

# Final Exam Review – Part 1

# Final Exam Logistics

---

As scheduled by Registrar:

- Date: Weds Dec 6, 2023 8:00 a.m. 10:00 a.m.
- Where: WH 113 - **NOT** regular classroom

Other details:

- Remote proctoring and alternate arrangements are ONLY possible via IIT-Online office or CDR
- Students in remote sections and who do not live in the Chicago metro area should contact Charles (Chuck) Scott at IIT online.
- If applicable: contact CDR regarding accommodations

# Final Exam Format

---

## Format:

- 110 minutes
- Multiple choice, matching, or similar questions
- One correct answer per question
- No penalty for incorrect answers (no penalty for guessing)
- Answer on bubble sheet – bring pencils!

## Material:

- Covers material before and after midterm (up to Session 27) with a focus on material after the midterm
- Some questions apply concepts from both 1st and 2nd half of class.

---

# MULTIPLE CHOICE STRATEGIES

# Multiple Choice Strategies

---

1. Read the questions carefully
2. Answer the question without looking at the options
3. Eliminate the incorrect options
4. Answer all the questions
5. Manage your time

LINK (from UT-Mississauga):

<https://www.utm.utoronto.ca/asc/media/2073/download?inline>

---

# MIDTERM EXAM QUESTIONS

# Multiple Choice Question

---

Which of the following describes an appropriate application of KL Divergence? Select the BEST answer.

- A. ...
- B. ...
- C. ...
- D. ...

# Multiple Choice Question

---

Which of the following describes an appropriate application of KL Divergence? Select the BEST answer.

- A. Generating a coding system given a reference distribution of unigram words
- B. Evaluating a text classifier by analyzing how it will perform on unseen data, that is, how frequently the classifier's labels match the true labels
- C. Comparing machine-generated texts to human-generated texts, to evaluate the quality of a machine translation system
- D. Outputting a prior distribution over document topic labels



# Multiple Choice Question

---

Which of the following describes a difference between L1 and L2-Regularization? Select the BEST answer.

- A. ...
- B. ...
- C. ...
- D. ...

# Multiple Choice Question

---

Which of the following describes a difference between L1 and L2-Regularization? Select the BEST answer.

- A. L1-regularization is intended to increase model complexity, and L2-regularization is applied to decrease model complexity.
- B. L2 regularization tends to push model weights to zero, removing those features from the model, and L1-regularization does not have this property.
- C. L1- and L2-regularization differ in the type of vector normalization applied to penalize model parameters.
- D. L1-regularization penalizes unigram weights, and L2-regularization penalizes bigram weights.

# Multiple Choice Question

---

You are developing a word-sense-disambiguation model. Your classmate suggests that you evaluate your tool by generating a dataset of **pseudo-words** by collapsing two semantically unrelated terms into a single token (e.g., “door” and “banana” into “door-banana”).

What is a disadvantage of this approach? Choose the best answer

- A. ...
- B. ...
- C. ...
- D. ...

# Multiple Choice Question

---

You are developing a word-sense-disambiguation model. Your classmate suggests that you evaluate your tool by generating a dataset of **pseudo-words** by collapsing two semantically unrelated terms into a single token (e.g., “door” and “banana” into “door-banana”).

What is a disadvantage of this approach? Choose the best answer

- A. This approach requires additional hand-annotated data, which is expensive and time-consuming to collect
- B. The resulting Cohen’s Kappa statistic is likely to be low.
- C. The resulting dataset may not be representative of the challenges for word sense disambiguation that occur in natural language.
- D. You cannot use bag of words (BOW) methods with this approach

# Multiple Choice Question

---

You are building a bag-of-words (BOW) text classifier over English texts. Which of the following are appropriate **word normalization** steps that you may want to consider? Select the BEST answer

- A. ...
- B. ...
- C. ...
- D. ...

# Multiple Choice Question

---

You are building a bag-of-words (BOW) text classifier over English texts. Which of the following are appropriate **word normalization** steps that you may want to consider? Select the BEST answer

- A. Using space characters as a delimiter to segment the input text into words.
- B. Applying an English language stemmer to remove suffixes, such as "-s" and "-ing"
- C. Removing words from your training set that do not appear in your test set
- D. Removing texts from the test set that contain out-of-vocabulary (OOV) words

# Multiple Choice Question

---

Which of the following is NOT true of sentiment analysis? Select the BEST answer.

- A. ...
- B. ...
- C. ...
- D. ...

# Multiple Choice Question

---

Which of the following is NOT true of sentiment analysis? Select the BEST answer.

- A. Lexicon-based methods that use positive and negative word lists have been shown to be effective for sentiment analysis.
- B. Word embeddings, such as word2vec, are often unable to represent positive and negative sentiment.
- C. Sentiment analysis is often not a strictly two-class problem and may be best expressed over a scale, or with a "neutral" label.
- D. A sentence may contain multiple aspects with different sentiment polarity.



# Multiple Choice Question

---

You have been asked to build a named-entity recognition (NER) tool for a specific biomedical sub-domain. Your teammate suggests using TF-IDF, because this method was very successful for a document clustering task. What is an issue you may encounter with this suggestion?

- A. ...
- B. ...
- C. ...
- D. ...

[NOTE: This was not a mid-term exam question]

# Multiple Choice Question

---

You have been asked to build a named-entity recognition (NER) tool for a specific biomedical sub-domain. Your teammate suggests using TF-IDF, because this method was very successful for a document clustering task. What is an issue you may encounter with this suggestion?

- A. It may be difficult to find document-level annotation for this specific sub-domain
- B. TF-IDF removes word order which is important for NER
- C. Lexical resources such as dictionaries may not exist for this sub-domain
- D. All of the above are true