**CS 5316 – Natural Language Processing**

**Quiz 4**

**(Time limit: 12 minutes)**

Instructions: (1) Please write legibly. Unreadable answers will NOT be graded; (2) Write in the spaces provided for the questions only.

1. (5 points) Consider the following four documents. (a) construct a term-document matrix using counts (lemmatize the vocabulary), (b) compute the relatedness between terms ‘beep’ and ‘beat’, and (c) compute the similarity between documents d1 and d2.

D1: beep beep beep emergency

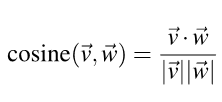
D2: heart attack emergency

D3: heart rate beat

D4: beep attack beat

(Continue this on the back-side)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Beep | Beat | Emergency | Heart | Attack | Rate |
| D1 | 3 | 0 | 1 | 0 | 0 | 0 |
| D2 | 0 | 0 | 1 | 1 | 1 | 0 |
| D3 | 0 | 1 | 0 | 1 | 0 | 1 |
| D4 | 1 | 1 | 0 | 0 | 1 | 0 |



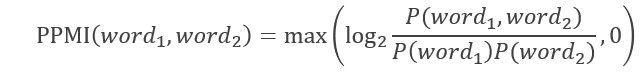
b) Relatedness=dotproduct (beep, beat)/|beep|\*|beat|=1/ (sqrt (9+1) \*sqrt(1+1))

=1/sqrt (10) \*sqrt (2)

c)Similarity= dotproduct (D1, D2)/|D1|\*|D2|=1/sqrt (10) \*sqrt (3)

=1/sqrt (10) \*sqrt (3)

1. (2 points) Consider the above corpus as running text. Determine the PPMI value of the term ‘attack’ in the context term ‘beep’. Assume context window is +1 and -1 terms/words.



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Beep | Beat | Emergency | Heart | Attack | Rate |
| Beep | 4 | 0 | 1 | 0 | 1 | 0 |
| Beat | 0 | 0 | 0 | 0 | 1 | 1 |
| Emergency | 1 | 0 | 0 | 0 | 1 | 0 |
| Heart | 0 | 0 | 0 | 0 | 1 | 1 |
| Attack | 1 | 1 | 1 | 1 | 0 | 0 |
| Rate | 0 | 1 | 0 | 1 | 0 | 0 |

=max(log\_2( (1/18) / (4/18 \* 6/18)),0)

1. (2 points) Suppose there are M features in a text classification problem. How many parameters need to be learned for a logistic regression model when there are (a) two classes (binary classification problem), and (b) 5 classes (multi-class classification problem).
2. M+1 if using mean square loss, with single output neuron and 2M+2 with two output units using cross entropy loss.
3. 5M +5 five output units