巨量資料分析

期中報告

教師: 毛敬豪博士

賴志偉

資工所一年級碩士在職專班

M10315902

1. RFC 2821 3

2. Kibana Screenshot 5

2.1 Dashboard 5

2.2 Sample Log entry 5

2.3 Spam distribution(Better Map) 6

2.4 Spam Distribution by Country 7

3. Top 100 Sender IP 7

4. Top 100 Spam Source by Country 10

5. Top 200 Terms in Subject 13

6. LogStash Parser 18

5.1 Notes 18

5.2Spam Mail Preprocessing Program 19

5.3 LogStash Parser Configuration 19

7. Five Chosen Features 20

6. K-Means Clustering Result 21

6.1 K = 2 21

6.2 K=5 22

6.3 K = 7 23

6.4 K = 9 24

# 1. RFC 2821

The spam archive contains emails in RFC 2822 format. In which a spam mail may look like below.

|  |
| --- |
| Delivered-To: [mailer-daemon@bruce-guenter.dyndns.org](mailto:mailer-daemon@bruce-guenter.dyndns.org)  Received: (qmail 1788 invoked from network); 1 Jan 2010 14:03:00 -0000  Received: from localhost (localhost [127.0.0.1])  by bruce-guenter.dyndns.org ([192.168.1.2]); 01 Jan 2010 14:03:00 -0000  Received: from zak.futurequest.net ([127.0.0.1])  by localhost ([127.0.0.1])  with SMTP via TCP; 01 Jan 2010 14:03:00 -0000  Received: (qmail 1714 invoked from network); 1 Jan 2010 14:02:59 -0000  Received: from 164-go3-2.acn.waw.pl (unknown [62.121.125.164])  by zak.futurequest.net ([69.5.6.152])  with ESMTP via TCP; 01 Jan 2010 14:02:58 -0000  From: VIAGRA R Online Shop <mailer-daemon@bruce-guenter.dyndns.org>  To: mailer-daemon@bruce-guenter.dyndns.org  Subject: Valued customer mailer-daemon@bruce-guenter.dyndns.org 80% OFF on Pfizer.  Content-Type: text/html; charset="ISO-8859-1"  MIME-Version: 1.0  Content-Length: 2172 |

As you can see, there might be several “Received” headers.

RFC 2821 dictates the details about “Received:” header field.

<http://www.rfc-base.org/txt/rfc-2821.txt>

|  |
| --- |
| 3.8.2 Received Lines in Gatewaying  When forwarding a message into or out of the Internet environment, a gateway MUST prepend a Received: line, but it MUST NOT alter in any  way a Received: line that is already in the header. |

To better determine the spam mail source, if there are multiple “Received:” header fields, we should choose the bottom one.

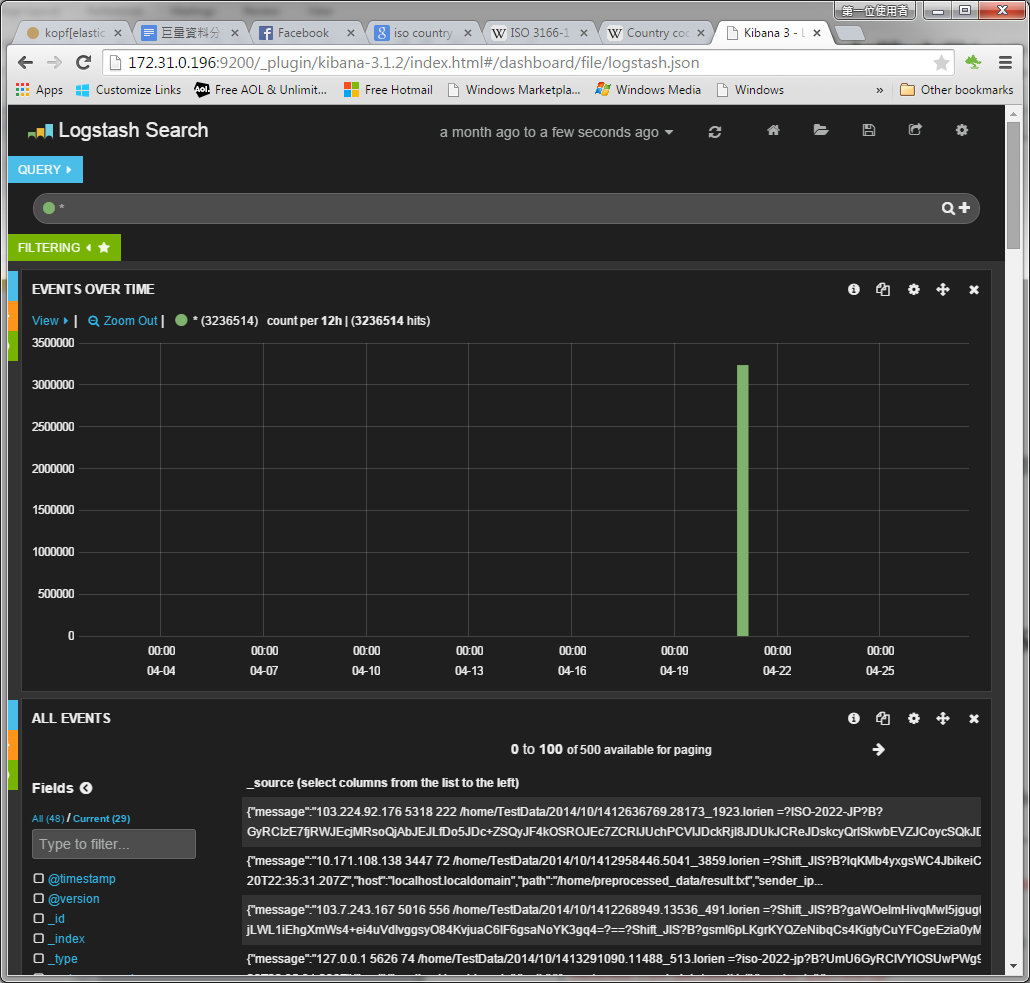
To prove this, I sent an email from my Hotmail account to my NTUST email account.

|  |
| --- |
| Received: from 140.118.31.97  by mail.ntust.edu.tw with Mail2000 ESMTP Server V6.00(12499:0:AUTH\_RELAY)  (envelope-from <rusty0831@hotmail.com>); Tue, 21 Apr 2015 19:41:02 +0800 (CST)  Return-Path: <rusty0831@hotmail.com>  X-MailGates: (RBL:2,PASS,0.000000e+00)(IPF:2,PASS)(bayesian:PASS,0.50,2)  (commtouch:2)(compute\_score:PASS,40,2)  Received: from 65.54.190.214  by mg2.ntust.edu.tw with MailGates ESMTP Server V4.0(20596:0:AUTH\_NONE)  (envelope-from <rusty0831@hotmail.com>); Tue, 21 Apr 2015 19:41:00 +0800 (CST)  Return-Path: <rusty0831@hotmail.com>  Received: from BAY172-W36 ([65.54.190.199]) by BAY004-OMC4S12.hotmail.com over TLS secured channel with Microsoft SMTPSVC(7.5.7601.22751);  Tue, 21 Apr 2015 04:40:54 -0700  X-TMN: [SUq1srI+y/pDEsm6P5ex28pJBRuFWFbsueZVtj2kMzI=]  X-Originating-Email: [rusty0831@hotmail.com]  Message-ID: <BAY172-W36693A9EA428AA9B10254EBCEF0@phx.gbl>  Return-Path: rusty0831@hotmail.com  Content-Type: multipart/alternative;  boundary="\_257779df-e82d-43bb-89b2-376f0be7c236\_"  From: =?big5?B?wbMgpXGnzA==?= <rusty0831@hotmail.com>  To: "m10315902@mail.ntust.edu.tw" <m10315902@mail.ntust.edu.tw>  Subject: Hello  Date: Tue, 21 Apr 2015 11:40:54 +0000  Importance: Normal  MIME-Version: 1.0  X-OriginalArrivalTime: 21 Apr 2015 11:40:54.0899 (UTC) FILETIME=[06D65430:01D07C28] |

Obviously, the bottom “Received” header BAY004-OMC4S12.hotmail.com is closest to the sender source.

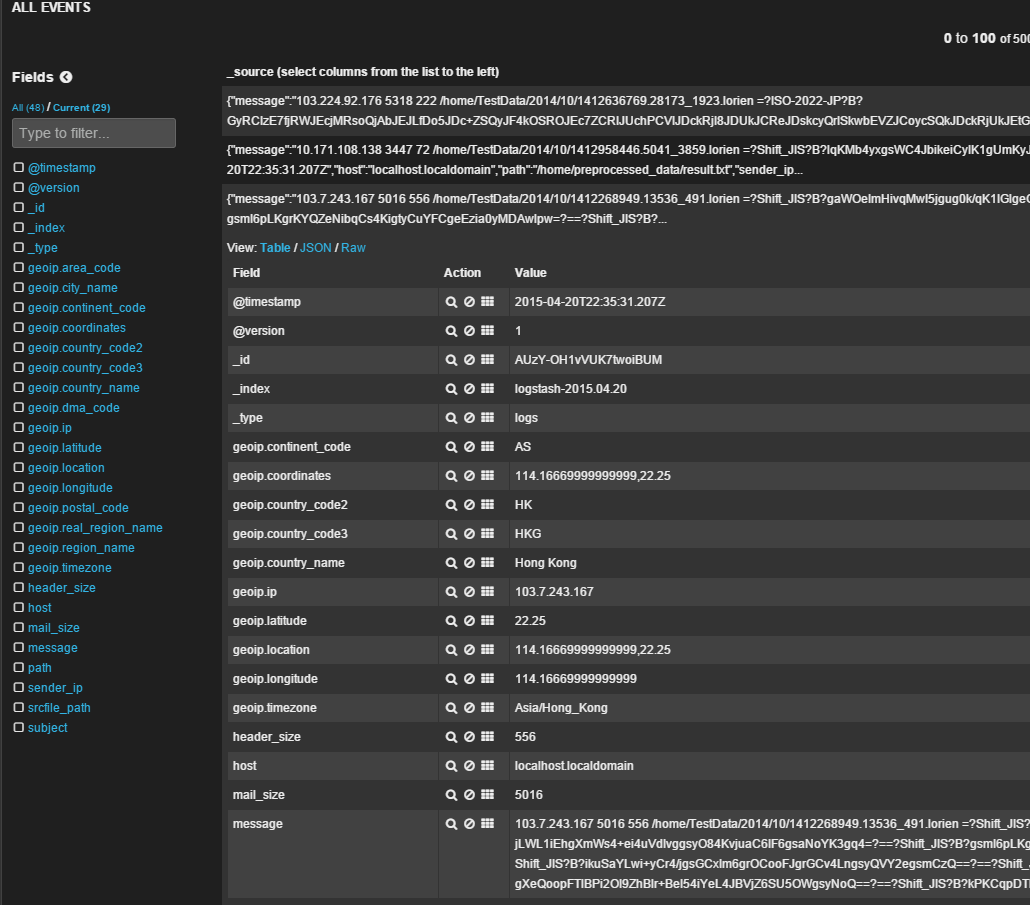
# Kibana Screenshot

## 2.1 Dashboard

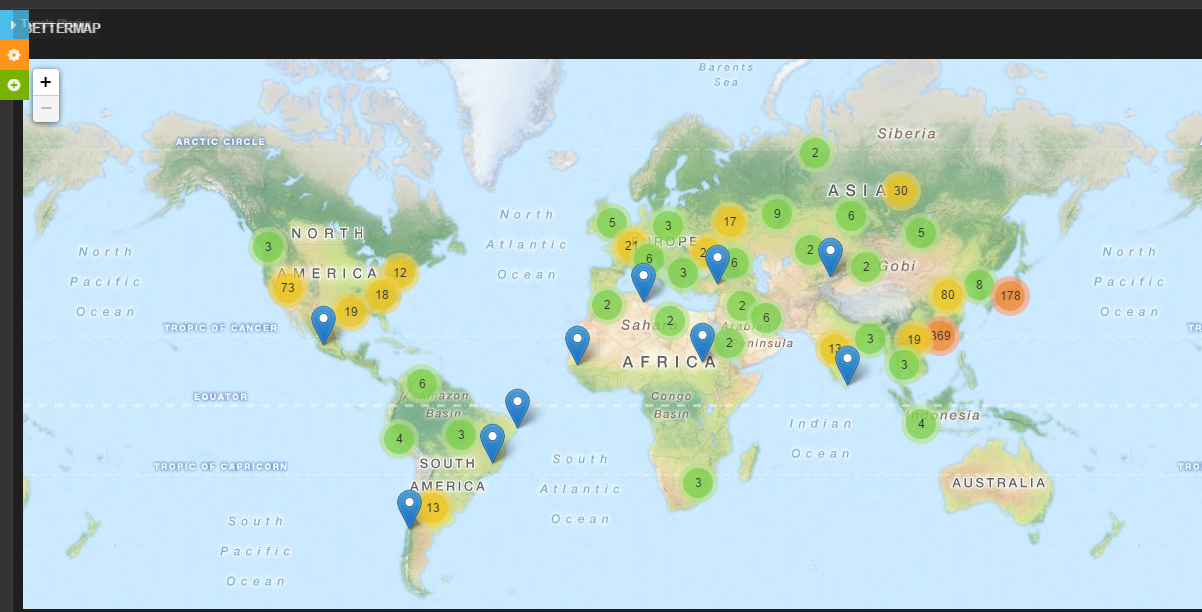


# 2.2 Sample Log entry

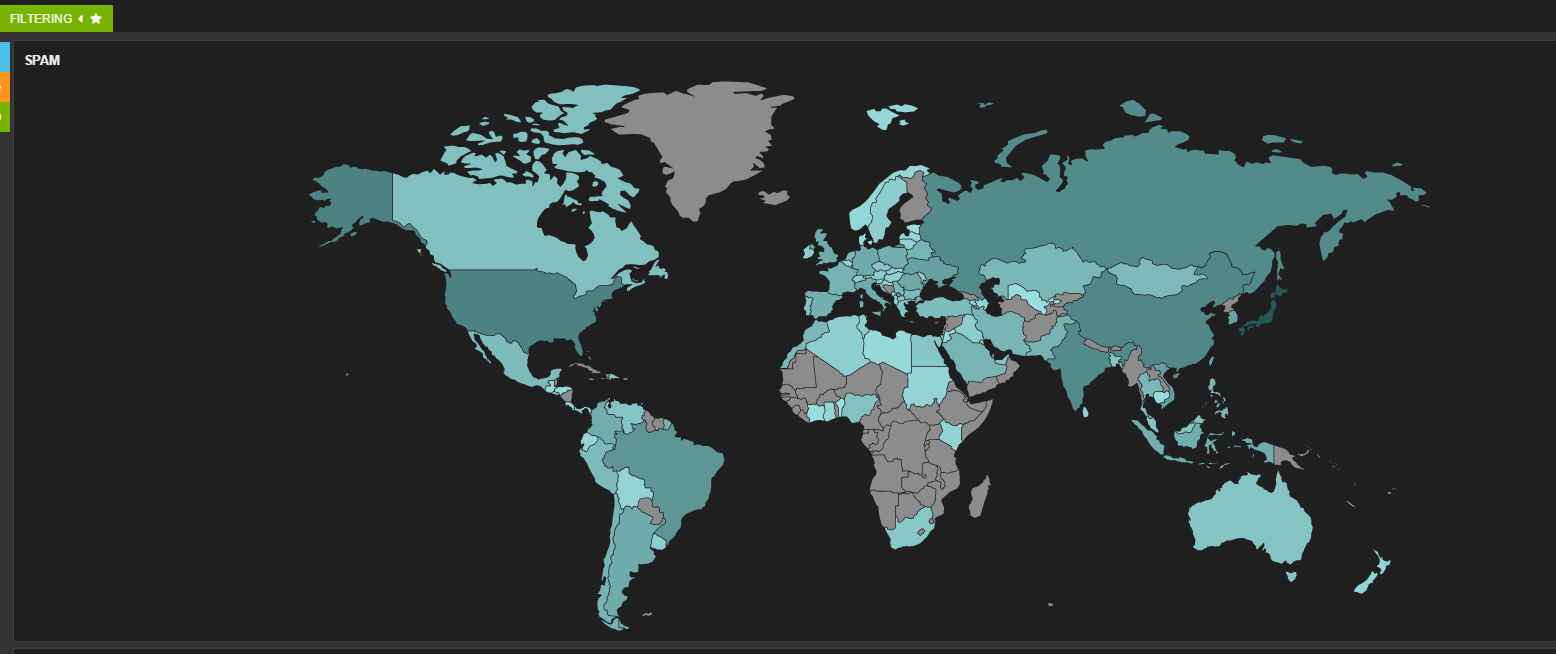
Here is a sample log entry. The source IP field is indexed with GeoIP.



## Spam distribution(Better Map)



## Spam Distribution by Country



# Top 100 Sender IP

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Sender IP** | **Count** | **GeoIP** |
| 1 | 127.0.0.1 | 226577 | <local host> |
| 2 | 103.3.16.11 | 46348 | Japan |
| 3 | 182.161.69.9 | 46309 | Japan |
| 4 | 182.161.71.128 | 32033 | Japan |
| 5 | 103.245.4.41 | 26969 | Japan |
| 6 | 103.245.4.42 | 22963 | Japan |
| 7 | 182.161.71.94 | 22768 | Japan |
| 8 | 103.3.18.229 | 20590 | Japan |
| 9 | 180.222.56.114 | 16711 | Japan |
| 10 | 111.223.202.13 | 16064 | Japan |
| 11 | 202.171.224.106 | 14028 | Japan |
| 12 | 67.198.152.82 | 12902 | USA |
| 13 | 103.6.63.65 | 11576 | Japan |
| 14 | 103.3.18.196 | 10401 | Japan |
| 15 | 192.168.169.84 | 10212 | <private IP> |
| 16 | 103.3.18.249 | 9036 | Japan |
| 17 | 182.236.110.1 | 8811 | Japan |
| 18 | 103.13.140.34 | 8297 | Japan |
| 19 | 116.214.95.86 | 7764 | Japan |
| 20 | 103.3.18.200 | 7573 | Japan |
| 21 | 192.168.108.105 | 7406 | <private IP> |
| 22 | 119.252.53.130 | 6998 | Japan |
| 23 | 182.161.71.22 | 6884 | Japan |
| 24 | 124.146.214.87 | 6708 | Japan |
| 25 | 103.245.6.8 | 6666 | Japan |
| 26 | 14.192.106.66 | 6330 | Japan |
| 27 | 103.249.212.156 | 6206 | N/A |
| 28 | 103.249.212.216 | 6068 | N/A |
| 29 | 103.249.212.186 | 6006 | N/A |
| 30 | 182.161.71.28 | 5913 | Japan |
| 31 | 103.3.18.76 | 5770 | Japan |
| 32 | 202.67.215.166 | 5757 | Hong Kong |
| 33 | 103.23.12.73 | 5664 | Japan |
| 34 | 103.3.18.245 | 5642 | Japan |
| 35 | 103.245.4.23 | 5469 | Japan |
| 36 | 119.252.44.32 | 5336 | Japan |
| 37 | 133.242.3.203 | 4944 | Japan |
| 38 | 115.30.22.239 | 4743 | Japan |
| 39 | 182.161.69.97 | 4709 | Japan |
| 40 | 111.223.202.36 | 4588 | Japan |
| 41 | 119.252.45.130 | 4528 | Japan |
| 42 | 14.192.117.2 | 4487 | Japan |
| 43 | 113.196.78.196 | 4437 | Taiwan |
| 44 | 192.168.169.81 | 4320 | <Private IP> |
| 45 | 14.192.113.66 | 4053 | Japan |
| 46 | 103.5.251.94 | 3957 | Japan |
| 47 | 211.129.99.202 | 3949 | Japan |
| 48 | 14.192.120.66 | 3840 | Japan |
| 49 | 14.192.117.160 | 3785 | Japan |
| 50 | 182.161.71.97 | 3646 | Japan |
| 51 | 114.112.20.249 | 3643 | China |
| 52 | 27.111.202.93 | 3282 | Hong Kong |
| 53 | 124.146.214.103 | 3250 | Japan |
| 54 | 119.252.53.226 | 3196 | Japan |
| 55 | 202.59.250.210 | 3126 | Taiwan |
| 56 | 182.161.71.106 | 3082 | Japan |
| 57 | 211.129.86.100 | 3070 | Japan |
| 58 | 61.63.30.66 | 3055 | Taiwan |
| 59 | 14.192.127.130 | 3029 | Japan |
| 60 | 173.45.112.195 | 3027 | USA |
| 61 | 119.252.62.130 | 2963 | Japan |
| 62 | 182.161.69.11 | 2953 | Japan |
| 63 | 103.12.217.40 | 2751 | Hong Kong |
| 64 | 119.252.62.131 | 2735 | Japan |
| 65 | 182.161.69.12 | 2661 | Japan |
| 66 | 219.101.197.122 | 2602 | Japan |
| 67 | 124.146.214.95 | 2544 | Japan |
| 68 | 103.240.80.6 | 2483 | N/A |
| 69 | 219.87.170.40 | 2442 | Taiwan |
| 70 | 111.223.202.133 | 2426 | Japan |
| 71 | 63.211.90.176 | 2415 | USA |
| 72 | 103.3.16.13 | 2379 | Japan |
| 73 | 14.192.124.224 | 2335 | Japan |
| 74 | 209.190.77.194 | 2187 | USA |
| 75 | 192.168.98.114 | 2186 | <Private IP> |
| 76 | 103.3.16.12 | 2144 | Japan |
| 77 | 10.10.10.18 | 2096 | <Private IP> |
| 78 | 111.223.202.141 | 1996 | Japan |
| 79 | 49.212.112.247 | 1892 | Japan |
| 80 | 10.10.10.11 | 1881 | <Private IP> |
| 81 | 210.150.128.202 | 1854 | Japan |
| 82 | 27.111.202.105 | 1799 | Hong Kong |
| 83 | 182.161.71.23 | 1788 | Japan |
| 84 | 119.252.44.194 | 1766 | Japan |
| 85 | 182.161.71.138 | 1707 | Japan |
| 86 | 183.60.143.121 | 1640 | China/Guangdong |
| 87 | 182.161.71.232 | 1579 | Japan |
| 88 | 117.18.68.86 | 1578 | Hong Kong |
| 89 | 49.212.112.225 | 1554 | Japan/Osaka |
| 90 | 119.252.44.224 | 1521 | Japan |
| 91 | 103.240.82.222 | 1500 | N/A |
| 92 | 219.101.197.92 | 1490 | Japan |
| 93 | 103.6.62.143 | 1482 | Japan |
| 94 | 103.240.82.224 | 1412 | N/A |
| 95 | 200.142.133.21 | 1159 | Brazil |
| 96 | 208.111.169.146 | 1096 | USA |
| 97 | 255.255.255.255 | 1073 | N/A |
| 98 | 221.214.208.226 | 1069 | China/Shanhai |
| 99 | 103.240.82.156 | 827 | N/A |
| 100 | 103.5.251.95 | 586 | Japan |

# Top 100 Spam Source by Country

|  |  |  |
| --- | --- | --- |
| **#** | **Country** | **Count** |
| 1 | Japan | 813495 |
| 2 | USA | 203658 |
| 3 | China | 165510 |
| 4 | Russia | 149340 |
| 5 | India | 149078 |
| 6 | Brazil | 90204 |
| 7 | Hong Kong | 68391 |
| 8 | Vietnam | 67388 |
| 9 | Ukraine | 58172 |
| 10 | South Korea | 53755 |
| 11 | Taiwan | 50245 |
| 12 | Romania | 36766 |
| 13 | Germany | 34565 |
| 14 | UK | 34280 |
| 15 | Argentina | 33936 |
| 16 | Italy | 32422 |
| 17 | Colombia | 29988 |
| 18 | Indonesia | 29790 |
| 19 | Poland | 26080 |
| 20 | Spain | 25876 |
| 21 | France | 22575 |
| 22 | Pakistan | 21671 |
| 23 | Chile | 20724 |
| 24 | Saudi Arabia | 19801 |
| 25 | Belarus | 19568 |
| 26 | Iran | 19085 |
| 27 | Kazakhstan | 17275 |
| 28 | Thailand | 16778 |
| 29 | Morocco | 16627 |
| 30 | Philippines | 16570 |
| 31 | Serbia | 15707 |
| 32 | Mongolia | 14595 |
| 33 | Bulgaria | 14384 |
| 34 | Singapore | 13897 |
| 35 | Mexico | 13719 |
| 36 | Peru | 13565 |
| 37 | Netherlands | 13458 |
| 38 | Turkey | 13350 |
| 39 | Israel | 11930 |
| 40 | Canada | 9424 |
| 41 | Dominican Republic | 8631 |
| 42 | Greece | 8590 |
| 43 | Nigeria | 8033 |
| 44 | Malaysia | 7855 |
| 45 | Portugal | 7512 |
| 46 | Australia | 7262 |
| 47 | Venezuela, Bolivarian Republic of | 7025 |
| 48 | Macedonia, the former Yugoslav Republic of | 6920 |
| 49 | Czech Republic | 6841 |
| 50 | United Arab Emirates | 6479 |
| 51 | Egypt | 6303 |
| 52 | Bangladesh | 6024 |
| 53 | South Africa | 5701 |
| 54 | Kuwait | 5001 |
| 55 | Austria | 3691 |
| 56 | Ireland | 3690 |
| 57 | Tunisia | 3651 |
| 58 | Switzerland | 3634 |
| 59 | Iraq | 3567 |
| 60 | Sweden | 3540 |
| 61 | Uruguay | 3468 |
| 62 | Croatia | 3200 |
| 63 | Algeria | 3185 |
| 64 | Montenegro | 3181 |
| 65 | Hungary | 3097 |
| 66 | Sri Lanka | 2954 |
| 67 | Albania | 2940 |
| 68 | Lithuania | 2779 |
| 69 | Moldova, Republic of | 2605 |
| 70 | Guatemala | 2312 |
| 71 | Slovakia | 2289 |
| 72 | Kenya | 2268 |
| 73 | Palestine, State of | 2168 |
| 74 | Belgium | 2141 |
| 75 | Sudan | 2139 |
| 76 | Armenia | 2135 |
| 77 | New Zealand | 2121 |
| 78 | Bolivia, Plurinational State of | 2053 |
| 79 | Azerbaijan | 2033 |
| 80 | Ghana | 1784 |
| 81 | Ecuador | 1759 |
| 82 | Latvia | 1676 |
| 83 | Norway | 1647 |
| 84 | Lebanon | 1570 |
| 85 | Costa Rica | 1541 |
| 86 | Denmark | 1522 |
| 87 | Panama | 1510 |
| 88 | El Salvador | 1487 |
| 89 | Mauritius | 1397 |
| 90 | Libya | 1362 |
| 91 | Slovenia | 1297 |
| 92 | Honduras | 1285 |
| 93 | Puerto Rico | 1259 |
| 94 | Jordan | 1259 |
| 95 | Cambodia | 1197 |
| 96 | Luxembourg | 1146 |
| 97 | Uzbekistan | 1101 |
| 98 | Côte d'Ivoire | 1066 |
| 99 | Cameroon | 1022 |
| 100 | Estonia | 1008 |

# Top 200 Terms in Subject

|  |  |  |
| --- | --- | --- |
| **#** | **Terms** | **Count** |
| 1 | b | 2065201 |
| 2 | iso | 1022165 |
| 3 | jp | 1017814 |
| 4 | 2022 | 1017212 |
| 5 | shift\_jis | 562721 |
| 6 | r | 383606 |
| 7 | koi8 | 380973 |
| 8 | viagra | 134602 |
| 9 | buy | 122349 |
| 10 | now | 113628 |
| 11 | for | 112130 |
| 12 | cialis | 105918 |
| 13 | your | 104261 |
| 14 | you | 91218 |
| 15 | the | 83244 |
| 16 | to | 81251 |
| 17 | discount | 79436 |
| 18 | off | 77881 |
| 19 | lists | 77198 |
| 20 | a | 70223 |
| 21 | and | 64618 |
| 22 | pharmacy | 54679 |
| 23 | x | 50840 |
| 24 | on | 50135 |
| 25 | sale | 48910 |
| 26 | of | 48060 |
| 27 | bruce | 47924 |
| 28 | watches | 46943 |
| 29 | online | 43168 |
| 30 | pills | 42180 |
| 31 | best | 41727 |
| 32 | free | 40547 |
| 33 | 8 | 39692 |
| 34 | with | 38944 |
| 35 | re | 38872 |
| 36 | is | 38263 |
| 37 | in | 38253 |
| 38 | windows | 37131 |
| 39 | get | 37035 |
| 40 | sjis | 37025 |
| 41 | today | 34623 |
| 42 | utf | 32763 |
| 43 | from | 32696 |
| 44 | 1251 | 32393 |
| 45 | 80 | 31661 |
| 46 | q | 30890 |
| 47 | guenter.dyndns.org | 30810 |
| 48 | new | 30476 |
| 49 | day | 29170 |
| 50 | prices | 28395 |
| 51 | vigara | 27979 |
| 52 | hi | 27926 |
| 53 | shipping | 26945 |
| 54 | we | 26711 |
| 55 | all | 26359 |
| 56 | have | 25668 |
| 57 | 1 | 25025 |
| 58 | at | 24553 |
| 59 | xae | 24418 |
| 60 | no | 23726 |
| 61 | delivery | 22254 |
| 62 | store | 22041 |
| 63 | 30 | 21965 |
| 64 | are | 21628 |
| 65 | this | 21548 |
| 66 | gb2312 | 21277 |
| 67 | more | 20006 |
| 68 | only | 19430 |
| 69 | 70 | 19099 |
| 70 | save | 18335 |
| 71 | our | 18263 |
| 72 | empty | 18018 |
| 73 | quality | 17890 |
| 74 | luxury | 17641 |
| 75 | cheap | 17496 |
| 76 | life | 17009 |
| 77 | rolex | 16906 |
| 78 | price | 16577 |
| 79 | good | 16119 |
| 80 | i | 16094 |
| 81 | penis | 15877 |
| 82 | be | 15565 |
| 83 | up | 15539 |
| 84 | replica | 15138 |
| 85 | here | 14768 |
| 86 | linux | 14179 |
| 87 | gz | 13937 |
| 88 | dear | 13901 |
| 89 | pill | 13701 |
| 90 | xc0 | 13552 |
| 91 | official | 13432 |
| 92 | can | 13323 |
| 93 | hello | 13208 |
| 94 | canadian | 13012 |
| 95 | gaakrpftlroxv4gmgsmcxjeyl3cjwprcgaey | 12815 |
| 96 | one | 12725 |
| 97 | need | 12635 |
| 98 | discounts | 12619 |
| 99 | usa | 12601 |
| 100 | special | 12514 |
| 101 | offer | 12093 |
| 102 | 3 | 12019 |
| 103 | site | 11999 |
| 104 | 75 | 11807 |
| 105 | xc1 | 11657 |
| 106 | xba | 11608 |
| 107 | it | 11551 |
| 108 | will | 11534 |
| 109 | 100 | 11510 |
| 110 | that | 11461 |
| 111 | xbf | 11430 |
| 112 | xb7 | 11424 |
| 113 | xb3 | 11079 |
| 114 | visa | 10861 |
| 115 | message | 10824 |
| 116 | vicodin | 10678 |
| 117 | watch | 10551 |
| 118 | ezmlm | 10493 |
| 119 | 8m | 10451 |
| 120 | super | 10352 |
| 121 | needed | 10344 |
| 122 | fast | 10248 |
| 123 | do | 10193 |
| 124 | love | 10190 |
| 125 | njl | 10094 |
| 126 | great | 10080 |
| 127 | xb0 | 9894 |
| 128 | svn | 9833 |
| 129 | notification | 9789 |
| 130 | 20 | 9725 |
| 131 | bruceg | 9665 |
| 132 | my | 9639 |
| 133 | 100mg | 9596 |
| 134 | enlargement | 9552 |
| 135 | available | 9522 |
| 136 | users | 9425 |
| 137 | 90 | 9420 |
| 138 | xb8 | 9395 |
| 139 | brand | 9390 |
| 140 | time | 9341 |
| 141 | 60 | 9339 |
| 142 | e | 9313 |
| 143 | levtira | 9298 |
| 144 | per | 9252 |
| 145 | 4 | 9121 |
| 146 | pfizer | 9076 |
| 147 | want | 9031 |
| 148 | her | 9004 |
| 149 | purchase | 8992 |
| 150 | cilais | 8981 |
| 151 | xb9 | 8974 |
| 152 | as | 8942 |
| 153 | me | 8866 |
| 154 | hydrocodone | 8834 |
| 155 | brands | 8762 |
| 156 | xbc | 8686 |
| 157 | 10 | 8642 |
| 158 | top | 8566 |
| 159 | kernel | 8548 |
| 160 | money | 8520 |
| 161 | customer | 8453 |
| 162 | xb1 | 8348 |
| 163 | high | 8331 |
| 164 | generic | 8272 |
| 165 | 7s | 8205 |
| 166 | xb5 | 8105 |
| 167 | 81 | 7992 |
| 168 | make | 7913 |
| 169 | xb2 | 7899 |
| 170 | user | 7858 |
| 171 | hey | 7854 |
| 172 | inches | 7842 |
| 173 | low | 7827 |
| 174 | k | 7789 |
| 175 | fwd | 7769 |
| 176 | russian | 7766 |
| 177 | xa6 | 7594 |
| 178 | us | 7592 |
| 179 | next | 7516 |
| 180 | xa1 | 7510 |
| 181 | 2010 | 7492 |
| 182 | mr | 7471 |
| 183 | by | 7322 |
| 184 | xc2 | 7239 |
| 185 | xc7 | 7224 |
| 186 | 77 | 7197 |
| 187 | gyrcjciksiq | 6929 |
| 188 | not | 6926 |
| 189 | 92 | 6906 |
| 190 | x81t | 6784 |
| 191 | xa9 | 5522 |
| 192 | pack | 5486 |
| 193 | gyhc | 5463 |
| 194 | sex | 5382 |
| 195 | gq | 5369 |
| 196 | 88 | 4215 |
| 197 | max | 4210 |
| 198 | 68 | 4182 |
| 199 | xe2 | 4136 |
| 200 | bags | 2058 |

# LogStash Parser

## 6.1 Notes

It appears that it’s inapplicable to parse correct sender IP address from RFC 2822 email format by using native LogStash parser directly.

To solve the problem, I wrote a program to pre-process the spam mail archive and summarize it to a single text file firstly.

The pre-processed format is as

<Sender IP>, <Mail Length>, <Subject Length>, <Source Spam Mail Path>, <Subject>

Then use LogStash to parse the pre-processed spam mail archive.

## 6.2Spam Mail Preprocessing Program



## 6.3 LogStash Parser Configuration

|  |
| --- |
| input {  file {  path => "/home/preprocessed\_data/result\_no\_subject.txt"  sincedb\_path => "/home/sincedb.db"  start\_position => beginning  }  }  filter {  grok {  match => [  "message",  "%{IPV4:sender\_ip} %{NUMBER:mail\_size} %{NUMBER:header\_size} %{DATA:srcfile\_path} %{GREEDYDATA:subject}"  ]  }  geoip {  source => "sender\_ip"  target => "geoip"  database =>"/root/logstash-1.4.2/vendor/geoip/GeoLiteCity.dat"  add\_field => [ "[geoip][coordinates]", "%{[geoip][longitude]}" ]  add\_field => [ "[geoip][coordinates]", "%{[geoip][latitude]}" ]  }  }  output {  elasticsearch {  protocol => "http"  index => "spam\_no\_subject"  }  } |

# Five Chosen Features

1. Mail length
2. Subject length
3. Hops

Take the following mail header for example; the number of Hops is 5.

|  |
| --- |
| Delivered-To: mailer-daemon@bruce-guenter.dyndns.org  Received: (qmail 1788 invoked from network); 1 Jan 2010 14:03:00 -0000  Received: from localhost (localhost [127.0.0.1])  by bruce-guenter.dyndns.org ([192.168.1.2]); 01 Jan 2010 14:03:00 -0000  Received: from zak.futurequest.net ([127.0.0.1])  by localhost ([127.0.0.1])  with SMTP via TCP; 01 Jan 2010 14:03:00 -0000  Received: (qmail 1714 invoked from network); 1 Jan 2010 14:02:59 -0000  Received: from 164-go3-2.acn.waw.pl (unknown [62.121.125.164])  by zak.futurequest.net ([69.5.6.152])  with ESMTP via TCP; 01 Jan 2010 14:02:58 -0000  From: VIAGRA R Online Shop <mailer-daemon@bruce-guenter.dyndns.org>  To: mailer-daemon@bruce-guenter.dyndns.org  Subject: Valued customer mailer-daemon@bruce-guenter.dyndns.org 80% OFF on Pfizer.  Content-Type: text/html; charset="ISO-8859-1"  MIME-Version: 1.0  Content-Length: 2172 |

1. Special character percentage in subject
2. Subject / Mail length ratio

# K-Means Clustering Result

All the clustering result as fellows use the following command line to invoke Mahout.

|  |
| --- |
| mahout kmeans   --input spamoct/data  --output spamoct/kmeans-output  -k <n>  --clusters irisoct/output  --maxIter 20  --method mapreduce  --distanceMeasure org.apache.mahout.common.distance.TanimotoDistanceMeasure  --clustering |

## K = 2

Raw:

|  |
| --- |
| VL-2191745{n=2767240 c=[4427.034, 81.444, 5.937, 0.132, NaN] r=[10266.118, 75.162, 1.542, 0.078, NaN]}  VL-1652576{n=469278 c=[1140.979, 51.948, 4.774, 0.121, 0.045] r=[174.831, 35.662, 1.480, 0.060, 0.029]} |

Table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Feature 1**  **Mail Size**  **(centroid)** | **Feature 2**  **Subject Length**  **(centroid)** | **Feature 3**  **Hops**  **(centroid)** | **Feature 4**  **Subject Special Character percentage**  **(centroid)** | **Feature 5**  **Subject Size/ Mail Size percentage**  **(centroid)** |
| VL-2191745 | 4427.034 | 81.444 | 5.937 | 0.132 | NaN |
| VL-1652576 | 1140.979 | 51.948 | 4.774 | 0.121 | 0.045 |

**Result Analysis:**

When K=2, the clustering result shows that there are significant difference in “Mail Size” (4427.034 vs 1140.979) and “Subject Length” (81.444 vs 51.948) between these 2 groups.

## K=5

|  |
| --- |
| VL-383827{n=150912 c=[2453.936, 74.156, 5.394, 0.144, NaN] r=[32838.377, 40.482, 1.049, 0.085, NaN]}  VL-2796729{n=621846 c=[4077.508, 88.962, 6.552, 0.133, 0.023] r=[775.395, 85.303, 1.701, 0.067, 0.024]}  VL-2617514{n=1177290 c=[2297.757, 81.965, 5.911, 0.137, 0.036] r=[375.372, 68.368, 1.453, 0.091, 0.028]}  VL-707045{n=681145 c=[1238.972, 53.828, 5.004, 0.125, 0.043] r=[207.652, 34.925, 1.374, 0.060, 0.027]}  VL-2793503{n=605331 c=[10459.026, 82.732, 5.639, 0.115, 0.010] r=[12754.556, 90.597, 1.613, 0.065, 0.012]} |

Table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Feature 1**  **Mail Size**  **(Centroid)** | **Feature 2**  **Subject Length**  **(Centroid)** | **Feature 3**  **Hops**  **(Centroid)** | **Feature 4**  **Subject Special Character percentage**  **(Centroid)** | **Feature 5**  **Subject Size/ Mail Size percentage**  **(Centroid)** |
| VL-383827 | 2453.936 | 74.156 | 5.394 | 0.144 | NaN |
| VL-2796729 | 4077.508 | 88.962 | 6.552 | 0.133 | 0.023 |
| VL-2617514 | 2297.757 | 81.965 | 5.911 | 0.137 | 0.036 |
| VL-707045 | 1238.972 | 53.828 | 5.004 | 0.125 | 0.043 |
| VL-2793503 | 10459.026 | 82.732 | 5.639 | 0.115 | 0.010 |

**Result Analysis:**

When K=5, the most significant difference are

1. Mail Size: From a maximum of 10459.026 to minimum of 1238.972.
2. Subject Length: From a maximum of 88.962 to 53.828. The group (VL-707045) having the smallest mail size also has the minimum of “Subject Length” (53.828) and Hops (5.004).

## K = 7

|  |
| --- |
| VL-1729446{n=335298 c=[1820.107, 65.455, 5.476, 0.133, NaN] r=[22037.727, 38.045, 0.952, 0.058, NaN]}  VL-719850{n=554196 c=[2429.576, 93.879, 5.972, 0.135, 0.038] r=[176.706, 85.780, 1.510, 0.103, 0.033]}  VL-532891{n=798378 c=[6064.805, 86.855, 6.192, 0.119, 0.015] r=[1488.212, 79.187, 1.608, 0.064, 0.013]}  VL-2061621{n=150696 c=[20519.999, 67.599, 4.748, 0.131, 0.004] r=[22703.421, 106.183, 1.476, 0.071, 0.010]}  VL-2889800{n=456099 c=[3168.700, 85.744, 6.568, 0.136, 0.027] r=[300.869, 85.450, 1.748, 0.071, 0.027]}  VL-1850970{n=388797 c=[1099.928, 48.650, 4.638, 0.119, 0.044] r=[164.258, 33.958, 1.535, 0.061, 0.029]}  VL-1442982{n=553064 c=[1848.337, 69.121, 5.543, 0.142, 0.038] r=[155.759, 39.366, 1.132, 0.082, 0.021]} |

Table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Feature 1**  **Mail Size**  **(Centroid)** | **Feature 2**  **Subject Length**  **(Centroid)** | **Feature 3**  **Hops**  **(Centroid)** | **Feature 4**  **Subject Special Character percentage**  **(Centroid)** | **Feature 5**  **Subject Size/ Mail Size percentage**  **(Centroid)** |
| VL-1729446 | 1820.107 | 65.455 | 5.476 | 0.133 | NaN |
| VL-719850 | 2429.576 | 93.879 | 5.972 | 0.135 | 0.038 |
| VL-532891 | 6064.805 | 86.855 | 6.192 | 0.119 | 0.015 |
| VL-2061621 | 20519.999 | 67.599 | 4.748 | 0.131 | 0.004 |
| VL-2889800 | 3168.700 | 85.744 | 6.568 | 0.136 | 0.027 |
| VL-1850970 | 1099.928 | 48.650 | 4.638 | 0.119 | 0.044 |
| VL-1442982 | 1848.337 | 69.121 | 5.543 | 0.142 | 0.038 |

**Result Analysis:**

When K=5, the most significant difference are

1. Mail Size: Ranging from a minimum of 1099.928 to a maximum of 20519.999.
2. The group “VL-1850970” has the smallest mail size (1099.928) and “Hops” (4.638).

## K = 9

|  |
| --- |
| VL-833968{n=222247 c=[2576.600, 68.023, 5.592, 0.139, NaN] r=[27054.444, 39.187, 1.321, 0.088, NaN]}  VL-1798512{n=520296 c=[2794.740, 95.999, 6.424, 0.131, 0.034] r=[214.454, 96.712, 1.636, 0.077, 0.034]}  VL-1159950{n=412674 c=[1694.584, 67.782, 5.499, 0.143, 0.040] r=[114.864, 37.977, 1.001, 0.074, 0.022]}  VL-1989531{n=319793 c=[2294.007, 84.358, 5.827, 0.141, 0.037] r=[97.164, 59.342, 1.464, 0.115, 0.026]}  VL-1303160{n=163943 c=[1333.813, 96.913, 5.492, 0.118, 0.073] r=[75.475, 31.333, 0.756, 0.049, 0.023]}  VL-517932{n=213596 c=[1370.703, 38.692, 5.466, 0.139, 0.028] r=[86.189, 20.055, 0.986, 0.059, 0.014]}  VL-1918993{n=510775 c=[4239.453, 85.739, 6.507, 0.133, 0.021] r=[686.016, 78.831, 1.703, 0.067, 0.020]}  VL-2580290{n=258031 c=[1018.490, 42.296, 4.216, 0.115, 0.041] r=[143.266, 28.505, 1.622, 0.065, 0.027]}  VL-2363321{n=615177 c=[10382.116, 82.708, 5.645, 0.115, 0.010] r=[12666.440, 90.272, 1.614, 0.065, 0.012]} |

Table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Feature 1**  **Mail Size**  **(Centroid)** | **Feature 2**  **Subject Length**  **(Centroid)** | **Feature 3**  **Hops**  **(Centroid)** | **Feature 4**  **Subject Special Character percentage**  **(Centroid)** | **Feature 5**  **Subject Size/ Mail Size percentage**  **(Centroid)** |
| VL-833968 | 2576.600 | 68.023 | 5.592 | 0.139 | NaN |
| VL-1798512 | 2794.740 | 95.999 | 6.424 | 0.131 | 0.034 |
| VL-1159950 | 1694.584 | 67.782 | 5.499 | 0.143 | 0.040 |
| VL-1989531 | 2294.007 | 84.358 | 5.827 | 0.141 | 0.037 |
| VL-1303160 | 1333.813 | 96.913 | 5.492 | 0.118 | 0.073 |
| VL-517932 | 1370.703 | 38.692 | 5.466 | 0.139 | 0.028 |
| VL-1918993 | 4239.453 | 85.739 | 6.507 | 0.133 | 0.021 |
| VL-2580290 | 1018.490 | 42.296 | 4.216 | 0.115 | 0.041 |
| VL-2363321 | 10382.116 | 82.708 | 5.645 | 0.115 | 0.010 |