

# Final Exam Review – Part 2

# 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!
- **Closed book – closed notes - no electronics**

## 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.

---

# TOPICS & CONCEPTS

# Sessions 1-12

---

- Vector-space model for text: BOW vs tf-idf
- Lexical representations: stems vs lemmas, n-grams, sub-word encoding (BPE, wordPiece)
- LSA (what is it based on?), word embeddings & word2vec
- Probability & Information theory concepts - conditional probability, independence, entropy, etc.
- Metrics: Precision, Recall, Accuracy
- GLMs and logistic regression, regularization, bias-variance tradeoff
- Text categorization, naïve bayes, logistic regression
- Supervised vs unsupervised learning, k-means, EM
- Generative vs discriminative methods
- FFNNs, gradient descent, cross-entropy loss
- NOT on exam: regular expressions

# Sessions 13-27

---

- Language models, ngram LMs, Markov assumption, smoothing
- Sequential tagging: NER, POS tagging
- Structured prediction (what is it?)
- HMMs and CRFs, Viterbi, EM for HMMs
- NN methods for sequential labeling: RNNs & LSTMs
- Syntax concepts: constituents, phrasal heads, structural ambiguity
- CFGs, production rules, CNF, terminal vs non-terminal
- Parsing constituents, treebanks, CKY, CFGs vs PCFGs
- Dependency grammars, dependency relationships, UD project, spanning & projective trees, free word order languages

# Sessions 13-27 (continued)

---

- Semantic role labeling, "shallow" parsing, FrameNet & PropBank, methods for SRL
- Machine translation: noisy-channel vs neural methods, word-alignment, parallel corpora, evaluation, BLEU
- Sequence-to-sequence models, attention
- Transformers, self-attention, fine-tuning, PLMs, LLMs
  - Reminder: Our guest lecture covered LLMs
- **Study question:** What topics from 1st half of class did we see again in 2nd half?

---

# SAMPLE QUESTIONS

# Sample Question

---

Which of the following is true regarding Naïve Bayes and Hidden Markov Models (HMMs)?

- A. HMMs are discriminative and NB are generative
- B. Naïve Bayes models are trained with Expectation Maximization\* and HMMs are trained with gradient descent
- C. HMMs encode transitions between hidden states and Naïve Bayes does not
- D. All of the above are true

(\*Is this ever true?)



# CFG Review: CKY Example

$S \rightarrow NP VP$   
 $NP \rightarrow ADJ NP$   
 $NP \rightarrow NP PP$   
 $VP \rightarrow V NP$   
 $VP \rightarrow VP PP$   
 $PP \rightarrow P N$   
 $ADJ \rightarrow \text{bright}$   
 $NP \rightarrow \text{night}$   
 $NP \rightarrow \text{shine}$   
 $NP \rightarrow \text{stars}$   
 $P \rightarrow \text{at}$   
 $V \rightarrow \text{shine}$   
 $V \rightarrow \text{stars}$   
 $VP \rightarrow \text{shine}$   
 $VP \rightarrow \text{stars}$

What values are in Cell A, B, C, D?

	$n$ (constituent start index)				
	bright	stars	shine	at	night
$m = \text{Constituent Length} - 1$	0	1	2	3	4
0	ADJ	NP, VP, V	NP, VP, V	P	NP
1	[CELL A]	[CELL C]	[CELL D]	PP	
2	S	NULL	VP		
3	[CELL B]	S, VP			
4	S				

# CFG Review: CKY Example

S → NP VP

NP → ADJ NP

NP → NP PP

VP → V NP

VP → VP PP

PP → P N

ADJ → bright

NP → night

NP → shine

NP → stars

P → at

V → shine

V → stars

VP → shine

VP → stars

What values are in Cell A, B, C, D?

CELL A: "bright stars: → NP

CELL B: "bright starts shine at" → NULL

CELL C: ...

CELL D: ...

	<i>n</i> (constituent start index)				
	bright	stars	shine	at	night
m=Constituent Length minus 1	0	1	2	3	4
0	ADJ	NP, VP, V	NP, VP, V	P	NP
1	[CELL A]	[CELL C]	[CELL D]	PP	
2	S	NULL	VP		
3	[CELL B]	S, VP			
4	S				