

Goal: extract all unigrams from elastic search and dump them into local files.

Step 0: first look on dataset.

Dataset: **spam trec 07**, which is a email messages dataset. Total number of email messages = 75,419.

```
[bingyu@fiji11 trec07p]$ ls
data delay full partial README.txt
[bingyu@fiji11 trec07p]$
```

There are two parts for this dataset:

1) file_name: **./full/index**, for example the first email, which is a **spam**, and the original email message is stored in **./data/inmail.1** file.

```
bingyu@fiji11:/huge1/people/bingyu/c
1:index
1 spam ../data/inmail.1
2 ham ../data/inmail.2
3 spam ../data/inmail.3
4 spam ../data/inmail.4
5 spam ../data/inmail.5
6 spam ../data/inmail.6
7 spam ../data/inmail.7
8 spam ../data/inmail.8
9 spam ../data/inmail.9
10 ham ../data/inmail.10
```

2) email content, for example: **./data/inmail.1**:

```
1:inmail.1
1 From RickyAmes@aol.com Sun Apr 8 13:07:32 2007
2 Return-Path: <RickyAmes@aol.com>
3 Received: from 129.97.78.23 ([211.202.101.74])
4 by speedy.uwaterloo.ca (8.12.8/8.12.5) with SMTP id 138H7G0I003017;
5 Sun, 8 Apr 2007 13:07:21 -0400
6 Received: from 0.144.152.6 by 211.202.101.74; Sun, 08 Apr 2007 19:04:48 +0100
7 Message-ID: <WYADCKPDFWTTWDXNFVUE@yahoo.com>
8 From: "Tomas Jacobs" <RickyAmes@aol.com>
9 Reply-To: "Tomas Jacobs" <RickyAmes@aol.com>
10 To: the00@speedy.uwaterloo.ca
11 Subject: Generic Cialis, branded quality@
12 Date: Sun, 08 Apr 2007 21:00:48 +0300
13 X-Mailer: Microsoft Outlook Express 6.00.2600.0000
14 MIME-Version: 1.0
15 Content-Type: multipart/alternative;
16 boundary="--8896484051606557286"
17 X-Priority: 5
18 X-MSMail-Priority: Normal
19 Status: RO
20 Content-Length: 988
21 Lines: 24
22
23 ----8896484051606557286
24 Content-Type: text/html;
25 Content-Transfer-Encoding: 7Bit
26
27 <html>
28 <body bgcolor="#ffffff">
29 <div style="border-color: #00FFFF; border-right-width: 0px; border-bottom-width: 0px; margin-bottom: 0px;" align="center">
30 <table style="border: 1px; border-style: solid; border-color:#000000;" cellpadding="5" cellspacing="0" bgcolor="#CCFFAA">
31 <tr>
32 <td style="border: 0px; border-bottom: 1px; border-style: solid; border-color:#000000;">
33 <center>
34 Do you feel the pressure to perform and not rising to the occasion??<br>
35 </center>
36 </td></tr><tr>
37 <td bgcolor=#FFFF33 style="border: 0px; border-bottom: 1px; border-style: solid; border-color:#000000;">
38 <center>
39
40 <b><a href='http://excoriationtuh.com/?lzmfnrdkleks'>Try <span>V</span><span>ia</span></span>gr<span>a</span>....</a></b></center>
41 </td></tr><td><center>your anxiety will be a thing of the past and you will<br>
42 be back to your old self.<br>
43 </center></td></tr></table></div></body></html>
44
45
46 ----8896484051606557286--
47
```

Step 1: Index all Spam Trec07 files into elastic search: For example:

```
8 }
9 "hits": {
10   "total": 1,
11   "max_score": 11.776807,
12   "hits": [
13     {
14       "_index": "spam_trec07p",
15       "_type": "document",
16       "_id": "0",
17       "_score": 11.776807,
18       "_source": {
19         "file_name": "inmail.1",
20         "label": "spam",
21         "body": "From RickyAmes@aol.com Sun Apr 8 13:07:32 2007 R
.78.23 ([211.202.101.74]) by speedy.uwaterloo.ca (8.12.8/8.12.5) with SM
2007 13:07:21 -0400 Received: from 0.144.152.6 by 211.202.101.74; Sun, 0
-ID: From: \"Tomas Jacobs\" Reply-To: \"Tomas Jacobs\" To: the00@speedy.u
Cialis, branded quality@ Date: Sun, 08 Apr 2007 21:00:48 +0300 X-Mailer:
.2600.0000 MIME-Version: 1.0 Content-Type: multipart/alternative; bounda
-Priority: 3 X-MSMail-Priority: Normal Status: RO Content-Length: 988 Li
Content-Type: text/html; Content-Transfer-Encoding: 7Bit Do you feel the
to the occasion?? Try Viagra..... your anxiety will be a thing of the pas
self. ----8896484051606557286--".
22       "split": "test"
23     }
24   ]
25 }
26 }
27 }
```

Each document includes: **label**(spam or ham), **body**(email contents), **split**(train or test [4:1 splitting rate.]).

-----> dump data from Step 1 into feature matrix: train.txt, test.txt.

Solve 3 main problems:

- 1) **red arrow**: each line in feature matrix stands for which email from elastic search.
- 2) **blue arrow**: the label here (1, 0) in feature matrix stands for what? (spam or ham).
- 3) **yellow line**: each sparse feature: e.g. 40:2 stands for what? (40 stands for each term index, 2 is the tf for the term in this email.)

```
bingyu@fii:~/IR
1: test.txt 2: train.txt
8 1 81:2 98:1 1574:4 1591:1 1677:4 1717:2 1737:2 1785:2 1810:1 1864:1 3408:1 3620:1 3720:3 4051:1 4066:1 4354:1 4470:1 47597:1
74484:1 80575:1 81025:1 81349:4 84128:1 90275:1 90424:4 94582:1 101022:1 101275:1 104350:7 114492:3 124753:3 187250:1 264119:1
274417:1 276023:1 277809:1 286578:1 337643:3 383721:1 421435:1 472066:4 485661:1 528026:1 528496:1 541488:1 553825:1 603105:1
616326:1 644289:1 645870:1 680184:1 700450:1 720349:1 775552:2 943449:1 1003731:1 1023691:1 1042922:1 1060599:3 1080784:1
1086336:1 1102329:4 1102332:1 1118168:1 1158734:1 1265155:1 1398141:1 1436597:1 1493639:2 1584742:1 1593879:1 1684847:1 1732186:1
1786498:1 1896588:1 2025331:1 2129918:2 2222785:1 2222821:1 2233160:1 2295776:5 2330276:1 2345031:2 2387452:1 2429203:1
2434076:1 2446132:2 2457272:1 2484842:5 2517385:1 2574110:1 2590900:1 2597504:1 2604615:1 2615397:1 2630876:1 2656273:2 2660088:1
1 693700:1 2706067:1 2740028:1 2741027:2 2835344:1 2920890:1 2972516:1
9 0 41:2 81:1 87:1 98:1 99:1 111:2 114:1 1574:3 1586:3 1593:1 1661:3 1677:5 1796:5 1854:1 3284:2 3496:1 3720:4 4073:1 4098:2 4558:2
36600:1 80477:2 81349:5 83614:2 84119:2 84977:1 88478:1 91571:2 96163:2 104926:2 107227:4 111716:1 114492:5 115768:2 116042:2
117869:1 120556:2 120575:1 121109:1 124753:7 141758:4 151764:2 168878:2 248627:2 253656:2 276023:3 277809:2 337643:3 379819:1
413261:1 415961:1 419900:2 528026:1 547544:2 578570:2 590530:1 642136:1 645870:2 651293:2 700450:1 720349:1 744915:1 758434:2
775552:4 794839:1 806652:1 837824:2 867004:1 1039599:1 1057701:3 1060599:1 1102329:5 1105762:1 1159715:2 1187645:2 1202769:1
1243380:1 1253477:1 1261560:2 1271849:1 1303490:3 1315927:1 1319485:1 1344269:1 1345707:1 1345754:1 1398141:1 1398518:4 1418578:1
3 1423134:1 1458609:2 1506819:1 1511223:1 1512507:1 1526482:3 1530157:2 1618616:1 1658456:1 1663722:2 1747123:1 1774074:1
1843285:1 1919326:2 1920720:2 1938019:1 1955359:1 2009693:1 2025331:1 2053188:1 2113299:1 2171113:1 2174144:2 2180123:1 2185503:1
1 2207221:1 2215566:3 2218441:1 2222785:1 2222821:1 2228009:1 2231624:1 2233160:1 2236029:2 2247275:4 2259232:5 2272163:2
2295776:9 2309279:1 2309820:1 2311217:2 2318176:1 2323369:1 2333211:1 2344788:1 2398171:1 2429203:1 2436178:1 2446132:3 2449149:2
2 2452312:2 2517291:1 2517385:1 2530193:1 2550310:1 2556030:1 2561752:2 2564101:1 2566588:2 2587904:2 2592987:1 2593145:2
2615397:2 2617577:7 2630876:3 2643988:1 2656273:6 2667023:1 2682851:2 2701278:2 2740045:4 2742798:1 2813943:2 2816067:1 2824920:2
2 854190:1 2871526:1 2889014:2
10 1 40:2 43:1 53:1 87:5 98:1 111:7 1538:2 1557:3 1573:1 1574:4 1588:1 1661:1 1677:9 1713:1 1717:3 1725:1 1737:1 1796:5 1802:1
1864:1 3223:2 3242:1 3620:1 3720:7 3750:1 4002:1 4270:2 4470:1 36600:1 36735:1 37114:1 38046:1 45087:1 80477:1 81025:1 81349:7
86776:1 86780:1 88971:5 91571:1 91576:1 92187:1 94088:4 102478:2 104880:1 108363:1 109017:1 114492:9 124753:1 135108:2 151474:1
155760:1 168878:1 169833:1 218898:1 259967:1 276023:4 277809:3 335151:1 337643:5 360271:1 380204:4 412046:1 413261:1 419950:1
436149:2 443084:17 475713:1 479106:1 493076:1 526430:1 528026:1 528159:1 558379:2 565467:1 616326:1 645870:1 655105:1 688166:4
700450:1 720349:1 736790:1 764211:1 775552:1 794839:1 800048:1 895274:1 917673:2 917721:1 954963:1 962433:1 967613:7 1028181:2
1042922:1 1043716:1 1060599:3 1062095:3 1083405:1 1091004:1 1091841:1 1102329:8 1104059:1 1125292:2 1187645:1 1202070:3 1217578:2
2 1243380:1 1271849:1 1320198:2 1354449:2 1390917:6 1398141:1 1421535:1 1458818:1 1474127:1 1493639:1 1515039:1 1521943:1
1593879:1 1596702:2 1599988:1 1603114:1 1611626:1 1641187:1 1660265:2 1663722:2 1687916:9 1689165:1 1713084:1 1719880:1 1719939:1
1 1737414:1 1768065:4 1773900:1 1807913:1 1820244:1 1822710:1 1831051:3 1912418:1 1952617:1 2003458:1 2025331:2 2075186:1
2075335:1 2093261:1 2110571:1 2163553:3 2218620:2 2222702:2 2222785:1 2222821:1 2223672:1 2223676:1 2223677:1 2233160:1 2233165:1
1 2244672:1 2274647:1 2276378:2 2282775:2 2295776:9 2323326:1 2329520:1 2333211:1 2333213:1 2333217:1 2345031:5 2357719:1
2392092:1 2429203:1 2429434:11 2433901:1 2446132:3 2452312:1 2457272:1 2477401:1 2495887:1 2517385:1 2527979:1 2532005:3
2550310:3 2566588:1 2590579:1 2590900:1 2596786:1 2597504:1 2605954:1 2610056:1 2615397:1 2617090:1 2630067:1 2630876:1 2656273:3
3 2682851:4 2694275:2 2811483:1 2813943:2 2831035:1 2870459:1 2870712:1 2879336:1 3002798:1 3016420:3 3036790:8 3066164:1
11 0 88:2 98:1 111:6 1538:1 1551:1 1556:2 1574:4 1663:3 1677:4 1713:2 1802:4 3223:2 3720:4 4470:1 35661:1 36600:1 36689:2 36894:2
37001:2 81025:1 81349:4 86387:2 87965:1 96185:1 97604:1 99216:1 113579:1 114492:4 120897:1 155220:1 168878:1 276023:2 277809:1
288027:1 292318:3 319902:1 337643:2 390070:1 413261:1 454023:4 524883:1 528026:2 547815:2 565467:1 590530:1 616326:1 642667:1
645870:1 700450:1 714605:1 720349:1 747215:1 755709:2 775552:1 794839:1 807950:2 844789:1 907937:1 909210:1 919095:1 1056227:1
1060599:3 1082492:1 1102329:4 1105174:1 1118813:1 1128314:1 1184889:1 1187645:1 1197073:1 1229354:1 1242672:1 1243380:1 1289310:1
1 1312764:1 1319485:1 1345707:1 1351503:1 1368732:1 1422027:1 1496331:1 1618616:1 1635533:1 1663722:2 1710589:1 1798366:1
1876459:1 1925139:3 2009693:1 2025331:1 2070841:1 2120848:1 2157013:1 2163801:1 2171605:1 2222785:1 2222821:1 2223672:2 2233160:1
1 2239282:2 2295776:4 2344788:1 2345031:1 2367302:1 2387972:1 2429203:2 2432077:1 2452312:1 2472077:1 2477401:1 2507798:1
2517385:1 2525218:4 2530035:1 2550310:3 2556030:1 2588835:1 2593670:1 2602472:1 2615397:1 2615808:1 2630876:2 2651104:1 2656273:4
4 2665572:2 2684264:4 2689563:1 2794839:2 2801836:2 2813943:2 2831899:1 2854190:1 2871526:1 2890368:1 2923015:1 2966455:1
3040685:1
```

Step 2: get ids for train and test:

```

9   "hits": {
10  "total": 1,
11  "max_score": 11.776807,
12  "hits": [
13    {
14      "_index": "spam_trec07p",
15      "_type": "document",
16      "_id": "0",
17      "_score": 11.776807,
18      "_source": {
19        "file_name": "inmail.1",
20        "label": "spam",
21        "body": "From RickyAmes@aol.com Sun Apr 8 13:07:32 2007 Re
.78.23 ([211.202.101.74]) by speedy.uwaterloo.ca (8.12.8/8.12.5) with SMT
2007 13:07:21 -0400 Received: from 0.144.152.6 by 211.202.101.74; Sun, 08
-ID: From: \"Tomas Jacobs\" Reply-To: \"Tomas Jacobs\" To: the00@speedy.u
Cialis, branded quality@ Date: Sun, 08 Apr 2007 21:00:48 +0300 X-Mailer:
.2600.0000 MIME-Version: 1.0 Content-Type: multipart/alternative; boundar
-Priority: 3 X-MSMail-Priority: Normal Status: R0 Content-Length: 988 Lin
Content-Type: text/html; Content-Transfer-Encoding: 7Bit Do you feel the
to the occasion?? Try Viagra.... your anxiety will be a thing of the pas
self. ----8896484051606557286--",
22      "split": "test"
23    }
24  ]
25 }
26 }
27 }

```

train_list = [id_1,id_3...]

test_list = [id_0,....]

Step 3: dump the train and test ids from elastic search into the local files:

For example, named: **train_ids_list.txt**, **test_ids_list.txt**.

```

1:train_ids_list.txt 2:test_ids_list.txt
7015 7014 9835
7016 7015 0
7017 7016 5
7018 7017 10
7019 7018 15
7020 7019 20
7021 7020 25
7022 7021 30
7023 7022 35
7024 7023 40
7025 7024 45
7026 7025 50
7027 7026 55
7028 7027 60
7029 7028 65
7030 7029 70
7031 7030 75

```

7015 is the line number of feature_matrix - 1.

0 is the id for inmail.1 in elastic search.

Step 4: dump labels(spam & ham) from elastic search into 0 & 1 in our feature

matrix.

1:label_list.txt

```
1 ham 1
2 spam 0
```

Step 5: build feature index list from training set.

This step is to map each term into unique index number, e.g.

```
bingyu@fiji11:~/IR
1:feature_list.txt
1 |jsny 1
2 3forof4ivqewo6f3fb4bo 2
3 2fconfig_mk.pm 3
4 edevmmt 4
5 arwjbvgsju8z6bswin 5
6 13jdbxpltewrwelerxfqtyh9gemhywriyqod1c0bs2 6
7 15rgso19029567 7
8 15rgso19029568 8
9 wu9gvji 9
10 0zk0pkcpb8y6qdkj7lshojuvzgb0re6j2vqhqblsui 10
11 37.144.25.232 11
12 gjrssi 12
13 Symuqjay9jboa57cgzyy2zntujyeqeor6h6jri0 13
14 oosi24ave7qozzpg2pwmnkzxsns1r9lmsdksoaeq1v7ogj9tppwampsnv 14
15 jaglnosedu8xgp0anqlchgoyeerusnfbf 15
16 afhwvdlbzuvobfgus3qc0doln3jv0up 16
17 anghcabjdclhaec3agloadhxc5tgalzibtmayabmabmbtspaghoayvrajtibmlnanxfaab 17
18 ssbrthxbba1aoue2wj4ssgzmgkoyoizr6w 18
19 i2sdwi crf5dkncmlhlf8pyteh11wf2ckaaaaa 19
20 verg9ap1ibnew4ej3rfarueqvgv9ibf 20
21 hajkmg 21
22 109.142.52.134 22
23 w6sylqzoab6mxtuomfmzrw 23
24 ac_cv_func_closefrom 24
25 alds2vu79y3apis1m0131aahgh0eaggr1su2xjpp 25
26 kcqqgfvcecgel8f 26
27 aelfhgtr 27
28 x8jrkdljjc3cutje85a3nyrtbivxy14vmnewfquoi 28
29 38d74b6482 29
30 outz3xx0efoutgouogdut 30
31 uglxmrcbruc6bxximbx56otvrrbcwztcpb2w3w1 31
32 4f6r78pxustwgdze7c8agmc78cdc0ztypg2to 32
33 wmh7o 33
34 18.12.2006 34
35 lgid5rcrhabmlcqbyagjfd7 35
36 4644832a 36
37 yrdkjweca6eywhdqwy6nmyzjbh9aqcw9eacsh6iuuarjw 37
38 15e2gv19015353 38
39 0 39
40 1 40
41 2 41
```

Step 6: dump train/test into feature matrix.


```
bingyu@fiji11:~/IR
1:test.txt
1 40:2 53:1 87:5 98:1 111:2 1538:2 1574:5 1586:1 1645:1 1651:1 1661:1 1663:1 1666:1 1669:1 1671:1 1677:6 1717:6 1737:1 1942:1
3223:3 3620:1 3720:6 4002:1 4270:1 4470:1 36600:1 36786:1 45087:1 80477:1 81025:1 81349:6 86776:1 86780:1 88971:4 91571:1 91576:
1 92030:1 92187:1 94088:4 109017:1 114492:6 124753:1 128370:1 135108:2 151474:1 155760:1 168878:1 213777:2 218898:1 259967:1
276023:4 277809:2 297753:3 337643:4 369727:1 380204:3 443084:13 526430:1 528026:1 528159:1 541488:1 550186:1 558379:2 616326:1
632556:1 645870:1 655105:2 688166:3 700450:1 720349:1 736790:1 764211:1 775552:1 800048:1 836812:1 917673:1 962433:1 967613:6
1042922:1 1043716:1 1060599:4 1079351:1 1091004:2 1091841:1 1102329:6 1104059:1 1125292:2 1187645:1 1202070:3 1217578:1 1271849:
1 1296639:1 1320198:1 1354449:2 1390917:5 1398141:1 1458818:1 1493639:1 1515039:2 1521943:1 1593879:2 1596702:2 1611626:1
1660265:2 1687916:7 1713084:1 1768065:3 1807913:1 1820244:1 1822710:1 1912418:1 1952617:1 2003458:1 2025331:1 2075335:1 2093261:
1 2109833:1 2110571:1 2150330:1 2218620:1 2222702:1 2222785:2 2222821:1 2223726:1 2233160:1 2244672:1 2274647:1 2276378:2
2295776:8 2300277:1 2323326:1 2329520:1 2333211:2 2333213:1 2333217:1 2345031:4 2357719:1 2363325:1 2392092:1 2429203:1 2429434:
10 2433901:1 2446132:2 2452312:1 2457272:1 2477401:1 2517385:1 2527979:1 2532005:2 2550310:2 2566588:1 2574732:1 2590579:1
2596786:1 2605954:2 2610056:1 2615397:1 2617090:1 2630067:1 2630876:1 2634698:1 2649391:1 2656273:3 2682851:4 2694275:2 2728760:
1 2831035:1 2870712:1 2879336:1 2906393:1 3002798:1 3016420:2 3036790:7 3066164:1
2 40:2 43:1 81:1 86:1 98:1 103:1 111:2 1574:2 1588:3 1675:3 1677:2 1713:3 1785:3 1793:1 1796:3 1816:1 2095:2 3225:2 3463:1 3720:
4 4015:1 4339:2 4558:4 36600:1 36897:2 37054:2 37111:2 81025:4 81166:2 81349:5 83943:2 84119:4 89702:2 96163:2 101325:2 103111:
3 107227:2 113165:2 114492:2 115004:2 120556:6 168878:2 248627:2 276023:3 277809:2 283971:1 337643:3 413261:1 435848:2 528026:1
534714:2 547544:2 578570:6 590530:1 598185:1 645870:2 700450:1 720349:1 775552:4 794839:1 806652:6 929728:1 1039599:1 1058350:3
1060599:1 1102329:5 1104248:1 1138130:2 1181210:1 1187645:2 1243380:1 1261560:2 1271849:1 1303490:2 1312764:1 1314983:1 1319485:
1 1345707:1 1398141:1 1436629:1 1454903:1 1506819:1 1526482:3 1530157:2 1618616:1 1663722:2 1704448:2 1712198:1 1713269:1
1773516:1 1803907:1 1820244:2 2009693:1 2022572:1 2025331:1 2075119:1 2082993:2 2139801:2 2163623:2 2185503:3 2222785:1 2222821:
1 2231624:1 2233160:1 2236029:2 2247275:4 2249547:2 2259232:3 2277905:1 2295776:5 2309820:1 2316283:2 2319444:2 2330276:2
2333211:1 2333282:2 2336049:4 2344788:1 2346365:1 2369532:1 2429203:1 2433750:3 2452312:2 2461706:2 2503357:2 2517291:1 2517385:
1 2550310:1 2556030:1 2561752:2 2566882:2 2587682:2 2590830:2 2598814:2 2610056:2 2614767:2 2615397:2 2617190:1 2617577:4
2630876:3 2643988:1 2656273:6 2682996:2 2688309:2 2705562:1 2778054:1 2813943:2 2824920:2 2854190:1 2865975:3 2871526:1 3066164:
2
```

index:value:

index is based on feature_list.txt

value is the tf count, you can use tf-idf or other score from elastic search.

Step 7: Running lib linear classification on train.txt, and test.txt

```
[bingyu@fiji11 liblinear-1.95]$ ./train ~/IR/train.txt ~/IR/linear.model
.....
optimization finished, #iter = 1000

WARNING: reaching max number of iterations
Using -s 2 may be faster (also see FAQ)

Objective value = -1.056825
nSV = 1171
[bingyu@fiji11 liblinear-1.95]$ ./predict ~/IR/test.txt ~/IR/linear.model ~/IR/linear.predict
Accuracy = 99.8674% (15064/15084)
```