Assignment 1
Methods in Computational Linguistics II
Spring 2015

Due February 25.

A. Probability (40 points)

1. Represent a joint occurrence between two categorical variables in python as a nested dictionary structure. Call this variable “joint”. (5)

2. Write a function inc(joint, x, y) that increments the count of x and y in the joint probability table. (5)

3. Write a function prob(joint, x, y) that calculates the joint probability of the events “x” and “y” as stored in the variable “joint”. (10)

4. Write two function marginalX(joint, x) and marginalY(joint, y) that calculates the marginal probability of the events “x”, and “y”, respectively. (10)

5. Write a function, condXY(joint, x, y), that calculates the conditional probability of p(xIy). (5)

6. Write a function, condYX(joint, y, x), that calculates the conditional probability of p(yIx). (5)

B. Open Ended Programming Problem. (60pts)

Use the nltk FreqDist object to compare the nltk texts 1-9. These include 1) Moby Dick, 2) Sense and Sensibility, 3) The Book of Genesis, 4) Inaugural Addresses, 5) Chat, 6) Monty Python and the Holy Grail, 7) Wall Street Journal, 8) Personals, 9) The Man Who Was Thursday.

1. Describe, in plain English, a way to describe the similarity between texts based on the frequency of linguistic qualities -- words, characters, word length. (10)
2. Write a function that takes two FreqDist objects and returns the similarity between two texts. (20)
3. Use this function to compare every pair of texts. Report this in a table. (5)
4. Identify the most similar and least similar texts using your function. (5)
5. Describe your results. Are the findings consistent with your intuition about the texts? Are there any surprising results? If so, try to explain how they are explained based on your similarity function. (10)
6. Present at least 2 ideas that could improve the calculation of similarity between two texts. (10)