Homework tips:

1. Use Zoran's template for your homework solutions.
2. Do not make homeworks into report formats. We don't need a title page, graphics, table of contents nor submissions that are 30+ pages. Be concise.
3. Always keep Problem statements in the document followed by steps. Prove your work – show enough screen captures that prove your results.
4. Always put your code in your doc as well as upload your code files in Canvas. We will deduct if missing.
5. We select solutions every week to showcase that are well described and easy to follow.
6. Keep your steps simple. Just bullet points are fine!
7. Upload your homework solution as .doc or .pdf. (Not both)
8. See sample homework problems below.
   p.s. You do not have to highlight/add boxes around results. Concise, concise, concise.

HU Extension Assignment xx

Handed out: 04/06/2017 Due by 11:59 PM EST on Saturday, 04/14/2017

Please, describe every step of your work and present all intermediate and final results in a Word document. Please, copy past text version of all essential command and snippets of results into the Word document with explanations of the purpose of those commands. We cannot retype text that is in JPG images. Please, always submit a separate copy of the original, working scripts and/or class files you used. Sometimes we need to run your code and retyping is too costly. Please include in your MS Word document only relevant portions of the console output or output files. Sometime either console output or the result file is too long and including it into the MS Word document makes that document too hard to read. PLEASE DO NOT EMBED files into your MS Word document. For issues and comments visit the class Discussion Board. You are not obliged to use Java or Eclipse. You are welcome to use any language and any IDE of your choice.

Problem 1:

On your Cloudera VM or any other VM you might be using install Kafka. Just in case, install one of the recent Kafka 0.8 versions. Demonstrate that you can create a topic, publish messages to that topic and consume messages sent to that topic. Use Kafka command line interface. (20%)

Untarred Kafka version 9.0.1 on my Cloudera Quickstart VM.

```
[cloudera@quickstart ~]$ tar -xzf kafka_2.11-0.9.0.1.tgz
[cloudera@quickstart ~]$ ls
access_log_1.txt          ebay.csv                    paragrapha.txt
all-bible                 ebaysmall.csv               paragraphbOLD.txt
all-shakespeare           eclipse                     paragraphb.txt
apache.access.2.log       emps.txt                    parcels
apache.access.log         enterprise-deployment.json  Pictures
cloudera-manager          express-deployment.json     prob1.py
cm_api.py                 kafka_2.11-0.9.0.1          prob1test.py
```

Stop Cloudera's version of Zookeeper

```
[cloudera@quickstart ~]$ sudo service zookeeper-server stop
JMX enabled by default
Using config: /etc/zookeeper/conf/zoo.cfg
Stopping zookeeper ... STOPPED
```

…

Problem 2. Remove the header of the attached engines_data.csv file and then import it into Spark. Randomly select 10% of you data for testing and use remaining data for training. Look initially at horsepower and displacement. Treat displacement as a feature and horsepower as the target variable. Use MLlib linear regression to identify the model for the relationship. Use test data to illustrate accuracy of your ability to predict the relationship. Create a diagram using D3 which presents the model (straight line), original test data and predictions of your analysis. Please label your axes and use different colors for original data and predicted data.

1. Delete the first line in Notepad (or using the sed command as described in the course note). In addition, I will open it in excel so I can see the data including the column names so I can count the index of the column used.
2. Import data into Spark (after putting my data into the shared folder with my VM
   $ pyspark:

   path="file:///mnt/hgfs/ShareFolder/HW11/engines_data.csv"

   raw_data=sc.textFile(path)

   …

Problem 3. Identify 10 most frequently used words longer than 7 characters in the entire corpus of Inaugural Addresses. Do not identify 10 words for every speech but rather 10 words for the entire corpus. Which among those words has the largest number of synonyms? List all synonyms for those 10 words. Which one of those 10 words has the largest number of

hyponyms? List all hyponyms of those 10 most frequently used “long” words. The purpose of this problem is to familiarize you with WordNet and concepts of synonyms and hyponyms.

(25%)

Your literature for Problems 1 and 2 are chapters 1 and 2 of Natural Language Processing with Python book by Steven Bird et al.

In [102]:

```
import nltk
from nltk.corpus import inaugural
import pandas as pd
from nltk.corpus import wordnet as wn
```

Loaded words from each address into a flat list called speeches

In [103]:

```
addressWords = [word for speech in inaugural.fileids() for word in inaugural.
words(speech)]
len(addressWords)
```

…