**Assignment 1**

**yuweic3**

**2017/9/7**

1. **Information Retrieval Background**
2. Technique of “raw term frequency with a dot product similarity” treats every word with equal importance. But in fact some less important words such as “is” and “the” appear more frequently than keywords in corpus and unavoidably decrease the weight of the keyword. Normalize the unrelated words can make those keywords stand out easily.
3. According to the definition of IDF:

when we add a copy of to the corpus,

i) it will increase the values for those words not in .

ii) it will decrease the values for those words in as

1. Precision = Recall =

=

average precision =

1. Cumulative Gain at 7 documents = 4
2. Normalized Discounted Cumulative Gain at 7 documents:
3. **Probabilistic Reasoning and Bayes Rule**

a)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 0 | 1/3 | 1/2 | 1/3 | 1/2 |
| 1 | 4/6 | 5/6 | 1/3 | 1/2 |

b)

Similarly,

Since

We conclude that message does not carry a virus.

c) If we use the sample data,

Clearly, we get a different result from problem b), maybe because and are not independent.

d) There is a rule to obey: prior + prior = 1

e) Yes. We can modify value and make greater than . For example, change to in sample 7.

1. To fully compute the conditional probability, we need to specify at least 14 probability values.
2. When 2 () x 2 () x 2 () – 1 (distribution sum to 1) = 7
3. When 2 () x 2 () x 2 () – 1 (distribution sum to 1) = 7

Therefore, we need at least 7+7 = 14 probability values.

g) Because in reality event A, L and K are probably correlative.

**3 Maximum Likelihood Estimation**

a) The log likelihood is maximized at

To solve :

Therefore, the MLE for the Poisson distribution is

b) The maximum a posteriori of is:

Similarly, we apply the logarithm

then compute derivate:

Therefore, the MAP for the Poisson distribution is