```
>>> df.select(split(col("email"),'@')[1].alias("domain"),"count").groupBy("domai
n").count().orderBy("count",ascending=False).show(50)
+--------------------+-----+

|              domain|count|
+--------------------+-----+
|           gmail.com| 7223|
|          apache.org|  361|
|          redhat.com|  241|
|         hotmail.com|  148|
|             163.com|  143|
|          debian.org|  115|
|    googlegroups.com|  105|
|           yahoo.com|  104|
|             gnu.org|   97|
|              qq.com|   92|
|         outlook.com|   88|
|          google.com|   87|
|         freebsd.org|   83|
|           intel.com|   61|
|              gmx.de|   61|
|      protonmail.com|   49|
|          oracle.com|   46|
|      googlemail.com|   37|
|             web.de|   36|
|        mirantis.com|   35|
|            gmx.net|   34|
|            126.com|   30|
|            free.fr|   29|
|    openjdk.java.net|   27|
|           posteo.de|   26|
|       linux.ibm.com|   25|
|          huawei.com|   25|
|          python.org|   25|
|           yandex.ru|   24|
|         comcast.net|   23|
|          icloud.com|   23|
|          kernel.org|   23|
|         t-online.de|   22|
|       ververica.com|   22|
|             gmx.com|   22|
|    lists.debian.org|   21|
|           pobox.com|   21|
|       canonical.com|   21|
|             arm.com|   20|
|            live.com|   20|
|          riseup.net|   20|
|         foxmail.com|   19|
|syzkaller.appspot...|   18|
|          vmware.com|   17|
|        ericsson.com|   17|
|     linux.intel.com|   17|
|        fastmail.com|   16|
|           nokia.com|   16|
|             isc.org|   16|
|          golang.org|   16|
+--------------------+-----+
only showing top 50 rows

>>> df.select(reverse(split(col("email"),'\.'))[0].alias("TLD"),"count").groupBy
```

```
("TLD").count().orderBy("count",ascending=False).show(50)
+----+-----+
| TLD|count|
+----+-----+
| com|11337|
| org| 1812|
| net|  643|
|  de|  630|
| edu|  247|
|  uk|  169|
|  fr|  166|
|  io|  126|
|  it|  102|
|  ca|   96|
|  ch|   90|
|  ru|   78|
|  au|   76|
|  nl|   71|
|  eu|   64|
|  at|   62|
|  se|   51|
|  pl|   47|
|  jp|   47|
|  in|   46|
|  us|   44|
|  cn|   42|
|  cz|   42|
|  br|   38|
|  nz|   33|
|  ai|   30|
|  be|   30|
| gov|   27|
|  co|   27|
|  me|   25|
|  es|   25|
|  fi|   24|
|name|   24|
|  no|   23|
|  dk|   22|
|info|   18|
|  cc|   17|
| dev|   13|
|  gr|   13|
| biz|   12|
|  ie|   12|
| xyz|   11|
|  tr|   11|
|  pt|   10|
|  hu|   10|
|  fm|   10|
|  sk|   10|
|  ro|    9|
|  is|    8|
|  mx|    8|
+----+-----+
only showing top 50 rows

>>> df.count()
16933
```

```
>>> df2.select("*").groupBy("word").count().orderBy("count",ascending=False).sho
w(100)
+-------------+-----+

|         word|count|
+-------------+-----+
|      message|24346|
|         send|24160|
|  unsubscribe|19066|
|         list|17044|
|         file|16183|
|         2021|15473|
|          jan|15255|
|         2022|14860|
|      mailing|14544|
|         mail|13304|
|        email|12632|
|           pm|11732|
|         code|11552|
|        group|11245|
|         data|11148|
|      problem|10853|
|       return|10772|
|        flink|10712|
|      version|10269|
|        error|10224|
|         time|10137|
|     received|10133|
|          set| 9880|
|       100644| 9853|
|   discussion| 9796|
|       google| 9735|
|         diff| 9708|
|         view| 9667|
|        visit| 9542|
|          add| 9499|
|       groups| 9491|
|   subscribed| 9390|
|          web| 9383|
|    receiving| 9182|
|      running| 8977|
|       emails| 8869|
|         case| 8860|
|          feb| 8652|
|         2020| 8446|
|         work| 8237|
|         test| 8207|
|        files| 8036|
|          job| 7817|
|         info| 7778|
|       apache| 7620|
|        check| 7551|
|         user| 7516|
|        issue| 7492|
|       change| 7354|
|      support| 7350|
|      address| 7305|
|         read| 7267|
|       struct| 7229|
|         type| 7210|
```

```
|       server| 7034|
|      request| 6781|
|          int| 6736|
|     function| 6692|
|          dec| 6679|
|       memory| 6130|
|     specific| 6063|
|         long| 5950|
|       python| 5887|
|          log| 5828|
|       create| 5818|
|         good| 5762|
|      release| 5625|
|        start| 5563|
|      package| 5447|
|           00| 5364|
|     messages| 5236|
|      default| 5193|
|       device| 5181|
|       source| 5162|
|       static| 5139|
|         void| 4979|
|      command| 4976|
|       failed| 4909|
|       cygwin| 4818|
|        build| 4813|
|configuration| 4649|
|       number| 4648|
|          nov| 4614|
|      contact| 4604|
|          bug| 4441|
|           16| 4426|
|          url| 4377|
|       update| 4344|
|      provide| 4328|
|      current| 4274|
|        guest| 4272|
|          git| 4197|
|        write| 4196|
|      working| 4193|
|          bit| 4154|
|       output| 4122|
|        table| 4113|
|          kvm| 3956|
|       client| 3934|
|        works| 3897|
+-------------+-----+
only showing top 100 rows

>>> df2.count()
3640497
```