Homework #4
406.424 Internet Applications
Spring, 2010

Due Date: May 13, 2010

1. (50 points) Modify the Galago tokenizer to remove the sequence of characters following the apostrophe from the words by using the given example code and text file. Describe the new rules your tokenizer implements. You're asked to submit the Java tokenizer program, the tokenized output, and the descriptions on your rules.

2. (20 points) Generate a list of records of the form <term frequency, ranking> using a given Zipf's Law Java code and text corpus. You're asked to submit the output list.

3. (30 points) Draw a plot for Zipf's distribution using the output list from Problem 2. You're asked to submit the plot.

All example Java codes and text files for these homework tasks will be available at course home page.

Submit your answers in MS-Word or PDF format via email to the instructor cc TA.