Due: 2022/04/29

Homework 10

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) Assume a situation where every document in the test collection has been assigned exactly one class, and that a classifier also assigns exactly one class to each document. This setup is called one-of classification. Show that in one-of classification

1. the total number of false positive decisions equals the total number of false negative decisions
2. micro-averaged F_1 and accuracy are identical.