California State Polytechnic University, Pomona

Project 1:

Web Crawling & Zipf’s Law

findone
us
coronavirus
years
county
march
day
going
time
people
school
state
trump
carolina
chat
told
even
campaign
know
voters
last
home
city
democratic
now
first
whats
said
around
university
search
happening
tuesday
according
may
president
tennessee
another
new
world
et
way
cnn
say
will
pm
health
three
two
messenger
facebook
white

Aaron Cervantes, Brent Tsuji, Theresa Van, and Federico Wang

CS 4990-02: Special Topics for Upper Division Students

Ben Steichen

04 March 2020

Overview:

This project consists of web crawls in three different languages. The language's chosen were all latin based; English, Spanish and Italian. The Zipf's Law analysis is done from the data scraped the web crawls. The source code is written in Python using the library Scrapy.

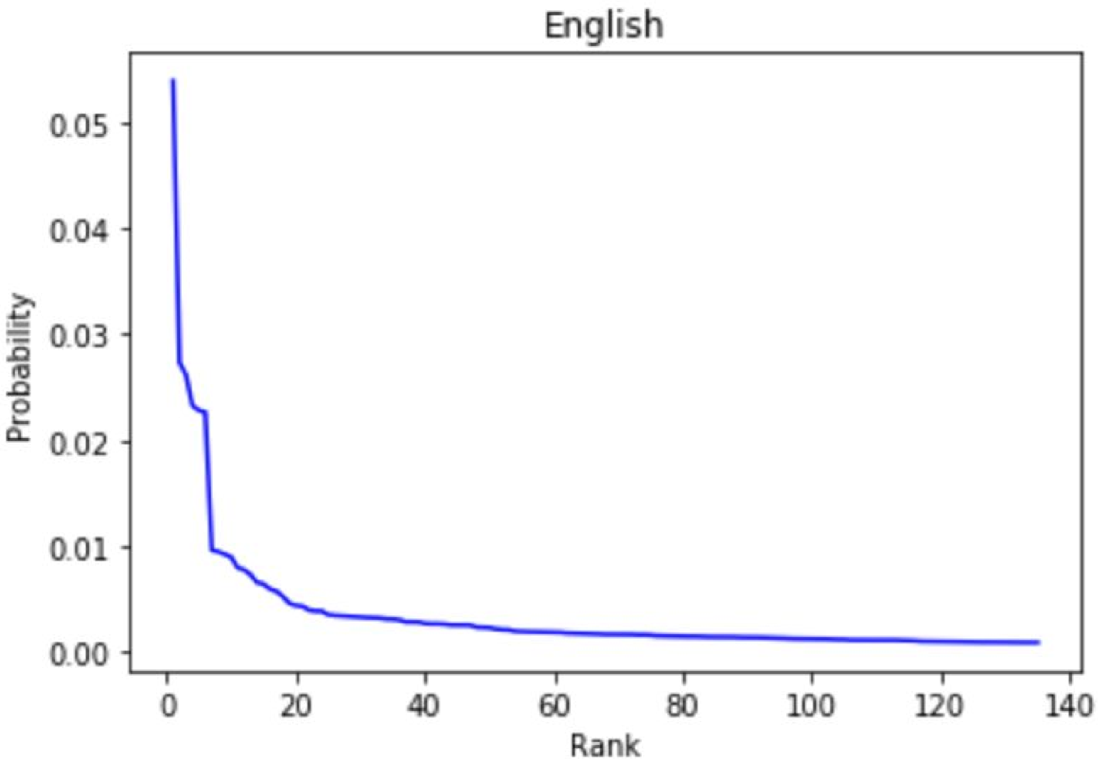
The Main Components:

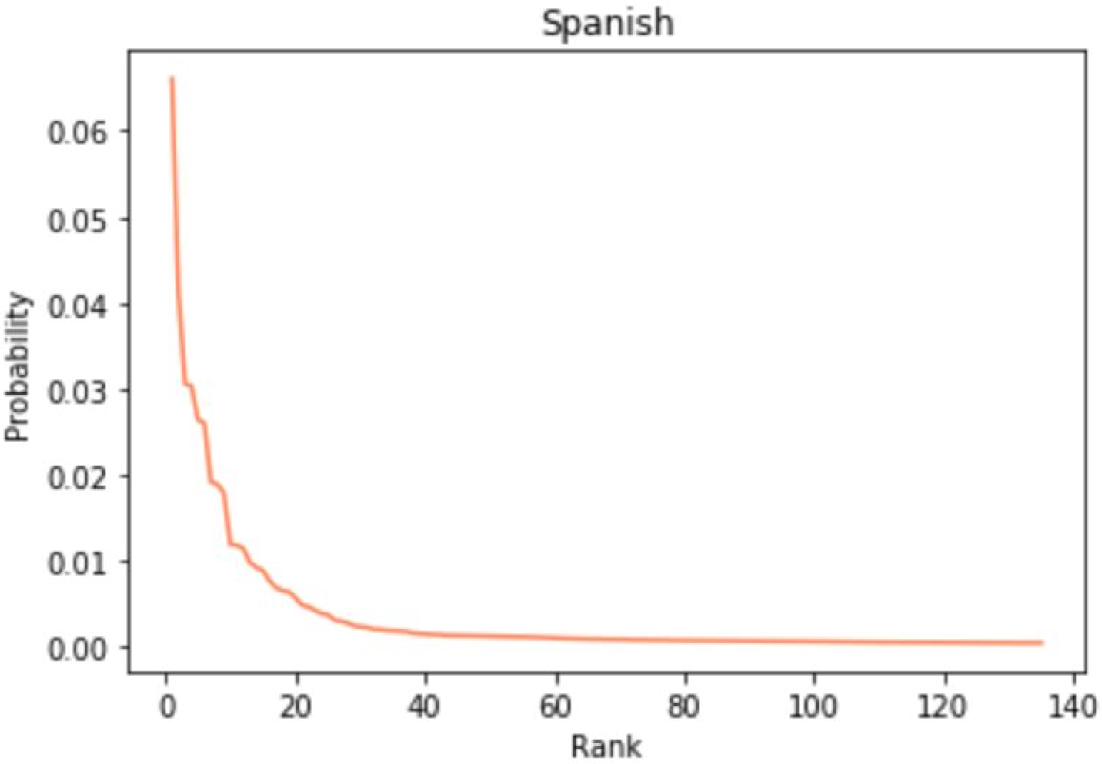
The program was written in Python utilizing the web scraping API: Scrapy. We used xpath extractions to extract text, outlinks and urls from each seed. We began with a starting url of CNN.com, a popular news website, we believed would be a dynamic conduit into a variety of outlinks. We parsed the text file which contained all html tags into our .csv file to have Rank, Probability, Word, and Count from our ZipfsLaw.py file. The ZipfsLaw.py file also conducted our analysis, to confirm that our vocabulary and frequency was accepted by the law. From this we analyzed further our probabilities with Zipf's law and made corresponding graphs to evaluate if the text scraped from the crawls followed consistently the curves discussed in class.

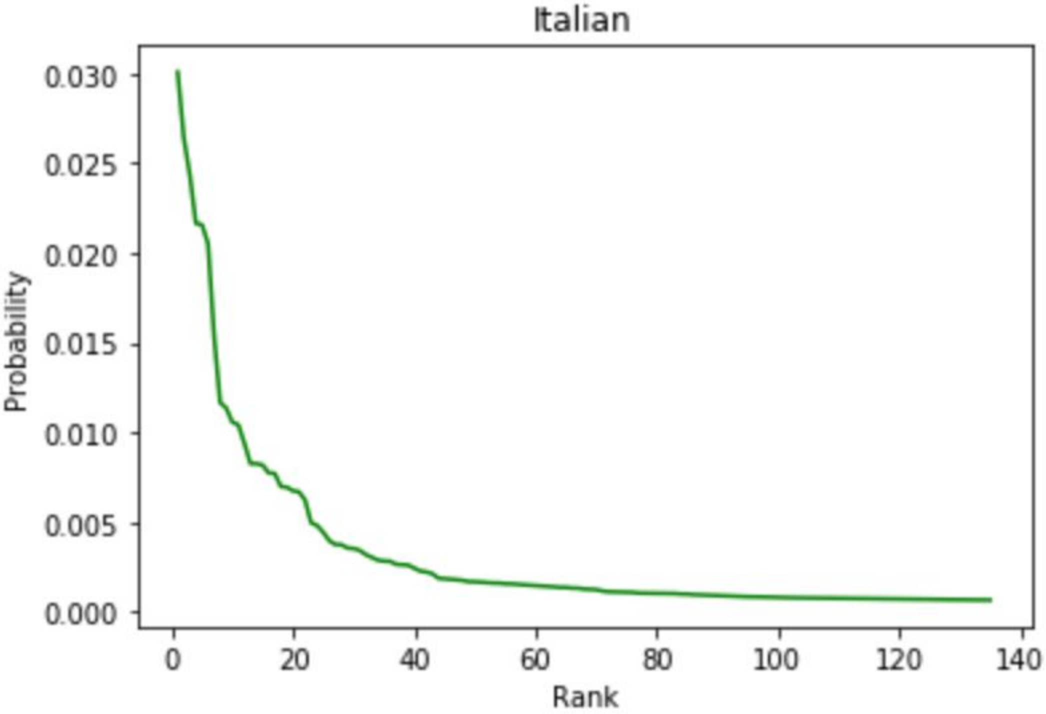
Challenges Faced:

Getting Scrapy to scrape a single page was straightforward, but understanding how to use Scrapy's infrastructure to its full potential was more in-depth. Because every search is done recursively in scrapy it was difficult to find a way to crawl the outlinks of a website without accidentally getting the same one over and over again but eventually was solved. Getting a proper xpath to filter out the texts and links was also a bit difficult because every website stores their data differently. Finding a third party parser was a challenge the team faced, but eventually we decided to create our own parser to suit our cleaning needs for removing the html tags and mess. This is shown in our source code title: Zipfs.py. From this we created a .csv file with the Rank, Probability, Count and Word. Lastly, the team faced issues with accuracy through false negatives on language detection. We had no way to improve on this, our crawler was able to detect the languages confidently.

Probability/Rank Plots: WordCloud: (most frequent words with stop words)




Spanish
Probability
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
20
40
60
80
100
120
140
Rank


Italian
Probability
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
140
Rank

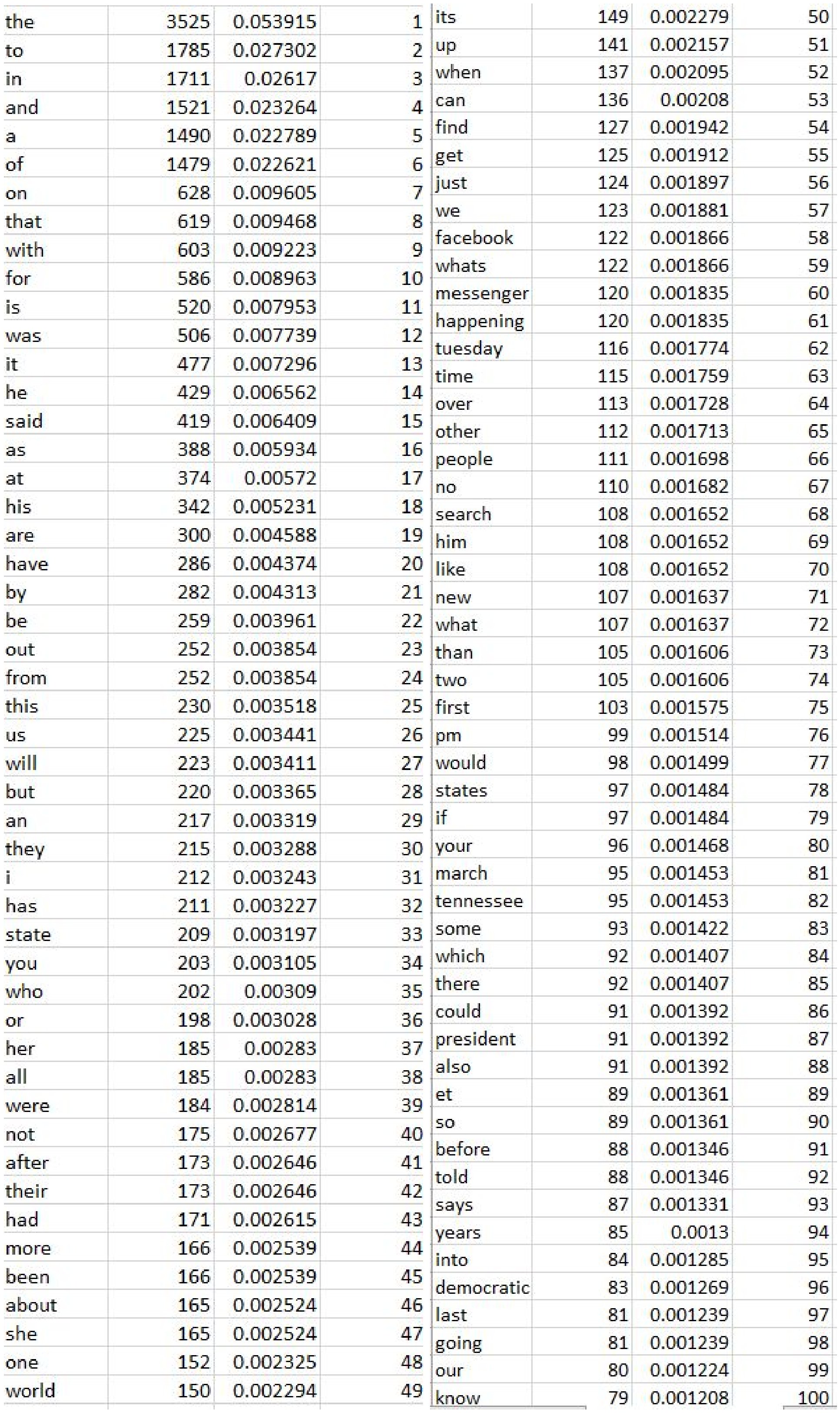
Zipf's law distribution analysis:

Zipf's law is a statistical distribution where words are ranked higher the more often a word pops up. According to Zipf's law, the frequency of words is approximately $\mathit{frequency}(\mathit{rank}) \cong 0.1/\mathit{rank}$. Therefore, the graph of the data should start from the top and continue to decrease. In a normal scenario where they follow the Zipf's law, they should ideally be similar to the graphs shown above due to the equation $0.1/\mathit{rank}$ being similar to $1/x$.

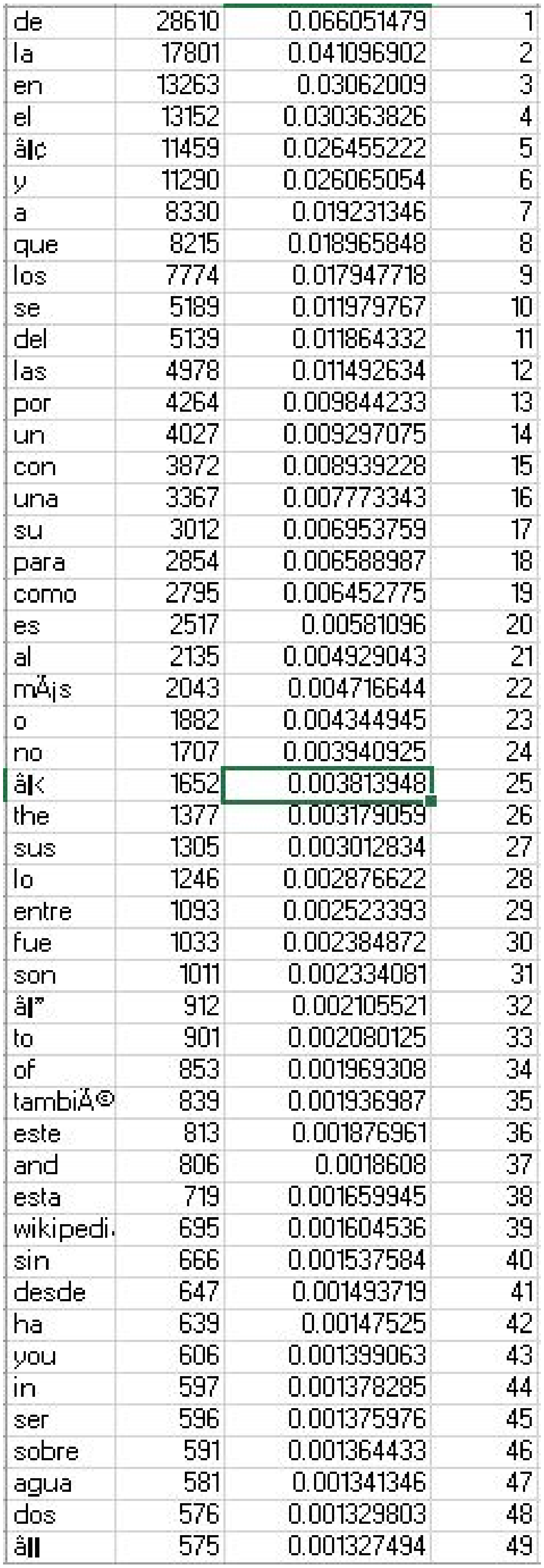
Contribution to project:

Aaron Cervantes and Brent Tsuji collaborated on the web crawler while Theresa Van and Federico Wang worked on Zipf's Law analysis.

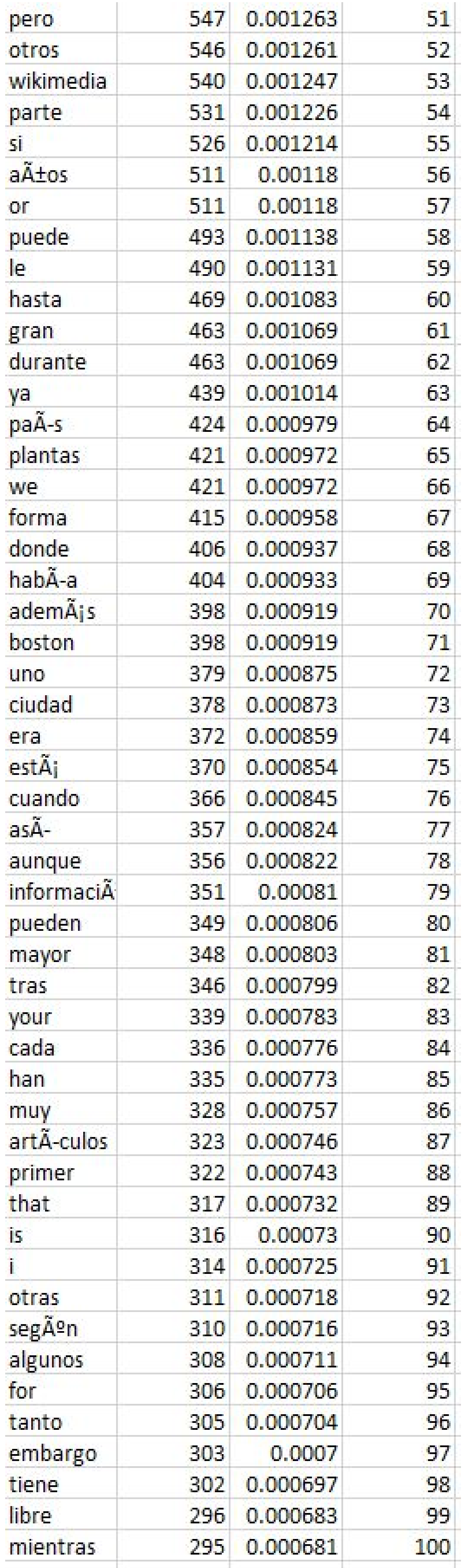
Appendix(seeds, domains): 100 most frequent words for each crawl:



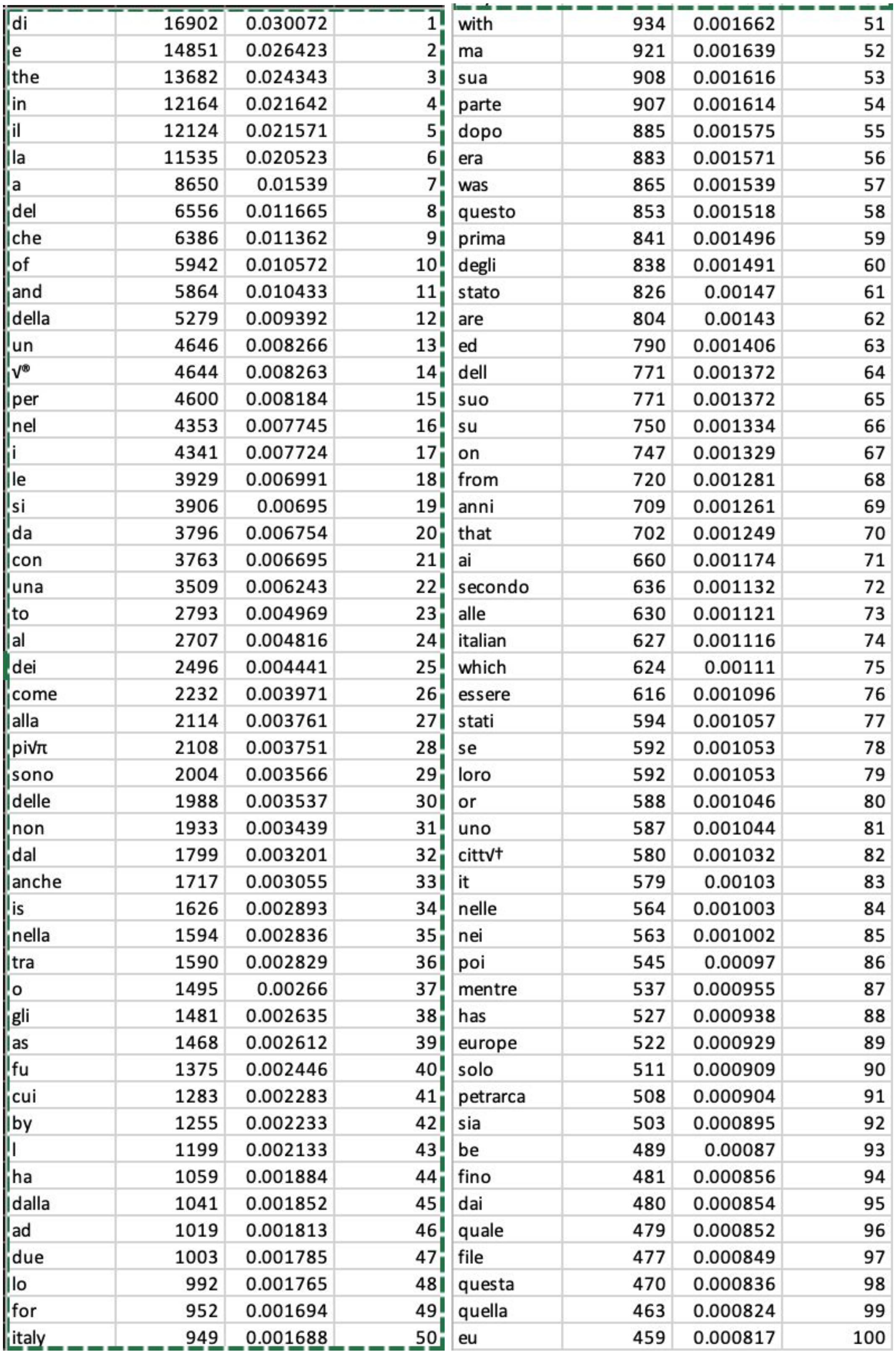
the	3525	0.053915	1
to	1785	0.027302	2
in	1711	0.02617	3
and	1521	0.023264	4
a	1490	0.022789	5
of	1479	0.022621	6
on	628	0.009605	7
that	619	0.009468	8
with	603	0.009223	9
for	586	0.008963	10
is	520	0.007953	11
was	506	0.007739	12
it	477	0.007296	13
he	429	0.006562	14
said	419	0.006409	15
as	388	0.005934	16
at	374	0.00572	17
his	342	0.005231	18
are	300	0.004588	19
have	286	0.004374	20
by	282	0.004313	21
be	259	0.003961	22
out	252	0.003854	23
from	252	0.003854	24
this	230	0.003518	25
us	225	0.003441	26
will	223	0.003411	27
but	220	0.003365	28
an	217	0.003319	29
they	215	0.003288	30
i	212	0.003243	31
has	211	0.003227	32
state	209	0.003197	33
you	203	0.003105	34
who	202	0.00309	35
or	198	0.003028	36
her	185	0.00283	37
all	185	0.00283	38
were	184	0.002814	39
not	175	0.002677	40
after	173	0.002646	41
their	173	0.002646	42
had	171	0.002615	43
more	166	0.002539	44
been	166	0.002539	45
about	165	0.002524	46
she	165	0.002524	47
one	152	0.002325	48
world	150	0.002294	49
its	149	0.002279	50
up	141	0.002157	51
when	137	0.002095	52
can	136	0.00208	53
find	127	0.001942	54
get	125	0.001912	55
just	124	0.001897	56
we	123	0.001881	57
facebook	122	0.001866	58
whats	122	0.001866	59
messenger	120	0.001835	60
happening	120	0.001835	61
tuesday	116	0.001774	62
time	115	0.001759	63
over	113	0.001728	64
other	112	0.001713	65
people	111	0.001698	66
no	110	0.001682	67
search	108	0.001652	68
him	108	0.001652	69
like	108	0.001652	70
new	107	0.001637	71
what	107	0.001637	72
than	105	0.001606	73
two	105	0.001606	74
first	103	0.001575	75
pm	99	0.001514	76
would	98	0.001499	77
states	97	0.001484	78
if	97	0.001484	79
your	96	0.001468	80
march	95	0.001453	81
tennessee	95	0.001453	82
some	93	0.001422	83
which	92	0.001407	84
there	92	0.001407	85
could	91	0.001392	86
president	91	0.001392	87
also	91	0.001392	88
et	89	0.001361	89
so	89	0.001361	90
before	88	0.001346	91
told	88	0.001346	92
says	87	0.001331	93
years	85	0.0013	94
into	84	0.001285	95
democratic	83	0.001269	96
last	81	0.001239	97
going	81	0.001239	98
our	80	0.001224	99
know	79	0.001208	100



de	28610	0.066051479	1
la	17801	0.041096902	2
en	13263	0.03062009	3
el	13152	0.030363826	4
â▌ç	11459	0.026455222	5
y	11290	0.026065054	6
a	8330	0.019231346	7
que	8215	0.018965848	8
los	7774	0.017947718	9
se	5189	0.011979767	10
del	5139	0.011864332	11
las	4978	0.011492634	12
por	4264	0.009844233	13
un	4027	0.009297075	14
con	3872	0.008939228	15
una	3367	0.007773343	16
su	3012	0.006953759	17
para	2854	0.006588987	18
como	2795	0.006452775	19
es	2517	0.00581096	20
al	2135	0.004929043	21
mÃ¡s	2043	0.004716644	22
o	1882	0.004344945	23
no	1707	0.003940925	24
â▌<	1652	0.003813948	25
the	1377	0.003179059	26
sus	1305	0.003012834	27
lo	1246	0.002876622	28
entre	1093	0.002523393	29
fue	1033	0.002384872	30
son	1011	0.002334081	31
â▌”	912	0.002105521	32
to	901	0.002080125	33
of	853	0.001969308	34
tambiÃ©	839	0.001936987	35
este	813	0.001876961	36
and	806	0.0018608	37
esta	719	0.001659945	38
wikipedi	695	0.001604536	39
sin	666	0.001537584	40
desde	647	0.001493719	41
ha	639	0.00147525	42
you	606	0.001399063	43
in	597	0.001378285	44
ser	596	0.001375976	45
sobre	591	0.001364433	46
agua	581	0.001341346	47
dos	576	0.001329803	48
â▌▌	575	0.001327494	49



pero	547	0.001263	51
otros	546	0.001261	52
wikimedia	540	0.001247	53
parte	531	0.001226	54
si	526	0.001214	55
aÃ±os	511	0.00118	56
or	511	0.00118	57
puede	493	0.001138	58
le	490	0.001131	59
hasta	469	0.001083	60
gran	463	0.001069	61
durante	463	0.001069	62
ya	439	0.001014	63
paÃ-s	424	0.000979	64
plantas	421	0.000972	65
we	421	0.000972	66
forma	415	0.000958	67
donde	406	0.000937	68
habÃ-a	404	0.000933	69
ademÃ¡s	398	0.000919	70
boston	398	0.000919	71
uno	379	0.000875	72
ciudad	378	0.000873	73
era	372	0.000859	74
estÃ¡	370	0.000854	75
cuando	366	0.000845	76
asÃ-	357	0.000824	77
aunque	356	0.000822	78
informaciÃ	351	0.00081	79
pueden	349	0.000806	80
mayor	348	0.000803	81
tras	346	0.000799	82
your	339	0.000783	83
cada	336	0.000776	84
han	335	0.000773	85
muy	328	0.000757	86
artÃ-culos	323	0.000746	87
primer	322	0.000743	88
that	317	0.000732	89
is	316	0.00073	90
i	314	0.000725	91
otras	311	0.000718	92
segÃºn	310	0.000716	93
algunos	308	0.000711	94
for	306	0.000706	95
tanto	305	0.000704	96
embargo	303	0.0007	97
tiene	302	0.000697	98
libre	296	0.000683	99
mientras	295	0.000681	100



di	16902	0.030072	1
e	14851	0.026423	2
the	13682	0.024343	3
in	12164	0.021642	4
il	12124	0.021571	5
la	11535	0.020523	6
a	8650	0.01539	7
del	6556	0.011665	8
che	6386	0.011362	9
of	5942	0.010572	10
and	5864	0.010433	11
della	5279	0.009392	12
un	4646	0.008266	13
√®	4644	0.008263	14
per	4600	0.008184	15
nel	4353	0.007745	16
i	4341	0.007724	17
le	3929	0.006991	18
si	3906	0.00695	19
da	3796	0.006754	20
con	3763	0.006695	21
una	3509	0.006243	22
to	2793	0.004969	23
al	2707	0.004816	24
dei	2496	0.004441	25
come	2232	0.003971	26
alla	2114	0.003761	27
pi√π	2108	0.003751	28
sono	2004	0.003566	29
delle	1988	0.003537	30
non	1933	0.003439	31
dal	1799	0.003201	32
anche	1717	0.003055	33
is	1626	0.002893	34
nella	1594	0.002836	35
tra	1590	0.002829	36
o	1495	0.00266	37
gli	1481	0.002635	38
as	1468	0.002612	39
fu	1375	0.002446	40
cui	1283	0.002283	41
by	1255	0.002233	42
l	1199	0.002133	43
ha	1059	0.001884	44
dalla	1041	0.001852	45
ad	1019	0.001813	46
due	1003	0.001785	47
lo	992	0.001765	48
for	952	0.001694	49
italy	949	0.001688	50
with	934	0.001662	51
ma	921	0.001639	52
sua	908	0.001616	53
parte	907	0.001614	54
dopo	885	0.001575	55
era	883	0.001571	56
was	865	0.001539	57
questo	853	0.001518	58
prima	841	0.001496	59
degli	838	0.001491	60
stato	826	0.00147	61
are	804	0.00143	62
ed	790	0.001406	63
dell	771	0.001372	64
suo	771	0.001372	65
su	750	0.001334	66
on	747	0.001329	67
from	720	0.001281	68
anni	709	0.001261	69
that	702	0.001249	70
ai	660	0.001174	71
secondo	636	0.001132	72
alle	630	0.001121	73
italian	627	0.001116	74
which	624	0.00111	75
essere	616	0.001096	76
stati	594	0.001057	77
se	592	0.001053	78
loro	592	0.001053	79
or	588	0.001046	80
uno	587	0.001044	81
citt√†	580	0.001032	82
it	579	0.00103	83
nelle	564	0.001003	84
nei	563	0.001002	85
poi	545	0.00097	86
mentre	537	0.000955	87
has	527	0.000938	88
europe	522	0.000929	89
solo	511	0.000909	90
petrarca	508	0.000904	91
sia	503	0.000895	92
be	489	0.00087	93
fino	481	0.000856	94
dai	480	0.000854	95
quale	479	0.000852	96
file	477	0.000849	97
questa	470	0.000836	98
quella	463	0.000824	99
eu	459	0.000817	100