

**Information Theory and Reliable Communications**  
**Computer Project: Source Modeling**

In class, we discussed the modeling of binary images using various Markov models. In this project, you are to model English text.

For this purpose, I have prepared for you a data file:

*Origin2.txt* This file contains the entire content of the book *The Origin of Species* by Charles Darwin. It contains approximately 1 million ASCII characters (about 160,000 words). The original file can be obtained from the website:

<http://www.literature.org/Works/Charles-Darwin/origin/>

The file *Origin2.txt* has been pre-processed to remove all non-alpha-numeric characters — except for the comma, the period, the space, and the carriage return. Also, all upper-case letters have been converted to lower-case. Thus, the alphabet size is 40 (26 letters, 10 numbers, 4 non-alpha-numeric characters).

*Task 1:* Find a mathematical model for this source. From the model, determine the entropy rate of the source (in bits/character).

*Task 2:* Now find a lossless compression algorithm which can compress the source. Compare the average rate of the compression algorithm with the entropy rate. What does this tell you about the model and/or the compression algorithm?

*Task 3:* Randomly generate a text file according to your model and compare it with the original file.

You can use any programming language (Matlab, Fortran, Pascal, C, C++, etc.) you want.

Use your imagination and creativity.

*Task 4:* Write up your project in a report and submit it along with your programs (on a floppy disk) to me by Thursday May 4, 2000 (5:10 pm). Assume that the reader is another bright student who has taken Information Theory. Do not assume that the reader knows what you are doing.

Grading: (60 pts Total) Your project grade will be based on the following:

1. Technical Content: (25 pts) How “good” is your model? How ‘good’ is the compression algorithm you used? How accurate and reliable are your results? Are your results repeatable? Did you say anything that is technically wrong? Did you say anything that is technically right? Did you make fair and meaningful comparisons?
2. Presentation: (25 pts) How clear is your written report? How clearly do you present your results? Does your report contain grammatical or typographical errors? Would another graduate student (who have taken ESE 535) understand and appreciate your work?
3. Originality: (10 pts) Does your project contain novel ideas? Did you take traditional approaches or unconventional ones? I encourage you to take risky and unconventional approaches. Even if these fail, you may include them in your report. But explain *why* you think they fail.

**Early-Completion Incentive Program:**

- If you turn in the project by 4/27/99 (5:10 pm), you will get 05% (03 pts) extra credit.
- If you turn in the project by 4/18/99 (5:10 pm), you will get 10% (06 pts) extra credit.
- If you turn in the project by 4/13/99 (5:10 pm), you will get 15% (09 pts) extra credit.
- If you turn in the project by 4/06/99 (5:10 pm), you will get 20% (12 pts) extra credit.
- If you turn in the project by 3/30/99 (5:10 pm), you will get 25% (15 pts) extra credit.
- If you don't turn in the project by 5/04/99 (3:50 pm), you will get no credit.