1. Sequence of words in the text matching the query is more relevant than subsequence of words in the text matching the query (that is, all the words in query appear in the relative order but some other words in between is a subsequence of words).
   2. A subsequence with lesser number of other words is more relevant than with more number of other words.
   3. All the words of the query appear in the same relative order is more relevant than they appearing in any other order in the document.
   4. Matching of non-stop-words is more relevant than matching of stop-words. In general, matching "important" words is more relevant than that of the other words.
   5. Matching all words in the query is more relevant than matching a subset of words.
   6. Matching exact words from the query is more relevant than matching a similar word (another word with same stem, possible minor typo in the word, synonym).

# Submission and Deadline:

* Deadline: 10-Oct-2017 at 1:30 PM.
* Write a report of of 3-6 double-column pages in IEEE Transactions paper format available at <http://ieeeauthorcenter.ieee.org/create-your-ieee-article/use-authoring-tools-and-ieee-article-templates/ieee-article-templates/templates-for-transactions/>. Let your name and SRN be at the header of every page.
* Submit a printed copy of your report to the instructor.
* Submit a zip file having a pdf copy of the report along with the source code as directed by your instructor. Let the name of the zip be like “UE15CS311\_A2\_01FB15ECS000.zip”.

# FAQs:

1. …