Joyce Woznica

jlwoznic@syr.edu

October 11, 2019

Joyce Woznica  
Homework 1 Submission

Contents

[Task 1 2](#_Toc21435655)

[Question 1 2](#_Toc21435656)

[Question 2 3](#_Toc21435657)

[Question 3 3](#_Toc21435658)

[Task 2 4](#_Toc21435659)

[Introduction 4](#_Toc21435660)

[Summaries 4](#_Toc21435661)

[First Article: “Google Flu Trends: The Limits of Big Data” 4](#_Toc21435662)

[Second Article: “In Defense of Google Flu Trends” 4](#_Toc21435663)

[Personal Thoughts 5](#_Toc21435664)

# Task 1

This section answers the questions 1 through 3 in the textbook section entitled 1.7 Exercises on page 16.

## Question 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.

Response: No. This is a simple database query and not a data mining task.

1. Dividing the customers of a company according to their profitability.

Response: No. This is a calculation which is not a data mining task.

1. Computing the total sales of a company.

Response: No. This is not a data mining task. It is a calculation which could then become a data element/variable/observation that is tracked in the data to be mined.

1. Sorting a student database based on student identification numbers.

Response: No. this is a simple sort and not a data mining task.

1. Predicting the outcomes of tossing a (fair) pair of dice.

Response: No. Prediction the outcome of tossing a fair pair of dice is probability not data mining. However, if the dice was not considered “fair” an estimate the probabilities of the outcomes could be determined which would align this activity with data mining.

1. Predicting the future stock price of a company using historical records.

Response: Yes. The prediction of future stock price from historical records is a data mining task. This could be done by creating a regression model using historical records to help build the prediction model. This is predictive modeling.

1. Monitoring the heart rate of a patient for abnormalities.

Response: Yes. monitoring for abnormalities is a data mining task called “anomaly detection.” A model can be built of the normal heart rate and then a notification when something other than the normal rate occurred.

1. Monitoring seismic waves for earthquake activities.

Response: Yes. A model could be created that looks at the different types of seismic waves that are associated with earthquake activities and notification could be done when one of these occurred. This would be an example of classification.

1. Extracting frequencies of a sound wave.

Response: No. This is signal processing not data mining.

## Question 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?

There are several areas in data mining that could be used to improve this company’s search engine. The application of clustering, classification, association rule mining and anomaly detection are noted below.

1. **Clustering** is grouping results that have a similar characteristic or something in common and then presenting this in a way that is more concise and easier to understand. In this context, using something like synonyms is a great way to group things. For example, if I search on “bank”, the results should also contain ones that include “financial institution” or “investment firm” or “brokerage.”
2. **Classification** is finding a set of criteria that describe or represent data. This classification can be used to help predict the class of an object that is not initially known. For example, this is common on news websites where news is grouped by local, breaking, and national. By classifying content with a specific news type classification, searches can return additional information related to the topic of interest.
3. **Association rule mining** is discovering rules that show a correspondence of attribute and value that occur together often. This is very common when doing shopping. For example, on Amazon, if searching on paint, the user is often presented with things that were purchased when paint was also purchased, like a paint brush or stir stick. Implementing association rule mining could help improve search results by anticipating what the user may be looking for and returning like items.
4. **Anomaly detection** is finding anomalies or unusual patterns in data that do not conform to the normal patterns. This is very important in search engines. The search engine can use this to avoid showing something that is not relevant or to change a marketing strategy based on new products that are searched that were not popular initially.

## Question 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.

Response: No. Data privacy is not an important issue with census data collected over this time frame.

1. IP addresses and visit times of Web users who visit your Website.

Response: Yes. This is private information that can be traced back to the user’s computer and data privacy is important with IP addresses.

1. Images from Earth-orbiting satellites.

Response: No. There are no privacy issues associated with these images.

1. Names and addresses of people from the telephone book.

Response: No. This is public information published in the phone book. Individuals can elect to have an unlisted number if they do not want to appear, so privacy is not an issue in this situation.

1. Names and email addresses collected from the Web.

Response: No. There are no privacy issues with information collected here because individuals elected to put this data into the web.

# Task 2

## Introduction

Google Flu Trends (GFT) was an idea that if Google monitored the health of millions of users by tracking their search queries, the data could be used to help predict if a flu epidemic or the disease was on the rise in a certain population. In this exercise, two articles were presented each providing a different evaluation on the success or failure of GFT.

## Summaries

### First Article: “Google Flu Trends: The Limits of Big Data”

The first article written by Steve Lohr appeared in the New York Times in March of 2014. This article criticizes Google Flu Trends using the fact that GFT overestimated the number of cases by exaggerating them by over 50% through the 2012-2013 flu season. Lohr summarizes the article from four data analysts that states that GFT overestimated its value and was questionable as a standalone flu prediction tool. In fact, the authors of the article summarized, “The Parable of Google Flu: Traps in Big Data Analysis,” state that the lag of two weeks to obtain the report from the Center for Disease Control (CDC) would have been more accurate than what GFT provided. These scientists felt that GFT did not use a wider range of data analysis tools combining the CDC data with the Google data to perform the analysis. Co-founder of GFT, Matt Mohebbi, states that GFT could and did provide advance warning of the 2007-2008 flu season and that this does add value. He also recognized that they were “only at the beginning of what’s possible with this big-data-style analysis.” However, the authors of the original article elected to write about how big data had failed instead.

### Second Article: “In Defense of Google Flu Trends”

The second article written by Alexis Madrigal during the same time frame give an alternative opinion of Google Flu Trends. After spending some time capturing the negative sentiment of most articles around GFT, the author notes that the paper summarized in the previous article actually stated that if the GFT data and the traditional monitoring done by the CDC were combined, the result is better than that either alone. GFT was never designed to replace the CDC data, but together with this information, it seems that helped provide a clearer picture of the status of the flu. GFT was always built as a complimentary tool to the CDC, not to replace it or do a better job. In fact, with the GFT data available combined with other data captured in hospitals, doctors’ offices, meteorological data and other information, better and more complete models can be built and have been built. Madrigal implies that not only did GFT set out to do what it originally intended, but it exceeded its own expectations.

### Personal Thoughts

Data analytics can be helpful and meaningful, but the impact of that assistance is often interpreted differently. Understanding the underlying goals of the data and analytics provided help to position the results in a true and open manner. The second article did a more complete job at summarizing Google Flu Trends and its overall goals and then backing that with how those goals were met. The critical article seemed hyper focused on debunking that data analysis in this manner solves anything instead of focusing on the benefits provided. The defense makes more sense and seems move valid than the criticism as the defense documents benefits seen and shown. GFT has significant influence in how different data sets can be used to enhance an overall picture of a situation. This has benefits within health verticals as well as other industries where presenting a more complete picture with more variables considered can provide a better model.