

CISC181 Fall 2006 Project 2

Statistical Text Generation

Submit electronically Tuesday Nov 14 midnight. Deliver paper copy in class Wednesday.

Description

Read a corpus from a file. Determine bigram frequency as discussed in class. Use the frequencies of the corpus to generate a text of specified length that has (probabilistically) the same bigram frequencies.

You may choose your implementation structure from the two discussed in class (parallel arrays vs. structs). See me during office hours if you want to consider some other structure.

See the FAQ for details we have discussed in class or questions I have received.

Special features

In addition to text generation, your program must have a facility for printing out the n most likely following words for a given word, along with their probabilities, and their frequency in the corpus.

Your program must be able to easily handle simple text corpora (for example, a single sentence corpus) as well as larger texts like Brown. For that reason, you will use code to convert Brown to a simple text format (rather than simply editing the file).

User interface

Use a simple menu or menus to allow the user to control the behavior of the program.

Grading

- 20% testing code and makefile(s): completeness and correctness
- 60% correct operation, coding, style
- 20% resubmission (points repaired / points lost)

Submission

Do not leave your scripting until the last minute. You will need to examine your output carefully. In particular, for each scripted output text you will use color to highlight three very nice sentences and three that are incoherent.

Submit all your working code and script. One day before the due date I will give you the lengths of the texts you will generate.