Midterm Exam (6:40pm – 8pm, April 9th) 20% of overall grade

20 total points

Part A (45 minutes) – open notes, closed laptop

1. **Naïve Bayes: Given a group of documents (4 points).**
    a. Joint, conditional, marginal probabilities
    b. Checking for independence
    c. Calculate likelihood, prior, evidence, and posterior
2. **Vectorization and Similarity: Given a group of documents (4 points).**
    a. Generate count vectorization, one-hot encoded vector, TF-IDF
    b. Question about cosine similarity
    c. Question about Euclidean distance
    d. Question about PMI
3. **Classification: Given predictions (y_pred) and actual results (y_test) (4 points)**
    a. Model evaluation: compute accuracy, precision, recall, F1 score, confusion matrix
    b. Question about interpreting model results

Part B (35 minutes) – open everything

1. **Regular expression, text preprocessing, classification: given a sample small text corpus (7 points):**
    a. Process data given certain constraints that will be provided at test time.
        i. numbers should not be included
        ii. proper names should be removed
        iii. only words longer than 3 characters should be included
    b. Find most similar documents
    c. Train a basic classifier on corpus and report model performance.

2. **Task involve likelihood of documents and perplexity (5 points).**