Hw1

Letian Chang

Student ID:445121

Problem 1:

(a) Log data:
amount： Huge amount data
Dimensionality: complex data entities
Structure: unstructured text.
Infinity: Infinite data
Labels: label information

(b) Wikipedia articles:
amount： Huge amount data growing every day.
Dimensionality: large features.
Structure: unstructured text.
Infinity: Infinite data
Labels: dependent variables.

(c) Database of chemical compounds:
amount： Huge amount of data points with complex data entities. Keeping increasing infinite.
Dimensionality: complex data entities
Structure: Structured with special rules.
Infinity: Infinite data
Labels: Have label information.

I think the main difference between them is the structure and dimensionality. (a) is unstructured and complex data entities (b) is unstructured and large features(c) is structured and complex data entities.

Problem 2:

The number of pairs of people is $\binom{10^9}{p}$

The number of pairs of days is $\binom{1000}{d}$

The chance that they will visit the same hotel on d given day is $(0.01^p/100000^{p-1})^d$

So f=$(0.01^p/100000^{p-1})^d * \binom{1000}{d} * \binom{10^9}{p}$

Problem 3:

(a) Labeled/annotated data may be slow and expensive to acquire and also difficult for experts to agree on. They always be classified and often used in supervised learning.

For an unlabeled data, it can be expressed in many different ways and the same expression can express many different things. Unlabeled data often used in unsupervised learning.

(b) The data-based approach used a big amount data in simple models and trump more elaborate models based on less data. Introduce general rules only when they improve translation over just memorizing particular phrases. Millions of specific features perform better than elaborate models that try to discover general rules. For many tasks, once we have a billion or so examples, we essentially have a closed set that represent what we need, without generative rules. Relying on big amount data has the further advantage that we can estimate models in an amount of time proportional to available data and can often parallelize them easily.

(c) Because of huge shared cognitive and cultural context, linguistic expression can be highly ambiguous. The same meaning can be expressed in many different ways, and the same expression can express many different meanings.

Ontology writing: There's a long tail of rarely used concepts that are too expensive to formalize with current technology.

Difficulty of implementation: The vast majority of small sites and individuals will find it too difficult, at least with current tools.

Competition: In some domains competing factions each want to promote their own ontology.

Inaccuracy and deception: We don't have an established methodology to deal with mistaken premises or with actors who lie, cheat, or otherwise deceive.

Problem 4:

Mapper inputs i= {15, 21, 24, 30, 49}

Mapper outputs: 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,[49])=(7,49)