Bigram Analysis

> Source of articles: Times of India Archives
>
> Start Date of articles: 1st January, 2016
>
> End Date of articles: 31st January, 2016

Objectives

- Get the articles and analyse the frequency of words used.
- Later we extended it to bigrams.

Procedure followed

- Downloaded articles from Times of India archives
- Tokenized the articles
- Stored all the words and pair of consecutive words for each category
- Calculated Pointwise Mutual Information (PMI) for all pairs of consecutive words

Technologies used

- NodeJS: for downloading articles and calculating PMI
- Python: for tokenizing and storing the words and the pair of consecutive words
- MongoDB: database used to store data

Formula for PMI

```
Porbability of finding the word W:
P(W) = count(W)/(sum of all frequencies of words)

Porbability of finding the bigram (Wi,Wi-1):
P(Wi,Wi-1) =
count(Wi,Wi-1)/(sum of all frequencies of consecutive words)

PMI:
PMI(Wi,Wi-1) = log( P(Wi,Wi-1)/( P(Wi)P(Wi-1) ) )
```

Challenges faced

- Redirection of links for the articles resulting in empty response.
  - Used `follow-redirects` module do fix redirection problem.
- Speed of downloading articles and parsing the words and bigrams for calculations.
  - For downloading articles, we ran 4 servers at a time with 2 threads on each server.
  - For parsing words and bigrams, we ran 4 servers at a time with one category on each server.
- Getting proper bigrams with PMI.
  - We set a threshold frequency for bigrams for each category.

Observations

- City
  - Most frequently used words and bigrams were related to
    - Politics
    - Crime
    - Money

- India
    - Most frequently used words and bigrams were related to
        - Politics
        - Crime
        - Terrorist attacks
- Life
    - Most frequently used words and bigrams were related to
        - Health
        - Diseases
        - Diet
- World
    - Most frequently occuring country/city names
        - China
        - United States
        - North Korea
        - Saudi Arabia
        - New York
    - Most frequently occuring words and bigrams were related to
        - Politics
        - Terrorism
- Business
    - Most frequently used words and bigrams were related to
        - Money
        - Stock Market
        - Banking
        - Petroleum
        - Development

Analysis details

Category	Number of articles
City	11,137
India	1,157
Life	932
World	563
Business	464

City

Bigrams

> Filters for the table
>
> Min PMI: 9
>
> Min frequency: 50

S. No.	Bigram	PMI
1	modus operandi	11.7213
2	prima facie	11.4713
3	saudi arabia	11.3766
4	wi fi	11.3463
5	smriti irani	11.3285
6	aam aadmi	11.1529
7	bullock cart	11.0430

S. No.	Bigram	PMI
8	bone marrow	10.9561
9	jawaharlal nehru	10.8396
10	bharatiya janata	10.7816
11	swine flu	10.6364
12	sri lanka	10.5937
13	oommen chandy	10.4838
14	makar sankranti	10.4699
15	chandrababu naidu	10.4163
16	pimpri chinchwad	10.4076
17	jd u	10.3519
18	mamata banerjee	10.3157
19	mehbooba mufti	10.2770
20	naveen patnaik	10.2601
21	devendra fadnavis	10.1470
22	penal code	10.1022
23	swachh bharat	10.0862
24	freedom fighter	10.0715
25	slum dweller	10.0346
26	rajya sabha	10.0135
27	shiv sena	9.9820
28	lok sabha	9.8566
29	rohith vemula	9.8273

S. No.	Bigram	PMI
30	sq ft	9.6569
31	western disturbance	9.6211
32	story offline	9.6129
33	arvind kejriwal	9.5935
34	chinese manjha	9.5800
35	dense fog	9.5551
36	birth anniversary	9.5337
37	renewable energy	9.4975
38	tribunal ngt	9.4840
39	mahatma gandhi	9.4766
40	tamil nadu	9.4739
41	manohar lal	9.4205
42	cctv footage	9.4136
43	appa rao	9.3603
44	vice chancellor	9.3048
45	j jayalalithaa	9.2945
46	cold wave	9.2788
47	writ petition	9.2346
48	sim card	9.2013
49	real estate	9.1993
50	boundary wall	9.1423
51	square yard	9.1422
52	stray dog	9.1354

S. No.	Bigram	PMI
53	narendra modis	9.1086
54	ration card	9.0813
55	cctv camera	9.0442
56	indira gandhi	9.0407
57	animal husbandry	9.0384

Frequencies

Bigrams:

S. No.	Bigram	Frequency
1	r crore	2593
2	state government	2026
3	chief minister	1982
4	year old	1835
5	police station	1679
6	r lakh	1350
7	official said	1334
8	high court	1314
9	police said	1064
10	told toi	1014
11	source said	1012
12	new delhi	948
13	municipal corporation	844

S. No.	Bigram	Frequency
14	civic body	765
15	prime minister	571
16	new year	545
17	police officer	508
18	tamil nadu	499
19	year ago	494
20	degree celsius	471

Words:

S. No.	Word	Frequency
1	said	33144
2	police	13682
3	year	10921
4	state	10302
5	government	9386
6	city	8307
7	r	7566
8	day	6818
9	official	6404
10	case	5638
11	people	5427
12	student	5423
13	minister	5291

S. No.	Word	Frequency
14	time	5148
15	road	5119
16	district	5007
17	area	4964
18	court	4775
19	department	4558
20	new	4474

India

Bigrams

> Filters for the table
>
> Min PMI: 9
>
> Min frequency: 10

S. No.	Bigram	PMI
1	wi fi	10.5464
2	saudi arabia	10.2781
3	barack obama	10.2781
4	lone wolf	10.1156

S. No.	Bigram	PMI
5	aam aadmi	10.1032
6	dipak misra	10.0716
7	ghulam nabi	10.0668
8	ardh kumbh	9.9690
9	bullock cart	9.8155
10	nobel laureate	9.8127
11	nicobar island	9.7972
12	mukul rohatgi	9.7857
13	shafi armar	9.7817
14	sitaram yechury	9.7441
15	terminally ill	9.7101
16	swami vivekananda	9.6485
17	sri lanka	9.6263
18	vikas swarup	9.5968
19	col niranjan	9.5850
20	bharatiya janata	9.5348
21	jawaharlal nehru	9.5232
22	kapil sibal	9.5204
23	passive euthanasia	9.4870
24	jet airway	9.4738
25	suresh prabhu	9.4452
26	madan gopal	9.3535

S. No.	Bigram	PMI
27	environmental clearance	9.3484
28	swachh bharat	9.3438
29	mamata banerjee	9.3337
30	arab league	9.3167
31	maulana masood	9.2634
32	lie detector	9.2066
33	nitin gadkari	9.1917
34	oommen chandy	9.1795
35	sexual harassment	9.1737
36	cook madan	9.1616
37	venkaiah naidu	9.1453
38	jd u	9.1411
39	sushma swaraj	9.1361
40	ford foundation	9.0731
41	nabam tuki	9.0266
42	masood azhar	9.0005

Frequencies

Bigrams:

S. No.	Bigram	Frequency
1	new delhi	710
2	prime minister	336

S. No.	Bigram	Frequency
3	chief minister	294
4	source said	237
5	r crore	236
6	narendra modi	223
7	supreme court	176
8	minister narendra	170
9	official said	162
10	terror attack	150
11	air force	147
12	state government	137
13	high court	121
14	told toi	118
15	pathankot attack	106
16	republic day	104
17	tamil nadu	95
18	chief justice	94
19	west bengal	92
20	security force	91

Words:

S. No.	Words	Frequency
1	said	3802
2	government	1702
3	india	1556
4	minister	1343
5	state	1231
6	delhi	1145
7	new	1088
8	year	1073
9	party	814
10	court	775
11	attack	759
12	day	746
13	congress	744
14	pakistan	720
15	police	716
16	country	713
17	bjp	711
18	chief	693
19	indian	666
20	people	618

Life

Bigrams

> Filters for the table
>
> Min PMI: 6.2
>
> Min frequency: 30

S. No.	Bigram	PMI
1	omega fatty	8.5011
2	bone marrow	8.3798
3	fatty acid	7.9130
4	olive oil	7.4450
5	zika virus	7.4041
6	social medium	7.1637
7	daily mirror	7.1179
8	basmati rice	7.1105
9	lucky colour	6.9773
10	brown rice	6.7669
11	dr jenkins	6.7401
12	heart attack	6.6881
13	vitamin d	6.6828
14	home remedy	6.5736
15	blood circulation	6.5348
16	blood pressure	6.4470

S. No.	Bigram	PMI
17	calorie intake	6.4232
18	vitamin c	6.3964
19	weight gain	6.3806
20	green tea	6.3775
21	weight loss	6.3195
22	long term	6.2300
23	junk food	6.2265

Frequencies

Bigrams:

S. No.	Bigram	Frequency
1	make sure	148
2	weight loss	100
3	blood pressure	94
4	year old	93
5	heart disease	90
6	health benefit	80
7	zika virus	73
8	say dr	63
9	vitamin c	63
10	fatty acid	62
11	new year	60

S. No.	Bigram	Frequency
12	type diabetes	59
13	brown rice	58
14	vitamin d	58
15	new study	51
16	year ago	50
17	social medium	48
18	long term	46
19	dont want	45
20	omega fatty	44

Words:

S. No.	Word	Frequency
1	time	1252
2	make	1199
3	like	1069
4	help	1037
5	say	997
6	people	918
7	body	915
8	year	873
9	said	835
10	day	812

S. No.	Word	Frequency
11	food	782
12	skin	772
13	woman	687
14	child	669
15	study	669
16	health	658
17	just	649
18	new	648
19	good	637
20	life	622

World

Bigrams

> Filters for the table
>
> Min PMI: 6
>
> Min frequency: 30

S. No.	Bigram	PMI
1	hong kong	8.9180
2	asylum seeker	8.8153
3	hydrogen bomb	7.4119

S. No.	Bigram	PMI
4	u s	7.3852
5	middle east	7.3512
6	prime minister	7.1277
7	hillary clinton	7.0899
8	john kerry	7.0754
9	saudi arabia	6.9955
10	fox news	6.9790
11	al qaida	6.9642
12	barack obama	6.9313
13	human right	6.9183
14	air strike	6.6930
15	white house	6.6814
16	zika virus	6.5728
17	news agency	6.5517
18	social medium	6.5504
19	told reuters	6.4148
20	told afp	6.4055
21	security council	6.3453
22	donald trump	6.3247
23	told reporter	6.3221
24	president barack	6.3162
25	foreign ministry	6.2998

S. No.	Bigram	PMI
26	south carolina	6.2739
27	presidential candidate	6.1728
28	new hampshire	6.0816
29	new york	6.0688

Frequencies

Bigrams:

S. No.	Bigram	Frequency
1	united state	240
2	north korea	235
3	saudi arabia	157
4	new york	156
5	islamic state	144
6	official said	115
7	year old	101
8	south korea	95
9	prime minister	87
10	white house	85
11	u s	79
12	human right	68
13	new year	64
14	donald trump	61

S. No.	Bigram	Frequency
15	news agency	56
16	north korean	55
17	new hampshire	55
18	nuclear test	53
19	security force	52
20	hillary clinton	51

Words:

S. No.	Word	Frequency
1	said	2361
2	state	886
3	year	791
4	people	687
5	new	631
6	country	524
7	north	465
8	official	453
9	attack	449
10	group	447
11	government	425
12	time	413
13	president	413

S. No.	Word	Frequency
14	china	410
15	trump	392
16	nuclear	386
17	korea	349
18	iran	330
19	told	326
20	force	317

Business

Bigrams

> Filters for the table
>
> Min PMI: 9
>
> Min frequency: 5

S. No.	Bigram	PMI
1	hero motocorp	10.6154
2	san francisco	10.6154
3	mscis broadest	10.4331
4	gen ze	10.4331
5	grama panchayat	10.4331
6	thomas cook	10.2789

S. No.	Bigram	PMI
7	jio infocomm	10.1454
8	silicon valley	10.1454
9	texas intermediate	10.1248
10	sukanya samriddhi	10.0276
11	tamil nadu	9.9222
12	rajya sabha	9.9222
13	nirmala sitharaman	9.9222
14	coca cola	9.8269
15	circuit breaker	9.8269
16	narayana hrudayalaya	9.7399
17	saudi arabia	9.5858
18	arundhati bhattacharya	9.5858
19	viral shot	9.5858
20	l ampt	9.5646
21	germany dax	9.5576
22	sq ft	9.5168
23	patanjali ayurved	9.5168
24	infinite analytics	9.4905
25	losing streak	9.3344
26	blue chip	9.2664
27	hang seng	9.2291
28	raw material	9.2291

S. No.	Bigram	PMI
29	jan dhan	9.2291
30	angel broking	9.2291
31	morgan stanley	9.1803
32	jp morgan	9.1803
33	intermediate wti	9.1803
34	somnath temple	9.1521
35	dedicated freight	9.1490
36	app click	9.1468
37	mercedes benz	9.1338
38	dual mode	9.1338
39	bharti airtel	9.1209
40	poll conducted	9.1158
41	aditya birla	9.1113
42	kongs hang	9.0956
43	shree cement	9.0893
44	sun pharma	9.0749
45	generic medicine	9.0749
46	latin america	9.0551
47	dhan yojana	9.0468
48	freight corridor	9.0313
49	electrified route	9.0059

Frequencies

Bigrams:

S. No.	Bigram	Frequency
1	r crore	443
2	new delhi	158
3	r lakh	98
4	oil price	92
5	stock market	78
6	u s	67
7	year ago	63
8	early trade	60
9	central bank	59
10	s ampp	58
11	lakh crore	55
12	net profit	49
13	managing director	49
14	long term	48
15	official said	47
16	crude oil	44
17	source said	43
18	mutual fund	42
19	emerging market	39
20	bse sensex	39

Words:

S. No.	Word	Frequency
1	said	1546
2	year	1105
3	india	997
4	market	898
5	company	775
6	r	741
7	bank	644
8	crore	532
9	new	510
10	government	500
11	growth	476
12	cent	420
13	price	411
14	investor	410
15	investment	378
16	global	360
17	rate	353
18	time	351
19	fund	329
20	china	326