Assignment 1 (W4242)

This assignment is due on Wednesday, October 2, 6:10pm before class. You can discuss with other people, but make sure write down their names on your homework.

Q1. Here is the link to the list of UFO sightings : `http://www.nuforc.org/webreports/ndxshape.html`. We are interested in UFOs corresponding to one of three shapes: circle, triangle and fireball.

1. This is a structured but dirty dataset. You should collect all sightings from that list up to and including September 9, 2013. Specifically, you should represent each sighting by these eight features: Data of sighting, Time of sighting, City, State, Shape, Duration, Summary, Posted date (when the sighting was posted to the website). You should do your best to convert all Durations to seconds, whenever possible. Turn in your cleaned dataset in csv format. Keep in mind a few guidelines:

   - If a duration has a"<" sign, you should simply ignore the"<" sign. For example if the duration is specified as "< 1 minute", consider the duration to be "1 minute". You should subsequently convert "1 minute" to "60 seconds".

   - If a duration has a range, use the upper limit as its value. For example, if the duration is listed as "5-8 minutes", you should consider the duration as "8 minutes". (Again, you will need to eventually convert minutes into seconds).

   - You may encounter some other oddities in the data. Do your best to extract maximum value from the messy data; be sure to explain to us the decisions you have made in terms of data extraction and cleaning.

2. Based on your cleaned data, answer the following questions using SQL queries :

   - How many UFO sightings in Alaska?
   - How many UFO sightings of durations less than 2 minutes in NY?
   - What's the average duration of UFO sightings of fireball?

- Which year has the maximum number of sightings?
- During the December, 2012, how many states witnessed UFO more than 30 times?

Q2. Download the first dataset from `http://acube.di.unipi.it/tmn-dataset/` and structure it. The final structured dataset is expected to be a matrix with five columns corresponding to date (in the format mmddyyyy), source, category, the number of words (not including punctuation) in the title, the number of commas in the description. Based on the structured dataset you get, answer the following answers:

- Which category of news is most popular?
- How many business events happened during April to May 2011?
- How many news events come from usatoday.com?
- Which newspaper website has the longest title in average?
- Which newspaper website uses least commas (in average) in description?

Q3. Pick one dataset from the list here `http://www.infochimps.com/datasets` and join it with the UFO dataset (joining on **date** or **location** is one place to start). Answer the following questions :

- Describe the dataset and why you chose it
- Create 5 questions based on the joined data set and answer them.
- Create one interesting visualization and explain what it shows

Q4. Here are two readings :

- `http://bits.blogs.nytimes.com/2013/06/01/why-big-data-is-not-truth/?_r=1&`
- `http://vldb.org/pvldb/vol5/p1674_tamraparnidasu_vldb2012.pdf`

Both these articles concern our attempts as data scientists to get to some notion of the ground truth. Please write 2-3 paragraphs with your thoughts after reading them, and connect your ideas to any specific examples from your field of study.