CSI 660 Topics in Computer Science: Social Computing

Homework 1 **Due: March 5th, 2019 11:59 pm**

In this homework, you will start to work with language data and apply algorithms and techniques that are taught in the class.

You are each going to be assigned a unique dataset. The dataset consists of two 90-minute chat room conversations between 4-8 people on some topic or task.

Question 1 (20 points): This part of the homework will get you familiar with existing NLP technology and provide hands-on experience with basic NLP tools.

Download Stanford CoreNLP suite from:
http://stanfordnlp.github.io/CoreNLP/#about

Download CoreNLP 3.9.2

This suite of tools includes the most widely used, basic NLP algorithms, like Tokenizing/Segmenting, Part-of-Speech tagging and Named Entity Tagging. The easiest way to run the program is to use the command line functionality of the software, documented here: http://stanfordnlp.github.io/CoreNLP/cmdline.html

The software produces output in .xml format as below:

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet href="CoreNLP-to-HTML.xsl" type="text/xsl"?>
<root>
 <document>
  <sentences>
   <sentence id="1">
    <tokens>
     <token id="1">
      <word>shelly</word>
      <lemma>shelly</lemma>
      <CharacterOffsetBegin>0</CharacterOffsetBegin>
      <CharacterOffsetEnd>6</CharacterOffsetEnd>
      <POS>NN</POS>
      <NER>O</NER>
     </token>
     <token id="2">
      <word>-LRB-</word>
```

```
        <lemma>-lrb-</lemma>
        <CharacterOffsetBegin>7</CharacterOffsetBegin>
        <CharacterOffsetEnd>8</CharacterOffsetEnd>
        <POS>-LRB-</POS>
        <NER>O</NER>
       </token>
        ...
        ...
    </sentence>
  </sentences>
 </document>
</root>
```

You will run the two conversations assigned to you in your dataset through this Stanford suite of tools and produce output for:

a) tokenizing
b) sentence splitting
c) part-of-speech tagging
d) lemmatizing
e) named entity recognition

All output for a) through e) will be in a single file and in the format above. Please send the two output files you generated to get points for this part of the homework.

Question 2 (10 points): While working with this conversational data, we have formed some hypotheses.

Our hypothesis is: Male chat participants have different language patterns when conversing in chat rooms than do female chat participants. They may speak in shorter sentences and use more pronouns while speaking.

What are the null hypotheses for the alternate hypotheses listed below:

2a. (5 points) **H_{2a}:** Male chat participants speak in significantly shorter sentences than do female chat participants.

2b. (5 points) **H_{2b}:** Male chat participants use significantly more pronouns than do female chat participants.

Question 3 (70 points): In this part of the homework, we will use the data set of conversations we have to test hypothesis H_{2b}.

H_{2b}: Male chat participants use significantly more pronouns than do female chat participants.

In Question 1, you will have produced the output files that have identified each pronoun in the conversation. The pronouns are identified by the following xml tags:

<POS>PRP</POS>
and
<POS>PRP$</POS>

You will write code to read the output xml files produced in Question 1 and count the number of pronouns for each participant in the conversation.

3a. (50 points) Produce a table of results as below by counting the number of pronouns each participant, either male of female, has made in both conversations. *(SAMPLE DATA)*

	# pronouns		# pronouns
Male 1	40	Female 1	33
Male 2	..	Female 2	..
Male 3	..	Female 3	..
Male 4	..	Female 4	..
..		..	

3b. (10 points)
What is the average number of pronouns used by males?
What is the average number of pronouns used by females?

3c. (10 points) We will use an online calculator to test whether the population of male and female participants in our sample differ significantly in their use of pronouns.

Please go to http://www.socscistatistics.com/tests/studentttest/Default.aspx
and input the number of pronouns you have determined for the two populations separately as Treatment 1 and Treatment 2. Let the significance level be 0.05 and the hypothesis be one-tailed. Use the calculator to compute the t and p values.

You will get a result such as below:
The *t*-value is -3.12149. The *p*-value is .020546. The result is significant at $p < .05$.
Please copy and paste the result you obtain into your report. Using the value of p you calculated, is the alternate hypothesis **H_{2b}** true for your dataset or not?