CSCI 7000-IR Fall 2008 Quiz 2

Name: ________________________

On my honor, as a University of Colorado at Boulder student, I have neither given nor received unauthorized assistance on this work. ________________________________

1. **(5 points)** True or False: Microsoft is in the search business

2. **(10 points)** Use naïve Bayes (multinomial w/ laplace smoothing) to classify the specified document into the appropriate category. Use the data on the accompanying sheet.

3. **(5 points)** True or false: A Bernoulli (binomial) approach to naïve Bayes text classification makes use of information about the words that don’t occur in a document.

4. **(5 Points)** In which one of the various text classification methods we’ve studied, does classification time dominate over training time?

5. **(10 Points)** Briefly describe two applications of clustering in information retrieval.

6. **(5 points)** True or false: Agglomerative clustering algorithms require as a parameter the desired number of clusters.

7. **(10 points)** Describe the basic steps involved in K-means clustering.

8. **(10 points)** One of the critical steps in K-means can be particularly expensive. Describe the step and why it’s expensive and suggest a way to mitigate the expense.