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

|              domain|count|
+--------------------+-----+
|           gmail.com| 5448|
|          apache.org|  217|
|          redhat.com|  184|
|             163.com|  124|
|         hotmail.com|  106|
|    googlegroups.com|   92|
|             gnu.org|   89|
|         freebsd.org|   77|
|           yahoo.com|   73|
|              qq.com|   71|
|         outlook.com|   66|
|          google.com|   65|
|          debian.org|   63|
|           intel.com|   56|
|              gmx.de|   39|
|        mirantis.com|   35|
|          oracle.com|   33|
|    openjdk.java.net|   27|
|             web.de|   26|
|      googlemail.com|   25|
|             126.com|   24|
|          python.org|   24|
|             gmx.net|   23|
|           yandex.ru|   23|
|          huawei.com|   22|
|      protonmail.com|   21|
|       ververica.com|   21|
|    lists.debian.org|   21|
|       linux.ibm.com|   20|
|             arm.com|   19|
|         t-online.de|   19|
|             free.fr|   18|
|         comcast.net|   18|
|       canonical.com|   17|
|          icloud.com|   17|
|     linux.intel.com|   17|
|           pobox.com|   16|
|          kernel.org|   16|
|          golang.org|   16|
|syzkaller.appspot...|   15|
|           posteo.de|   15|
|             gmx.com|   15|
|          cisco.com|   14|
|          vmware.com|   14|
|       openstack.org|   14|
|        ericsson.com|   14|
|          nokia.com|   13|
|            hpe.com|   13|
|          linaro.org|   13|
|          cygwin.com|   13|
+--------------------+-----+
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| 8616|
| org| 1395|
| net|  462|
|  de|  428|
| edu|  154|
|  uk|  121|
|  io|  101|
|  fr|   96|
|  ca|   71|
|  it|   69|
|  ru|   66|
|  au|   58|
|  ch|   57|
|  nl|   52|
|  eu|   46|
|  at|   45|
|  jp|   40|
|  in|   39|
|  cn|   37|
|  se|   37|
|  us|   35|
|  pl|   32|
|  cz|   31|
|  br|   30|
|  nz|   21|
|  ai|   21|
|  co|   20|
|  es|   20|
|  me|   19|
|  be|   19|
|name|   18|
|  fi|   16|
|  no|   15|
|  cc|   13|
| gov|   13|
|  dk|   13|
|info|   12|
| dev|   10|
|  gr|   10|
|  ie|   10|
|  sk|    9|
| biz|    9|
| xyz|    9|
|  hu|    9|
|  fm|    8|
|  tr|    7|
|  pt|    7|
|  kr|    7|
|  mx|    7|
|  za|    6|
+----+-----+
only showing top 50 rows

>>> df.count()
12730
```

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

|         word|count|
+-------------+-----+
|         send|20987|
|      message|17516|
|  unsubscribe|15541|
|         2021|15221|
|         list|13017|
|      mailing|12402|
|         mail|11647|
|         file|11133|
|        flink|10114|
|        email| 9919|
|           pm| 9248|
|        group| 8865|
|      problem| 8853|
|         code| 8659|
|         data| 8657|
|         2020| 8398|
|     received| 8246|
|       google| 7921|
|   discussion| 7920|
|          jan| 7893|
|       groups| 7730|
|   subscribed| 7726|
|        visit| 7716|
|       return| 7685|
|         view| 7623|
|         time| 7541|
|          web| 7514|
|    receiving| 7413|
|       emails| 7213|
|        error| 7201|
|          job| 7002|
|      version| 6881|
|          set| 6801|
|          dec| 6599|
|         info| 6549|
|       100644| 6530|
|         2022| 6351|
|         diff| 6323|
|      running| 6317|
|         case| 6223|
|      address| 6154|
|          add| 6072|
|         work| 5997|
|         type| 5840|
|        files| 5545|
|       server| 5311|
|         user| 5243|
|       python| 5187|
|       struct| 5131|
|         read| 5023|
|       change| 4936|
|        issue| 4734|
|      support| 4694|
|          int| 4670|
```

```
|          nov| 4544|
|     messages| 4530|
|       cygwin| 4471|
|       memory| 4383|
|        start| 4376|
|     function| 4341|
|           00| 4335|
|       create| 4274|
|        check| 4212|
|         test| 4078|
|         good| 4066|
|      command| 3946|
|         long| 3907|
|       source| 3855|
|       apache| 3807|
|      default| 3711|
|      request| 3698|
|      release| 3597|
|      package| 3508|
|       static| 3490|
|         void| 3370|
|configuration| 3356|
|       number| 3339|
|   connection| 3280|
|       device| 3218|
|        write| 3190|
|      provide| 3162|
|        class| 3153|
|     specific| 3149|
|      freebsd| 3063|
|       client| 3057|
|      working| 3055|
|        table| 3022|
|          log| 3020|
|       output| 2986|
|      current| 2968|
|        build| 2965|
|        works| 2950|
|       failed| 2928|
|      process| 2923|
|         task| 2913|
|          lot| 2903|
|          bit| 2895|
|         port| 2888|
|       import| 2886|
|      windows| 2877|
+-------------+-----+
only showing top 100 rows

>>> df2.count()
2542744
```