Name: Hao Zhang

WustlKey: h.zhang633

ID: 452003

Name: Hanming Li

WustlKey: lihanming

ID: 451802

Homework 1

Problem 1

(a) The log data has the big data property of amount of records, because the data is updated very fast and the new records come up every second; and the property of infinity, because of the data comes every second and it is time dependency; and the property of structure, because the records in the data all have the same regular pattern; and the property of labels, because every record has a label of time. But it does not have the property of size of each record, because the records of the data are simple and the data points are one-dimensional.

(b) The Database of Wikipedia articles has the data property of amount of records, because the great amount of articles in the Wikipedia; and the property of size of each record, because every article has many words and information in it, and it is complex; And they all have the property of label, because each article has a title as its label. But it does not has the property of infinity, because the article will not be added fast; and it does not has the property of structure, because the articles are different from each other.

(c) The Database of chemical compounds has the big data property of amount of records, because the great amount of kinds of compounds; and the property of size of each record, because every record in the database is complex and high-dimensional, which has several different kinds of information; and the property of structure, because every record has a same structure; and the property of labels, because each data has labels such as bond ID, atom ID. But it does not has the property of infinity, because the amount of compounds in the database will not change fast with time.

(d) The data set described in (a) use text to record the information directly, such like writing the time and the data. But the data set described in (b) use text to tell something in language, it uses sentence to express the information rather than give the information through data directly. This differences make the analysis tasks different, for the data set in (a), the information can be extracted directly from the records, and the search can be done depends on the time, but for the data set in (b), the information need to be get by understanding the sentence and article, and the search need to be done base on the key words in the text.

(e) For (a), the data points are the record of each time, they are represented by writing the records of data depend on time. For (b), the data points are each article, they are represented as articles. For (c), each data point is the data of a kind of chemical compound, it is represented as the information of each kind of compound.

Name: Hao Zhang

WustlKey: h.zhang633

ID: 452003

Name: Hanming Li

WustlKey: lihanming

ID: 451802

Problem 2

A group of p people visit a hotel on any given day’s probability is (0.01)^p, because the probability for each people to visit hotel on any given day is (0.01).

The probability that they visit the same hotel is (0.01)^p/((10)^5)^(p-1), because the probability of p people visit a same hotel of 10^5 hotels is (1/10^5)^(p-1).

And the number of possible groups of p people is (10^9 p) ≈ ((10^9)^p)/p!.

The number of d days is (1000 d) ≈ (1000^d)/d!.

So the number of evil-doers is:

f=((0.01)^p/((10)^5)^(p-1))*(((10^9)^p)/p!)*((1000^d)/d!)=((10)^(p+3d+5))/p!d!

Name: Hao Zhang

WustlKey: h.zhang633

ID: 452003

Name: Hanming Li

WustlKey: lihanming

ID: 451802

Problem 3

(a) The unlabeled data cannot be extracted directly according to their attributes, semantic or parameters, but can be get through applying some statistical approach to analysis the basic parameters, the unlabeled data are much more plentiful and have more detail information in them. The labeled data are the data can be extracted directly by searching their attributes, semantic or parameters, which are easy to be get, but not that much.

(b) The data-based approach first makes all the data structured into tables, and each column in the tables has a herder, which is the shared attribute or parameter of all the data in this column. Then it finds out the relationships between the columns through some analysis, such as when the data appears, where are the data are used, and use this relationship to find out the information and detail data from the data in the tables.

(c) For the great amount of tables and columns in them, the data are put in them in different ways according to different criterion, so it is possible that the same meaning can be expressed in different ways, and the same expression can express different meanings, which make it difficult to infer relationships between column headers and make connection between data.