Nick Videtti

IST-707 Summer 2022

Homework 1

**Task 1:**

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.

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.
      1. **Yes**
   2. Dividing the customers of a company according to their profitability.
      1. **Yes**
   3. Computing the total sales of a company.
      1. **Yes**
   4. Sorting a student database based on student identification numbers.
      1. **No**
   5. Predicting the outcomes of tossing a (fair) pair of dice.
      1. **Yes**
   6. Predicting the future stock price of a company using historical trends.
      1. **Yes**
   7. Monitoring the heart rate of a patient for abnormalities.
      1. **Yes**
   8. Monitoring seismic waves for earthquake activities.
      1. **Yes**
   9. Extracting the frequencies of a sound wave.
      1. **Yes**
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.

**Data mining could help the Internet search engine company in many ways. Association rule mining would help give better results by ordering results in terms of popularity of past searches for the same query. Other algorithms could help aid in misspelled queries to direct a user to the closest correctly spelled query. Anomaly detection would be good for identifying any query results that are clicked on with very high or very low frequency to help with the sorting of results. Text mining would also be a good technique to use for identifying the most relevant query results.**

1. For each of the following data sets, explain whether or not data privacy is an important issue.
   1. Census data collected from 1900-1950.
      1. **Yes**
   2. IP addresses and visit times of Web users who visit your Website.
      1. **Yes**
   3. Images from Earth-orbiting satellites.
      1. **Yes**
   4. Names and addresses of people from the telephone book.
      1. **No**
   5. Names and email addresses collected from the Web
      1. **Yes**

**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 criticism of the Google Flu Trend was mainly in its accuracy. It was proven that simply looking at the current trends was far more accurate as a predictor than Google Flu Trend was. As of the end of the 2013-2014 Flu season, the last few Flu seasons consecutively had been much less substantial than Google Flu Trend had predicted. It was said that 100 weeks were over-predicted in a 108 week span. The creators of Google Flu Trend also admitted consistently throughout the implementation of Google Flu Trend that it was not made to be used without being supplemented with lab studies/research and other Flu trend indicators that we already had.**

**The support of Google Flu Trends had to do more with looking back at how it performed against its own initial goals. If you think about Google Flu Trends as an additional tool rather than an end-all-be-all, it actually was successful. The Atlantic contributes a lot of the criticism of Google Flu Trends to the fact that it did not reach the unrealistic expectations, or live up to the over-hype of its critics.**

**I personally would side with Google Flu Trends being considered a success. I do not think that the criticism is fully valid, because Google Flu Trends did a good job at what it was supposed to, even if it didn’t really do anything more than that. Another thing to consider is that Big Data was an up and coming field and that something like Google Flu Trends was a novel innovation. Even though “Back to The Future 2” had flying cars in 2015, that doesn’t mean that we are at all disappointed with all the significant advancements that have been made in the auto industry since 1985. I would be interested to hear how the Google Flu Trend co-creator’s new company is doing, given the advancements in the Data Science world over the last decade.**