

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

20 total points

Part A (45 minutes)

- 1. Naïve Bayes: Given a group of documents (3 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}) (3 points)**
 - a. Model evaluation: compute accuracy, precision, recall, F1 score, confusion matrix
 - b. Question about interpreting model results
- 4. Word Embeddings (2 points):**
 - a. Question about CBOW versus skipgram high-level architecture
 - b. Question about interpreting embedding vectors.

Part B (35 minutes)

- 1. Regular expression, text preprocessing, classification: given a sample small text corpus (6 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. Find most similar words using word embeddings
 - d. Train a basic classifier on corpus and report model performance.
- 2. Task involve likelihood of documents and perplexity (4 points).**