# DATA.STAT.840 Statistical Methods for Text Data Analysis

## Exercises for Lectures 8-9: Probabilistic Context-free Grammars

### Exercise 8.1: grammar for equations.

While equations are not natural language, they are featured in many scientific publications and are a useful example for grammars. Propose a context-free grammar that can create equations of the kind used in simple calculators: nonnegative integer numbers, a limited set of variables (x, y, z), operators (+, -, *, /, ^), parentheses, a limited set of functions (log, exp, sin, cos, tan). Example equations that your grammar should be able to generate:

- 1+2*3^10
- 24-(3*x-(z^y))/(3-x)
- sin(2*x-cos(y))^(2-10)

Then, show how an example equation is derived from your grammar: show the derivation tree of applying rules to arrive at the final equation.
Note: in this exercise you do not need to assign probabilities to the rules of your grammar.
Report the definition of your grammar, and the derivation of your example equation.

### Exercise 8.2: Chomsky normal form.

The probabilistic context free grammar below is a simplified version of the one used to generate "hmm_sentences.txt" in exercise 7.4. Transform the grammar into Chomsky normal form, so that the possible sentences and their probabilities are the same as in the grammar below. Report the definition of your resulting Chomsky normal form grammar (rules and their probabilities).

| | |
|---|---|
| S --> STMANY | 1.0 |
| STMANY --> S1 . | 0.6 |
| STMANY --> S1 , but STMANY | 0.4 |
| S1 --> SUBJ QVERB1 QVERB2 OBJ | 1.0 |
| SUBJ --> ARTICLE DESC NOUN | 1.0 |
| DESC --> ADJECTIVE | 0.7 |
| DESC --> ADJECTIVE DESC | 0.3 |
| OBJ --> ARTICLE DESC NOUN | 1.0 |
| QVERB1 --> can | 0.2 |
| QVERB1 --> will | 0.5 |
| QVERB1 --> may | 0.3 |
| ARTICLE --> a | 0.6 |
| ARTICLE --> the | 0.4 |
| QVERB2 --> explain | 0.4 |
| QVERB2 --> help | 0.2 |
| QVERB2 --> answer | 0.4 |
| ADJECTIVE --> wise | 0.3 |
| ADJECTIVE --> friendly | 0.5 |
| ADJECTIVE --> insightful | 0.2 |
| NOUN --> cat | 0.7 |
| NOUN --> dog | 0.2 |
| NOUN --> fox | 0.1 |

### Exercise 8.3: Inside-outside algorithm.

Use the inside-outside algorithm for the Chomsky normal form grammar you produced in exercise 8.2, to calculate the probability of the sentence "a wise fox can help the friendly insightful cat". Report your computation and the resulting probability. Hint: the resulting probability should be the same as it is in the original grammar given in exercise 8.2.

# DATA.STAT.840 Statistical Methods for Text Data Analysis

## Exercises for Lectures 9-10: Information retrieval

### Exercise 9.1: Retrieval using TF-IDF and unigram language models.

Consider these four artificial documents:

**d1:** the robot is insightful but you are strong and i may answer and the wise fox is insightful and you are insightful and i am insightful but i will explain the insightful bird

**d2:** the bird is insightful

**d3:** when will they explain the friendly insightful strong insightful bird and is the bird strong and is a strong robot insightful

**d4:** a cat is strong but you are cautious and i may help but a fox is insightful but are they strong and when may you answer

In total, these four documents have the following vocabulary of 25 words:

'a', 'am', 'and', 'answer', 'are', 'bird', 'but', 'cat', 'cautious', 'explain', 'fox', 'friendly', 'help', 'i', 'insightful', 'is', 'may', 'robot', 'strong', 'the', 'they', 'when', 'will', 'wise', 'you'

Consider the **query: "**insightful bird".

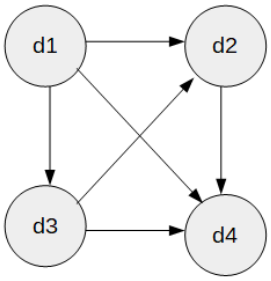
a) Compute TF-IDF vectors for d1 and d2, using "raw count" for term frequency and "logarithmic inverse document frequency" for inverse document frequency, as described on Lecture 3 slides 37-38. Compute also a TF-IDF vector for the query. Then compute the cosine similarity between the query and d1 and d2. Which document is closer to the query?

b) Compute the unigram probability for the query given by d1 and d2, as discussed on Lecture 10 slide 14. In this exercise you do not need to apply smoothing to the probabilities. Which document gives larger probability to the query?

Report your computations and results.

### Exercise 9.2: Pagerank.

Suppose the four documents d1, d2, d3, d4 of exercise 9.1 are webpages that link to each other with hyperlinks as shown in the picture below.




Use the equations of Lecture 10, slides 18-21 to solve a Pagerank prior for the document probabilities. Report your computation and the resulting prior probabilities of the four documents.