

Introduction to Web Science: Assignment #3

Dr. Nelson

Alexander Nwala

Thursday, April 2, 2015

Contents

Problem 1	3
Problem 2	13

Problem 1

For the text you saved for the 10000 URIs from A1, Q2: Use the “boilerpipe” software to remove the HTML templates from all HTML pages (document how many pages link from the tweets were non-HTML and had to be skipped)

<https://code.google.com/p/boilerpipe/>
 WSDM 2010 paper: <http://www.13s.de/~kohlschutter/boilerplate/>

For how many of the 10000 URIs was boilerpipe successful? Compare the total words, unique words, and byte sizes before and after use of boilerpipe. For what classes of pages was it successful? For what classes of pages was it unsuccessful? Provide examples of both successful and unsuccessful removals and discuss at length.

SOLUTION 1

The solution for this problem is outlined by the following steps:

Remove boilerplate from HTML pages: Due to Listing 1, the HTML templates derived after dereferencing the URIs were removed with justtext [1]

Listing 1: Remove Boilerplate

```
#Remove boilerplate snippet
def extractText(listOfURIs):
    ...
    for i in range(0, len(listOfURIs) ):
        ...
        try:
            URI = listOfURIs[i].split(',', )[1].strip()
            print i, URI

            page = urllib2.urlopen(URI).read()
            paragraphs = justtext.justtext(page, justtext.get_stoplist('English'))

            for paragraph in paragraphs:
                if paragraph['class'] == 'good':
                    processedText = paragraph['text']

                    processedText = processedText.encode('ascii', 'ignore')
                    outputFile.write(URI + ':\\n' + processedText + '\\n\\n')

                else:
                    print 'class:', paragraph['class']

                    if( URI in alreadyAddedList ):
                        pass
                    else:
                        statOutputFile.write(URI + ' bad\\n\\n')
                        alreadyAddedList.append(URI)

            print
```

```

30     except:
31         statOutputFile.write(URI + ': ERROR\n\n')
32         exc_type, exc_obj, exc_tb = sys.exc_info()
33         fname = os.path.split(exc_tb.tb_frame.f_code.co_filename)[1]
34         print(fname, exc_tb.tb_lineno, sys.exc_info() )
35         print
36
37     outputFile.close()
38     statOutputFile.close()
39
40 ...

```

Consider the following remarks: Given the list of 10,000 URIs, only 2,164 were unique URIs from A2 - I believe this was caused due to overlap in my search results from Twitter (in retrospect, my algorithm to retrieve URIs should not have used search).

Given my initial list of 2,164 URIs, 1,132 were skipped due to ERRORS ranging from 404s to corrupt URIs as seen by the following examples of HTTP response codes:

```

$ curl -I https://api.twitter.com/oauth/authorize?oauth_token
=8bwrg25dbghw8aszsf0wul3q6gxswey
HTTP/1.1 403 Forbidden

$ curl -I http://click.linksynergy.com/link?id=vdg03e619pw&offerid=
210072.9001235091&type=2&murl=
http://www.animate-onlineshop.jp/products/detail.php%3fproduct_id%3d1235091
HTTP/1.1 400 Bad Request

$ curl -I http://p0p.pw/erh
HTTP/1.1 404 Not Found

$ curl -I http://raumu.ciao.jp/yy/
HTTP/1.1 404 Not Found

$ curl -I http://butta.info/u1f3rd/
curl: (6) Could not resolve host: butta.info

```

There were 60 successful text extraction operations from the list of URIs.

Because of the low success rate, I began investigating a possible cause by randomly sampling 15 URIs. I consider the following results below, a justification for the low success rate.

```

5 URIs: links of Google sites with no content which required a
click to redirect
5 URIs: Japanese sites ranging from blogs to sites with no content
3 URIs: 2 Spanish news sites which required clicks to see content,
1 Facebook site
2 URIs: 1 image and one was unavailable (404)

```

Consider the following statistic collected from 60 successful text extraction operations:

Total words before removal: 549,788
Total words after removal: 31,135

Total unique words before removal: 51,044
Total unique words after removal: 7,819

Total size (bytes) before removal: 31,010,041
Total size (bytes) after removal: 192,437

The classes of pages for which boilerplate removal was successful includes: organized pages with well defined/extensive text areas such as blogs and pages with comment sections. Overall, pages in which the template was on the left and right with text in the middle were successful. Consider the following examples (charts 1 - 3) of successful removals (green boxes) based on the format described.

Chart Example 1 URI:

https://www-304.ibm.com/connections/blogs/socialbusiness/entry/you_are_sitting_on_a_volcano_that_is_ready_to_blow?lang=en_us

Chart Example 2 URI

<http://www.thedailybeast.com/articles/2015/02/03/average-soldiers-don-t-trust-their-generals-and-they-have-a-point.html>

Chart Example 3 URI

http://prod.www.giants.clubs.nfl.com/news-and-blogs/article-1/know-your-giants-te-larry-donnell-/dbaf4b75-cb7c-40f3-b326-22a5600aebbd?utm_source=dlvr.it&utm_medium=twitter

SocialBusiness
Insights Blog

This Blog Search

My Blogs Public Blogs My Updates

Tags

- #20questions
- 20_questions
- 20questions
- analytics askjim
- ben_martin
- business cloud
- collaboration
- community
- connected
- connections data
- digital email
- employees
- engagement event
- femke_goechhart hr
- ibm ibm_connect
- ibm_connections
- ibm_redbooks ibm_connections
- ibmconnect
- ibmredbooks
- ibmverse ics
- innovation meeting
- millennials
- mobile
- newwaytowork
- notes podcast
- questions
- reimagine our wo

Social Business Insights Blog

News and thoughts on becoming a business that is engaged, transparent and nimble. Join us in cultivating a spirit of collaboration and community. Managed by Daniel Davis [@danielkdtwt](#) and Samantha Klein [@samjoyk](#). We're following IBM Social Computing Guidelines.

The short URL for this blog is <http://ibm.com/blogs/socialbusiness>

You Are Sitting on a Volcano That Is Ready to Blow

Warren Whitlock | Jan 21 | Tags: newwaytowork ibm_verse new_way_to_work verse ibmverse warren_whitlock

0 Comments | 8,367 Visits

The changes coming to the workplace are huge and can't be overstated. They remind me of Joe's journey from one of my favorite films of all time, *Joe Versus the Volcano*.

In the movie, Tom Hanks's character Joe Banks works in a comic rendition of a traditional 20th century workplace where the office workers are treated like cogs in a machine, not human beings with ideas and creative input to help reach objectives. In this dysfunctional setting, communication is only the boss telling the workers what to do, and collaboration is "do it or you screw up the whole operation."

Joe is understandably miserable. He used to have a good life, but now he's settled to become a drone in a thankless job. He's convinced life is over and he's ready to accept that.

Joe takes an assignment that will send him to a faraway island where he will jump into a volcano to sacrifice himself for the company. Since Joe thinks he's going to die, he lets loose and begins to experience new things. Along the way, he learns how he's been asleep, is reminded of his true value and realizes that he could be contributing so much more and enjoying the experience.

And then Joe climbs the mountain and jumps into the volcano.

the Atlantic

Do your clients understand the **Power of Delivering Exceptional Customer Experiences?**

IBM

IBM Social Business Elsewhere

- IBM Smarter Planet
- IBM Social Business
- IBMSocialBiz on Facebook
- @IBMSocialBiz on Twitter
- Social software for business
- IBM Collaboration Solutions Community
- Social Business on Tumblr
- IBM Center for Social Research
- @ctr4socialsoft on Twitter

Blog Authors

 Samantha Klein

Chart Example 1: Successful removal

The screenshot shows the official website of the New York Giants. At the top, there's a navigation bar with links for NEWS, TEAM, VIDEOS, PHOTOS, HISTORY, SCHEDULES, FAN ZONE, and TICKETS. Below this, a banner for 'UP NEXT LIVE' features a video player showing a game highlight of Eli Manning throwing a pass to Larry Donnell. The main article, titled 'Know Your Giants: TE Larry Donnell', is highlighted with a green box. It includes a photo of Michael Eisen, the author, and a video thumbnail below it. To the right, there's a sidebar with 'LATEST NEWS' and 'LATEST VIDEOS' sections.

LATEST NEWS

- What to watch for: Giants 2015 Draft storylines
- NFL Network's 5 prospects worth trading up for
- Eisen's Mailbag: Who will be the Giants toughest opponent in 2015?
- Eli Manning is primed for second season under Ben McAdoo's offense
- NFC East Mock Draft Roundup

[MORE NEWS »](#)

LATEST VIDEOS

- Big Blue Kickoff Live (4/1)
- Best Giants Trick Plays
- Video Mailbag: What will Damontre Moore's role be?
- Big Blue Kickoff Live (3/31)
- Cruz's 99-yard TD voted best

Chart Example 2: Successful removal

THE DAILY BEAST POLITICS ENTERTAINMENT WORLD U.S. NEWS TECH + HEALTH BEASTSTYLE BOOKS



Brendan McDermid/Reuters

BETRAYAL 02.03.15

 **Rep. Duncan Hunter**

Average Soldiers Don't Trust Their Generals and They Have a Point

A survey last year showed only 27% of the military felt senior leaders looked out for their best interests. To fix the morale crisis generals need to stop acting like politicians.

Through a decade-plus of war, America's military men and women, and the families that support them, have experienced their share of hardships. Separations through multiple deployments and the inherent dangers of combat are enough to press the emotional and physical limits of even the strongest individuals.

For some of these faithful defenders of America's interests, there have been difficulties far beyond the battlefield—difficulties not imposed by any enemy or the distance and time that separates them from their loved ones. Most ironically, the assault against them—intended or not—can sometimes come from within the military institution for which they fought, bled and sacrificed so much.

It's no wonder why there's concern for morale in today's force. Just recently, outgoing Secretary of Defense Chuck Hagel said he too is worried about the decline in enthusiasm, and he believes it will take some time to reversing the mindset and perspective of the force.

READ THIS. *list*

- 1 Corruption In The Military's Top Ranks
- 2 We Found Iran's Secretive Drone Base
- 3 Elvis Is The Declaration Of Independence
- 4 Gay Vet Shamed Indiana From The Grave
- 5 Money Can Buy You A Bigger Brain
by Charlotte Lytton

Chart Example 3: Successful removal

The classes of pages for which boilerplate removal was unsuccessful includes: Pages with disorganized format (no well defined template) in which text was spread accross template as well as pages with unconventional layouts:

Chart Example 4 URI:

<http://www.stocktradepartner.com/>

Chart Example 5 URI:

http://www.fandango.com/dawnoftheplanetoftheapes_156265/movieoverview?wssaffid=11838&wssac=123&cjid=cj_10576763_7644471_

Chart Example 6 URI:

<http://www.yesasia.com/us/you-who-came-from-the-stars-blu-ray-vol-2-japan-version/1037741202-0-0-0-en/info.html>

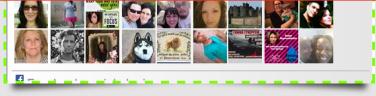
Hot Trading News
ALWAYS up to date

New Trader Update and Market Status Notification System
Mon, Feb 2, 2015, 3:06:45 PM
As previously stated, as of today, Trader Update and Market Status notices for the NYSE, NYSE MKT, NYSE Arca Equities, NYSE ...

Reminder: New Trader Update and Market Status Notification System
Fri, Jan 30, 2015, 3:07:22 PM
This is a reminder that effective Monday, February 2, 2015, Trader Update and Market Status notices for the NYSE, NYSE MKT, ...

NYSE Cash Equities Trading Collars - Symbol Rollout Beginning Monday, February 2, 2015
Fri, Jan 30, 2015, 1:57:24 PM
As previously stated, the NYSE will be introducing the new Trading Collars functionality starting Monday, February 2, ...

Announcing Pillar: Our New Tra NYSE GROUP inc.7



Hot News From Bloomberg
ALWAYS up to date

- Bloomberg Intelligence: Waiting for the Job's Report (Audio)
Thu, Apr 2, 2015, 7:15:40 AM
- Bloomberg - The First Word: Riccadonna on Jobs
Thu, Apr 2, 2015, 7:10:38 AM
- Bloomberg - The First Word: Samra on Markets
Thu, Apr 2, 2015, 6:07:56 AM
- Bloomberg Law Brief: Religious-Rights Legislation (Audio)
Thu, Apr 2, 2015, 6:02:38 AM

MARKE WATCH NEWS
ALWAYS UP TO DATE AND RELEVANT TO THE TRADER

- Jobless claims fall 20,000 to 268,000 and near post-recession low
Thu, Apr 2, 2015, 8:31:29 AM
Jobless claims fall 20,000 to 268,000 and near post-recession low
- U.S. stocks end lower after weak jobs, factory data
Wed, Apr 1, 2015, 4:08:28 PM
U.S. stocks end lower after weak jobs, factory data
- New Jersey Sen. Robert Menendez indicted on corruption charges
Wed, Apr 1, 2015, 3:58:50 PM
New Jersey Sen. Robert Menendez indicted on corruption charges
- McDonald's to raise pay for U.S. restaurant workers by 10%: WSJ
Wed, Apr 1, 2015, 3:29:17 PM

Stock Trade Partner

JOIN XM TODAY BY CLICKING BELOW

- XM Trading
- DAILY SIGNALS
- NO-REQOUTES
- 50% DEPOSIT BONUS
- LOYALTY REWARDS
- XM POINTS
- STATUS UPGRADES
- FREE \$30 TO START
- MQL5 COMMUNITY
- SIGNAL SUBSCRIPTION

NASDAQ STOCK NEWS FEED HERE
YOU WILL FIND MORE RELEVANT INFO RELATED TO STOCKS

ESPN for \$36? Analyst Shows True Cost of A La Carte Cable
Wed, Apr 1, 2015, 9:48:28 AM
Americans think they want Comcast, Time Warner Cable, DIRECTV, DISH Network and the rest of the pay television providers to ...

Chart Example 4: Unsuccessful removal

Shop » All Departments » Japanese » Korean » Chinese » World » Other
[International Shipping](#)

Korea TV Series & Dramas | New & Future Releases | Preorder | Bestsellers | [Search](#)



You Who Came From the Stars (Blu-ray) (Vol. 2) (Japan Version) Blu-ray Region A
 Park Hae Jin | Kim Soo Hyun | Jeon Ji Hyun

Our Price: **US\$207.99** [~£143.51]
 | Save for later

Availability: Usually ships within 7 to 14 days

Like Be the first of your friends to

Related promotions:
Get a FREE Japan Mini capsule toy!
This item is eligible for Free International Shipping

Important information about purchasing this product:
This product cannot be cancelled or returned after the order has been placed. For more details, please refer to our [return policy](#).
Blu-ray Discs are exclusively compatible with Blu-ray Disc players, and cannot be played on conventional DVD players or HD DVD players.
This product will not be shipped to Hong Kong.

[Sign in](#) to rate and write review
[Write a Review](#)

Bookmark & Share


<http://www.yesasia.com>

YESASIA Editorial Description
 The Korean drama sensation that swept Asia! Returning to the small screen for her first TV drama in over a decade, Korea's original sassy girl Jeon Ji Hyun reteams with her *The Thieves* co-star Kim Soo Hyun in the SBS fantasy romance *You Who Came From the Stars* (a.k.a. *My Love From the Star*). Filled with heart, humor and suspense, the romantic drama revolves around the unlikely love between a bumbling actress and an emotionally detached alien who landed in Korea during Joseon times and stuck around for over 400 hundred years till the present day. Co-starring Park Hae Jin, Yoo In Na and Shin Sung Rok, *You Who Came From The Stars* not only lit up the ratings in Korea, it became a huge trend-setting hit throughout Asia.



COLORS miss A



2014 HONG KONG MOVIES SALE



EXODUS



Daikaiju Super Robot Taisen



A ONE Ayumi Hamasaki

Chart Example 5: Unsuccessful removal

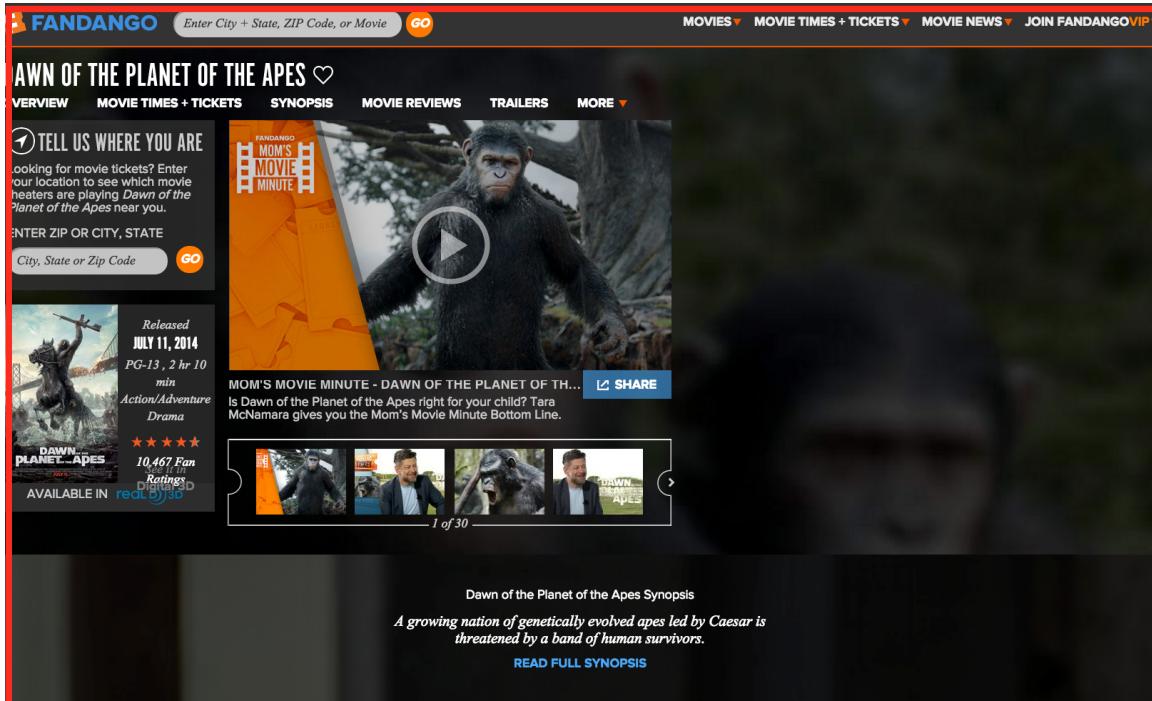


Chart Example 6: Unsuccessful removal

Problem 2

Collection1: Extract all the unique terms and their frequency from the 10000 files*

Collection2: Extract all the unique terms and their frequency of the 10000 files* after running boilerpipe

Construct a table with the top 50 terms from each collection. Find a common stop word list. How many of the 50 terms are on that stop word list?

For both collections, construct a graph with the x-axis as word rank, and y-axis as word frequency.

Do either follow a Zipf distribution? Support your answer.

SOLUTION 2

The solution for this problem is outlined by the following steps:

Extract unique terms: Due to Listing 2, the unique terms were retrieved by splitting the string of all the HTML files on the space character. Subsequently Listing 2. kept track of the term and term frequencies in a dictionary.

Listing 2: Get Unique Terms

```
# Get unique terms from HTML files and plaintext (after boilerplate removal)
def getTermsFromPlaintext():

    page = ''
    wordFrequencyDict = {}
    try:
        inputFile = open('plaintext.txt', 'r')
        page = inputFile.read()
        inputFile.close()
    except:
        exc_type, exc_obj, exc_tb = sys.exc_info()
        fname = os.path.split(exc_tb.tb_frame.f_code.co_filename)[1]
        print(fname, exc_tb.tb_lineno, sys.exc_info())

    15   getWordFrequencyDict(page, wordFrequencyDict)
    wordFrequencyDict = sorted(wordFrequencyDict.items(), key=lambda x:x[1], reverse=True)

    print
    print 'begin'

    20   #total words
    totalWords = 0
    for tup in wordFrequencyDict:
        totalWords = tup[1] + totalWords

    25   print 'totalWords:', totalWords
    print 'totalWordsUnique:', len(wordFrequencyDict)
    '''

# top 50 terms - start
```

```
30     count = 0
31     for tup in wordFrequencyDict:
32         count = count + 1
33
34         print tup[0].strip(), tup[1]
35
36         if count == 50:
37             break
38     # top 50 terms - end
39     '''
40
41 def getUniqueTermsFromHTML():
42
43     lines = []
44     try:
45         inputFile = open('justURLs.txt', 'r')
46         lines = inputFile.readlines()
47         inputFile.close()
48     except:
49         exc_type, exc_obj, exc_tb = sys.exc_info()
50         fname = os.path.split(exc_tb.tb_frame.f_code.co_filename)[1]
51         print(fname, exc_tb.tb_lineno, sys.exc_info())
52
53     wordFrequencyDict = {}
54     instanceCount = 0
55     for l in lines:
56         print instanceCount, len(wordFrequencyDict)
57         l = l.strip()
58         page = ''
59         try:
60             page = urllib2.urlopen(l).read()
61         except:
62             #print 'Error: ', l
63             exc_type, exc_obj, exc_tb = sys.exc_info()
64             fname = os.path.split(exc_tb.tb_frame.f_code.co_filename)[1]
65             print(fname, exc_tb.tb_lineno, sys.exc_info())
66
67             getWordFrequencyDict(page, wordFrequencyDict)
68             #print len(wordFrequencyDict)
69             instanceCount = instanceCount + 1
70
71             wordFrequencyDict = sorted(wordFrequencyDict.items(), key=lambda x:x[1], reverse=True)
72             print
73             print 'begin'
74
75             #total words
76             totalWords = 0
77             for tup in wordFrequencyDict:
78                 totalWords = tup[1] + totalWords
79
80             print 'totalWords:', totalWords
81             '''


```

```

# top 50 terms - start
count = 0
#print wordFrequencyDict
85 for tup in wordFrequencyDict:
    count = count + 1

    print tup[0].strip(), tup[1]

90 if count == 50:
    break
# top 50 terms - end
'''

```

Construct table: Consider Table 1 and 2.

Find terms which coincide in Stopwords list: **43** terms from Table 2. (Plaintext files) coincided with a stopwords list derived from <http://www.ranks.nl/stopwords>. Also **16** terms from Table 1. (HTML files) coincided with the same list.

The file `stopwords.txt` contains the stopwords.

Plot chart: Due to Listing 3, the terms and term frequencies pre/post boilerplate removal was plotted. As seen in Chart Example 7, both charts follow the Zipfian distribution [2]: The frequency of each word is approximately inversely proportional to its rank in the frequency table - “power law.” This means the most frequent word occurs approximately twice as often as the second most frequent word and so on.

Listing 3: Plot Term Frequencies

```

#!/usr/bin/env Rscript

TermTermFrequencyP <- read.table('plaintextTop50Terms.txt', header=T)
5 TermTermFrequencyH <- read.table('rawHTMLTop50Terms.txt', header=T)

plot(TermTermFrequencyP$Frequency, type='o', col='blue', xlab='Word Rank', ylab='Word
Frequency', main='Distribution of terms in HTML files pre boilerplate \nremoval (
red) and post boilerplate removal (blue)')
10 #plot(TermTermFrequencyH$Frequency, type='o', col='blue', xlab='Word Rank', ylab='Word
Frequency', main='Distribution of terms in HTML files')
lines(TermTermFrequencyH$Frequency, type='o', col='red')

```

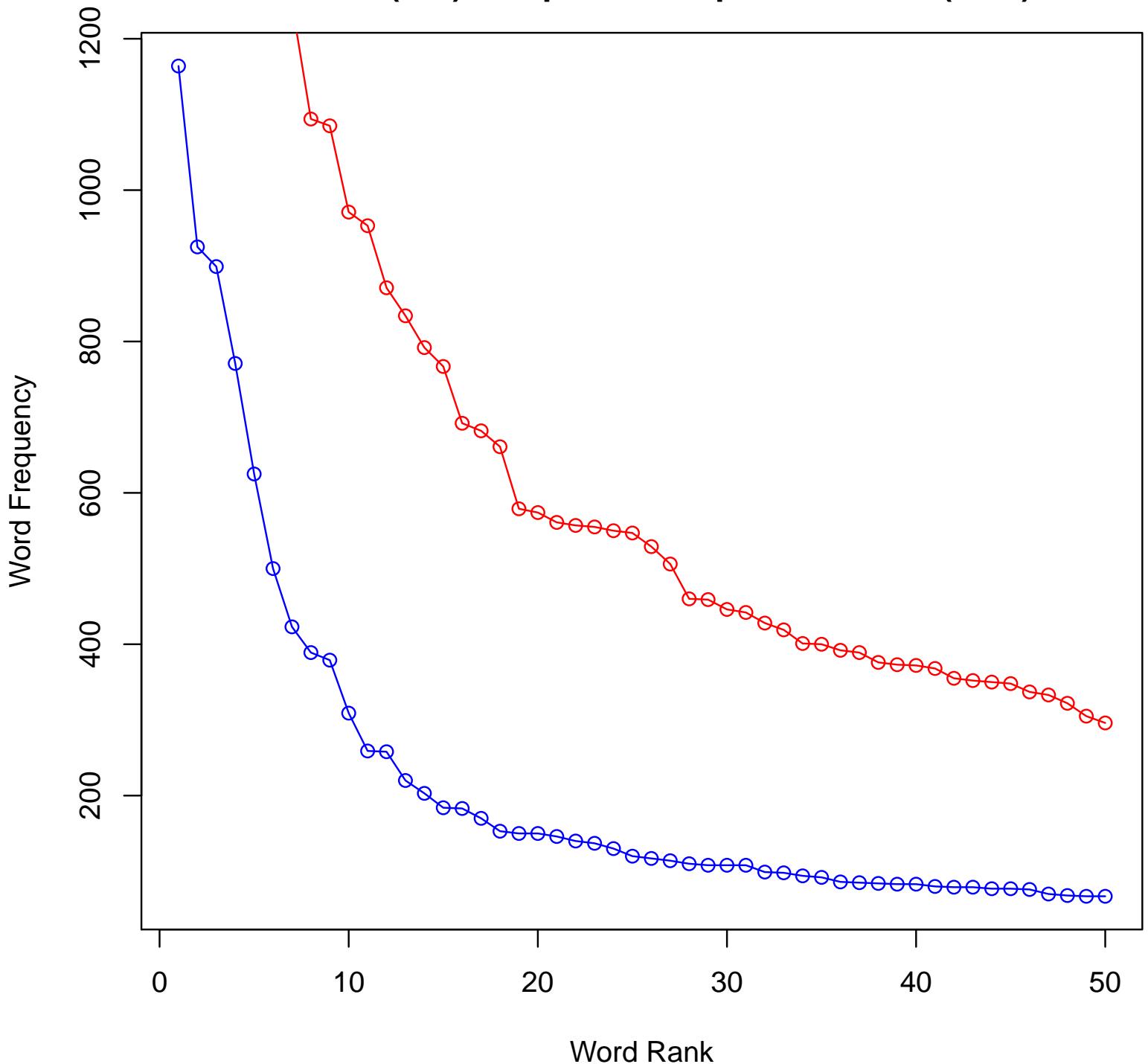
Table 1: Top 50 Terms Extracted From HTML Files

Item	Term	Frequency
1	<div	2924
2	=	2033
3	the	1944
4	<a	1819
5	to	1556
6	and	1503
7	a	1247
8	of	1094
9	{	1085
10	<a	971
11	<span	953
12	</div>	871
13	in	834
14	{	792
15	for	767
16	-	692
17	<li	682
18	at	661
19	var	579
20	point:false,	574
21	on	561
22	+	557
23	</div>	555
24	is	550
25	target=_blank	547
26	B	529
27	position:left"},	506
28	you	460
29	your	459
30		446
31	class=b-link	442
32	with	428
33	I	419
34	blog-admin	401
35	target=_self	400
36	&	392
37	2014	389
38		376
39	December	373
40	0	372
41	rel=nofollow	368
42	<img	355
43	onclick=return	352
44	if	350
45	nofollow:false,tab_target:false,	348
46	class=b-ico	337
47	type=hidden	333
48	The	322
49	}	305
50	class=b-link	296

Table 2: Top 50 Terms Extracted Post Boilerplate Removal

Item	Term	Frequency
1	the	1164
2	and	925
3	to	899
4	a	771
5	of	625
6	in	500
7	is	423
8	you	389
9	I	379
10	for	309
11	with	259
12	that	258
13	your	220
14	on	203
15	was	184
16	are	183
17	or	170
18	it	153
19	at	150
20	be	150
21	as	146
22	this	140
23	by	137
24	will	130
25	have	120
26	can	117
27	from	114
28	her	110
29	not	108
30	my	108
31	all	108
32	an	99
33	when	98
34	The	94
35	get	92
36	we	86
37	about	85
38	domain	84
39	so	83
40	our	83
41	like	80
42	he	79
43	but	79
44	up	77
45	his	77
46	just	76
47	one	70
48	who	68
49	has	67
50	if	67

Chart Example 7: Distribution of terms in HTML files pre boilerplate removal (red) and post boilerplate removal (blue)



References

- [1] jusText. <https://pypi.python.org/pypi/jusText/2.0.0>. Accessed: 2015-04-01.
- [2] The mystery of Zipf. <https://plus.maths.org/content/mystery-zipf>. Accessed: 2015-04-01.