

Information Retrieval CS317 Fall 2016

Homework 1

Assigned: 31 Aug 2016

Due: 7th Sep 2016 (at beginning of class)

Problem 1 (2 points)

Which of the documents in Table 1 will be retrieved given the Boolean query below?

((chaucer OR milton) AND (NOT swift)) OR ((NOT chaucer) AND (swift OR shakespeare))

Table 1 Term-document weight in Boolean model

	Chaucer	Milton	Shakespeare	Swift
D1	0	0	0	0
D2	0	0	0	1
D3	0	0	1	0
D4	0	0	1	1
D5	0	1	0	0
D6	0	1	0	1
D7	0	1	1	0
D8	0	1	1	1
D9	1	0	0	0
D10	1	0	0	1
D11	1	0	1	0
D12	1	0	1	1
D13	1	1	0	0
D14	1	1	0	1
D15	1	1	1	0
D16	1	1	1	1

Problem 2 (6 points)

A. In this homework problem, you will write a quick program (You can use any programming language of your choice) to explore Zipf's Law. **(4 points)**

Go to the [Project Gutenberg website](#) and download [Alice in Wonderland](#) by Lewis Carol (Plain Text UTF-8 format). Strip off the header, and thus consider only the text starting at "ALICE'S ADVENTURES IN WONDERLAND", just preceding "CHAPTER 1"; also, strip off the footer, eliminating the license agreement and other extraneous text, and thus consider only the text up through, and including, "THE END". Use this Perl script



parse.pl

to strip the text of punctuation obtaining the original text as a list of words. For example, on a unix based systems, you should run a command like

```
parse.pl alice30.txt > output
```

Write a quick program or script that counts word frequencies. For the most frequent 25 words and for the most frequent 25 *additional* words that start with the letter *f* (a total of 50 words), print the word, the number of times it occurs, its rank in the overall list of words, the probability of occurrence, and the product of the rank and the probability. Also indicate the total number of words and the total number of unique words that you found. Discuss whether this text satisfies Zipf's Law. Feel free to use other parts of the ranked list of terms.

B. Suppose that while you were building retrieval index, you decided to omit all words that occur fewer than five times (i.e., one to four times). According to Zipf's Law, what proportion of the *total* words in the collection would you omit? (Justify your answer.) What proportion would *actually* be omitted in the *Alice in Wonderland* text above?

(2 points)

Problem 3 (6 points)

Prof. Aslam has a collection of songs on his iPod, each song played a number of times as



Song Frequencies

indicated in this ranked list

Figure 1(a) shows the frequency-rank plot for these songs (the blue dots) along with a "best fit" Zipfian model (red curve). Figure 1 (b) shows this same data on a log-log scale.

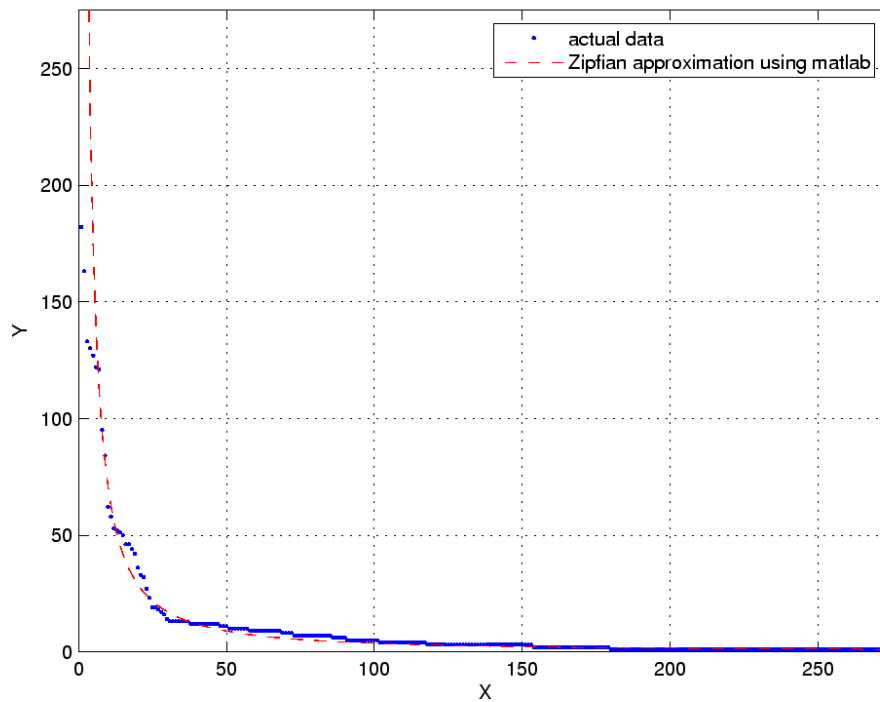


Figure 1(a): Frequency-rank plot and best fit Zipfian model.

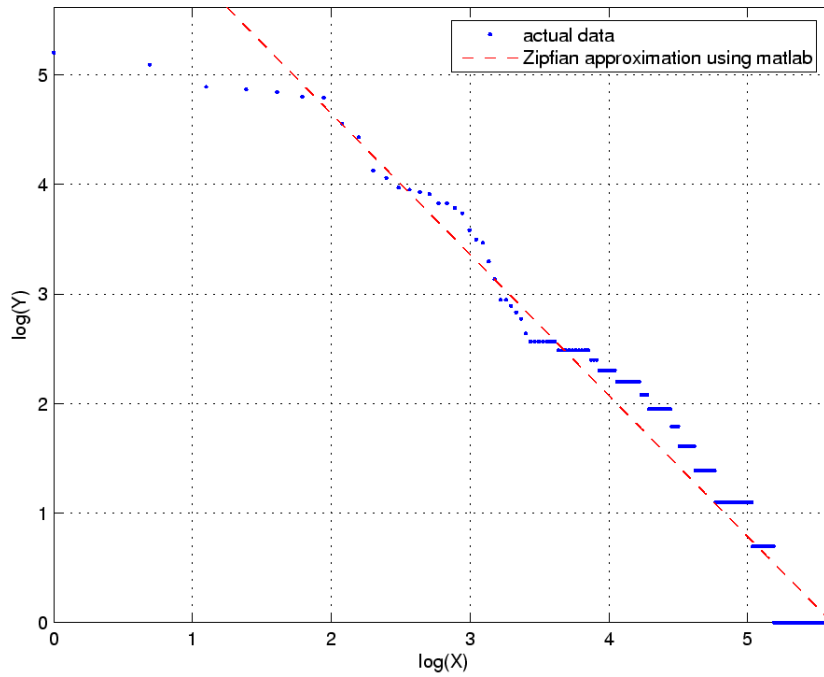


Figure 1(b) Log-log plot of data from Figure 1(a)

Your task is to find the best fit Zipfian model for this data, i.e., the parameters of the red curve, by first finding the best fitting straight line for the log-log data using the least squares technique. As a reminder, if the log-log data points are represented by

$d_i = (x_i, y_i); i = 1..n$ and you are looking for the linear function (straight line) given by

$$y = mx + b$$

then

$$b = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{\bar{y} (\sum_{i=1}^n x_i^2) - \bar{x} \sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

$$m = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{(\sum_{i=1}^n x_i y_i) - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

For the given data, compute the fitting coefficients m and b , and from m and b , compute the parameters for the best fitting Zipfian model (i.e., the red curve in Figure 1(a)).

Problem 4 (6 points)

A. According to Heaps' Law, what proportion of a collection of text must be read before 90% of its vocabulary has been encountered? You may assume that $\beta = 0.5$. Hint: to solve this problem you don't need to know the value of K . **(3 points)**

B. Verify Heap's Law on the *Alice in Wonderland* text. Process each word of the text, in order, and compute the following pairs of numbers: (number of words processed, number of unique words seen). These pairs of numbers, treated as (x,y) values, should satisfy Heaps Law. Appropriately transform the data and use least squares to determine the model parameters K and β , in much the same manner as Zipf's Law example from class. **(3 points)**