```
>>> df.select(split(col("email"),'@')[1].alias("domain")).groupBy("domain").coun
t().withColumn("rank",rank().over(windowSpec)).show(50)
+--------------------+-----+----+

|              domain|count|rank|
+--------------------+-----+----+
|           gmail.com| 7659|   1|
|          apache.org|  395|   2|
|          redhat.com|  252|   3|
|         hotmail.com|  156|   4|
|             163.com|  145|   5|
|          debian.org|  126|   6|
|           yahoo.com|  112|   7|
|    googlegroups.com|  105|   8|
|             gnu.org|   98|   9|
|              qq.com|   98|   9|
|          google.com|   95|  11|
|         outlook.com|   92|  12|
|         freebsd.org|   83|  13|
|      protonmail.com|   63|  14|
|           intel.com|   62|  15|
|              gmx.de|   62|  15|
|          oracle.com|   47|  17|
|              web.de|   39|  18|
|      googlemail.com|   38|  19|
|        mirantis.com|   35|  20|
|             gmx.net|   35|  20|
|             126.com|   32|  22|
|             free.fr|   32|  22|
|           posteo.de|   28|  24|
|    openjdk.java.net|   27|  25|
|          kernel.org|   26|  26|
|          icloud.com|   26|  26|
|       linux.ibm.com|   26|  26|
|          huawei.com|   25|  29|
|          python.org|   25|  29|
|         comcast.net|   24|  31|
|           yandex.ru|   24|  31|
|             gmx.com|   23|  33|
|         t-online.de|   23|  33|
|       ververica.com|   22|  35|
|syzkaller.appspot...|   22|  35|
|          riseup.net|   21|  37|
|    lists.debian.org|   21|  37|
|       canonical.com|   21|  37|
|             arm.com|   21|  37|
|           pobox.com|   21|  37|
|            live.com|   21|  37|
|         foxmail.com|   20|  43|
|          vmware.com|   19|  44|
|           nokia.com|   17|  45|
|          amazon.com|   17|  45|
|             cern.ch|   17|  45|
|        ericsson.com|   17|  45|
|     linux.intel.com|   17|  45|
|        fastmail.com|   16|  50|
+--------------------+-----+----+
only showing top 50 rows

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

```
count().withColumn("rank",rank().over(windowSpec)).show(50)
+----+-----+----+
| TLD|count|rank|
+----+-----+----+
| com|12021|   1|
| org| 1905|   2|
| net|  686|   3|
|  de|  680|   4|
| edu|  265|   5|
|  fr|  182|   6|
|  uk|  181|   7|
|  io|  143|   8|
|  it|  118|   9|
|  ca|  109|  10|
|  ch|   97|  11|
|  ru|   84|  12|
|  au|   78|  13|
|  nl|   75|  14|
|  eu|   68|  15|
|  at|   66|  16|
|  se|   54|  17|
|  jp|   49|  18|
|  pl|   49|  18|
|  us|   47|  20|
|  in|   46|  21|
|  cn|   44|  22|
|  cz|   42|  23|
|  br|   40|  24|
|  nz|   36|  25|
|  be|   34|  26|
|  ai|   31|  27|
|  es|   30|  28|
| gov|   29|  29|
|  co|   28|  30|
|  me|   26|  31|
|  fi|   26|  31|
|name|   25|  33|
|  no|   25|  33|
|  dk|   23|  35|
|  cc|   19|  36|
|info|   18|  37|
| dev|   14|  38|
|  ie|   14|  38|
| biz|   14|  38|
|  gr|   13|  41|
|  tr|   12|  42|
| xyz|   12|  42|
|  pt|   11|  44|
|  ro|   10|  45|
|  sk|   10|  45|
|  fm|   10|  45|
|  hu|   10|  45|
|  sg|    8|  49|
|  is|    8|  49|
+----+-----+----+
only showing top 50 rows

>>> df.count()
17998
```

```
>>> df2.groupBy("word").count().withColumn("rank",rank().over(windowSpec)).show(
100)
+-------------+-----+----+

|         word|count|rank|
+-------------+-----+----+
|      message|26486|   1|
|         send|25081|   2|
|  unsubscribe|20100|   3|
|         list|18398|   4|
|         file|17989|   5|
|         2022|17366|   6|
|         2021|15512|   7|
|          jan|15323|   8|
|      mailing|15188|   9|
|         mail|13740|  10|
|        email|13455|  11|
|           pm|12530|  12|
|       100644|12497|  13|
|         code|12460|  14|
|       return|12406|  15|
|         diff|12377|  16|
|         data|12102|  17|
|        group|11995|  18|
|          feb|11888|  19|
|      problem|11466|  20|
|        error|11164|  21|
|      version|11088|  22|
|         time|10868|  23|
|        flink|10842|  24|
|          set|10835|  25|
|     received|10665|  26|
|          add|10524|  27|
|   discussion|10306|  28|
|       google|10247|  29|
|         view|10199|  30|
|        visit|10064|  31|
|       groups| 9969|  32|
|          web| 9861|  33|
|   subscribed| 9828|  34|
|         case| 9682|  35|
|    receiving| 9676|  36|
|      running| 9508|  37|
|       emails| 9348|  38|
|        files| 9002|  39|
|         test| 8904|  40|
|         work| 8884|  41|
|       apache| 8776|  42|
|       struct| 8764|  43|
|         2020| 8451|  44|
|        issue| 8366|  45|
|        check| 8322|  46|
|         user| 8191|  47|
|      support| 8189|  48|
|       change| 8101|  49|
|         read| 8019|  50|
|          job| 7986|  51|
|         info| 7979|  52|
|          int| 7893|  53|
|      request| 7835|  54|
```

```
|      address| 7789|  55|
|         type| 7724|  56|
|       server| 7650|  57|
|     function| 7313|  58|
|     specific| 7007|  59|
|       memory| 6933|  60|
|          log| 6722|  61|
|          dec| 6689|  62|
|         long| 6569|  63|
|       create| 6435|  64|
|       static| 6191|  65|
|         good| 6185|  66|
|      release| 6132|  67|
|       python| 6090|  68|
|         void| 5947|  69|
|       device| 5895|  70|
|        start| 5886|  71|
|      package| 5855|  72|
|      default| 5590|  73|
|           00| 5587|  74|
|       source| 5559|  75|
|      contact| 5500|  76|
|     messages| 5395|  77|
|       failed| 5382|  78|
|          url| 5319|  79|
|      command| 5268|  80|
|        build| 5239|  81|
|          bug| 5114|  82|
|          git| 5082|  83|
|       number| 5024|  84|
|configuration| 4943|  85|
|       cygwin| 4888|  86|
|       update| 4836|  87|
|        guest| 4763|  88|
|      current| 4694|  89|
|      provide| 4693|  90|
|          nov| 4619|  91|
|           16| 4597|  92|
|       github| 4595|  93|
|          bit| 4569|  94|
|        write| 4540|  95|
|        table| 4520|  96|
|      working| 4515|  97|
|          kvm| 4501|  98|
|       kernel| 4420|  99|
|       output| 4396| 100|
+-------------+-----+----+
only showing top 100 rows

>>> df2.count()
3985991
```