University of Lethbridge

Department of Mathematics & Computer Science

FIAT LUX

CPSC 4310/5310/7310 – Spring 2020

Natural Language Processing (NLP)

Assignment 1 [60 points]

Due on January 24th, 2020

The problems are adopted from the textbook.

1. **[5 points]**

   Give two sentences that are ambiguous. Specify their different interpretations.

2. **[10 points]**

   Give regular expressions to the following. Check your expressions using any regular expression tool.

   (a) the set of all alphabetic strings.

   (b) the set of all lower case strings ending with a e.

   (c) the set of all lower case strings ending with the string $ation$.

   (d) the set of all whole numbers between 250 and 350, inclusive.

   (e) the set of all percentage rates such as 2%, 2.5%, 10.15%, etc.

   (f) the set of all strings with two consecutive repeated words.

3. **[5 points]**

   Consider the following data sets comprising of 3 boolean input attributes and 1 boolean output.

Example	A_1	A_2	A_3	$output$
x_1	1	0	0	0
x_2	1	0	1	0
x_3	0	1	0	0
x_4	1	1	1	1
x_5	1	1	0	1

Show the decision tree learned from these data. Show also the best attribute to split on.

4. **[10 points]**

   Compute the minimum edit distance by hand, and figure out whether *case* is closer to *crane* or *care*. Align them.

5. **[30 points]**

   Given the Brown corpus available from NLTK (Natural Language Processing Tool Kit), it consists of several categories/genres of texts, write a program that computes:

   - How many word tokens does each category/genre have?
   - How many word types does each category/genre have?
   - What is the vocabulary size of the whole corpus?

   Your program should compute with the following variations:

   (a) with stopwords,

   (b) without stopwords,

   (c) without stopwords and lemmatization, and

   (d) without stopwords and stemming.

   You may display them by category. On our lab computers, you have to run python3.6 because if you just type in python it defaults to python2.7 and NLTK is not available in that version.