DALHOUSIE UNIVERSITY
Inspiring Minds

Winter 2017 (Jan9-Apr10)
Faculty of Computer Science

CSCI 4152/6509 — Natural Language Processing

Assignment 3

Due: *by midnight, Friday, Mar 24, 2017*
Worth: 107 marks
Instructor: Vlado Keselj, CS bldg 432, 902.494.2893, vlado@dnlp.ca
TA: Magdalena Jankowska, jankowsk@cs.dal.ca

Assignment Instructions:

All answers must be submitted through the SVN by the duedate.

The Lab questions must be submitted in the appropriate directories as specified in the labs (and questions).

The answer files for the other questions, should be submitted in the SVN directory *csuserid*/`a3`, similarly to the previous assignment.

All files must be plain-text files, unless specified differently by the question.

1) (25 marks, files in *csuserid*/`lab3`) Complete the Lab 3 as instructed. In particular, you will need to properly:

a) (5 marks) Submit the file '`matching.pl`' as instructed,

b) (5 marks) Submit the file '`matching-data.pl`' as instructed,

c) (5 marks) Submit the file '`word_counter.pl`' as instructed,

d) (5 marks) Submit the file '`replace.pl`' as instructed, and

e) (5 marks) Submit the file '`ngram-output.txt`' as instructed.

2) (17 marks, files in *csuserid*/`lab4`) Complete Lab 4 as instructed. In particular, you will need to properly:

a) (3 marks) Submit the example file '`array-examples.pl`' as instructed (Step 2)

b) (3 marks) Submit the example file '`test-hash.pl`' as instructed (Step 3)

c) (5 marks) Submit the program file '`letter_counter_blanks.pl`' and the output file '`out_letters.txt`' as instructed (Step 4)

d) (6 marks) Submit the program '`word_counter.pl`' and the output file '`out_word_counter.txt`' as instructed (Step 5)

Notice that the examples from (a) and (b) need to compile: if a syntax error got introduced to an example program by your typing mistake or by introducing incorrect characters through copying and pasting from a pdf file, so that the example program does not compile, it will not be accepted. That follows from the lab instructions that programs are to be tested before submitting.

3) (30 marks, files in *csuserid*/`lab5`) Complete the Lab 5 as instructed. In particular, you will need to properly:

a) (4 marks) Submit the file '`gcd.prolog`' as instructed,

b) (4 marks) Submit the file '`prog1.prolog`' as instructed,

c) (4 marks) Submit the file '`factorial.prolog`' as instructed,

d) (18 marks) Submit the files `task.prolog` and `task-queries.txt` as instructed (Step 9 task).

4) (35 marks, file `a3q4.txt`, or `a3q4.pdf`) Let us assume that you work on a task of analyzing Twitter data. You need to detect whether a Twitter business account belongs to B2B or B2C category, and you decided to test if this can be done based on certain keywords used in company name and using the Naïve Bayes method. To test the method, you decided to use the following features in classification:

- The feature $B \in \{t, f\}$, which is set to 't' (true) if the word '**B**ank' appears in the company name, and otherwise it is set to 'f' (false).
- The feature $C \in \{t, f\}$, which is set to 't' (true) if the word '**C**onsulting' appears in the company name, and otherwise it is set to 'f' (false).
- The feature $S \in \{t, f\}$, which is set to 't' (true) if the word '**S**ervices' appears in the company name, and otherwise it is set to 'f' (false).

The class itself is modeled using the variable $T \in \{b, c\}$, where T stands for Business **T**ype, and b representes a B2**B** business, and c represents B2**C** business.

The training data is presented in the following table:

tweets	B	C	S	T
35	f	f	f	b
16	f	f	f	c
22	f	f	t	b
1	f	f	t	c
35	f	t	f	b
2	f	t	f	c
12	f	t	t	b
8	t	f	f	b
43	t	f	f	c
4	t	f	t	b
3	t	f	t	c
7	t	t	f	b
4	t	t	f	c
8	t	t	t	b

a) (15 marks) Calculate the conditional probability tables (CPTs) for the Naïve Bayes model.

b) (5 marks) Calculate $P(T = b \,|\, B = t, C = t, S = f)$ using the Naïve Bayes model and briefly describe what this conditional probability represents.

c) (5 marks) What is the most likely value of the class variable T for the partial configuration $(B = t, C = t, S = f)$?

d) (5 marks) What is $P(T = b \,|\, B = t, C = t, S = f)$ if we use the Joint Distribution Model?

e) (5 marks) What is $P(T = b \,|\, B = t, C = t, S = f)$ if we use the Fully Independent Model?

Note: In assignments, always include intermediate results and sufficient details about the way the results are obtained.