

Metrics of successful websites and companies

Danai Avratoglou

January 2017

Contents

1	Introduction	3
2	Data gathering	4
2.1	Data Source - Fortune 500	4
2.2	Metrics	5
2.3	Python	7
2.4	Scripts	7
2.4.1	Fortune data	7
2.4.2	Html download	10
2.4.3	Social media existence	10
2.4.4	Readability index	10
2.4.5	10
3	Data Analysis	11
3.1	Data loading	11
3.2	R	11
3.3	Scripts	11
3.3.1	Statistical Analysis	11
3.3.2	Regression Model	11
3.3.3	Clustering	11
4	Conclusions	12
5	Bibliography	13
A		
	Appendix A: Fortune 500 Companies	14

1 Introduction

An on-line presence of a company was not an important factor of its success until a few years ago. Taking although into account the vast spread of the impact that internet has on consumers regarding their choices this hypothesis is not valid any more. Companies are obliged by the trends to be active on-line and to maintain a website that depicts the image they want their consumers to perceive. The purpose of this paper is to understand the relationship that exists between the website of a company and its success. Trying to comprehend this relationship a comparison will take place between specific metrics of the websites of the Fortune's 500 more successful companies and their financial status. By performing regression models and statistical analysis this paper will try to explore which metrics of a company's website influence its success.

2 Data gathering

The first step in order to contact this research is to conclude to the companies that are going to be examined. Since the purpose of this paper is to see if the website metrics that will be examined are influencing the success of the company it is a good idea to examine websites of some already successful companies and try to find out what they have in common.

2.1 Data Source - Fortune 500

The Fortune 500 is an annual list compiled and published by Fortune magazine that ranks 500 of the largest U.S. corporations by total revenue for their respective fiscal years. The list includes public companies, along with privately held companies for which revenues are publicly available.[1, 2]

For the purposes of this paper we will use this list of companies and we will examine their websites in order to understand if they indeed have something in common or if their success is irrelevant with their on-line presence.

The first thing that we will need is a list of the names that are include in the fortune 500. The easiest way to obtain this list is from the following article: <http://www.zyxware.com/articles/4344/list-of-fortune-500-companies-and-their-websites>. The way that we will obtain the list will be explained in the 2.4 Scripts section of this chapter. In the following table one can see the 50 most succesful companies that are included in the Fortune 500 during the period this paper is taking place. The rest of the list is available in the Appendix A.A

Table 1: Fortune 500 - 50 first companies

1. Walmart	2. Exxon Mobil	3. Apple
4. Berkshire Hathaway	5. McKesson	6. UnitedHealth Group
7. CVS Health	8. General Motors	9. Ford Motor
10. AT&T	11. General Electric	12. AmerisourceBergen
13. Verizon	14. Chevron	15. Costco
16. Fannie Mae	17. Kroger	18. Amazon.com
19. Walgreens Boots Alliance	20. HP	21. Cardinal Health
22. Express Scripts Holding	23. J.P. Morgan Chase	24. Boeing
25. Microsoft	26. Bank of America Corp.	27. Wells Fargo
28. Home Depot	29. Citigroup	30. Phillips 66
31. IBM	32. Valero Energy	33. Anthem
34. Procter & Gamble	35. State Farm Insurance Cos.	36. Alphabet
37. Comcast	38. Target	39. Johnson & Johnson
40. MetLife	41. Archer Daniels Midland	42. Marathon Petroleum
43. Freddie Mac	44. PepsiCo	45. United Technologies
46. Aetna	47. Lowe's	48. UPS
49. AIG	50. Prudential Financial	

2.2 Metrics

Now that we have declared the companies that we are going to use we will also need to decide which metrics are we going to examine for each site. Since we cannot have access to metrics such as traffic we will have to examine metrics that are more related to how the site is structure and what exactly does the initial page of each site includes. Initially we can divide the metrics in two major categories:

- What we see?
- What lays behind of what we see?

In the first category we are referring to metrics that can easily be conceived by the naked eye as well. For example the images that a website is using in its landing page. How many there are and if they are big or small.

The second category is not so obvious and it includes informations that usually is visible only to the web developer or the creator of the page. The information here are being given from the html code of a site. For instance we can see if the html code has any errors, that can lead to some malfunction in the site for example.

Now that we have a first understanding of the two main categories that the metrics we will use are divided in, we can see in detail the metrics that will be examined in this paper:

Loading time: One aspect of a website that is crucial is the time it takes for it to load. Nowadays that the internet speed is going higher and higher most people do not have the patient to wait for a page to load. That is why we think that a metrics that should be definitely included in this research in the loading time of the initial page of a site.

Number of links: When someone is browsing through the internet in many cases they are not completely sure what exactly they are looking for and that is why in a website it would be wise to have some links that can direct the user to find what he wants. These links can either direct the user in another page of the same site, in which case the link will be characterized as an internal link. Or they can lead the user to another site, where in that case the link is characterized as an external link. For the purpose of this paper since it is not so clear which type of links are more important to a user we will examine both the internal and the external links.

Social media: Our era is marked by the social media wave that has changed our lifestyle and our daily habits. So it would be considered an overview if we didn't take under consideration the number of social media that the company chooses to participate in. Even though they can also be considered as external links of the website we will examine them separately in order to see if any particular social medium effects the company's revenues.

Number and size of images: Since the site is the first thing that a user will see for the company and there is a famous quote that says that "*First impressions counts*" we should also examine how the companies decide to visualize their landing page. In other words to see how many images they include in it and more on that what sizes are those images.

It is completely different to see only one huge image in the landing page of a site with not many words or descriptions than to see many small images with different information. The purpose is to examine if these type of diversities between the examined websites are actually related to how they are doing profit wise.

Type of images: This kind of information is a little more complicated for a simple user to understand, but in many cases it plays a very important role. For example some websites are using specific type of images or banners that are not compatible with all the browsers, leading the user to see some break points in the website and even stop visiting it. That is why we believe is important to review this metric as well.

Number of words: They say a picture worth a thousand words but that is not enough in our case. After exploring the number, sizes and type of pictures that a website is using we should also explore the number of words it is using to accompany the images and complete the outcome that a user will come across. The metrics we will use will be two. The first one will be the total words that are being used in the landing page and the other one will be the total unique words that are being used. When we are referring to unique words we mean words that are not so commonly use such as "a" or "and" and they give an air of individuality to the text. By using this metric we would try to see if the words that are being used are just as important as the actual content and if the words can make a difference.

Readability index: In order for the previous metrics to be completed we will also have to take under consideration how comprehending is the text used in the websites for the users. This information can be obtained by calculating the readability index of the website. How the readability index is being calculating will be furthered explained in another section.

HTML Validation: Moreover we will have to check the quality of the html code behind the website we are seeing. Are there any mistakes in the code for example any brackets that opened and never closed or any links that do not work. We will examine again two different metrics here the number of errors and the number of warnings. The warning are parts of the code that even though they work at the time there is a good chance to malfunction if any changes or addition are to be made to the html code.

2.3 Python

After having a first look into the variables/ metrics we will use in order to contact this research we should also see how we are going to obtain all this information.

Since all of this information can be subtract from the html code of a company's website we should use a programming language in order to download the html pages and then to extract the specific metrics we want to examine from them. For the purposes of this paper the programming language that will be used for downloading the metrics from the websites is Python. More specifically the version of Python that will be used is the 2.7 one. Python is a widely used high-level programming language used for general-purpose programming, created by Guido van Rossum and first released in 1991. An interpreted language, Python has a design philosophy which emphasizes code readability (notably using white space indentation to delimit code blocks rather than curly braces or keywords), and a syntax which allows programmers to express concepts in fewer lines of code than possible in languages such as C++ or Java.

The language provides constructs intended to enable writing clear programs on both a small and large scale. Furthermore the way that Python allows a user to programming is common to all users which gives this language a leverage as a program build in Python can be easily understood from another user without any difficulty.

The environment that is going to be used is from the Anaconda package which is a free open source distribution of the Python and R programming languages for large-scale data processing, predictive analytics, and scientific computing, that aims to simplify package management and deployment. To be more precise from this package we are going to use the Jupyter Notebook.[4]

2.4 Scripts

In order to gather all the needed metrics we had to create many small scripts so as to collect them. In this section we will present in detail the code that was used in order to extract the information that later will help us contact the analysis of the relationship between those metrics and the company's status.

2.4.1 Fortune data

The first step is to download and gather the names of the companies that we are going to examine and also the url of their websites and finally their ranking. Those information can be found in the following url as he have already mentioned in the beginning. <http://www.zyxware.com/articles/4344/list-of-fortune-500-companies-and-their-websites>.

The way to keep only the those three informations as separate variables is by separating from the html code of this page the needed information.

The first step is to create 3 empty lists where we will include the informations

we are going to extract. The first list will contain the rank of each site, the second one will contain the name of the company and the 3rd one will contain the actual link of the company's site:

```
list_company_number = []  
list_company_name = []  
list_company_website = []
```

The second step is to upload some libraries that will help us create this function but also the ones that are going to follow.

```
import urllib  
import urllib2  
import time  
import os  
from bs4 import BeautifulSoup  
import re  
import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt
```

Finally the third step is to create the function that will firstly download the html code of the url at hand, secondly keep only the part of the code that we need to examine and thirdly save this part into the empty lists we created above. This function is called `websites` and takes as variable to work only the url of the site we need to examine:

```
def websites (url):  
    from time import time  
    start = time ()  
    browser = urllib2.build_opener()  
    browser.addheaders = [( 'User-agent ', 'Mozilla/5.0 ')]  
    response = browser.open(url)  
    myHTML = response.read()  
    soup = BeautifulSoup(myHTML,"lxml")  
    o = 0  
    td_list = []  
    for row2 in soup.html.body.findAll( 'td '):  
        td_list.insert(o, row2)  
        o = o + 1  
    a = 0  
    b = 1  
    c = 2  
    list_numbering = 0  
    for i in range (0,500):  
        num = str(td_list[a])  
        company = str(td_list[b])  
        site = str(td_list[c])
```



```

c_num = re.findall('>(.*?)</td>', num)
c_num = str(c_num[0])
c_name = re.findall('>(.*?)</td>', company)
c_name = str(c_name[0])
c_site = re.findall('">(.*?)</a>', site)
c_site = str(c_site[0])
list_company_number.insert(list_numbering, c_num)
list_company_name.insert(list_numbering, c_name)
list_company_website.insert(list_numbering, c_site)
a = a + 3
b = b + 3
c = c + 3
list_numbering = list_numbering + 1
end = time ()
duration = round (end - start , 1)
minutes = round (duration /60, 1)
print 'The_lists_are_ready_in_', duration , '_seconds'
print 'The_lists_are_ready_in_', minutes , '_minutes'

```

The steps we followed to create the following function are the following:

Step 1 We create a fake browser that we are going to use in order to open the page and downloaded. The reason we do that is that many sites do not allow us to download their page because they are afraid of stealing important information. Since we are not using any private information we use this method to avoid issues while trying to open the html page at hand.

Step 2 We open the url and we read it while saving it in the variable "my-HTML".

Step 3 With the help of the BeautifulSoup library we read the page as a lxml file and then for each row of this file we are looking for the "td" parts of the code where the informations we want are included.

Step 4 Since we need the names and the urls of all the 500 sites we created a loop from 0 to 500 where for each i we try to isolate the part of the code that contains the information that we want. Moreover even though it seems that with this loop we calculate 501 numbers since in Python the second bracket is always open we actually count from zero to 499.

Step 5 We use reg expressions¹ in order to state precisely what part of the already selected code we want to keep.

¹A regular expression is a special sequence of characters that helps you match or find other strings or sets of strings, using a specialized syntax held in a pattern.

Step 6 We insert with a specific order the names, the ranking and the url to the corresponding lists and finally we create a text that will appear when the function is completed. Here we have also calculated the time that this function took to be completed and we will appear it as well along with the text.

2.4.2 Html download

With what codes did I downloaded the metrics I needed

2.4.3 Social media existence

With what codes did I downloaded the metrics I needed

2.4.4 Readability index

With what codes did I downloaded the metrics I needed

2.4.5 ...

With what codes did I downloaded the metrics I needed

Note Write the sites I needed and see where to put the codes

3 Data Analysis

To see the other commands in action, suppose at this point of text I type

3.1 Data loading

What metrics I needed to download

3.2 R

A few words about the language i used and why

3.3 Scripts

With what codes did I downloaded the metrics I needed

3.3.1 Statistical Analysis

With what codes did I downloaded the metrics I needed

3.3.2 Regression Model

With what codes did I downloaded the metrics I needed

3.3.3 Clustering

With what codes did I downloaded the metrics I needed

4 Conclusions

5 Bibliography

References

- [1] *<https://en.wikipedia.org/wiki/Fortune500>*
- [2] *<http://beta.fortune.com/fortune500>*
- [3] *<http://www.tablesgenerator.com/>*
- [4] *<https://www.continuum.io/downloads>*

A

Appendix A: Fortune 500 Companies

Table 2: Fortune 500 - Companies Ranked: 51 - 100

51. Intel	52. Humana	53. Disney
54. Cisco Systems	55. Pfizer	56. Dow Chemical
57. Sysco	58. FedEx	59. Caterpillar
60. Lockheed Martin	61. N.Y. Life Insurance	62. Coca-Cola
63. HCA Holdings	64. Ingram Micro	65. Energy Transfer Equity
66. Tyson Foods	67. American Airlines Group	68. Delta Air Lines
69. Nationwide	70. Johnson Controls	71. Best Buy
72. Merck	73. Liberty Mutual I.G.	74. Goldman Sachs Group
75. Honeywell International	76. Massachusetts Mutual L.I.	77. Oracle
78. Morgan Stanley	79. Cigna	80. U.C. Holdings
81. Allstate	82. TIAA	83. INTL FCStone
84. CHS	85. American Express	86. Gilead Sciences
87. Publix Super Markets	88. General Dynamics	89. TJX
90. ConocoPhillips	91. Nike	92. World Fuel Services
93. 3M	94. Mondelez International	95. Exelon
96. Twenty-First Century Fox	97. Deere	98. Tesoro
99. Time Warner	100. Northwestern Mutual	

Table 3: Fortune 500 - Companies Ranked: 101 - 150

101. DuPont	102. Avnet	103. Macy's
104. Enterprise Products Partners	105. Travelers Cos.	106. Philip Morris International
107. Rite Aid	108. Tech Data	109. McDonald's
110. Qualcomm	111. Sears Holdings	112. Capital One Financial
113. EMC	114. USAA	115. Duke Energy
116. Time Warner Cable	117. Halliburton	118. Northrop Grumman
119. Arrow Electronics	120. Raytheon	121. Plains GP Holdings
122. US Foods Holding	123. AbbVie	124. Centene
125. Community Health Systems	126. Alcoa	127. International Paper
128. Emerson Electric	129. Union Pacific	130. Amgen
131. U.S. Bancorp	132. Staples	133. Danaher
134. Whirlpool	135. Aflac	136. AutoNation
137. Progressive	138. Abbott Laboratories	139. Dollar General
140. Tenet Healthcare	141. Eli Lilly	142. Southwest Airlines
143. Penske Automotive Group	144. ManpowerGroup	145. Kohl's
146. Starbucks	147. Paccar	148. Cummins
149. Altria Group	150. Xerox	

Table 4: Fortune 500 - Companies Ranked: 151 - 200

151. Kimberly-Clark	152. Hartford F.S.G.	153. Kraft Heinz
154. Lear	155. Fluor	156. AECOM
157. Facebook	158. Jabil Circuit	159. CenturyLink
160. Supervalu	161. General Mills	162. Southern
163. NextEra Energy	164. Thermo Fisher Scientific	165. American Electric Power
166. PG&E Corp.	167. NGL Energy Partners	168. Bristol-Myers Squibb
169. Goodyear Tire & Rubber	170. Nucor	171. PNC F.S.G.
172. Health Net	173. Micron Technology	174. Colgate-Palmolive
175. Freeport-McMoRan	176. ConAgra Foods	177. Gap
178. Baker Hughes	179. Bank of N.Y. Mellon C.	180. Dollar Tree
181. Whole Foods Market	182. PPG Industries	183. Genuine Parts
184. Icahn Enterprises	185. Performance Food Group	186. Omnicom Group
187. DISH Network	188. FirstEnergy	189. Monsanto
190. AES	191. CarMax	192. National Oilwell Varco
193. NRG Energy	194. Western Digital	195. Marriott International
196. Office Depot	197. Nordstrom	198. Kinder Morgan
199. Aramark	200. DaVita HealthCare Partners	

Table 5: Fortune 500 - Companies Ranked: 201 - 250

201. Molina Healthcare	202. WellCare Health Plans	203. CBS
204. Visa	205. Lincoln National	206. Ecolab
207. Kellogg	208. C.H. Robinson Worldwide	209. Textron
210. Loews	211. Illinois Tool Works	212. Synnex
213. Viacom	214. HollyFrontier	215. Land O'Lakes
216. Devon Energy	217. PBF Energy	218. Yum Brands
219. Texas Instruments	220. CDW	221. Waste Management
222. Marsh & McLennan	223. Chesapeake Energy	224. Parker-Hannifin
225. Occidental Petroleum	226. Guardian Life I.C.A.	227. Farmers Ins. Exchange
228. J.C. Penney	229. Consolidated Edison	230. Cognizant Tech. Solutions
231. VF	232. Ameriprise Financial	233. Computer Sciences
234. L Brands	235. Jacobs Engineering Group	236. Principal Financial
237. Ross Stores	238. Bed Bath & Beyond	239. CSX
240. Toys R Us	241. Las Vegas Sands	242. Leucadia National
243. Dominion Resources	244. United States Steel	245. L-3 Communications
246. Edison International	247. Entergy	248. ADP
249. First Data	250. BlackRock	

Table 6: Fortune 500 - Companies Ranked: 251 - 300

251. WestRock	252. Voya Financial	253. Sherwin-Williams
254. Hilton Worldwide Holdings	255. R.R. Donnelley & Sons	256. Stanley Black & Decker
257. Xcel Energy	258. Murphy USA	259. CBRE Group
260. D.R. Horton	261. Estee Lauder	262. Praxair
263. Biogen	264. State Street Corp.	265. Unum Group
266. Reynolds American	267. Group 1 Automotive	268. Henry Schein
269. Hertz Global Holdings	270. Norfolk Southern	271. Reinsurance G. of America
272. Public Service E. G.	273. BB&T Corp.	274. DTE Energy
275. Assurant	276. Global Partners	277. Huntsman
278. Becton Dickinson	279. Semptra Energy	280. AutoZone
281. Navistar International	282. Precision Castparts	283. Discover F. S.
284. Liberty Interactive	285. W.W. Grainger	286. Baxter International
287. Stryker	288. Air Products & Chemicals	289. Western Refining
290. Universal Health Services	291. Owens & Minor	292. Charter Communications
293. Advance Auto Parts	294. MasterCard	295. Applied Materials
296. Eastman Chemical	297. Sonic Automotive	298. Ally Financial
299. CST Brands	300. eBay	

Table 7: Fortune 500 - Companies Ranked: 301 - 350

301. Lennar	302. GameStop	303. Reliance Steel & Aluminum
304. Hormel Foods	305. Celgene	306. Genworth Financial
307. PayPal Holdings	308. Priceline Group	309. MGM Resorts International
310. Autoliv	311. Fidelity National Financial	312. Republic Services
313. Corning	314. Peter Kiewit Sons'	315. Univar
316. Mosaic	317. Core-Mark Holding	318. Thrivent F. for Lutherans
319. Cameron International	320. HD Supply Holdings	321. Crown Holdings
322. EOG Resources	323. Veritiv	324. Anadarko Petroleum
325. Laboratory C. of A.	326. Pacific Life	327. News Corp.
328. Jarden	329. SunTrust Banks	330. Avis Budget Group
331. Broadcom	332. American Family I. G.	333. Level 3 Communications
334. Tenneco	335. United Natural Foods	336. Dean Foods
337. Campbell Soup	338. Mohawk Industries	339. BorgWarner
340. PVH	341. Ball	342. O'Reilly Automotive
343. Eversource Energy	344. Franklin Resources	345. Masco
346. Lithia Motors	347. KKR	348. Oneok
349. Newmont Mining	350. PPL	

Table 8: Fortune 500 - Companies Ranked: 351 - 400

351. SpartanNash	352. Quanta Services	353. XPO Logistics
354. Ralph Lauren	355. Interpublic Group	356. Steel Dynamics
357. WESCO International	358. Quest Diagnostics	359. Boston Scientific
360. AGCO	361. Foot Locker	362. Hershey
363. CenterPoint Energy	364. Williams	365. Dick's Sporting Goods
366. Live Nation Entertainment	367. Mutual of Omaha Ins.	368. W.R. Berkley
369. LKQ	370. Avon Products	371. Darden Restaurants
372. Kindred Healthcare	373. Weyerhaeuser	374. Casey's General Stores
375. Sealed Air	376. Fifth Third Bancorp	377. Dover
378. Huntington Ingalls Industries	379. Netflix	380. Dillard's
381. EMCOR Group	382. Jones Financial	383. AK Steel Holding
384. UGI	385. Expedia	386. salesforce.com
387. Targa Resources	388. Apache	389. Spirit AeroSystems H.
390. Expeditors Inter. of Washington	391. Anixter International	392. Fidelity N. Inf. S.
393. Asbury Automotive Group	394. Hess	395. Ryder System
396. Terex	397. Coca-Cola Eur. P.	398. Auto-Owners Insurance
399. Cablevision Systems	400. Symantec	

Table 9: Fortune 500 - Companies Ranked: 401 - 450

401. Charles Schwab	402. Calpine	403. CMS Energy
404. Alliance Data Systems	405. JetBlue Airways	406. Discovery Communic.
407. Trinity Industries	408. Sanmina	409. NCR
410. FMC Technologies	411. Erie Insurance Group	412. Rockwell Automation
413. Dr Pepper Snapple Group	414. iHeartMedia	415. Tractor Supply
416. J.B. Hunt Transport Services	417. Commercial Metals	418. Owens-Illinois
419. Harman Inter. Ind.	420. Baxalta	421. American F. G.
422. NetApp	423. Graybar Electric	424. Oshkosh
425. Ameren	426. A-Mark Precious Metals	427. Barnes & Noble
428. Dana Holding	429. Constellation Brands	430. LifePoint Health
431. Zimmer Biomet H.	432. Harley-Davidson	433. PulteGroup
434. Newell Brands	435. Avery Dennison	436. Jones Lang LaSalle
437. WEC Energy Group	438. Marathon Oil	439. TravelCenters of A.
440. United Rentals	441. HRG Group	442. Old Republic Inter.
443. Windstream Holdings	444. Starwood Hotels & Resorts	445. Delek US Holdings
446. Packaging Corp. of A.	447. Quintiles Transnational H.	448. Hanesbrands
449. Realogy Holdings	450. Mattel	

Table 10: Fortune 500 - Companies Ranked: 451 - 500

451. Motorola Solutions	452. J.M. Smucker	453. Regions Financial
454. Celanese	455. Clorox	456. Ingredion
457. Genesis Healthcare	458. Peabody Energy	459. Alaska Air Group
460. Seaboard	461. Frontier Communic.	462. Amphenol
463. Lansing Trade Group	464. SanDisk	465. St. Jude Medical
466. Wyndham Worldwide	467. Kelly Services	468. Western Union
469. Envision Healthcare H.	470. Visteon	471. Arthur J. Gallagher
472. Host Hotels & Resorts	473. Ashland	474. Insight Enterprises
475. Energy Future Holdings	476. Markel	477. Essendant
478. CH2M Hill	479. Western & Southern F.G.	480. Owens Corning
481. S&P Global	482. Raymond James Financial	483. NiSource
484. Airgas	485. ABM Industries	486. Citizens F.G.
487. Booz Allen Hamilton H.	488. Simon Property Group	489. Domtar
490. Rockwell Collins	491. Lam Research	492. Fiserv
493. Spectra Energy	494. Navient	495. Big Lots
496. Telephone & Data Systems	497. First American Financial	498. NVR
499. Cincinnati Financial	500. Burlington Stores	
