CSE427S

HOMEWORK 1

M. Neumann

Due: THU 6 SEPT 2018 4PM

Getting Started

Update your SVN repository.

When needed, you will find additional materials for *homework x* in the folder `hwx`. So, for the current assignment the folder is `hw1`.

Hint: You can **check your submission** to the SVN repository by viewing `https://svn.seas.wustl.edu/repositories/<yourwustlkey>/cse427s_fl18` in a web browser.

SUBMISSION INSTRUCTIONS

WRITTEN:

- all written work needs to be submitted electronically in *pdf format*[1] via GRADESCOPE
- provide the following information on *every* page of your `pdf` file:
  - name
  - student ID
  - wustlkey (your `wustlkey` indicates the location (SVN repository) for your code submissions; without this information we will not be able to grade your code!)
- start every problem on a *new page*
- **FOR GROUPS**: make a **group submission** on GRADESCOPE and provide the following information for **all group members** on every page of your `pdf` file: names, student IDs, and location of code submissions (**one** student's `wustlkey`)[2]

CODE:

- code needs to be submitted via SVN repository commit (detailed submission instructions are provided whenever code submissions are required)
- make sure to always use the required *file name(s)* and *submission format(s)*
- **comment your code** to receive maximum credit

Find instructions on how to get your **GRADESCOPE account**, **submit your work**, and **add your group member** on the course webpage.

[1] Please, **type your solutions** or use **clear hand-writing**. If we cannot read your answer, we cannot give you credit nor will we be able to meet any regrade requests concerning your writing.

[2] It is sufficient to commit code to one SVN repository. If you do not specify the repository for your group's code submission clearly, we will only check the first repository, sorting wustlkeys alphabetically.

PIAZZA

All course related announcements will be made there. Ask all questions about course materials, logistics, and homework assignments on Piazza using the appropriate tags.

GRADING RESULTS AND REGRADES

Grades will be maintained on Canvas. Grading comments will be provided via GRADESCOPE . We will make a Piazza announcement when the grading is completed. All regrade requests need to made via GRADESCOPE **within one week** of the grade announcement.

SVN REPOSITORY

Check out your SVN repository. Find instructions on how to checkout, update, and commit to your SVN repository here: `http://sites.wustl.edu/neumann/resources/cse427s_resources/`

Note: Code will have to be submitted exclusively via your SVN repository. To **check if your submission to the SVN repository was successful** view your repository in a web browser by entering this url (mind browser caching):
`https://svn.seas.wustl.edu/repositories/<yourwustlkey>/cse427s_fl18`

PROBLEM 1: Big Data Characteristics (50%)

Describe the *characteristics* of the following Big datasets. There may not be one correct answer, so explain your decisions by **including an argument** why each of the characteristics *is* or *is not* present.

(a) **Access Log Data**: all HTTP logs from a popular online vendor's webserver(s) collected over the duration of 5 years

The following example shows *four* example entries of this log data:

```
2013-03-15 12:39 - 74.125.226.230 /common/logo.gif 1231ms - 2326
2013-03-15 12:39 - 157.166.255.18 /catalog/cat1.html 891ms - 1211
2013-03-15 12:40 - 65.50.196.141 /common/logo.gif 1992ms - 1198
2013-03-15 12:41 - 64.69.4.150 /common/promoex.jpg 3992ms - 2326
...
```

(b) **Wikipedia articles**: text (no images, info boxes etc.) of all articles of the English Wikipedia and the links between them

The following example shows (part of) *one* example article with the links indicated in blue:

> Big data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy. There are three dimensions to big data known as Volume, Variety and Velocity. Lately, the term "big data" tends to refer to ...

(c) **Chemical compounds**: data set of all existing chemical compounds (e.g. the ZINC database) commonly used for drug design

The following example shows (part of) the representation of *one* example chemical compound:

bonds in form of (bond ID: atom ID, atom ID):

```
(1:  2,1) (2:  14,1) (3:  3,2) (4:  4,3) (5:  12,3) (6, 3,4) (7:  5,4) (8:  6,5)
(9:  5,6) (10:  7,6) (11:  11,6) (12:  6,7) (13:  8,7) (14:  21,7) (15:  7,8) ...
```

bond types as (bond ID: bond type):

```
(1:  47) (2:  47) (3:  47) (4:  50) (5:  47) (6:  47) (7:  47) (8:  50) ...
```

atom types as (atom ID: atom type):

```
(1:  7) (2:  3) (3:  3) (4:  6) (5:  3) (6:  3) (7:  6) (8:  3) (9:  22) ...
```

property:

```
mutagenic
```

Bonds are atomID-atomID pairs indicating that two atoms are connected. Bond types indicate the type of bond (47 = single, 50 = double, ...). Atom types are for example 7=oxygen, 3=chlorine, 6=carbon, 22=hydrogen,... The property of a chemical can be either *mutagenic* or *non-mutagenic*.

(d) Both data sets described in (a) and (b) are represented as *text*.

- What is the main difference between those data sets?
- What does that imply for data analysis tasks such as information extraction or pattern recognition?

(e) For the example data sets described in (a)-(c), what are the **data points** and what **data types/data structures** do we use to represent those data points programmatically?

PROBLEM 2: Bonferroni's Principle (25%)

"Be careful with what you mine in Big data – it could be random!"

This question generalizes the example of "evil-doers" visiting hotels, as in Section 1.2.3 of the MMDS book. Suppose (as described there) that there are one billion people being monitored for 1000 days. Each person has a 1% probability of visiting a hotel on any given day, and hotels hold 100 people each, so there are $100,000$ hotels. However, our test for evil-doers is different. We consider a group of p people evil-doers if they all stayed at the same hotel on d different days. Derive the formula for the (approximated) expected number of false accusations f (that is, the expected number of sets of p people that will be suspected of evil-doing), assuming that in fact there are no evil-doers, but all people behave at random, following the conditions stated in this problem (1% probability of visiting a hotel, etc.).

Note: You may assume that d and p are sufficiently small and thus, $\binom{1000}{d} \approx \frac{1000^d}{d!}$, and similarly for p.

Hint: you can use this table to check your formula (the values for f are rounded):

d	p	f
2	2	2.5×10^5
2	3	$0.83 \approx 1$
3	2	10^{-1}
3	3	3×10^{-14}

PROBLEM 3: The Unreasonable Effectiveness of Data (25%)

Read the article "The Unreasonable Effectiveness of Data" by Alon Halevy, Peter Norvig, and Fernando Pereira. Based on this article, answer the following questions.

(a) What are the differences between unlabeled and labeled/annotated data?

(b) Summarize the *data-based approach* described in the article.

(c) What are the limits of this approach?

Bonus Problem (5% up to a max. of 100%) - no group work!

We will be doing a little experiment in the course. We will collect your **emotional description** and a **star rating** for each homework assignment and hopefully, you will be able to use this data for sentiment analysis at the end of the semester!

Sentiment Analysis tries to assess the *emotion* or *mood* in a text document. Essentially it can be computed by comparing the set of words in each document to an existing dictionary of positive words, negative words, and neutral words. In our experiment, we give you the chance to voice your opinions on each homework by writing a couple of sentences as a review for the homework. You will not be graded on what your review says, but rather solely the completion of it. At the end of the year, given that we have enough data for each homework, you will perform sentiment analysis on this data to see which homeworks you and your peers regarded as "positive" and which as "negative".

So, please write a review for this homework and store it in the file `hw1_review.txt` provided in your SVN repository.

This file should only include the review, **no other information** such as name, wustlkey, etc. Remember that you are not graded for the content of your review, solely it's completion.

To be able to evaluate your sentiment analysis approach we need the ground truth. So, in addition to your textual review, take 1 minute to provide a star rating for `hw1` via this link: `https://wustl.az1.qualtrics.com/jfe/form/SV_aaAc0ILmBleyKzj`.

You can only earn bonus points if you write a meaningful review of **at least 50 words** and provide the corresponding star rating. Bonus points are given to the **owner of the repository only** (no group work!).

Submit your review in the file **`hw1_review.txt`** provided in the `hw1` folder in your SVN repository. **To commit the file run:**

```
$ svn commit -m 'hw1 review submission' .
```