CSE427S – Homework 1

M. Neumann

Due THU 01/28/2016 10am

The solutions to this homework can be submitted by handing in a paper copy (we prefer typed answers over hand-written answers) or by committing an electronic version to your SVN repository (cf. Problem 02). If you try the latter and make sure your SVN commit was successful (Hint: `svn add` the file before committing!).

If you work in a group of 2, indicate **both names and wustl keys** on your submission!

Problem 01 (not for credit)

If you haven't done so already, sign up to Piazza (`piazza.com/wustl/spring2016/cse427`). All course related announcements will be made there. Ask all questions about the course, readings, and homework assignments on Piazza.

Problem 02 (not for credit)

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/`

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

In the folder `hw1` you will find the file `hw1_P02.txt`. Answer the questions and commit the file back into your SVN repository.

Note: Programming assignments will be submitted exclusively via your SVN repository, so make sure you are able to update and commit to it by committing `hw1_P02.txt`.

Problem 1: Big Data Properties (Dimensions) (25%)

Describe the properties (also referred to as *dimensions*) of the following Big Data examples (there is not one correct answer, so **justify** your thoughts.) Both datasets described in (a) and (b) are represented as text. What is the main difference between those datasets?

(a) Log data

Example entries in a log file look like this:

```
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

A part of the text file could be:

```
Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making.
```

(c) Database of chemical compounds (e.g. the ZINC database)

One chemical compound could be represented like this:

```
bonds:
(2,1) (14,1) (1,2) (3,2) (2,3) (4,3) (12,3) (3,4) (5,4) (6,5) (5,6) (7,6) (11,6) (6,7) (8,7) (21,7) (7,8) ...
bond IDs:
47 47 47 50 47 47 47 47 50 47 47 50 117 117 117 117 50 117 ...
atom IDs:
3 3 3 3 3 3 3 3 3 3 6 7 7 3 3 3 3 ...
costly to observe property:
mutagenic
```

The bond ID indicates the type of bond (single, double, ...). The atom ID indicated the type of atom (oxygen, chlorine, carbon, hydrogen,...). The property of a chemical can be either *mutagenic* or *non-mutagenic*. Assessing this property involves expensive and time consuming experiments in a chemical laboratory.

Problem 2: Bonferroni's Principle (25%)

"Be careful with what you mine in your 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:

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?

Problem 4: MapReduce (25%)

Suppose the input data to a MapReduce operation consists of integer values (the keys are not important). The map function takes an integer i and produces the list of pairs (p, i) such that p is a prime divisor of i. For example, $map(12) = [(2, 12), (3, 12)]$. The reduce function is addition. That is, $reduce(p, [i_1, i_2, ..., i_k])$ is $(p, i_1 + i_2 + ... + i_k)$.

Compute the output for the following integer inputs $i = \{15, 21, 24, 30, 49\}$. Include all Mapper inputs, Mapper outputs, Reducer inputs, and Reducer outputs in your answer.

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 for each homework assignment and hopefully, we 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 homework you and your peers regarded as "positive" or "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.

You can only earn bonus points if you write **at least 50 words**. Bonus points are given to the **owner of the repository only** (no group work!).