Name: Ren Wan
WUSTL key: wanren

# Homework 1

1. (a) Dimensions of log data

   i. Date

   ii. Time

   iii. IP address

   iv. File path

   v. Loading time

   vi. File size

   (b) Dimensions of Wikipedia articles
   The number of dimensions would be the number of occurrences of unique words in the text.

   (c) Dimensions of chemical compounds

   i. Bonds

   ii. Bond IDs

   iii. Atom IDs

   iv. Property of the chemical

   Difference between those datasets:
   *Amount*: They are all big, but chemical compounds dataset should be much larger.
   *Dimensionality*: Log data and chemical compounds are fixed dimension. Wikipedia articles is not.
   *Infinity*: They are all infinity.
   *Structure*: They all have relations between data points. Maybe graphs.
   *Label*: Log and chemical compounds can be labeled by one of the dimensions. Wikipedia can be labeled by machine learning models.

2. The probability for a group of p people go to the same hotel on d different days is $(\frac{0.01^p}{100000^{p-1}})^d$

   The number of groups of p people is $\binom{10^9}{p} = \frac{10^{9p}}{p!}$

   The number of d days is $\binom{1000}{d} = \frac{1000^d}{d!}$

   $f = (\frac{0.01^p}{100000^{p-1}})^d * \frac{10^{9p}}{p!} * \frac{1000^d}{d!} = \frac{10^{(-2p-5(p-1))d+9p+3d}}{p!d!} = \frac{10^{-7pd+9p+8d}}{p!d!}$

3. (a) Differences between unlabeled and labeled/annotated data: Annotation is not only slow and expensive to acquire but also difficult for experts to agree on. Unlabeled data is much more plentiful than labeled data.

   (b) Data-based approach: Models counts the number of occurrences of each n-gram sequence from a corpus of billions or trillions of words and automatically learn useful semantic relationships from the corresponding results or from the accumulated evidence of Web-based text patterns and formatted table.

   (c) Limitation: This approach needs very large corpus, otherwise the results are poor.

4. Map outouts:

   $map(15) = [(3, 15), (5, 15)]$
   $map(21) = [(3, 21), (7, 21)]$
   $map(24) = [(2, 24), (3, 24)]$
   $map(30) = [(2, 30), (3, 30), (5, 30)]$
   $map(49) = [(7, 49)]$

   Reducer inputs and outputs:

   $reduce(2, [24, 30]) = (2, 54)$
   $reduce(3, [15, 21, 24, 30]) = (3, 90)$
   $reduce(5, [15, 30]) = (5, 45)$
   $reduce(7, [21, 49]) = (7, 70)$