**INFORMATION RETRIEVAL \_ ASSIGNMENT 7**

**Content**: Text Classification

**1. PREPARING:**

1.1. Go to page

<https://archive.ics.uci.edu/ml/machine-learning-databases/20newsgroups-mld/>

1.2. Download and extract 20\_newsgroups.tar.gz

**2. TRAINING & CLASSIFYING**

2.1. Choose randomly from each category 5000 documents for testing.

2.2. Choose n0 documents from the 15000 remaining for training.

2.2. Perform training and classifying by the following methods:

- K-Nearest Neighbors

- Naïve Bayes for Text

with the recommended vocabulary size is 2000 (randomly selected from the training data)

2.3. Perform the above procedure with n0 = 1000, 3000, 5000,…, 15000 and draw a curve for displaying the performance. *(Note: When n0 varies, the smaller training set should be a subset of the larger training set).*