Midterm Exam for Natural Language Processing
March 9, 2017

Name: ________________________________
Net ID _______________________

Instructions
There are 7 questions, each will be worth 15 points for a total of 105 points. You will have approximately 1:15 minutes
to complete this test.
The test materials will include this printout and, optionally, one blank blue-covered booklet. I suggest that you fill in
all answers directly on this printout. The blue booklet is available if you would like scrap paper, or if you run out of
space on the test and need somewhere else to write your answers. You also have the option of putting all your answers
in the blue booklet, if you prefer. Whatever you do, please make it clear. Write me notes about where to find particular
questions if necessary. As long as I can find and understand your answers, it’s fine. If you use the blue booklet, please
write your name on this as well, so I can identify it as yours if it gets separated from this printout.

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.

Question 1. Write a regular expression to identify names of newspapers. The expression should match all the examples
below and generalize to cover some other newspaper names. It should also be specific enough that it does not match
very many non-newspaper names, e.g., it should not match Fred Smith or Mary Quincy-Jones.

1. Chicago Sun-Times
2. Los Angeles Times
3. Tampa Bay Times
4. New York Times
5. Washington Post
6. New York Post
7. Denver Post
8. Chicago Tribune
9. San Diego Union-Tribune
10. Star Tribune
11. New York Daily News
12. San Jose Mercury News
13. Dallas Morning News
14. Orange County Register
15. Philadelphia Inquirer

(([A-Z][a-z]+)[ -])+(Times|Post|Tribune|News|Register|Inquirer)
Question 2. Manually process the following sentence in two ways, filling in the columns in the chart below:

\textit{The plant regenerates from bushfire by growing from its base or from buds within its trunk.}\footnote{From Wikipedia Article about Banksia attenuata, a type of tree.}

I have tokenized this sentence and placed the tokens in the first column in the table below. Fill in the second column with PENN TREEBANK parts of speech (POS) tags, as per Table 1 (unlike conventional Penn Treebank Tags, all punctuation is marked \textit{PU}). In the third column, enter a BIO tag indicating whether a token is beginning a noun group (B), inside a noun group (I) or outside of a noun group (O). Remember not to include right modifiers as noun groups are not full NPs. If you are uncertain about any part of speech assignment, include a short note why you chose the tag you did.

<table>
<thead>
<tr>
<th>Token</th>
<th>POS Tag</th>
<th>BIO Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>DT</td>
<td>B</td>
</tr>
<tr>
<td>plant</td>
<td>NN</td>
<td>I</td>
</tr>
<tr>
<td>regenerates</td>
<td>VBJ</td>
<td>0</td>
</tr>
<tr>
<td>from</td>
<td>IN</td>
<td>0</td>
</tr>
<tr>
<td>bushfire</td>
<td>NN</td>
<td>B</td>
</tr>
<tr>
<td>by</td>
<td>IN</td>
<td>O</td>
</tr>
<tr>
<td>growing</td>
<td>VBG</td>
<td>O</td>
</tr>
<tr>
<td>from</td>
<td>IN</td>
<td>O</td>
</tr>
<tr>
<td>its</td>
<td>PRPS</td>
<td>B</td>
</tr>
<tr>
<td>base</td>
<td>NN</td>
<td>I</td>
</tr>
<tr>
<td>or</td>
<td>CC</td>
<td>O</td>
</tr>
<tr>
<td>from</td>
<td>IN</td>
<td>O</td>
</tr>
<tr>
<td>buds</td>
<td>NNS</td>
<td>B</td>
</tr>
<tr>
<td>within</td>
<td>IN</td>
<td>O</td>
</tr>
<tr>
<td>its</td>
<td>PRPS</td>
<td>B</td>
</tr>
<tr>
<td>trunk</td>
<td>NN</td>
<td>I</td>
</tr>
<tr>
<td>.</td>
<td>PU</td>
<td>O</td>
</tr>
</tbody>
</table>

\begin{table}
\centering
\begin{tabular}{|c|c|}
\hline
Tag & Description \hline
CC & Coordinating conjunction \hline
CD & Cardinal number \hline
DT & Determiner \hline
EX & Existential there \hline
FW & Foreign word \hline
IN & Preposition or subordinating conjunction \hline
JJ & Adjective \hline
JJR & Adjective, comparative \hline
JJS & Adjective, superlative \hline
LS & List item marker \hline
MD & Modal \hline
NN & Noun, singular or mass \hline
NNS & Noun, plural \hline
NNP & Proper noun, singular \hline
NNPS & Proper noun, plural \hline
PDT & Predeterminer \hline
POS & Possessive ending \hline
PRP & Personal pronoun \hline
PRPS & Possessive pronoun \hline
RB & Adverb \hline
RBR & Adverb, comparative \hline
RBS & Adverb, superlative \hline
RP & Particle \hline
SYM & Symbol \hline
TO & to \hline
UH & Interjection \hline
VB & Verb, base form \hline
VBD & Verb, past tense \hline
VBG & Verb, gerund or present participle \hline
VBN & Verb, past participle \hline
VBZ & Verb, 3rd person singular present \hline
VBP & Verb, non-3rd person singular present \hline
WDT & Wh-determiner \hline
WP & Wh-pronoun \hline
WP$ & Possessive wh-pronoun \hline
WRB & Wh-adverb \hline
PU & Punctuation \hline
\end{tabular}
\caption{Penn Treebank POS tags}
\end{table}
Question 3. Draw a Phrase Structure Tree analyzing the same sentence you analyzed in Question 2, i.e.,

*The plant regenerates from bushfire by growing from its base or from buds within its trunk.*

You can optionally assume the same POS tags you used in Question 2. Assume that a coordinate conjunction (e.g., *and*, *or*) combines two of the same type of constituents, i.e., $X \rightarrow X$ and $X$. The period should be a child of the sentence that it ends.

---

Subject to interpretation

*Growing words act like nouns (NN) and verbs (VBG) at the same time. So this phrase could be an NP or VP.*
Question 4: Fill in the CKY chart below for sentence *Leaves fall down* assuming the rules below. Remember that the rows of the chart represent start positions and the columns represent end positions.

1. $S \rightarrow NP \ VP$
2. $NP \rightarrow NNS$
3. $NP \rightarrow NN$
4. $NP \rightarrow NN\ NN$
5. $VP \rightarrow VBP$
6. $VP \rightarrow VBZ$
7. $VP \rightarrow VBP\ RB$
8. $VP \rightarrow VBP\ NP$
9. $NNS \rightarrow leaves$
10. $VBZ \rightarrow leaves$
11. $NN \rightarrow fall$
12. $VBP \rightarrow fall$
13. $RB \rightarrow down$
14. $NN \rightarrow down$

<table>
<thead>
<tr>
<th></th>
<th>leaves</th>
<th>fall</th>
<th>down</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>NNS, VBZ, NP, VP</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>1</td>
<td>XXXXXXXXXXXXXXXXXXXX</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>XXXXXXXXXXXXXXXXXXXX</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>XXXXXXXXXXXXXXXXXXXX</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>XXXXXXXXXXXXXXXXXXXX</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>XXXXXXXXXXXXXXXXXXXX</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>XXXXXXXXXXXXXXXXXXXX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>XXXXXXXXXXXXXXXXXXXX</td>
<td>XXXXXXXXXXXXXX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>XXXXXXXXXXXXXXXXXXXX</td>
<td>XXXXXXXXXXXXXX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>XXXXXXXXXXXXXXXXXXXX</td>
<td>XXXXXXXXXXXXXX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>XXXXXXXXXXXXXXXXXXXX</td>
<td>XXXXXXXXXXXXXX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>XXXXXXXXXXXXXXXXXXXX</td>
<td>XXXXXXXXXXXXXX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>XXXXXXXXXXXXXXXXXXXX</td>
<td>XXXXXXXXXXXXXX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>XXXXXXXXXXXXXXXXXXXX</td>
<td>XXXXXXXXXXXXXX</td>
<td></td>
</tr>
</tbody>
</table>

*Leaves fall down*
**Question 5.** Below is one query and two documents (1 and 2), completely written in lowercase characters.  

This question concerns retrieving documents for ad hoc information retrieval (IR) using feature vectors, based on TFIDF values for words contained in the query. Create a total of 3 feature vectors, one for the query, one for Document 1 and one for Document 2. Each vector should consist of TFIDF values for each of the following terms: construct, dentist, bridge, tooth. Assume that there is a “stemming” procedure that normalizes the word forms on the right-hand side of the following table to the stems on the left-hand side. Calculate the IDF for each of the terms based on a set of 10,000 documents in which: 600 contain (some form of) the word construct; 40 contain the word dentist, 200 contain the word bridge, and 100 contain the word teeth. Please use natural logarithms for computing IDF. Round all numbers to two decimal places. The next step for IR, calculating cosine similarity, is being omitted due to time constraints of the test.

<table>
<thead>
<tr>
<th>Stem</th>
<th>Word Forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>construct</td>
<td>construct, constructs, constructed, constructing, construction, constructions</td>
</tr>
<tr>
<td>dentist</td>
<td>dentist, dentists, dentistry</td>
</tr>
<tr>
<td>bridge</td>
<td>bridge, bridges, bridged, bridging</td>
</tr>
<tr>
<td>tooth</td>
<td>tooth, teeth</td>
</tr>
</tbody>
</table>

What is the Vector for the Query? \([\text{construct,dentist,bridge,teeth}]= [2.81,5.52,3.91,4.61]\)

What is the Vector for Document 1? \([2\times2.81,0,8\times3.91,0] = [5.62, 0, 31.28, 0]\)

What is the Vector for Document 2? \([2.81,5.52,2\times3.91,6\times4.61] = [2.81, 5.52, 7.82, 27.66]\)

- construct IDF = \(\log(10000/600) = 2.81\)
- dentist IDF = \(\log(10000/40) = 5.52\)
- bridge IDF = \(\log(10000/200) = 3.91\)
- teeth IDF = \(\log(10000/100) = 4.61\)

**Query:** how will the **dentist construct** a **bridge** to fix my **teeth**?

**[Document 1:]** the first permanent **bridge** to cross the chang jiang section of the river was the wuhan yangtze river **bridge**, **constructed** from 1955 to 1957. the dual-use road-rail **bridge** was a major **construction** project in the early years of the people’s republic and was completed with soviet assistance. the second **bridge** was a single-track railway **bridge** built in chongqing in 1959. the nanjing yangtze river **bridge**, also a road-rail **bridge**, was the first **bridge** to cross the lower reaches of the yangtze.

**Document 2:** a **bridge** is **constructed** by reducing the **teeth** on either side of the missing **tooth**. in other words, the abutment **teeth** including portions which are otherwise perfectly healthy are “reduced” in size using a high-speed rotary tool. to determine whether or not the abutment **teeth** can support a **bridge** without failure from lack of support from remaining root structures, the **dentist** employs ante’s law: "the root surface area of the abutment **teeth** has to equal or surpass that of the **teeth** being replaced with pontics."

Note: To convert to base 2, multiply all values in vectors by 1.443. To convert to base 10, multiply all values by .434.

---

2 Both documents are based on excerpts from Wikipedia entries for types of bridges.
Question 6. Using the Viterbi algorithm with the transition and likelihood probabilities below: (a) calculate the probability that the sequence of words *duck down* will be assigned the parts of speech: VB RB (as in the command telling someone to duck in a downward direction); (b) the probability that the sequence will be assigned the parts of speech: NN NN, as in the noun phrase that refers to the feathers of a duck. **Hint:** You may not have to fill out the entire table to calculate these two probabilities.

(a) NN NN has a probability score of $1.44 \times 10^{-9}$
(b) VB RB has a probability score of $4.5 \times 10^{-7}$

<table>
<thead>
<tr>
<th>Likelihood Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>POS</strong></td>
</tr>
<tr>
<td>NN</td>
</tr>
<tr>
<td>VB</td>
</tr>
<tr>
<td>VBP</td>
</tr>
<tr>
<td>RB</td>
</tr>
</tbody>
</table>

| Transition Probabilities |

<table>
<thead>
<tr>
<th>START</th>
<th>duck</th>
<th>down</th>
<th>END</th>
</tr>
</thead>
<tbody>
<tr>
<td>START</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td>(from Start) $1 \times .6 \times .001$</td>
<td>(from NN) $1 \times .6 \times .001 \times .4 \times .00001$</td>
<td></td>
</tr>
<tr>
<td>VB</td>
<td>(from Start) $1 \times .4 \times .0003$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VBP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RB</td>
<td></td>
<td>(from VB) $1 \times .4 \times .0003 \times .25 \times .05$</td>
<td></td>
</tr>
</tbody>
</table>
| END   |     |     | (from NN) $1 \times .6 \times .001 \times .4 \times .00001 \times .6 = 1.44 \times 10^{-9}$
|       |     |     | (from RB) $1 \times .4 \times .0003 \times .25 \times .05 \times .3 = 4.5 \times 10^{-7}$ |
Question 7. This is a two part question. The first part is annotation and the second part is evaluation. For the annotation, choose word sequences referring to Geo-political entities (GPEs) in the following paragraph from Wikipedia about Cichlids a family of species of fish. Word choice may be indicated by making a separate list or by circling the words in the text.

GPEs are words or word sequences that name specific locations that have governments (New York City, New York, the United States, Greenwich Village). GPEs also include land masses that consist of groups of such entities, e.g., North America because it is made up of 3 GPEs: Canada, the United States and Mexico. GPEs are usually proper nouns, but can also be proper adjectives (adjectives like American which are typically capitalized).

Assume that your annotation is the answer key. For the second part of the question, evaluate the System Output for precision, recall and f-measure, using the answer key you created. Please round to the nearest 1/10 of a percent.

Recall = 100%

Precision = \frac{12}{25} = 48%

F-measure = \frac{2}{1+\frac{1}{48}} = 64.9%

Notes: Rio Grande is the name of a river. The latin words/phrases in the quote mostly refer to species of fish.

Cichlids are one of the largest vertebrate families in the world. They are most diverse in Africa and South America, Africa alone is estimated to host at least 1,600 species. Central America and Mexico have about 120 species, as far north as the Rio Grande in southern Texas. Madagascar has its own distinctive species (Katria, Oxylapia, Paratilapia, Paretroplus, Ptychochromis, and Ptychochromoides), only distantly related to those on the African mainland. Native cichlids are largely absent in Asia, except for 9 species in Israel, Lebanon, and Syria (Astatotilapia flavijosephi, Oreochromis aureus, O. niloticus, Sarotherodon galilaeus, Coptodon zillii, and Tristramella spp.)

System Output:

- Cichlids ×
- Africa ✓
- South America ✓
- Africa ✓
- Central America ✓
- Mexico ✓
- Rio Grande ×
- Texas ✓
- Madagascar ✓
- Katria ×
- Oxylapia ×
- Paratilapia ×
- Paretroplus ×
- Ptychochromis ×
- Ptychochromoides ×
- African ✓
- Asia ✓
- Israel ✓
- Lebanon ✓
- Syria ✓
- Astatotilapia ×
- Oreochromis ×
- Sarotherodon ×
- Coptodon ×
- Tristramella ×