INF4820: Algorithms for AI and NLP (Fall 2013)
— Final Exam —

General Instructions

- Please read through the complete exam once before starting to answer questions. About thirty minutes into the exam, the instructor will come around to answer any questions of clarification (including English terminology).
- As discussed in class, the exam is only given in English, but you are free to answer in any of *Bokmål*, English, or *Nynorsk*.
- To give you an idea about the relative weighting of different questions during grading, we've assigned points to each of them (summing to 100).

1 Common Lisp (15 points total)

(a) When working with vector space representations, we're often dealing with vectors that are simultaneously both *sparse* (i.e., they have a low ratio of non-zero elements) and extremely *high-dimensional*. Discuss some relevant choices for data structure when implementing such vectors in Lisp (e.g., based on lists, arrays, or hash-tables), highlighting the advantages and disadvantages of different approaches (in terms of storage efficiency, look-up performance, etc). *[8 points]*

(b) Write *two* versions of a function `swap`; one based on recursion and one based on iteration. The function should take three parameters – `x`, `y` and `list` – where the goal is to replace every element matching `x` with `y` in the list `list`. This is an example of the expected behavior;

```
> (swap "foo" "bar" '("zap" "foo" "foo" "zap" "foo"))
→ ("zap" "bar" "bar" "zap" "bar")
```

Try to avoid using destructive operations if you can. *[7 points]*

2 Classification and Clustering (30 points total)

(a) A common "pre-processing" step when working with vector space models is to apply *length normalization* to the vectors. Explain what is meant by this and the various effects it can have. *[10 points]*

(b) Describe the method of kNN classification. Briefly discuss its advantages and disadvantages. *[10 points]*

(c) Describe the method of k-Means clustering. Briefly discuss its advantages and disadvantages. *[10 points]*

3 Linear Structures (21 points total)

(a) How exactly does an n-gram model compute the probability $P(s)$ for a string $s = w_1^n$, assuming a bigram model? State the central assumption made in this modelling approach. *[5 points]*

(b) In a few sentences, explain how Hidden Markov Models extends on simple n-gram models. *[4 points]*

$P(N\|\langle S\rangle) = \frac{6}{8}$	$P(V\|\langle S\rangle) = \frac{1}{8}$	$P(P\|\langle S\rangle) = \frac{1}{8}$	
$P(N\|N) = \frac{1}{8}$	$P(V\|N) = \frac{2}{8}$	$P(P\|N) = \frac{1}{8}$	$P(\langle/S\rangle\|N) = \frac{4}{8}$
$P(N\|V) = \frac{2}{3}$	$P(V\|V) = 0$	$P(P\|V) = \frac{1}{3}$	$P(\langle/S\rangle\|V) = 0$
$P(N\|P) = \frac{4}{4}$	$P(V\|P) = \frac{1}{4}$	$P(P\|P) = 0$	$P(\langle/S\rangle\|P) = 0$

(a) Transition probabilities

$P(\textit{flies}\|N) = \frac{1}{4}$	$P(\textit{cats}\|N) = \frac{2}{4}$	$P(\textit{fruit}\|N) = \frac{1}{4}$	$P(\textit{like}\|N) = 0$	$P(\textit{as}\|N) = 0$
$P(\textit{flies}\|V) = \frac{1}{3}$	$P(\textit{cats}\|V) = 0$	$P(\textit{fruit}\|V) = 0$	$P(\textit{like}\|V) = \frac{2}{3}$	$P(\textit{as}\|V) = 0$
$P(\textit{flies}\|P) = 0$	$P(\textit{cats}\|P) = 0$	$P(\textit{fruit}\|P) = 0$	$P(\textit{like}\|P) = \frac{1}{4}$	$P(\textit{as}\|P) = \frac{3}{4}$

(b) Emission probabilities

Table 1: Hidden Markov Model probabilities

(c) Give the general formula used to calculate the values in the trellis for the *Forward algorithm*. Using the transition and emission probabilities given in Tables 1a and 1b and an observation sequence of *flies like fruit*, show the calculations made by the *Forward algorithm* for the first two columns of the trellis. (You don't need to solve the calculation — we just want to see that you know which formulae are used, and which numbers go in to them.) *[7 points]*

(d) Briefly state the differences between the Forward and Viterbi algorithms, both in terms of what they calculate, and in the details of their implementations. *[5 points]*

4 Hierarchical Structures (25 points total)

(a) Given the following treebank, list the rules necessary to derive the trees, their counts across all the trees and the maximum likelihood estimation of their conditional probabilities. *[7 points]*

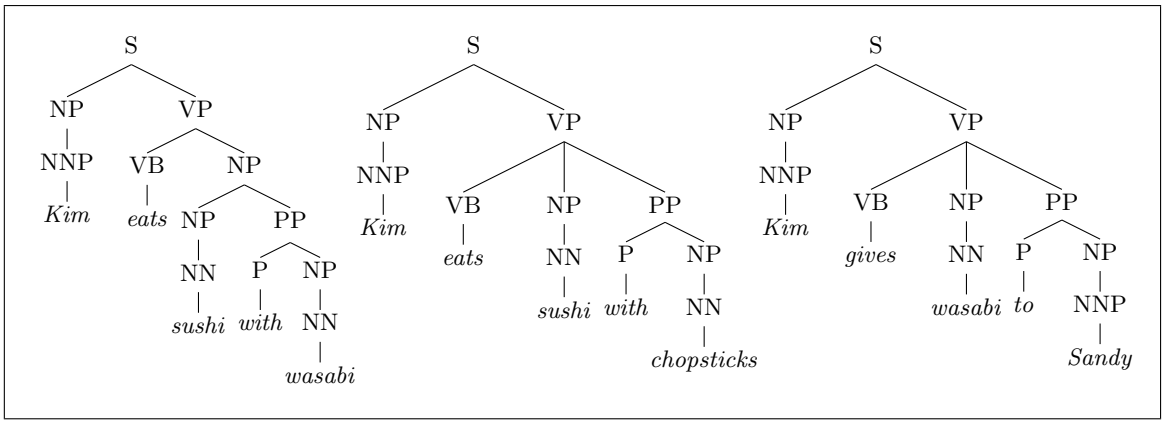

Figure 1: Treebank for MLE calculations

(b) In a few sentences, explain the fundamental rule of chart parsing, including its purpose, the structures it operates over, and the features they have. *[9 points]*

(c) In a few sentences, discuss the concept of *local ambiguity* in parsing. What process does our generalised chart parser use for efficiently recording local ambiguities? Briefly sketch out the ideas behind the implementation of this process. *[9 points]*

5 Structured Probabilistic Modelling (9 points total)

(a) What is dynamic programming and why is it used? What properties of a problem make it suitable to use a dynamic programming algorithm to solve it? *[5 points]*

(b) Write 3–4 sentences about the differences between generative and discriminative models, including the advantages and disadvantages of each. *[4 points]*