

Homework 9

Student Number:

Name:

Problem 1. (20 points) The rationale for the positional independence assumption is that there is no useful information in the fact that a term occurs in position k of a document. Find exceptions and give explanations.

Problem 2. (40 points) Assume the number of documents in the training set is N , the vocabulary size of used terms are D (dimension of the vector space model), and the number of classes is C .

1. What is the time complexity of KNN in testing (assume $K=1$, no preprocessing and optimization on distance calculation) ?
2. What is the time complexity of Naive Bayes in testing ?
3. Under which conditions will you prefer Naive Bayes to KNN for deployment (list at least two conditions) ?

Problem 3. (40 points) Which of the following documents have identical and different bag of words representations for (i) the Bernoulli model (ii) the multinomial model? If there are differences, describe them.

- 1) He moved from London, Ontario, to London, England.
- 2) He moved from London, England, to London, Ontario.
- 3) He moved from England to London, Ontario.