Big Data Analytics
Stony Brook University
CSE538 - Spring 2023

Exam 1 Review Guide
Exam Time: 3/5, 4:00pm

Suggestions for study:

1. Review the readings in syllabus schedule
2. Review your notes associated with the slides and topics below – use lecture recordings where unclear.
3. Review assignment 1 and quizzes (In Brightspace click Weekly Quiz -> title of the specific quiz -> in the quiz page you will see a down arrow besides the title of the quiz.)
4. Practice example questions below and exercises in the book. (only those related to the topics we've covered).
5. Ask questions on Piazza

Topics for the exam:

- Introduction to NLP
  - NLP's goal
  - Common Applications
  - Interpretation of garden-path sentences
  - Regular Expressions
    - character classes, character ranges
    - not characters
    - words, whitespace, and word boundaries
  - Probability
    - Sample space and events
    - Independence
    - Conditional Probability
  - Tokenization
    - Parts of speech
    - word tokenizer
    - BPE
    - word piece tokenization (conceptually)
- Maximum Entropy / Supervised Classification
  - Model definition (compared to linear function)
  - Likelihood function to Loss function
  - Separating hyperplane
  - Multivariate Features / loss function
  - Overfitting, L1 and L2 regularization
  - Feature extraction: one-hot, multi-hot representations

    - Train, dev, test split
- Syntax and Dependency Parsing
    - Relations (core universal dependency relations)
    - head and dependent
    - Transition-based dependency parsing
    - Projectivity
    - Idea of semantic roles and verbal predicates
- Lexical and Verbal Semantics
    - terminology (lemmas, homonymy, etc…)
    - different types of word sense disambiguation
    - Lesk algorithm
    - distributional hypothesis
    - concept of vector semantics
    - word2vec - skip-gram model
    - topic modeling – LDA
- Intro to LMs
    - 2 task versions and their equivalence
    - applications
    - chain rule, markov assumption
    - unigram, bigram LMs

Exam Day Procedure:

Student remain outside while exam is being setup. The door will open to the classroom approximately 5 minutes before start time.

Once entering, students should sit at a seat with an exam in front of it.

No items are permitted outside one's bag, backpack, or briefcase except:
(1) pen or pencil,
(2) SBU student id, and
(3) optionally, a bottle of water
No calculators or other materials are permitted on one's desk. The last page of the exam is scrap paper that can be torn off after the exam begins. This page must be turned in with a name on it.

Once the exam begins, put your name on every page of the exam.

Students will be given 5 minute and 1 minute warnings before the end of exam, which will be announced as "Exam is over. Pens and pencils down". **Once the exam ends, all exams must be closed and all writing utensils down. Students who do not close the exam or who do not put their pen down will receive a 50 point deduction off their exam grade.**

FAQ:

- **Will there be programming on the exam?**

    - Yes, as we have gone over and in the syllabus: Coding specific algorithms is a great way to both learn and demonstrate understand of concepts / algorithms.
    - A good rule of thumb is to just use python code, though pseudocode will be allowed.
    - As long as the key pieces of the algorithm are clear, small syntax errors won't count off (for example, missing a colon after a "for" will not count against you as long as the cope of the loop is clear. If you miss the colon and don't indent clearly, then it's unclear and can reduce points).
- **How much of the readings will be covered?**
    - The readings are considered required. To fully understand a topic, one should approach from multiple perspectives and the readings are different perspective, following the same concepts of the class. The readings also cover some concepts in more depth than we get to in class. Some of these are even pointed out in the slides. Therefore, the readings are very helpful toward mastery of the class concepts.
    - That said, typically, there will be around 1 question that covers material from the readings which wasn't covered as extensively in class.

Please post any other questions to Piazza by Sunday.

Example Questions.

1. Circle "T" if the statement below is true, circle "F" if the statement is false, and leave blank if you are completly unsure.

A. **T** / **F** The regular expression '[a-Z]+' will match 1 or more numbers or letters..

B. **T** / **F** Period disambiguation is the task of deciding if a period is marking the end of a sentence or is a part of the word.

C. **T** / **F** Regularization is the process of rebalancing classes so that they are uniform.

D. **T** / **F** The generalized markov assumption holds that for some k <= n,

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | X_{i-k}, X_{i-(k-1)}, ..., X_i)$$

2. Pseudocode: Assume LogisticRegression is a class that takes in "penaltyType" and "penalty". It further has a methods "fit(X, y)" to train the model (returns nothing; saves model in object) and "test(X, y)" to run predictions on X based on the model and return an accuracy. Further, X_train is a matrix of trainining data features and y_train is a matrix of training data clases. Please implement a method that finds the best penaltyType and penalty value over the training data.

3. Multiple Choice. In byte-pair encoding (BPE), assume we have a corpus, D=["hug", "pug", "pun", "bunk", "bun", "bunk"] the vocabulary after the first iteration is:

A. {"hug", "pug", "pun", "bun", "bunk"}
B. {"b", "g", "h", "n", "p", "k", "u", "un"}
C. {"b", "g", "h", "p", "k", "un"}
D. {"b", "g", "h", "n", "p", "k", "un", "ug", "pu", "bu"}
E. {"b", "g", "h", "n", "p", "k", "u", "bun"}

4. problem solving: Given the following bigram counts (Xi-1 as rows, Xi as columns). Total represents a total, including additional unigrams not shown. (Please show your work for all questions.)

	i	**love**	**my**	**iphone**	**UNIGRAMTOTAL**
i	**0**	**10**	**5**	**1**	**100**
love	**1**	**0**	**10**	**2**	**20**
my	**1**	**5**	**2**	**5**	**40**
Iphone	**0**	**1**	**1**	**1**	**10**

TOTAL UNIGRAM INSTANCES: 1000
TOTAL UNIQUE WORDS: 100

a) Assuming an bigram model, what is P(Xi = 'love' | Xi-1 = 'I')? 0.1

b) Assuming an bigram model, what is P(X1 = 'i', X2 = 'love', X3 = 'iphone')? 0.1*0.1*2/20 = 0.001

c) Assuming an add-one smoothed bigram model, what is P(Xi = 'love' | Xi-1 = 'I')? ← outside scope of exam

ANSWER KEY BELOW