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
```

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:

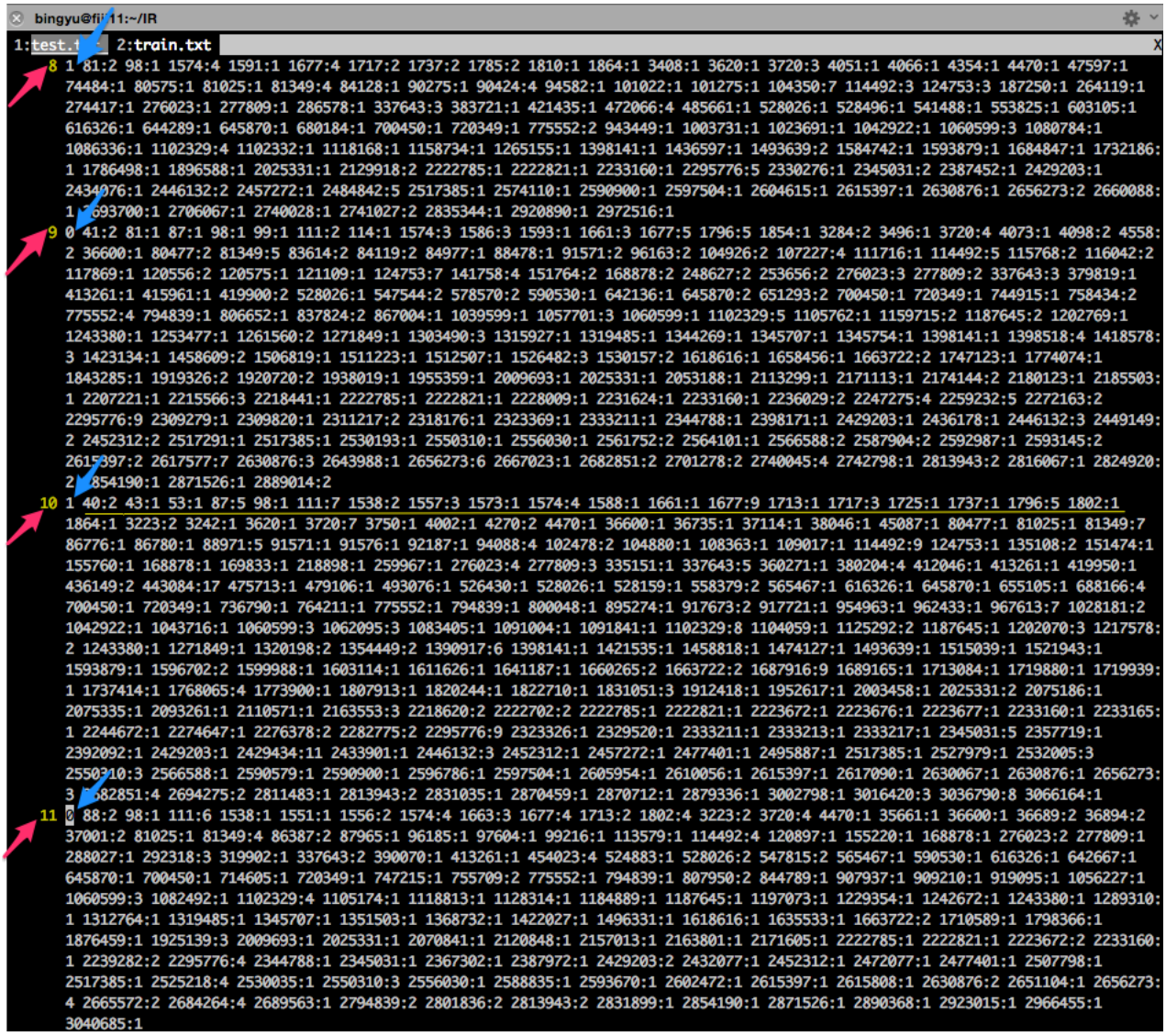
```
1:inmail.1
  1 From RickyAmes@aol.com  Sun Apr  8 13:07:32 2007
  2 Return-Path: <RickyAmes@aol.com>
  3 Received: from 129.97.78.23 ([211.202.101.74])
  4     by speedy.uwaterloo.ca (8.12.8/8.12.5) with SMTP id l38H7G0I003017;
  5     Sun, 8 Apr 2007 13:07:21 -0400
  6 Received: from 0.144.152.6 by 211.202.101.74; Sun, 08 Apr 2007 19:04:48 +0100
  7 Message-ID: <WYADCKPDFWWTWTXNFVUE@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: 3
 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.
 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:


```
    },
    "hits": {
        "total": 1,
        "max_score": 11.776807,
        "hits": [
            {
                "_index": "spam_trec07p",
                "_type": "document",
                "_id": "0",
                "_score": 11.776807,
                "_source": {
                    "file_name": "inmail.1",
                    "label": "spam",
                    "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.
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 pa
self. ----8896484051606557286--",
                    "split": "test"
                }
            }
        ]
    }
}
```


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.)



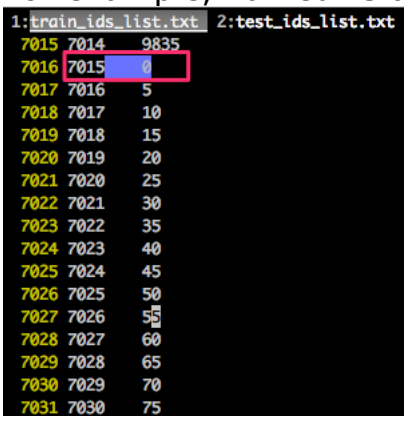
Step 2: get ids for train and test:

```
    "hits": {
      "total": 1,
      "max_score": 11.776807,
      "hits": [
        {
          "_index": "spam_trec07p",
          "_type": "document",
          "_id": "0",
          "_score": 11.776807,
          "_source": {
            "file_name": "inmail.1",
            "label": "spam",
            "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: RO 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--",
            "split": "test"
          }
        }
      ]
    }
}
```


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.**



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.

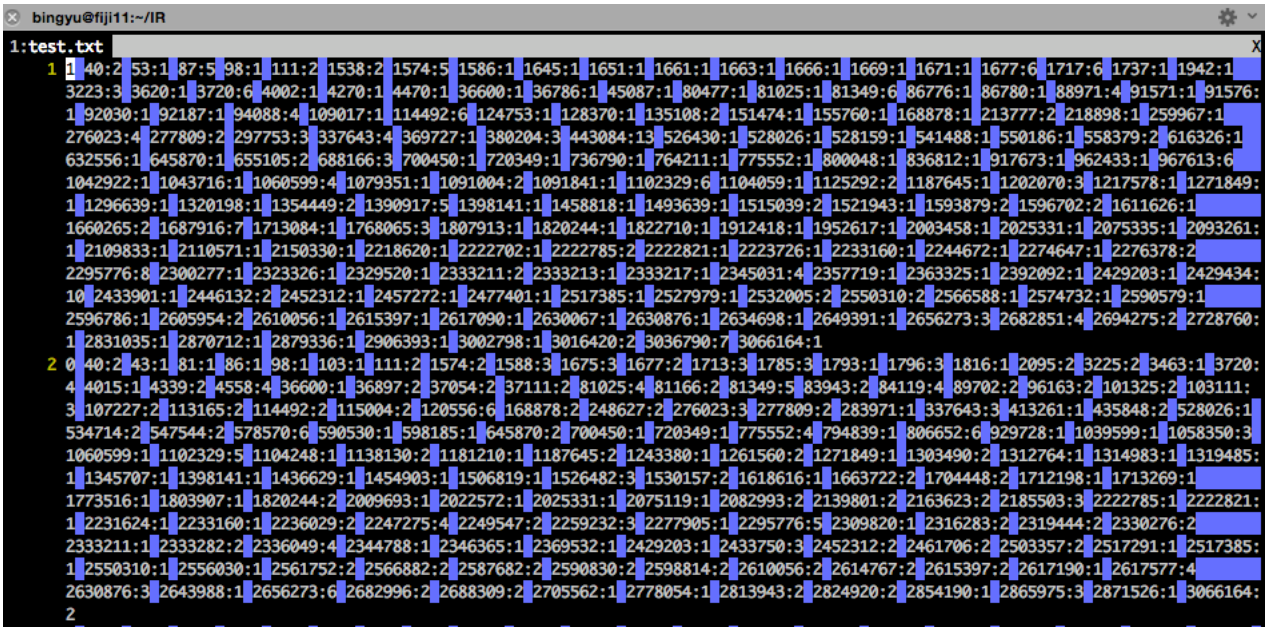
```
1:label_list.txt
 1 ham 1
 2 spam    0
```

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 njsny   1
     2 3forof4ivqewo6f3fb4bo   2
     3 2fconfig_mk.pm  3
     4 edevmmt 4
     5 arwjbvgsju8z6bswin  5
     6 13jdbxpltewrwelerxfqtyh9gemhywriyqod1c0bs2  6
     7 l5rgsol9029567  7
     8 l5rgsol9029568  8
     9 wu9gvji 9
    10 0zk0pkcpb8y6qdkj7lshojujvzgb0re6j2vqhqblsui 10
    11 37.144.25.232   11
    12 gjrssji 12
    13 5ymuqjqy9jboa57cgzyy2zntujyeqeor6h6jri0 13
    14 oosi24ave7qozzpg2pwmnkzxhsns1r9lmxsdksoaueq1v7ogj9tppwampsnv    14
    15 jaglnosedu8xgp0anqlchgoyeerusnfbf   15
    16 afhwvdnlbzuvobfgus3qc0doln3jv0up    16
    17 anghcabjdclhaec3agloahyxc5tgalzibtmyaabmabrmbtpsaghoayvrajtibmlnanxfaab 17
    18 ssbrthxbbq1aoue2wjd4ssgzmgkoyojzrg6w    18
    19 i2sdwicrf5dknamihlf8pytehl1wf2ckaaaaa   19
    20 verg9qp1ibnew4ej3rfqrueqvgv9ibf 20
    21 hajkmg  21
    22 109.142.52.134  22
    23 w6sylqzoab6mxetuomfmzrw 23
    24 ac_cv_func_closefrom    24
    25 alds2vu79y3apislm0l3laahgh0eqggrlsu2xjjp    25
    26 kcqqgfvcecgel8f 26
    27 aelfhgtr    27
    28 x8jrkdljjc3cutje85a3nyrtbivxy14vmnewfqruoi  28
    29 38d74b6482  29
    30 outz3xx0efoutgouogdut   30
    31 uglxmrcbruwc6bxximbx56otvwrrbcwztcpb2w3w1   31
    32 4f6r78pxustwwgdze7c8agmc78cdc0ztypg2to  32
    33 wmh7o   33
    34 18.12.2006  34
    35 lgid5rcrihabmlcqbyqgjfd7    35
    36 4644832a    36
    37 yrdkjweca6eywhdqywy6nmyzjbh9aqcw9eacsh6iuuarjw  37
    38 l5e2gvl9015353  38
    39 0   39
    40 1   40
    41 2   41
```

Step 6: dump train/test into feature matrix.



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