DALHOUSIE UNIVERSITY
Inspiring Minds

Winter 2019 (Jan7-Apr8)
Faculty of Computer Science

CSCI 4152/6509 — Natural Language Processing

Assignment 4

Due: *Monday Apr 8, 2019 by midnight*
Worth: 160 marks (= 31 + 24 + 50 + 20 + 35)
Instructor: Vlado Keselj, CS bldg 432, 902.494.2893, vlado@dnlp.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 *CSID*/`a4`, similarly to the previous assignment.

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

1) (31 marks, submit files via SVN in *csuserid*/`lab8`) Complete the Lab 8 as instructed. In particular, you will need to properly:

a) (5 marks) Submit the file '`gcd.prolog`' as instructed.

b) (5 marks) Submit the file '`prog1.prolog`' as instructed.

c) (5 marks) Submit the file '`factorial.prolog`' as instructed.

d) (8 marks) Submit the file '`task.prolog`' as instructed.

e) (8 marks) Submit the file '`task-queries.txt`' as instructed.

2) (24 marks, files in *csuserid*/`lab9`) Complete the Lab 9 as instructed. In particular, you will need to properly:

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

b) (4 marks) Submit the file '`dcg.pl`' as instructed.

c) (4 marks) Submit the file '`dcg-ptree.prolog`' as instructed.

d) (4 marks) Submit the file '`dcg-agr.prolog`' as instructed.

e) (4 marks) Submit the file '`dcg-pcfg.prolog`' as instructed.

f) (4 marks) Submit the file '`dcg-agr2.prolog`' as instructed.

3) (50 marks, one or more files in SVN: *csuserid*`/a4/a4q3.txt`, *csuserid*`/a4/a4q3.pdf`, or *csuserid*`/a4/a4q3.jpg`.)
Consider the following four sentences tagged with POS tags:

```
swat V flies N like P ants N
time N flies V like P an D arrow N
flies N like V arrow N
flies N like V ants N
```

a) (20 marks) Calculate necessary CPTs (Conditional Probability Tables) for the HMM model. *Do not apply any smoothing.*

b) (10 marks) Consider tagging the sentence: "`arrow flies like ants`"

Draw the factor graph for this problem and mark all messages that need to be computed. It is probably hard to do this in plain text, so you can make it a part of a pdf file, or jpg file with a name as specified at the beginning of the question. You can can even draw the graph by hand and submit a photo of it.

c) (20 marks) Calculate all necessary messages and find the optimal values for T_1, T_2, T_3, and T_4. Show values of all messages and calculation needed to find optimal values of the variables.

Of course, you need to use unsmoothed tables from part a). There are no unseen words in the given sentence, so smoothing or some other way of handling unseen words is not necessary.

You are encouraged to write a program for some of these tasks, but you must submit output of the program with sufficient details as required because you will be marked on the output.

4) (20 marks, one or more files in SVN: *csuserid*`/a4/a4q4.txt`, *csuserid*`/a4/a4q4.pdf`, or *csuserid*`/a4/a4q4.jpg`.)

Consider the following two sentences:

She read the book with a great passion.
and
She read the book with a great title.

They have different structure because "great passion" in the first sentence refers to way she read the book, while "great title" in the second sentence refers to the book that has a "great title". This difference can be seen if we properly construct parse trees of the two sentences.

a) (10 marks) Draw the parse trees of the sentences using the standard tags and rules.
b) (5 marks) Annotate the first parse tree with head information and dependencies. (as done in the example in class)
c) (5 marks) Re-write the first parse tree using the bracketed notation.

5) (35 marks, one or more files in SVN: *csuserid*`/a4/a4q5.txt`, *csuserid*`/a4/a4q5.pdf`, or *csuserid*`/a4/a4q5.jpg`.)
a) (15 marks) Extract the context-free grammar induced from the following parse trees. If we express this grammar as a PCFG, what are the probabilities that can be inferred from the trees?

```
(S (WNP (WDT What)
        (NN courses))
   (VP (BE are)
       (VP (VBN offered)
           (PP (IN in)
               (NN fall)))))

(S (WNP Who)
   (VP (VBZ teaches)
       (NP (NN CSCI)
           (NN 1100))))
```

b) (15 marks) Using CYK algorithm for marginalization and the grammar obtained in the part a), parse the sentence:
What CSCI teaches fall fall

Remember that the CYK algorithm for marginalization keeps track of probabilities for all non-terminals, and adds probabilities if the same span of words can be derived in more than one way from the same non-terminal.

c) (5 marks) What is the final parse tree obtained from parsing in the part b) and what is the probability of this tree?