



DSO 560 - Text Analytics & Natural Language Processing

Instructor: Yu Chen

Midterm Exam Outline

80 minutes (6:40pm - 8:00pm PST)

PICK 4 of the 5 SECTIONS TO COMPLETE

___/16pts

1. Classification & Model Evaluation (4 pts, recommended 20 minutes)

- a. **Given classification/actual results from two different models, compute**
 - i. Accuracy (0.25 pts)
 - ii. Precision (0.25 pts)
 - iii. F1 Score (0.25 pts)
- b. **Compute one of the following probabilities (likelihood, prior, posterior, marginal):**
 - i. (0.25pts)
- c. **Correctly construct confusion matrix (1 pt).**
- d. **Qualitative question and metrics to use for evaluation given business use case. (0.25 pts for correct metric, 0.75 pts for explanation)**

2. Naïve Bayes (4 pts, recommended 20 minutes)

You will be provided a labelled dataset.

- A. Qualitative question discussing text preprocessing steps needed (**1 pts** - 0.25 for correctly identifying steps involved, 0.75 for explaining why they are needed)
- B. Calculate the following probabilities for the dataset (**2 pts**):
 - a. Prior (0.5 pts)
 - b. Likelihood (0.5 pts)
 - c. Evidence (0.5 pts)
 - d. Posterior (0.5 pts)
- C. Question about independence (**1 pt - 0.5 pts** for correct answer, **0.5 pts** for explanation).

3. Vectorization and Similarity (4 pts, recommended 20 minutes)

You will be provided with a text dataset as well as the equations for term frequency and inverse document frequency to use.

- a. Generate from dataset TF-IDF vectors (**1pt - 0.25 for IDF, 0.25 for correct TF-IDF, 0.5 for TFs**)
- b. Business question using a similarity/ distance metric and computed vectors (**1 pt** - 0.5 correct answer computed, 0.5 for explanation)
- c. Word2vec business question (**1 pt** - 0.5 for a correct answer, 0.5 for business explanation)
- d. Similarity/ distance question about different documents (**1 pt**)

4. N-Gram Language Models (4 pts, recommended 20 minutes)

Given the following documents (these are examples of what may appear on the midterm):

1. *Schools are open.*
2. *He is late today.*
3. *I went to school late today.*
4. *He went to school late.*
5. *They went home after school.*

- A. Question computing transition matrix probabilities. (0.5 for correct preprocessing steps, 0.5 for transition frequency, 1 pt for correct transition matrix)
- B. Question about a specific n-gram probability of seeing a test document, may involve perplexity (**1pt**)
- C. Question about a specific n-gram probability of seeing a test document, may involve perplexity (**1pt**)

5. True/False (4 pts, recommended 20 minutes)

For each of the statements below, indicate if it is true or false. **Whether it is true or false, to earn full credit, provide an explanation and simple real-life example.** Each question is 1pt (0.5 for correct answer, 0.5 for valid explanation and example).