Midterm Exam (6:40pm – 8:05pm, April 9th) 20% of overall grade
20 total points
Note that I may ask either conceptual or calculation questions!

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

1. **Naïve Bayes: Given a group of documents (4 points).**
    i. .25 conditional probability
    ii. .25 marginal probability
    iii. .25 joint probability
    iv. .25 conditional probability

    a. **Checking for independence (1 pts):**
        i. .5 correct definition of independence
        ii. .5 correct explanation and answer

    b. **Calculate likelihood, prior, evidence, and posterior:**
        i. .50 correct prior
        ii. .50 correct evidence
        iii. .50 correct likelihood
        iv. .50 correct posterior

2. **Vectorization and Similarity: Given a group of documents (4 points).**
    a. **Generate count vectorization, one-hot encoded vector, TF-IDF:**
        i. .25 for correct word count vector
        ii. .25 for correct one-hot encoded vector
        iii. 1 for TF IDF vector (.5 for correct IDF, .5 for correct TF)
    b. **Question about cosine similarity**
        i. .25 correct calculation of norms
        ii. .25 correct calculation of dot products
        iii. .50 correct answer to question
    c. **Question about Euclidean distance**
        i. .50 correct calculation of distances
        ii. .50 correct answer to question
    d. **Qualitative question about distance / similarity in NLP**
        i. .25 for valid reason 1
        ii. .25 for valid reason 2

3. **Classification: Given predictions (y_pred) and actual results (y_test) (4 points)**
    a. **Model evaluation:**
        i. .25 correct accuracy
        ii. .25 correct precision
        iii. .25 correct recall
        iv. .25 correct F1 score
        v. 1.0 correct confusion matrix
    b. **Qualitative question about interpreting model results**
        i. 0.5 Correct metric chosen
        ii. 0.5 Explanation is valid
    c. **Qualitative question about interpreting model results**
        i. 0.5 Correct answer
        ii. 0.5 Explanation is valid

Part B (40 minutes) – open everything

1. **Regular expression, text preprocessing, classification: given a sample small text corpus (7 points):**
        i. 1 data loaded, encoding correct
        ii. 1 creating correct target variable
        iii. 1 regex, stopwords, tokenization applied according to instructions
        iv. 1 similarities are calculated correctly
        v. 1 Uses correct model evaluation methodology
        vi. 1 Model metrics are interpreted correctly
        vii. 1 TF-IDF business recommendation is explained correctly

2. **Task involve likelihood of documents and perplexity (5 points):**
        i. 1 Likelihood for a single word is correctly calculated
        ii. 1 Likelihood for a word beginning a sentence is correct
        iii. 1 Likelihood of a particular bigram is calculated correctly
        iv. 1 Likelihood of two test sentences are calculated correctly
        v. 1 Perplexity is calculated correctly