Homework- 1  
Shivani S Mahaddalkar

**Task 1: review data mining concepts and tasks**

Answer the exercise questions 1-3 in Textbook 1.7. For Question 2, feel free to change the question scenario from “an Internet search engine company” to any organization that you would like to think of. It can be a company, government office, NGO, etc.

See questions 1-3 as below:

1. Discuss whether or not each of the following activities is a data mining task.
   1. Dividing the customers of a company according to their gender  
      -It is not a data mining task because it is a trivial extraction and the information presented is not a previously unknown one.
   2. Dividing the customers of a company according to their profitability  
      -It depends whether it is calculated using simple arithmetic formulae or requires prediction based on the customer’s history. In the former case, it will not be a data mining task, whereas, in the latter case it is a data mining task.
   3. Computing the total sales of a company  
      -It is not a data mining task as it requires basic mathematic functions to compute the total sales of a company.
   4. Sorting a student database based on student identification numbers  
      -It is not a data mining task as it requires a basic order by command to sort.
   5. Predicting the outcomes of tossing a (fair) pair of dice  
      -It is not a data mining task as it is a straight-forward probability distribution and it is a trivial task to determine that the probability of either heads or tails is 0.5.
   6. Predicting the future stock price of a company using historical records  
      -It is a data mining task, because it requires understanding past behavior of the stock price of a company and predict the future prices. It is not a trivial task and requires machine learning algorithms to identify the trends.
   7. Monitoring the heart rate of a patient for abnormalities  
      -It is a data mining task as it requires anomaly detection to distinguish the abnormal heart rate of a patient from the normal behavior.
   8. Monitoring seismic waves of earthquake activities  
      -It is again a case of anomaly detection, where we can use the past seismic behavior and determine the trends to monitor earthquake activities. It is a non-trivial extraction and that is why it is a data mining task.
   9. Extracting the frequencies of a sound wave  
      -It is not a data mining task because it requires a processor which can determine the sound frequencies manually.
2. Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.  
   -In an Internet search engine company *clustering* can be used to club search results together in clusters of similar thematic categories. This could ease the user’s experience by only picking a cluster which is thematically closer to what they are looking for.  
   -*Classification* can be used to assign the search results to pre determined categories like “science”, “entertainment”, “politics”, “sports”, etc.  
   -*Association rule mining* can be used to assess which search queries are followed by other search queries and they can be later classified to broader categories. To optimize the user experience, recommended suggestions can be given using association rule mining.  
   - *Anomaly detection* can be used to understand the content that is going viral. When the user traffic is high, it can be understood that the piece of content is going viral and can be pushed to the top of recommendations to increase visibility for such content.
3. For each of the following data sets, explain whether or not data privacy is an important issue.
   1. Census data collected from 1900-1950  
      -No, the data collected was consensually given and the person understood that the data given was used for government use.
   2. IP addresses and visit times of Web users who visit your website  
      -Yes, IP addresses are sensitive and require careful handling to avoid data breaches.
   3. Images from Earth-orbiting satellites  
      -No, they do not pose a threat to data security as they do not target individual users.
   4. Names and addresses of people from the telephone book  
      -No, the information in a telephone book is consensually given and the people understand that their address and names are going to be in a public domain.
   5. Names and email addresses collected from the Web  
      -No, the information is given out in a public domain and it does not involve a data breach as long as more personal information is not extracted from the two data fields.

**Task 2: practice your critical thinking and writing**

Read the following two news articles. One criticized Google Flu Trend, and the other defended it. Write one paragraph to summarize the criticism, and another paragraph for the defense. Write the third paragraph to offer your own thought, e.g. is the criticism valid? Does the defense make sense? What other problems or benefit do you see in Google Flu Trend or similar big data applications?

<http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/>

<http://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/>

The NY Times article criticizes the Google Flu trends for the inflated numbers as compared to the CDC numbers. It reasons that it has consistently overstated the numbers by a big margin of 50% of the CDC findings even after deliberate efforts to fine tune the algorithm. A team led by Northeastern political scientist David Lazer claimed that the stand alone data from CDC findings was a better predictor than Google Flu trends. The authors reason that the Flu trends suffers from *big data hubris* and state that big data sets do not trump traditional reporting solely on the basis of that size of the data. While the authors seem to be in favor of big data and analytics, they are skeptical of the Flu trends algorithm.

The Atlantic article counters the claims made by the team led by David Lazer and highlights the point that the Google Flu trends when combined with the CDC findings provide better results than the CDC findings alone. It also contends that Google Flu trends was not meant to be an all encompassing solution and as long as it proved to be beneficial to the CDC, it was a good outcome. It provides great value to the CDC and the media hype initially created when Google Flu trends worked proved counteractive to its own success.

The biggest enemy in the whole situation was the high expectations of people from big data analytics. Big data and its analysis requires a lot of fine tuning and multiple iterations. Although efforts are underway to make the algorithms more efficient, there will always be room for error. The intention to solve a particular problem is also very important and needs to be conveyed directly. Since Google Flu Trends was never meant to be a replacement for the CDC findings, and was rather meant to be a tool to validate and elevate the findings. Not to excuse the highly inflated values, but it was rather prematurely disregarded by the Northeastern scientist and his team.

Given that we are in the middle of a pandemic and expert epidemiologists say that the risk of another zoonotic viral outbreak is inevitable, a tool like Google Flu trends which tracks the trends of common symptoms can prove to be very beneficial. Data breaches and data security are of utmost importance when it comes to healthcare related areas. Such data in the wrong hands can be used to exploit people of both their rational and irrational fears regarding health. I would place the regulation of data laws and strict penalties for data breaches just as highly as improving the algorithms.