Programming and Data Structures in Python

Duration: 1 Hour and 15 minutes
Maximum Marks: 20

1. Write a python function `checkA` which takes a string and identifies every incorrect occurrence of the articles `a` and `an`. That is, it should return every occurrence of the word `a` which is followed by a word beginning with a vowel as well as every occurrence of the word `an` followed by a word beginning with a consonent. It returns a list of locations (as in the starting and ending position) of each such incorrect occurrence. You are not obliged to use the `re` library. *(5 marks)*

2. You are given 4 csv files named `test1.csv, test2.csv, test3.csv` and `test4.csv`. Each contains the login-id and mark of each student in a class. The list of students are the same across all the 4 tests and you may assume that all lists are sorted by login-ids.

   Your aim is to output four lists: the first list should contain the list of login-ids of students who obtained their maximum across the four tests in Test 1, the second list contains the login-ids of students who obtained their maximum in Test 2, and so on. If a student obtained his or her highest marks in multiple tests then his or her login-id must appear in all those lists. *(7 marks)*

3. We wish to maintain a database of stories and tags. For our purposes a story consists of a pair of strings giving the title and the body of the story. We also have a list of tags which are just words. Users are allowed to insert stories into the database and also insert new tags. They will also query the database by providing a list of tags and the database should return a list containing pairs, the title and an identifier, for all the stories that contain all the tags in the list. The users may also ask the database to retrieve a story by providing an identifier and the database then returns the story with that identifier. Thus your story database class, `Stories` must support the following methods:

   (a) `insert(title,story)` which inserts a story.

   (b) `insertTag(tag)` which inserts a tag.

   (c) `query(ls)` which returns a list of pairs of the form `(t,id)` where `t` is the title of a story and `id` is a unique identifier of this story in the database (which may then be used to retrieve it).

   (d) `retrieve(id)` which returns the story associated with the tag `id`.

   Here is an example:

```
new = Stories()
new.insert("First", "This is my first Story")
new.insert("First", "I am not imaginative with titles")
new.insert("Second", "My Second story")
new.insertTag("not")
```

```
new.insertTag("my")
new.insertTag("Story")
z = new.query(["Story",my"])
for i in z:
  (title,id) = i
  print(title):
  print(new.retrieve(id))
```

will result in the following output:

```
First
This is my First Story
Second
My Second story
```

You know that typical usage would involve at most 50 tags in all, not more than 1000 stories where each story may be very long say hundreds of thousands of words and there will be millions of queries and a few thousand retrieves. What data-structures will you use and why? [You don't have write out code. You just need to explain how you what data structures you will use to store whatever data you want to and any relevant detail on how each of these methods will access or modify that data]

Suppose, we would also like the insert function to add a story to the database only if it is not already there. How would you modify your implementation to handle this ? [Again no code is necessary, just in formation on what datastructures you will use and how they will accessed or updated in each operation] *(8 marks)*
