INF 551 – Fall 2017

Homework #2

Due: October 6, Friday
100 points

1. [70 points] Consider a question-answering data set "qa.json". The file contains a list of topics, e.g., Super_Bowl_50. For each title, it has a number of paragraphs, and for each paragraph, it shows the content/context of the paragraph, a list of question-answers pairs, where the answers of questions all appear in the paragraph. Note that there may be multiple answers for a question. In this assignment, we consider only the first answer (index being "0"). Note also that "answer_start" indicates the position of the first character that starts the answer. Each question also has a unique identifier ("id").

   The data set comes from the Stanford Squad project (https://rajpurkar.github.io/SQuAD-explorer/, dev set) where you may look up for more details.

   a. [30 points] Write a Python script "stats.py" that takes the "qa.json" as the input data set and produces a list of different types of questions in the data set and for each type, how many questions belong to the type.

      Here, we categorize the questions by the beginning words/phrases of the questions. We consider the following types:

      ```
      how
      how many
      how much
      what
      when
      where
      which
      who
      whom
      ```

      For example, the question "Which NFL team represented the AFC at Super Bowl 50?" is a type "which" question since it starts with "Which".

      Run your script this way:

      python stats.py qa.json

      Your output should be a JSON document in the following format:

      {"how": 5, "how many": 2, …}

   b. [40 points] Write a Python script "search.py" that takes the "qa.json" and a list of keywords as the input. It returns a list of questions that each contains all the keywords.

For each question, it also output its id and first answer. Note that keyword matching is NOT case sensitive.

Example execution of the script:

python search.py qa.json “super bowl”

The output should be a JSON document, formatted as follows:

```
[
  {"id": "...",
   "question": "...",
   "answer": "..."},

  {"id": "...",
    ...
  },

  ...
]
```

Submission: Submit two scripts as indicated above, but also prepend your name to facilitate grading. For example, <FirstName>_<LastName>_stats.py.

2. [30 points] Consider the provided nutrion.xml which records the nutrition facts of a number of food. The xml file is from here: https://alistapart.com/d/usingxml/xml_uses_a.html. Answer the following questions using XPath or XQuery as indicated. For all questions except for the last question, return the values only (not XML elements or attributes).
   a. (XPath) How many calories (per serving) does the food whose name contains "Chicken" have?
   b. (XPath) List names of all food which contains Vitamin C.
   c. (XPath) List names of all food which contains Vitamin A and C.
   d. (XPath) List names of all food where at least 50% of the total fat is saturated (ignore the food with no fat).
   e. (XQuery) Which food (name) has the highest amount of cholesterol per serving?
   f. (XQuery) List the name and the total fat per serving for each food that contains Calcium (ca).
   Your output should come in the following format:

```
<result>
   <food>
      <name>Bagels, New York Style </name>
      <total-fat>4</total-fat>
   </food>
    ...
</result>
```

Submission: Submit a text file that contains a list of XPath expressions or XQuery. Include the answer for each question right after the expression or query for the question.