Practice Midterm Exam for Natural Language Processing

Name: _____
Net ID _____

Instructions

In the actual midterm there will be 7 questions, each will be worth 15 points. You also get 10 point for signing your name on all test materials, seriously, because when students forget to sign their names, I have to somehow figure out whose test a particular piece of paper belongs to. The maximum score on the test will be 115. You will have approximately 1:15 minutes to complete this test.

This practice test will have a different number of problems that are intended to be of the same basic type of question as on the actual midterm. **THE PRACTICE TEST IS DESIGNED TO TAKE LONGER TO COMPLETE THAN THE ACTUAL TEST WOULD (AROUND 2 HOURS, RATHER THAN 1:15).** The test materials will include this printout and one blank test booklet. I suggest that you fill in all answers directly on this printout and use the blank test booklet as scrap paper. However, if you run out of space, you have the option of using the test booklet. If you do this, please include a clear note on the test so I know where to look for your answer.

This test is an open book/open notes test: Please feel free to bring your text book, your notes, copies of class lectures and other reading material to the test. A calculator is also permitted and it is OK to look at materials on the web in order to read helpful information, being mindful of the time limit. Just don't use a program that solves a problem for you, e.g., do not find a part of speech tagger and run it if asked to manually annotate mark parts of speech – that WOULD be cheating.

Answer all questions on the test. If you show your work and you make a simple arithmetic mistake, but it is clear you knew how to do it, you will get partial credit.

- William R. Breakey M.D.
- Pamela J. Fischer M.D.
- Leighton E. Cluff M.D.
- James S. Thompson, M.D.
- C.M. Franklin, M.D.
- Atul Gawande, M.D.
- Dr. Talcott
- Dr. J. Gordon Melton
- Dr. Etienne-Emile Baulieu
- Dr. Karl Thomae
- Dr. Alan D. Lourie
- Dr. Xiaotong Fei
- Doctor Dre
- Doctor Dolittle
- Doctor William Archibald Spooner
- Doctor No

Figure 1: Correct Instances of Doctors in Our Corpus

Question 1. Write a regular expression for identifying names of doctors in text. Your regular expression should match the examples in figure 1, but should not recognize either non-names (words lacking capital letters) or names that do not include the identifying title information (*Dr.*, *Doctor*, *M.D.*). Do your best to include information about spaces, hyphens, commas and periods, as per the examples.

```
((Doctor|Dr\.)( [A-Z][a-z\.]+)+)|(([A-Z][a-z\.]+ )+M\.D\.)
```

Tag	Description	Tag	Description
CC	Coordinating conjunction	RB	Adverb
CD	Cardinal number	RBR	Adverb, comparative
DT	Determiner	RBS	Adverb, superlative
EX	Existential there	RP	Particle
FW	Foreign word	SYM	Symbol
IN	Preposition or subordinating conjunction	TO	to
JJ	Adjective	UH	Interjection
JJR	Adjective, comparative	VB	Verb, base form
JJS	Adjective, superlative	VBD	Verb, past tense
LS	List item marker	VBG	Verb, gerund or present participle
MD	Modal	VBN	Verb, past participle
NN	Noun, singular or mass	VBP	Verb, non-3rd person singular present
NNS	Noun, plural	VBZ	Verb, 3rd person singular present
NNP	Proper noun, singular	WDT	Wh-determiner
NNPS	Proper noun, plural	WP	Wh-pronoun
PDT	Predeterminer	WP$	Possessive wh-pronoun
POS	Possessive ending	WRB	Wh-adverb
PRP	Personal pronoun	PU	Punctuation
PRP$	Possessive pronoun		

Table 1: **Penn Treebank POS tags**

Question 2. Assign Penn parts of speech tags (as per Table 1) to all the words in the following two sentences using the notation word/POS:

a. *John/NNP and/CC Mary/NNP bought/VBD a/DT refrigerator/NN with/IN three/CD doors/NNS ./PU*

b. *It/PRP was/VBD purchased/VBN from/IN a/DT very/RB small/JJ store/NN near/IN their/PRP$ house/NN ./PU*

Question 3. Mark the noun groups in the following sentence using BIO (beginning, intermediate, other) tags.

Mary	B
has	O
a	B
room	I
with	O
a	B
view	I
and	O
a	B
bottle	I
of	O
beer	B

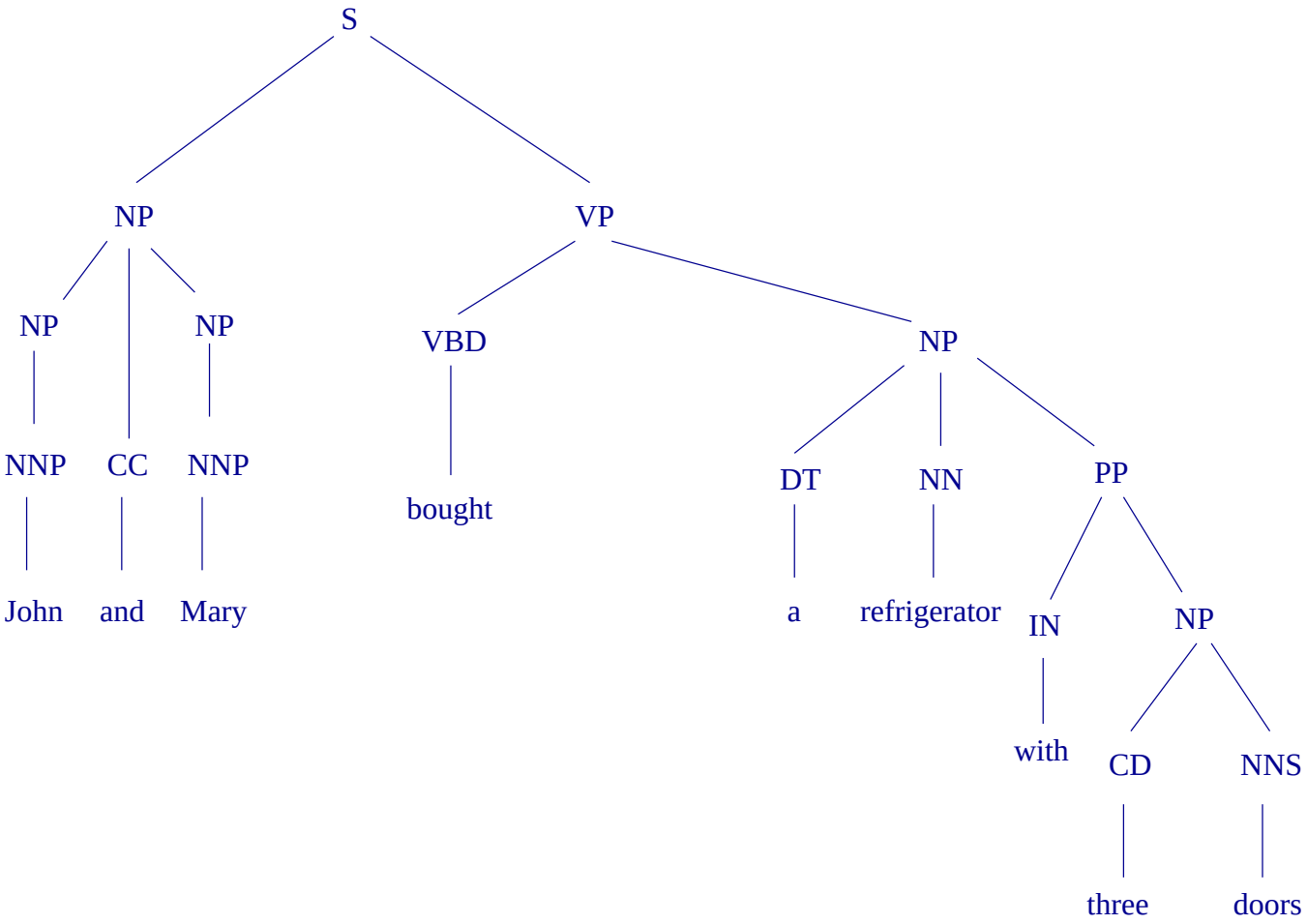

Figure 2: Possible Answer to Question 4

Question 4. Draw a Phrase Structure Tree representing one parse of the following sentence. Make a list of the phrase structure rules that you are assuming.

John and Mary bought a refrigerator with three doors .

1. $S \rightarrow NP\ VP$
2. $NP \rightarrow NP\ CC\ NP$
3. $NP \rightarrow DT\ NN\ PP$
4. $NP \rightarrow CD\ NNS$
5. $NP \rightarrow NNP$
6. $VP \rightarrow VBD\ NP$
7. $PP \rightarrow IN\ NP$
8. $NNP \rightarrow John$
9. $NNP \rightarrow Mary$
10. $NN \rightarrow refrigerator$
11. $NNS \rightarrow doors$
12. $CC \rightarrow and$
13. $VBD \rightarrow bought$
14. $DT \rightarrow a$
15. $CD \rightarrow three$
16. $IN \rightarrow with$

Question 5. Calculate precision, recall and f-measure in order to score the following system against the answer key. Assume any item reported by the system and found in the answer key is correct.

The system reports that the following strings of words describing *attack* events:

1. *Jay Leno attacked Conan O'brien.*
2. *attacks by the U.S.-backed rebels* Correct
3. *the latest in a series of attacks in the 10-year-old civil war.* Correct
4. *Mr. Baldwin is also attacking the greater problem: lack of ringers.*
5. *the criminals were convicted for bombings.* Correct
6. *The broadway musical "Bridges of Madison County" bombed.*
7. *Groupon fires CEO Andrew Mason.*

The answer key includes the following strings of words describing *attack* events:

1. *the martians bombarded the Earth with death rays*
2. *attacks by the U.S.-backed rebels* Found by System
3. *the latest in a series of attacks in the 10-year-old civil war.* Found by System
4. *the criminals were convicted for bombings.* Found by System
5. *the allies launched a missile at the enemy stronghold.*

Precision = 3/7 $\approx$ .429
Recall = 3/5 = .6
F-measure = $\frac{2}{\frac{1}{3/7}+\frac{3}{5}} = \frac{2}{7/3+5/3} = .5$

Question 6. Fill in the CKY chart below for sentence *The rain rains down* assuming the following rules:

1. **S** $\rightarrow$ **NP VP**
2. **NP** $\rightarrow$ **N**
3. **NP** $\rightarrow$ **DT N**
4. **VP** $\rightarrow$ **V ADVP**
5. **VP** $\rightarrow$ **V**
6. **ADVP** $\rightarrow$ **ADV**
7. **DT** $\rightarrow$ **the**
8. **N** $\rightarrow$ **rain**
9. **N** $\rightarrow$ **rains**
10. **V** $\rightarrow$ **rain**
11. **V** $\rightarrow$ **rains**
12. **ADV** $\rightarrow$ **down**

	The	*rain*	*rains*	*down*
	1	2	3	4
0	DT	NP	S	S
1		N, V, NP, VP	S	S
2			N, V, NP, VP	VP
3				ADV, ADVP

Question 7. Some defining characteristics of organization and facility as per the ACE guidelines are as follows:

- *An Organization entity must have some formally established association. Typical examples are businesses, government units, sports teams, and formally organized music groups. Industrial sectors and industries are also treated as Organization entities.* (ACE Entity Guidelines v6.6, page 7)
- *A facility is a functional, primarily man-made structure. These include buildings and similar facilities designed for human habitation, such as houses, factories, stadiums, office buildings, gymnasiums, prisons, museums, and space stations; objects of similar size designed for storage, such as barns, parking garages and airplane hangars; elements of transportation infrastructure, including streets, highways, airports, ports, train stations, bridges, and tunnels. Roughly speaking, facilities are artifacts falling under the domains of architecture and civil engineering.* (ACE Entity Guidelines v6.6, page 22)

In the following text from the May 3, 2012 New York Times (*A House Tour: Yes, That House*) mark the organizations by underlining them and writing an **ORG** immediately above them; mark the facilities by underlining them and writing **FAC** immediately above. If a particular piece of text is difficult to mark only **ORG** or only **FAC**, mark it **ORG/FAC**. Mark noun groups ignoring determiners including both names and common nouns representing FAC and ORG constiuents. Do not mark pronouns.

```
After the 9/11 attacks, the system changed radically.  Now, anyone who
wants to tour the White House/FAC must apply through the office/ORG of his
or
her representative in Congress/ORG, which forwards the names to the White
House/ORG for clearance...
Once they get the green light, visitors show up at the appointed time
on 15th Street/FAC between E/FAC and F Streets/FAC and join the line to enter
through the southeast gate/FAC.
Anyone who has flown on an airline/ORG in recent years will recognize the
familiar territory of identity checks and electronic scans, although
here you do get to keep your shoes on.  At the head of the line,
rangers from the National Park Service/ORG check photo IDs against a list
of names.
```

Question 8. Assuming that the following sentence is at the beginning of a file, fill in the table below listing each token (word and punctuation), along with its start character offset and its end character offset. Note that there are more blank lines in the table than there are tokens. So it is expected that you will leave one or more line blank.

This sentence contains words, characters, spaces and punctuation.

Token	Start Offset	End Offset
This	0	4
sentence	5	13
contains	14	22
words	23	28
,	28	29
characters	30	40
,	40	41
spaces	42	48
and	49	52
punctuation	53	64
.	64	65

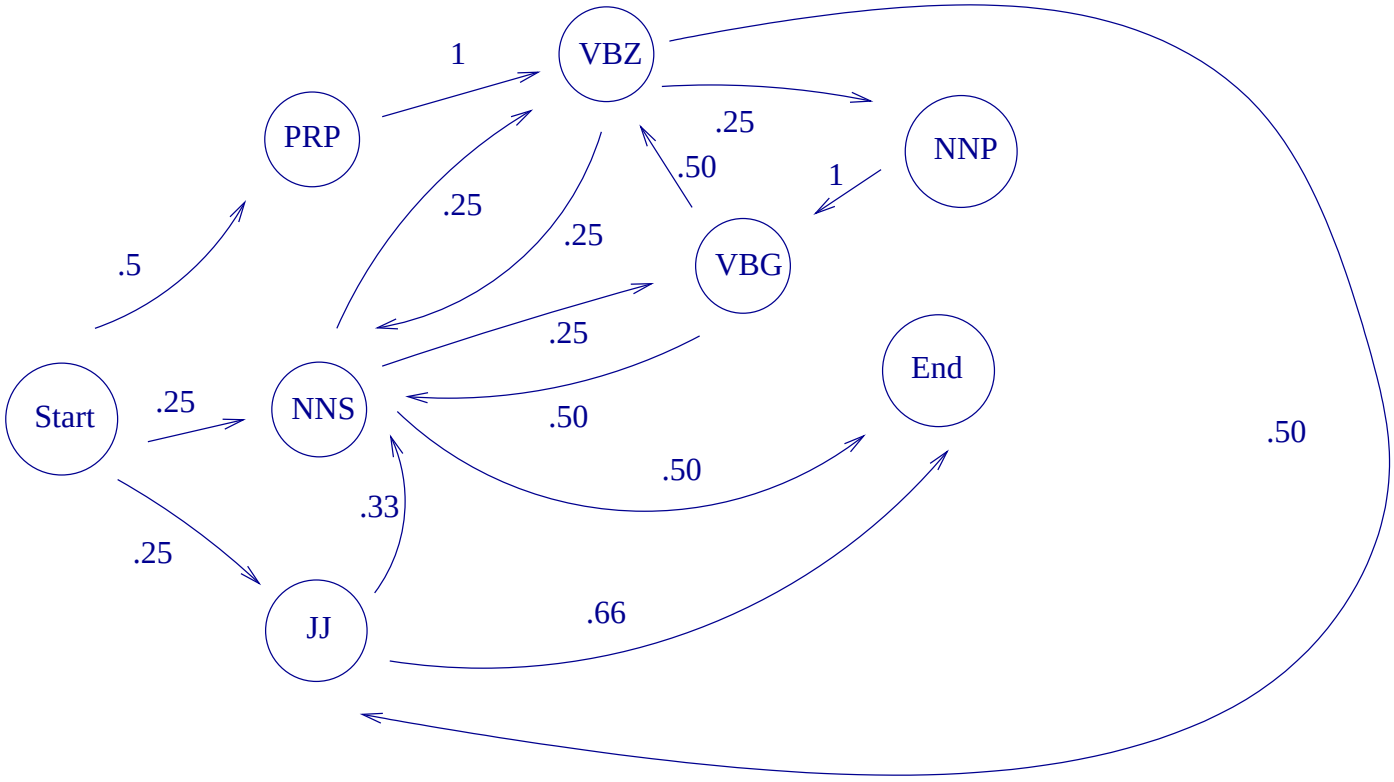

Figure 3: Prior Probability for Question 9

Question 9. Given the training data below, execute the following 3 steps: (a) calculate the likelihood probabilities for each word given each POS; (b) draw a finite state machine where states are POS and edges are labeled with transition probabilities; (c) draw a chart where the columns are positions in the sentence and the rows are names of states (start, end, POS tags) and fill in the probability scores assigned by the Viterbi algorithm assigning POS tags to the string *flying planes*.

Training Data:

- *buffalo/NNS flying/VBG is/VBZ dangerous/JJ*
- *flying/JJ planes/NNS are/VBZ numerous/JJ*
- *I/PRP saw/VBZ Mary/NNP flying/VBG planes/NNS*
- *He/PRP planes/VBZ shelves/NNS*

Likelihood for Question 9

JJ	dangerous: .33	flying: .33	numerous: .33	
NNP	Mary: 1			
NNS	buffalo: .25	planes: .5	shelves: .25	
PRP	I: .5	he: .5		
VBG	flying: 1			
VBZ	is: .25	are: .25	saw: .25	planes: .25

	BEGIN	flying	planes	END
BEGIN	1.0			
JJ		.33 * .25		
NNP				
NNS			(from JJ) .33 * .25 * .5 * .33 (from VBG) 0	
PRP				
VBG		1 * 0		
VBZ			(from JJ) .33 * .25 * .25 * 0 (from VBG) 0	
END				(from NNS) .33 * .25 * .5 * .33 * .5 ≈ .0068

Figure 4: Viterbi for Question 9

Question 10. Calculate the TFIDF for the terms listed below for documents 1 to 4. There are 10,000 documents in a collection. The number of times each of these terms occur in documents 1 to 4 as well as the number of documents in the collections are listed below. Use this information to fill in the TFIDF scores in the table below.

Number of Documents Containing Terms:

- **reverse cascade**: 3 IDF = log(10000/3) ≈ 8.11
- **full shower**: 50 IDF = log(10000/50) ≈ 5.30
- **half bath**: 10 IDF = log(10000/10) ≈ 6.91
- **multiplex**: 3 IDF = log(10000/3) ≈ 8.11

Term Frequencies				
	Documents			
	Doc 1	Doc 2	Doc 3	Doc 4
reverse cascade	8	10	0	0
full shower	3	1	2	2
half bath	0	0	8	7
multiplex	2	2	2	9

TFIDF for terms in documents				
	Documents			
	Doc 1	Doc 2	Doc 3	Doc 4
reverse cascade	8.11 * 8 = 64.88	8.11 * 10 = 81.10	0	0
full shower	5.30 * 3 + 15.90	5.30 * 1 = 5.30	5.30 * 2 = 10.60	5.30 * 2 = 10.60
half bath	0	0	6.91 * 8 = 55.28	6.91 * 7 = 48.37
multiplex	8.11 * 2 = 16.22	8.11 * 2 = 16.22	8.11 * 2 = 16.22	8.11 * 9 = 72.99