Task 1

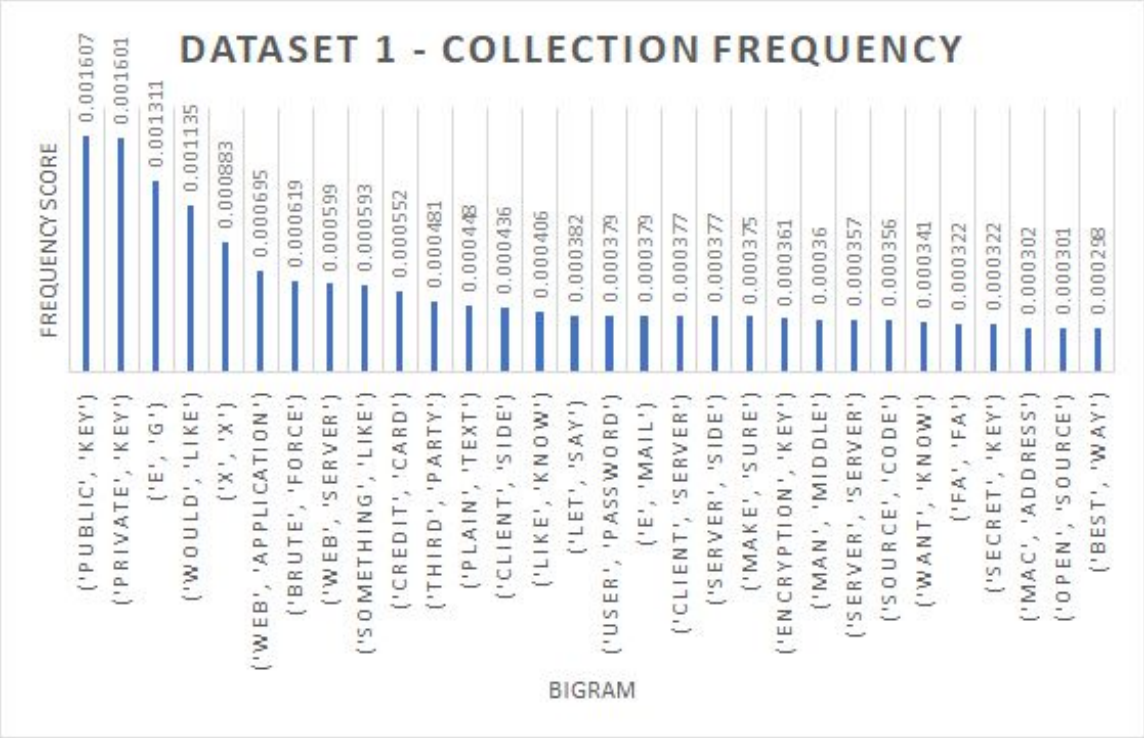
Scores of Dataset 1 (StackOverflow Questions)

Bigram	Dataset 1 - Collection Frequency
('public', 'key')	0.001607
('private', 'key')	0.001601
('e', 'g')	0.001311
('would', 'like')	0.001135
('x', 'x')	0.000883
('web', 'application')	0.000695
('brute', 'force')	0.000619
('web', 'server')	0.000599
('something', 'like')	0.000593
('credit', 'card')	0.000552
('third', 'party')	0.000481
('plain', 'text')	0.000448
('client', 'side')	0.000436
('like', 'know')	0.000406
('let', 'say')	0.000382
('user', 'password')	0.000379
('e', 'mail')	0.000379
('client', 'server')	0.000377
('server', 'side')	0.000377
('make', 'sure')	0.000375
('encryption', 'key')	0.000361
('man', 'middle')	0.00036
('server', 'server')	0.000357
('source', 'code')	0.000356
('want', 'know')	0.000341
('fa', 'fa')	0.000322
('secret', 'key')	0.000322
('mac', 'address')	0.000302
('open', 'source')	0.000301
('best', 'way')	0.000298

Bigram	Dataset 1 - Collocation Score
('antagonist','fortnightly')	20.728809
('appeaser','inmate')	20.728809
('arthritis','deplete')	20.728809
('barnard','libra')	20.728809
('bedazzle','incompletion')	20.728809
('benzene','pedal')	20.728809
('bonfire','melted')	20.728809
('cacao','hilarity')	20.728809
('californium','scandium')	20.728809
('cancer','herpes')	20.728809
('celebration','arthritis')	20.728809
('cholera','impost')	20.728809
('clapped','glee')	20.728809
('clockwork','muse')	20.728809
('cockle','cobby')	20.728809
('colonel','captain')	20.728809
('countryside','greenery')	20.728809
('crept','bitterroot')	20.728809
('crust','cockle')	20.728809
('cynic','miserable')	20.728809
('dean','pierce')	20.728809
('deplete','vestry')	20.728809
('destroyer','bourgeoisie')	20.728809
('deuterium','luckless')	20.728809
('developmental','singularity')	20.728809
('diseased','pervert')	20.728809
('dominance','resiliency')	20.728809
('emu','lation')	20.728809
('fairy','tales')	20.728809
('ferrite','bead')	20.728809

Analysis & Observations of StackOverflow Dataset

Collection Frequency Score



Inferences from the plot

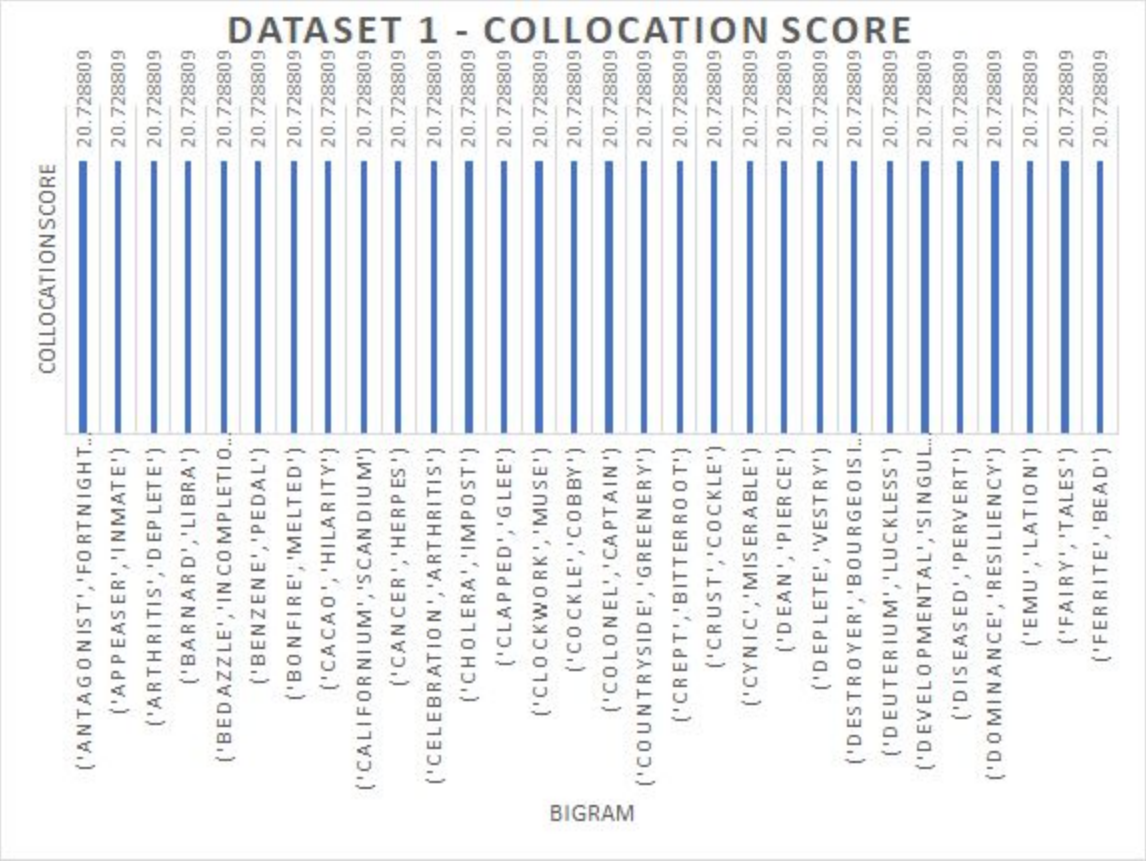
By looking at the Collection Frequency stats of Dataset 1(stackoverflow questions) we notice that the bigrams with top 30 highest frequency scores are related to Computer Security and Encryption and other software architecture related topics.

This can be observed with the bigrams that indicate

Computer Security & Encryption:
<public, key>
<private, key>
<secret, key>
<encryption, key>
Access Credentials:
<credit, card>
<e, mail>
<user, password>
<mac, address>
Attacks:
<man, middle>
<brute, force>
Software Application Architecture:
<web, application>
<client, side>
<client, server>

<server, side>
<source, code>
<open, source>

Collocation Score



Inferences from the plot

We notice that the collocation score of all the bigrams is same across the dataset 1 (stackoverflow questions). Moreover we also notice that, though we anticipate some technical topics to be scored in top bigrams but the collocation has generated pointwise mutual informational bigrams. As the score of each bigram in the collection is same, it is hard to correlate the dataset with what the dataset is all about.

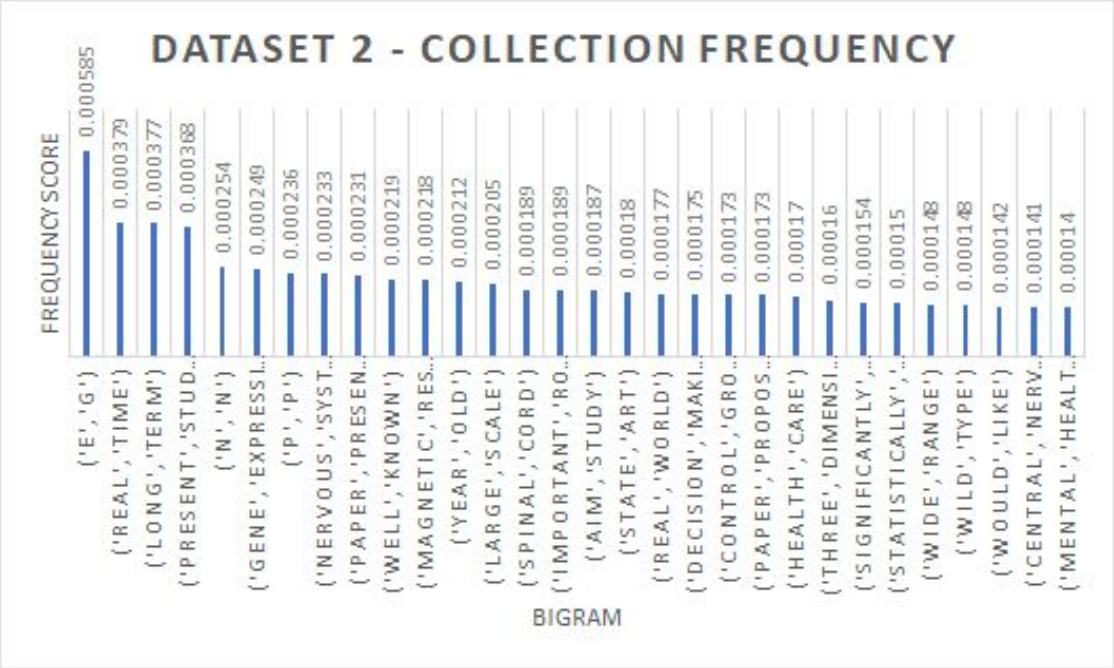
Scores of Dataset 2 (Abstracts of Research Papers)

Bigram	Dataset 2 - Collection Frequency
('e','g')	0.000585
('real','time')	0.000379
('long','term')	0.000377
('present','study')	0.000368
('n','n')	0.000254
('gene','expression')	0.000249
('p','p')	0.000236
('nervous','system')	0.000233
('paper','present')	0.000231
('well','known')	0.000219
('magnetic','resonance')	0.000218
('year','old')	0.000212
('large','scale')	0.000205
('spinal','cord')	0.000189
('important','role')	0.000189
('aim','study')	0.000187
('state','art')	0.00018
('real','world')	0.000177
('decision','making')	0.000175
('control','group')	0.000173
('paper','propose')	0.000173
('health','care')	0.00017
('three','dimensional')	0.00016
('significantly','higher')	0.000154
('statistically','significant')	0.00015
('wide','range')	0.000148
('wild','type')	0.000148
('would','like')	0.000142
('central','nervous')	0.000141
('mental','health')	0.00014

Bigram	Dataset 2 - Collocation Score
('adjectival','determinative')	21.300589
('algaecide','sanative')	21.300589
('amoebiasis','trichomoniasis')	21.300589
('amusing','avocation')	21.300589
('amusingly','lurid')	21.300589
('aneroid','manometer')	21.300589
('anophthalmos','microphthalmos')	21.300589
('antilithic','emmenagogue')	21.300589
('apport','serge')	21.300589
('arcuated','frown')	21.300589
('arrogant','rude')	21.300589
('ashen','felter')	21.300589
('attractively','repulsively')	21.300589
('aviator','trespasser')	21.300589
('bade','heller')	21.300589
('barbecue','spit')	21.300589
('bellows','pelting')	21.300589
('belted','clearwing')	21.300589
('biocoenosis','photophilic')	21.300589
('brassy','ringy')	21.300589
('breeze','whipping')	21.300589
('brewer','cotman')	21.300589
('caderas','dourine')	21.300589
('cannibalism','spadefoot')	21.300589
('caryopsis','spikelet')	21.300589
('cashmere','sweater')	21.300589
('ceremony','innocence')	21.300589
('chiro','practic')	21.300589
('chondroma','durra')	21.300589
('clandestine','paramilitary')	21.300589

Analysis & Observations of Abstracts of Research Papers

Collection Frequency Score



Inferences from the plot

By looking at the Collection Frequency stats of Dataset 2 (Abstracts of Research Papers) we notice that the bigrams with top 30 highest frequency scores are related to Medical Field in the *neuroscience specialization* related topics.

This can be concluded with the bigrams that indicate topics related to

Neuroscience:

<gene, expression>

<nervous, system>

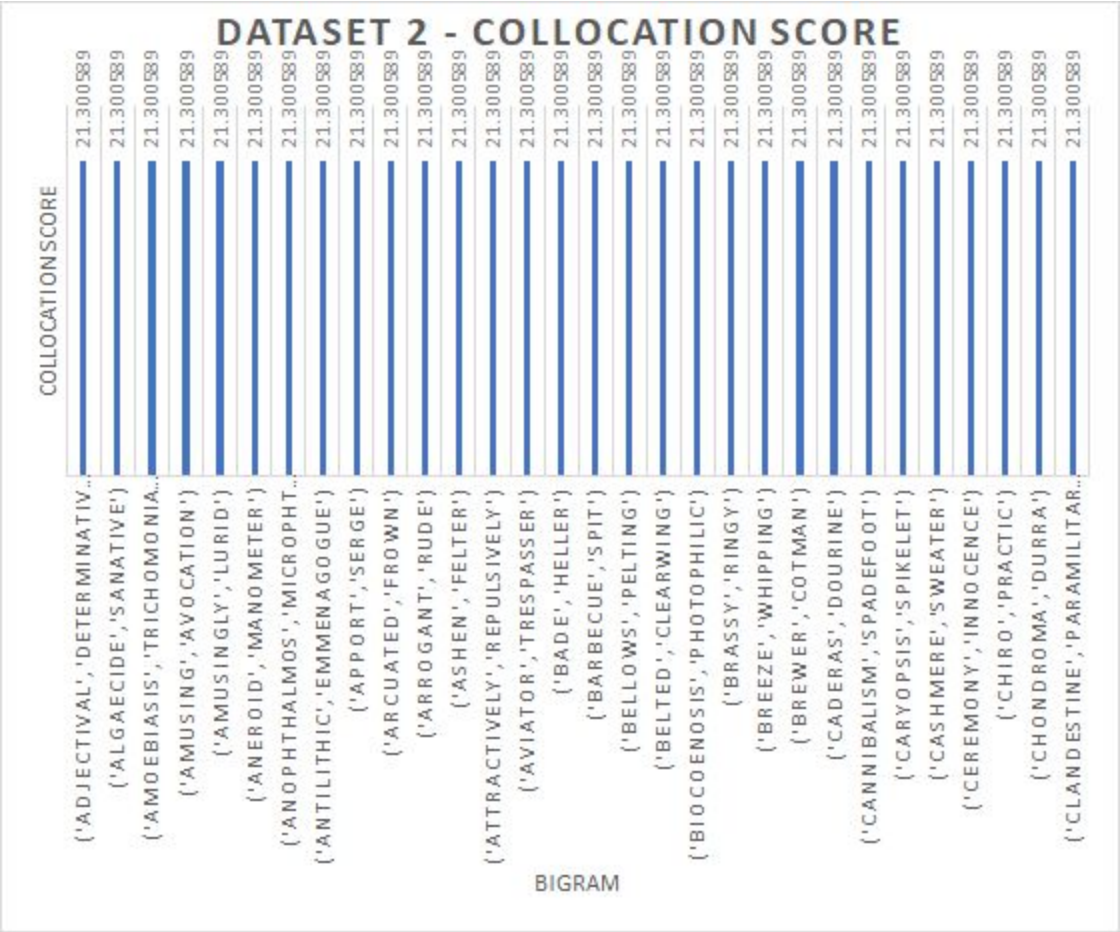
<spinal, cord>

<magnetic, resonance>

<health, care>

<central, nervous>

Collocation Score

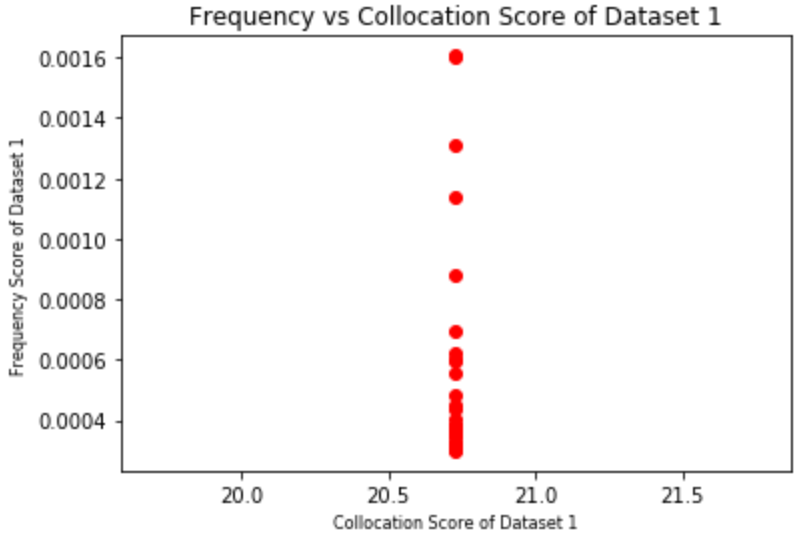


Inferences from the plot

We notice that the collocation score of all the bigrams is same across the dataset 2 (Abstracts of Research Papers). Moreover we also notice that, though we anticipate some technical topics to be scored in top bigrams but the collocation has generated pointwise mutual informational bigrams. As the score of each bigram in the collection is same, it is hard to correlate the dataset with what the dataset is all about.

Task 2

Dataset 1 (StackOverflow Questions)

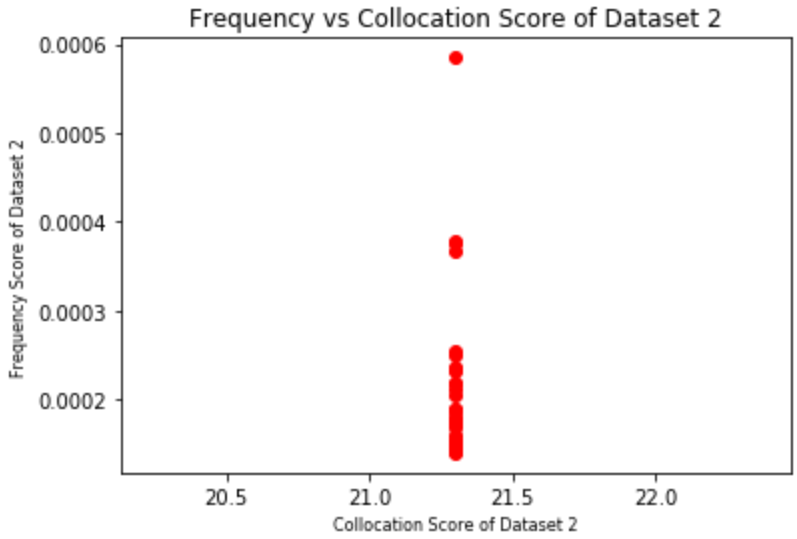


Inferences from the plot

The above graph for dataset 1 (stackoverflow questions) shows the collocation scores of bigrams on x axis which is evident that all the bigrams has same collocation score vs the collection frequency scores of most frequent bigrams on y axis.

One this is evident that, out of 30 most frequent bigrams the frequency scores for the top 10 bigrams is above 0.0005 and rest all fall below the score 0.0005

Dataset 2 (Abstracts of Research Papers)



Inferences from the plot

The above graph for dataset 2 (Abstracts of Research Papers) shows the collocation scores of bigrams on x axis which is evident that all the bigrams has same collocation score vs the collection frequency scores of most frequent bigrams on y axis.

It is evident that, out of 30 most frequent bigrams the frequency scores for the top 13 bigrams is above 0.0002 and rest all fall below the score 0.0002

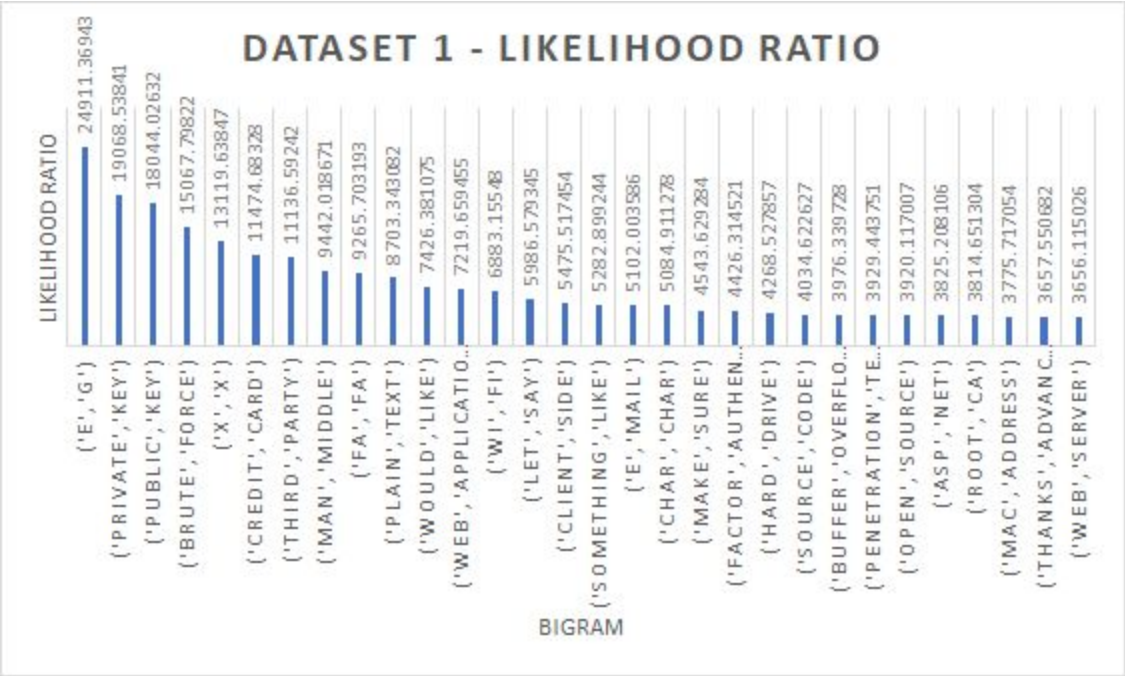
Task 3

I would like to evaluate Likelihood Ratio Hypothesis Bigram measure and analyse the Datasets 1 and 2.

Dataset 1 (StackOverflow Questions)

Bigram	Dataset 1 - Likelihood Ratio
('e','g')	24911.36943
('private','key')	19068.53841
('public','key')	18044.02632
('brute','force')	15067.79822
('x','x')	13119.63847
('credit','card')	11474.68328
('third','party')	11136.59242
('man','middle')	9442.018671
('fa','fa')	9265.703193
('plain','text')	8703.343082
('would','like')	7426.381075
('web','application')	7219.659455
('wi','fi')	6883.15548
('let','say')	5986.579345
('client','side')	5475.517454
('something','like')	5282.899244
('e','mail')	5102.003586
('char','char')	5084.911278
('make','sure')	4543.629284
('factor','authentication')	4426.314521
('hard','drive')	4268.527857
('source','code')	4034.622627
('buffer','overflow')	3976.339728
('penetration','testing')	3929.443751
('open','source')	3920.117007
('asp','net')	3825.208106
('root','ca')	3814.651304
('mac','address')	3775.717054
('thanks','advance')	3657.550682
('web','server')	3656.115026

Likelihood Ratio Score



Inferences from the plot

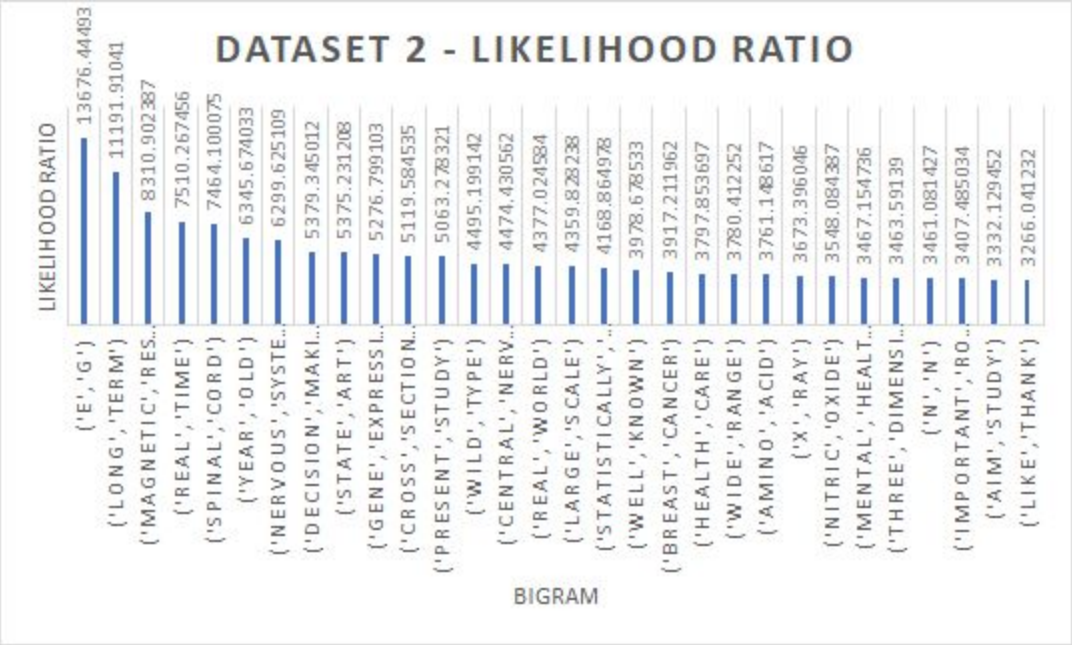
By looking at the Likelihood Ratio hypothesis scores of Dataset 1(stackoverflow questions) we notice that the bigrams with top 30 highest frequency scores are related to Computer Security and Encryption and other software architecture related topics.

This can be observed with the bigrams that indicate

Computer Security & Encryption:
<public, key>
<private, key>
Access Credentials:
<credit, card>
<e, mail>
<mac, address>
<factor, authentication>
Attacks:
<man, middle>
<brute, force>
<penetration, testing>
<buffer, overflow>
Software Application Architecture:
<web, application>
<web, server>
<client, side>
<source, code>
Tools & Technologies:
<third, party>
<open, source>
<asp, net>

Dataset 2 (Abstracts of Research Papers)

Bigram	Dataset 2 - Likelihood Ratio
('e','g')	13676.44493
('long','term')	11191.91041
('magnetic','resonance')	8310.902387
('real','time')	7510.267456
('spinal','cord')	7464.100075
('year','old')	6345.674033
('nervous','system')	6299.625109
('decision','making')	5379.345012
('state','art')	5375.231208
('gene','expression')	5276.799103
('cross','sectional')	5119.584535
('present','study')	5063.278321
('wild','type')	4495.199142
('central','nervous')	4474.430562
('real','world')	4377.024584
('large','scale')	4359.828238
('statistically','significant')	4168.864978
('well','known')	3978.678533
('breast','cancer')	3917.211962
('health','care')	3797.853697
('wide','range')	3780.412252
('amino','acid')	3761.148617
('x','ray')	3673.396046
('nitric','oxide')	3548.084387
('mental','health')	3467.154736
('three','dimensional')	3463.59139
('n','n')	3461.081427
('important','role')	3407.485034
('aim','study')	3332.129452
('like','thank')	3266.041232



Inferences from the plot

By looking at the Likelihood Ratio hypothesis scores of Dataset 2 (Abstracts of Research Papers) we notice that the bigrams with top 30 highest frequency scores are related to Medical Field in the *neuroscience specialization* related topics.

This can be concluded with the bigrams that indicate topics related to

Neuroscience:

<magnetic, resonance>
<spinal, cord>
<nervous, system>
<decision, making>
<gene, expression>
<cross, sectional>
<central, nervous>
<health, care>

Cancer:

<breast, cancer>

Study of Diagnosed Health Reports:

<cross, sectional>
<x, ray>
<three, dimensional>
<aim, study>

Conclusion

On an overall basis, The Collection Frequency Score and the Likelihood Score Hypothesis helped in the identification of what the majority of the dataset is concentrated about where as the Collocation Scores though had a bigram separation, but it failed to provide distinguished scores for each bigram there by difficult for any reviewer to come to a final judgement about what the dataset is more concentrated about.