Assignment 2

Fall 2017
CS834 Introduction to Information Retrieval
Dr. Michael Nelson

Orkun Krand

October 14, 2017

1 Question 4.1

1.1 Question

Plot rank-frequency curves (using a log-log graph) for words and bigrams in the Wikipedia collection available through the book website (http://www.search-engines-book.com). Plot a curve for the combination of the two. What are the best values for the parameter c for each curve?

1.2 Methodology

I wrote a Python script `pagevisitor.py` that uses NLTK (Natural Language Toolkit) and BeautifulSoup to access a folder and its subfolders (looking for 'en' as default as it is the root folder of the Wikipedia corpus from the book website), and process the documents found to collect word, bigram, and inlink information. By the time I realized I should've converted everything to lowercase to avoid duplicates, it was too late. `q1graphs.r` was created to visualize the data collected by `pagevisitor.py`

1.3 Results

`pagevisitor.py` found 232,919 words and 1,662,253 bigrams. Visualizing these numbers, we receive the following graphs:

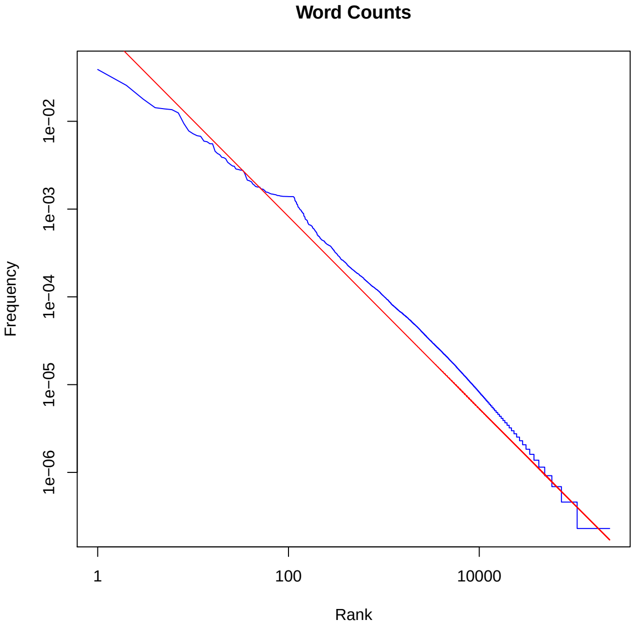


Figure 1: Word Count

Maximum likelihood estimation of words

Call: mle(minuslogl = ll, start = list(s = 1))

Coefficients: Estimate Std. Error s 1.002823 0.0001287333

-2 log L: 73354417

C = 1.269302e-01

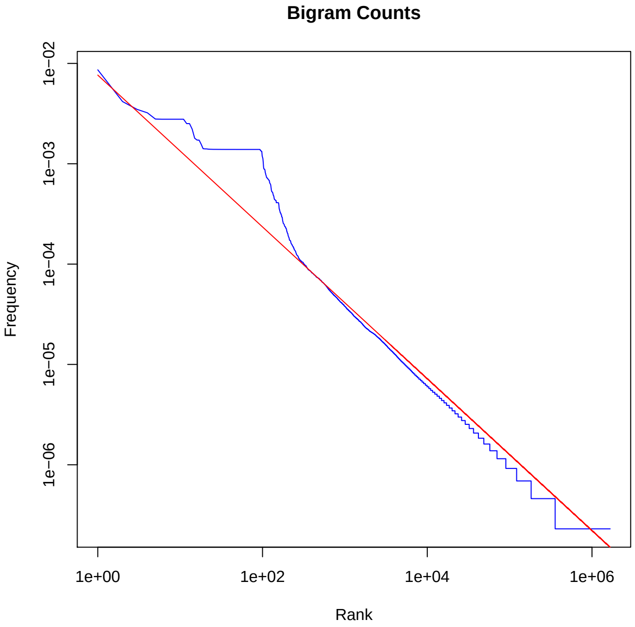


Figure 2: Bigram Count

Maximum likelihood estimation of bigrams

Call: mle(minuslogl = ll, start = list(s = 1))

Coefficients: Estimate Std. Error s 0.8292686 0.0001292527

-2 log L: 106198894

C = 7.650163e-03

Maximum likelihood estimation of combined

Call: mle(minuslogl = ll, start = list(s = 1))

Coefficients: Estimate Std. Error s 0.9250651 8.066917e-05

-2 log L: 191232526

C = 3.019035e-02

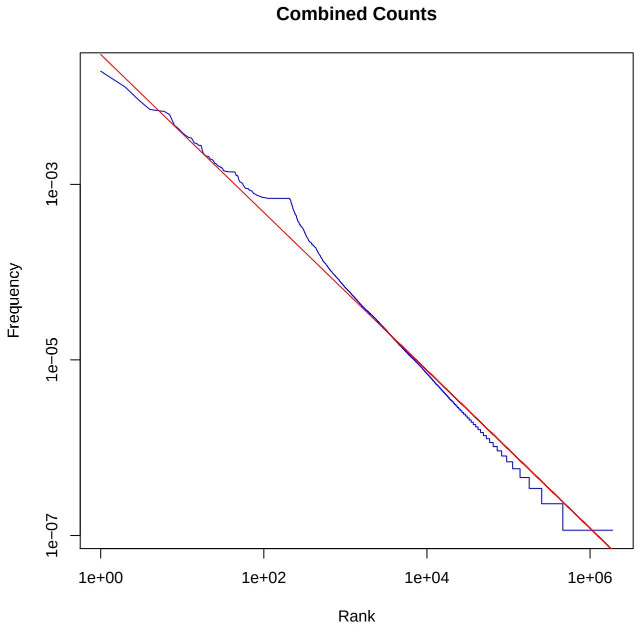


Figure 3: Both Word and Bigram Counts

2 Question 4.2

2.1 Question

Plot vocabulary growth for the Wikipedia collection and estimate the parameters for Heaps' law. Should the order in which the documents are processed make any difference?

2.2 Methodology

Modifying `pagevisitor.py` to keep track of vocabulary (unique words) and total words and logging them after accessing each document enabled me to create the required plot for this question. Using `q2graphs.r`, I was able to create visualizations.

2.3 Results

The order in which the documents are processed is important for Heap's Law because we compute the sum of all words in the dictionary after processing each document.

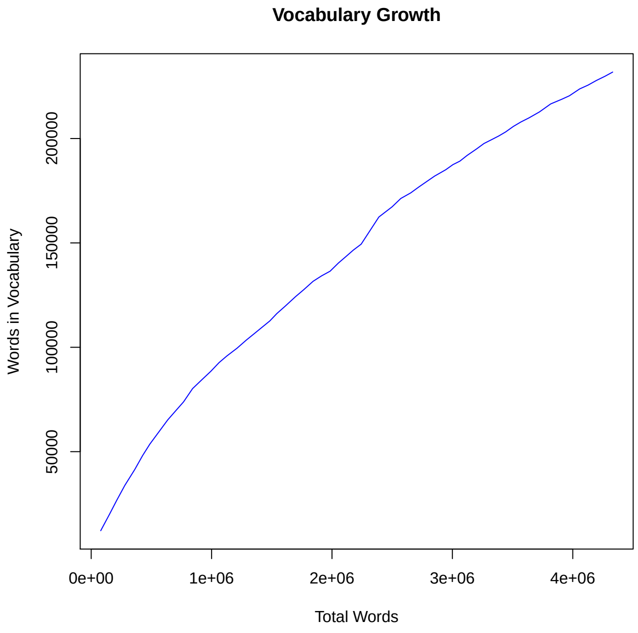


Figure 4: Vocabulary Growth

3 Question 4.3

3.1 Question

Try to estimate the number of web pages indexed by two different search engines using the technique described in this chapter. Compare the size estimates from a range of queries and discuss the consistency (or lack of it) of these estimates.

3.2 Methodology

According to the book[1], assuming the two terms are independent of each other,

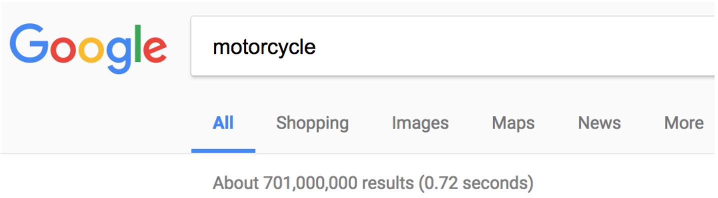
$$f_{ab} = N \cdot f_a/N \cdot f_b/N \tag{1}$$

$$N = (f_a \cdot f_b)/f_{ab} \tag{2}$$

I will use Google and Yahoo! for this assignment. The queries I will use are "motorcycle cake" and "yamaha quinoa".

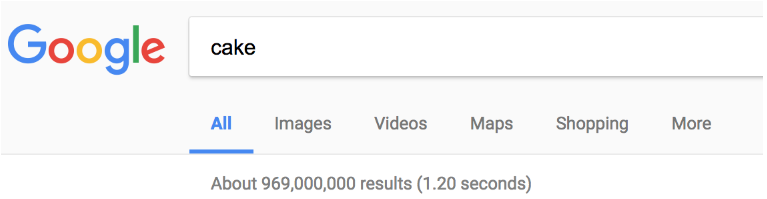
3.3 Results

Working with Google first:


motorcycle

About 701,000,000 results (0.72 seconds)

Figure 5: Google: motorcycle

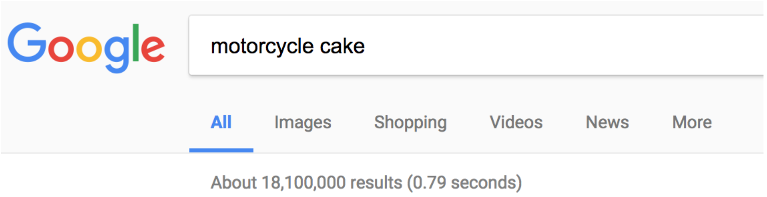

cake

About 969,000,000 results (1.20 seconds)

Figure 6: Google: cake

According to the first query, the size of Google is:

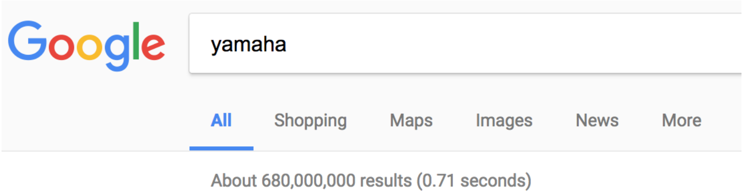
$$N = \frac{(701,000,000 \cdot 969,000,000)}{18,100,000} = 37,528,674,033 \tag{3}$$



motorcycle cake

About 18,100,000 results (0.79 seconds)

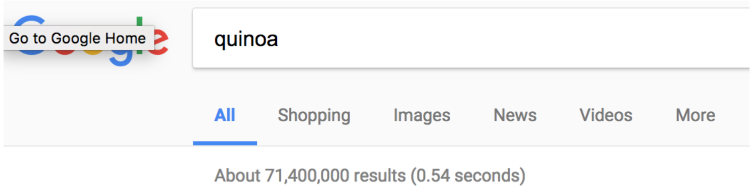
Figure 7: Google: motorcycle cake



yamaha

About 680,000,000 results (0.71 seconds)

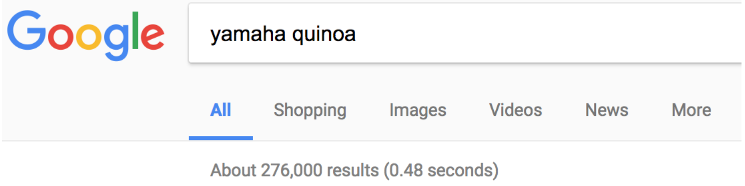
Figure 8: Google: yamaha



quinoa

About 71,400,000 results (0.54 seconds)

Figure 9: Google: quinoa



yamaha quinoa

About 276,000 results (0.48 seconds)

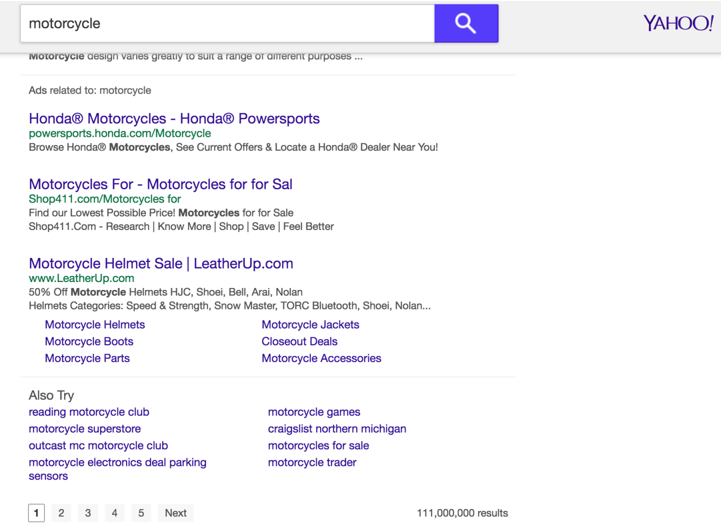
Figure 10: Google: yamaha quinoa

The second query, using the equation in the book gives us:

$$N = \frac{(680,000,000 \cdot 71,400,000)}{276,000} = 175,913,043,478 \tag{4}$$

These two numbers are very far away from each other. This may be due to *motorcycle* and *cake* both being generic terms whereas *yamaha* and *quinoa* are more specific terms that are not related. So if I had to pick between these two results, I would pick the larger (176B) as the better estimation for the size of Google.

Let's do the same for Yahoo!:



motorcycle

Motorcycle design varies greatly to suit a range of different purposes ...

Ads related to: motorcycle

Honda® Motorcycles - Honda® Powersports
powersports.honda.com/Motorcycle
Browse Honda® **Motorcycles**, See Current Offers & Locate a Honda® Dealer Near You!

Motorcycles For - Motorcycles for for Sal
Shop411.com/Motorcycles for
Find our Lowest Possible Price! **Motorcycles** for for Sale
Shop411.Com - Research | Know More | Shop | Save | Feel Better

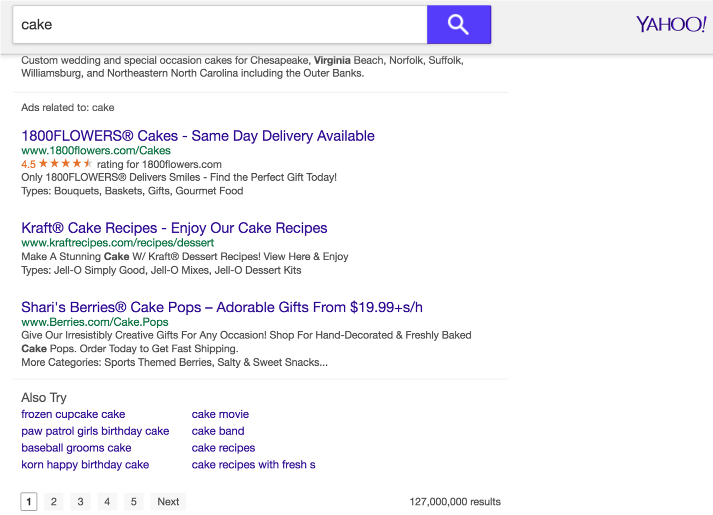
Motorcycle Helmet Sale | LeatherUp.com
www.LeatherUp.com
50% Off **Motorcycle** Helmets HJC, Shoei, Bell, Arai, Nolan
Helmets Categories: Speed & Strength, Snow Master, TORC Bluetooth, Shoei, Nolan...

Motorcycle Helmets
Motorcycle Boots
Motorcycle Parts
Motorcycle Jackets
Closeout Deals
Motorcycle Accessories

Also Try
reading motorcycle club
motorcycle superstore
outcast mc motorcycle club
motorcycle electronics deal parking sensors
motorcycle games
craigslist northern michigan
motorcycles for sale
motorcycle trader

111,000,000 results

Figure 11: Yahoo: motorcycle



cake

Custom wedding and special occasion cakes for Chesapeake, **Virginia** Beach, Norfolk, Suffolk, Williamsburg, and Northeastern North Carolina including the Outer Banks.

Ads related to: cake

1800FLOWERS® Cakes - Same Day Delivery Available
www.1800flowers.com/Cakes
4.5 ★★★★★ rating for 1800flowers.com
Only 1800FLOWERS® Delivers Smiles - Find the Perfect Gift Today!
Types: Bouquets, Baskets, Gifts, Gourmet Food

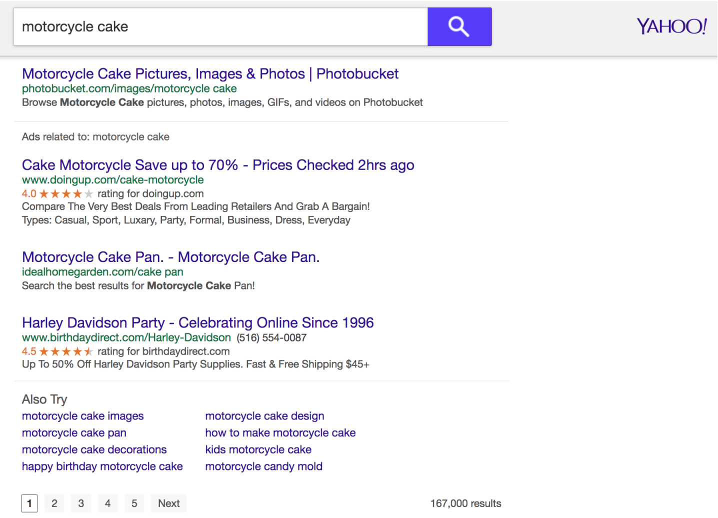
Kraft® Cake Recipes - Enjoy Our Cake Recipes
www.kraftrecipes.com/recipes/dessert
Make A Stunning **Cake** W/ Kraft® Dessert Recipes! View Here & Enjoy
Types: Jell-O Simply Good, Jell-O Mixes, Jell-O Dessert Kits

Shari's Berries® Cake Pops – Adorable Gifts From $19.99+s/h
www.Berries.com/Cake.Pops
Give Our Irresistibly Creative Gifts For Any Occasion! Shop For Hand-Decorated & Freshly Baked **Cake** Pops. Order Today to Get Fast Shipping.
More Categories: Sports Themed Berries, Salty & Sweet Snacks...

Also Try
frozen cupcake cake
paw patrol girls birthday cake
baseball grooms cake
korn happy birthday cake
cake movie
cake band
cake recipes
cake recipes with fresh s

127,000,000 results

Figure 12: Yahoo: cake



motorcycle cake

YAHOO!

Motorcycle Cake Pictures, Images & Photos | Photobucket
photobucket.com/images/motorcycle cake
Browse **Motorcycle Cake** pictures, photos, images, GIFs, and videos on Photobucket

Ads related to: motorcycle cake

Cake Motorcycle Save up to 70% - Prices Checked 2hrs ago
www.doingup.com/cake-motorcycle
4.0 ★★★★☆ rating for doingup.com
Compare The Very Best Deals From Leading Retailers And Grab A Bargain!
Types: Casual, Sport, Luxary, Party, Formal, Business, Dress, Everyday

Motorcycle Cake Pan. - Motorcycle Cake Pan.
idealhomegarden.com/cake pan
Search the best results for **Motorcycle Cake** Pan!

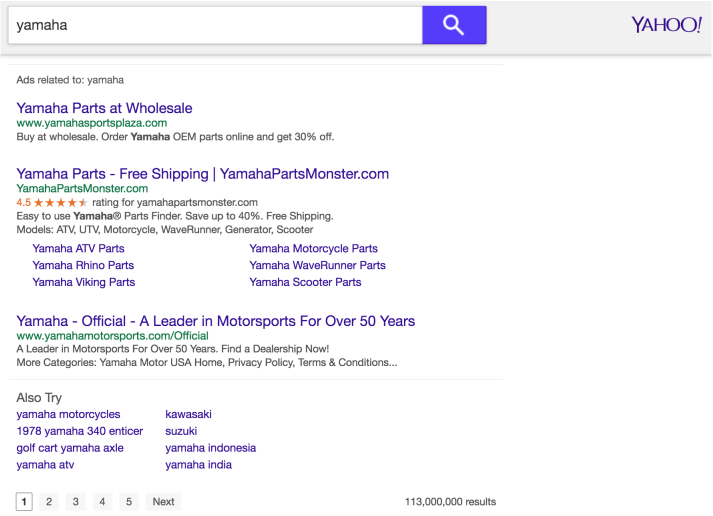
Harley Davidson Party - Celebrating Online Since 1996
www.birthdaydirect.com/Harley-Davidson (516) 554-0087
4.5 ★★★★★ rating for birthdaydirect.com
Up To 50% Off Harley Davidson Party Supplies. Fast & Free Shipping $45+

Also Try

motorcycle cake images	motorcycle cake design
motorcycle cake pan	how to make motorcycle cake
motorcycle cake decorations	kids motorcycle cake
happy birthday motorcycle cake	motorcycle candy mold

1 2 3 4 5 Next 167,000 results

Figure 13: Yahoo: motorcycle cake



yamaha

YAHOO!

Ads related to: yamaha

Yamaha Parts at Wholesale
www.yamahasportsplaza.com
Buy at wholesale. Order **Yamaha** OEM parts online and get 30% off.

Yamaha Parts - Free Shipping | YamahaPartsMonster.com
YamahaPartsMonster.com
4.5 ★★★★★ rating for yamahapartsmonster.com
Easy to use **Yamaha**® Parts Finder. Save up to 40%. Free Shipping.
Models: ATV, UTV, Motorcycle, WaveRunner, Generator, Scooter

Yamaha ATV Parts	Yamaha Motorcycle Parts
Yamaha Rhino Parts	Yamaha WaveRunner Parts
Yamaha Viking Parts	Yamaha Scooter Parts

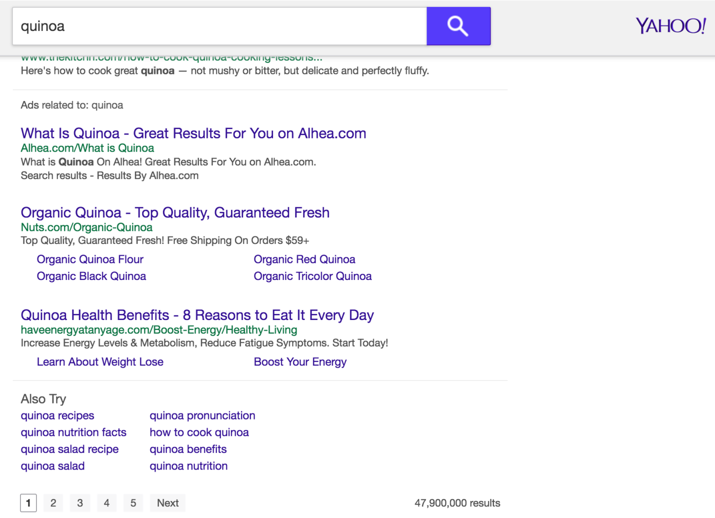
Yamaha - Official - A Leader in Motorsports For Over 50 Years
www.yamahamotorsports.com/Official
A Leader in Motorsports For Over 50 Years. Find a Dealership Now!
More Categories: Yamaha Motor USA Home, Privacy Policy, Terms & Conditions...

Also Try

yamaha motorcycles	kawasaki
1978 yamaha 340 enticer	suzuki
golf cart yamaha axle	yamaha indonesia
yamaha atv	yamaha india

1 2 3 4 5 Next 113,000,000 results

Figure 14: Yahoo: yamaha



quinoa

YAHOO!

www.thekitchn.com/how-to-cook-quinoa-cooking-lessons...
Here's how to cook great **quinoa** — not mushy or bitter, but delicate and perfectly fluffy.

Ads related to: quinoa

What Is Quinoa - Great Results For You on Alhea.com
Alhea.com/What is Quinoa
What is **Quinoa** On Alhea! Great Results For You on Alhea.com.
Search results - Results By Alhea.com

Organic Quinoa - Top Quality, Guaranteed Fresh
Nuts.com/Organic-Quinoa
Top Quality, Guaranteed Fresh! Free Shipping On Orders $59+

Organic Quinoa Flour | Organic Red Quinoa
Organic Black Quinoa | Organic Tricolor Quinoa

Quinoa Health Benefits - 8 Reasons to Eat It Every Day
haveenergyatanyage.com/Boost-Energy/Healthy-Living
Increase Energy Levels & Metabolism, Reduce Fatigue Symptoms. Start Today!

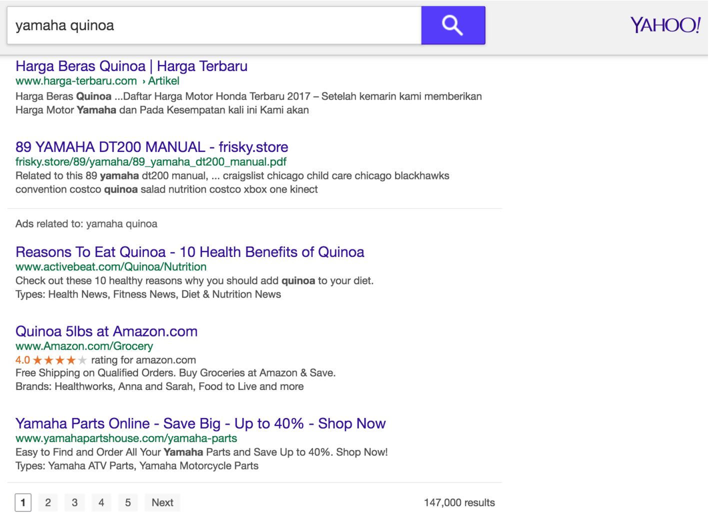
Learn About Weight Lose | Boost Your Energy

Also Try

quinoa recipes	quinoa pronunciation
quinoa nutrition facts	how to cook quinoa
quinoa salad recipe	quinoa benefits
quinoa salad	quinoa nutrition

47,900,000 results

Figure 15: Yahoo: quinoa



yamaha quinoa

YAHOO!

Harga Beras Quinoa | Harga Terbaru
www.harga-terbaru.com › Artikel
Harga Beras **Quinoa** ...Daftar Harga Motor Honda Terbaru 2017 – Setelah kemarin kami memberikan Harga Motor **Yamaha** dan Pada Kesempatan kali ini Kami akan

89 YAMAHA DT200 MANUAL - frisky.store
frisky.store/89/yamaha/89_yamaha_dt200_manual.pdf
Related to this 89 **yamaha** dt200 manual, ... craigslist chicago child care chicago blackhawks convention costco **quinoa** salad nutrition costco xbox one kinect

Ads related to: yamaha quinoa

Reasons To Eat Quinoa - 10 Health Benefits of Quinoa
www.activebeat.com/Quinoa/Nutrition
Check out these 10 healthy reasons why you should add **quinoa** to your diet.
Types: Health News, Fitness News, Diet & Nutrition News

Quinoa 5lbs at Amazon.com
www.Amazon.com/Grocery
4.0 ★★★★☆ rating for amazon.com
Free Shipping on Qualified Orders. Buy Groceries at Amazon & Save.
Brands: Healthworks, Anna and Sarah, Food to Live and more

Yamaha Parts Online - Save Big - Up to 40% - Shop Now
www.yamahapartshouse.com/yamaha-parts
Easy to Find and Order All Your **Yamaha** Parts and Save Up to 40%. Shop Now!
Types: Yamaha ATV Parts, Yamaha Motorcycle Parts

147,000 results

Figure 16: Yahoo: yamaha quinoa

According to the first query, the size of Yahoo! is:

$$N = \frac{(111,000,000 \cdot 127,000,000)}{167,000} = 84,413,173,652 \quad (5)$$

The second query, using the equation in the book gives us:

$$N = \frac{(113,000,000 \cdot 47,900,000)}{147,000} = 36,821,088,435 \quad (6)$$

Unlike the first case, the second equation gave a smaller number than the first. But I still find the second number more believable because of the reason I explained earlier. This would mean that Google is almost 5 times bigger than Yahoo!. Checking out the real numbers surely would be interesting.

4 Question 4.8

4.1 Question

Find the 10 Wikipedia documents with the most inlinks. Show the collection of anchor text for those pages.

4.2 Methodology

Modifying `pagevisitor.py` to find the inlinks for every page accessed was required for to find the top 10. Anchor text was also added to the output. Sorting them based on the number of inlinks gave me the top 10 pages.

4.3 Results

Number of Inlinks	URI	Anchor Text
2264	articles/2/0/0/2007.html	2007, As of 2007
1896	articles/s/m/a/User%7ESmackBot_cc7a.html	SmackBot
1770	articles/2/0/0/2008.html	2008
1363	articles/u/n/i/United_States_09d4.html	United States Of America, USA, Union, Thirteen Colonies, U.S., United States of America, US, United States, American, Americans, America, American nation, American citizen
982	articles/2/0/0/2006.html	2006
791	articles/a/l/a/User%7EAlaibot_de3d.html	Alaibot
676	articles/c/y/d/User%7ECydebot_38a6.html	Cydebot
675	articles/l/i/v/Category%7ELiving_people_7259.html	Living people
663	articles/b/l/u/User%7EBluebot_e595.html	Bluebot
655	articles/g/e/o/Geographic_coordinate_system.html	coordinates, Location, Coordinates

Table 1: Top Ten Pages With the Highest Number of Inlinks

5 Question 5.8

5.1 Question

Write a program that can build a simple inverted index of a set of text documents. Each inverted list will contain the file names of the documents that contain that word. Suppose the file A contains the text “the quick brown fox”, and file B contains “the slow blue fox”.

```
The output of your program would be:
% ./your-program A B
blue B
brown A
fox A B
quick A
slow B
the A B
```

5.2 Methodology

Modifying the `pagevisitor.py` script allowed me to create an inverted index of the Wikipedia corpus. I wasn’t sure what you meant by examples from the data set so I went ahead and created an inverted index for the whole thing. It ended up being a big file (104MB) but oh well.

References

[1] Croft, William Bruce, et al. *Search Engines: Information Retrieval in Practice*. Pearson, 2010.