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

|              domain|count|rank|
+--------------------+-----+----+
|           gmail.com| 8653|   1|
|          apache.org|  491|   2|
|          redhat.com|  295|   3|
|         hotmail.com|  186|   4|
|             163.com|  170|   5|
|          debian.org|  155|   6|
|              qq.com|  127|   7|
|           yahoo.com|  122|   8|
|          google.com|  110|   9|
|         outlook.com|  108|  10|
|     googlegroups.com|  106|  11|
|             gnu.org|  105|  12|
|         freebsd.org|   85|  13|
|              gmx.de|   77|  14|
|      protonmail.com|   70|  15|
|           intel.com|   65|  16|
|          oracle.com|   56|  17|
|              web.de|   48|  18|
|      googlemail.com|   47|  19|
|             126.com|   40|  20|
|             gmx.net|   40|  20|
|             free.fr|   36|  22|
|        mirantis.com|   35|  23|
|           posteo.de|   32|  24|
|         foxmail.com|   32|  24|
|          icloud.com|   31|  26|
|         comcast.net|   30|  27|
|          huawei.com|   29|  28|
|           yandex.ru|   28|  29|
|        linux.ibm.com|   28|  29|
|          kernel.org|   27|  31|
|    openjdk.java.net|   27|  31|
|         t-online.de|   26|  33|
|             gmx.com|   25|  34|
|          riseup.net|   25|  34|
|          python.org|   25|  34|
|syzkaller.appspot...|   24|  37|
|             arm.com|   24|  37|
|           pobox.com|   24|  37|
|       canonical.com|   23|  40|
|          vmware.com|   23|  40|
|        cloudera.com|   23|  40|
|       ververica.com|   22|  43|
|        ericsson.com|   22|  43|
|            live.com|   22|  43|
|    lists.debian.org|   21|  46|
|         disroot.org|   20|  47|
|             isc.org|   18|  48|
|          amazon.com|   18|  48|
|          golang.org|   18|  48|
+--------------------+-----+----+
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|13713|   1|
| org| 2146|   2|
| net|  798|   3|
|  de|  780|   4|
| edu|  319|   5|
|  fr|  231|   6|
|  uk|  219|   7|
|  io|  163|   8|
|  it|  140|   9|
|  ca|  133|  10|
|  ch|  109|  11|
|  ru|   96|  12|
|  au|   86|  13|
|  nl|   86|  13|
|  eu|   76|  15|
|  at|   75|  16|
|  se|   63|  17|
|  pl|   57|  18|
|  cn|   57|  18|
|  jp|   55|  20|
|  us|   51|  21|
|  in|   51|  21|
|  br|   50|  23|
|  cz|   49|  24|
|  ai|   48|  25|
|  nz|   42|  26|
|  be|   39|  27|
|  me|   36|  28|
|  es|   35|  29|
| gov|   34|  30|
|  co|   34|  30|
|  no|   30|  32|
|  fi|   29|  33|
|name|   28|  34|
|  dk|   27|  35|
|  cc|   23|  36|
|info|   23|  36|
| dev|   16|  38|
| biz|   16|  38|
|  gr|   15|  40|
|  ie|   14|  41|
|  pt|   13|  42|
|  tr|   13|  42|
|  hu|   13|  42|
|  ro|   12|  45|
| xyz|   12|  45|
|  fm|   11|  47|
|  sk|   11|  47|
|  is|   10|  49|
|  mx|   10|  49|
+----+-----+----+
only showing top 50 rows

>>> df.count()
20625
```

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

|          word|count|rank|
+--------------+-----+----+
|       message|34422|   1|
|          send|27438|   2|
|          2022|25255|   3|
|          file|23255|   4|
|   unsubscribe|22992|   5|
|          list|21938|   6|
|          diff|17547|   7|
|        100644|17408|   8|
|       mailing|16969|   9|
|        return|15979|  10|
|         email|15665|  11|
|          2021|15648|  12|
|           jan|15483|  13|
|          code|15191|  14|
|          mail|15118|  15|
|           add|14824|  16|
|            pm|14737|  17|
|         error|14454|  18|
|          data|14310|  19|
|         group|14188|  20|
|       version|13881|  21|
|        apache|13708|  22|
|           set|13707|  23|
|       problem|13267|  24|
|          time|13236|  25|
|           feb|12566|  26|
|      received|12246|  27|
|       request|11979|  28|
|        struct|11916|  29|
|          case|11783|  30|
|    discussion|11735|  31|
|         files|11686|  32|
|          view|11679|  33|
|        google|11662|  34|
|         check|11597|  35|
|         issue|11576|  36|
|         visit|11532|  37|
|          test|11401|  38|
|           web|11366|  39|
|       running|11334|  40|
|        groups|11319|  41|
|         flink|11303|  42|
|    subscribed|11136|  43|
|     receiving|11129|  44|
|           log|11050|  45|
|      specific|11050|  45|
|          work|10853|  47|
|        change|10853|  47|
|       support|10678|  49|
|        emails|10650|  50|
|          read|10424|  51|
|           int|10385|  52|
|          user|10048|  53|
|        server| 9287|  54|
|           url| 9277|  55|
|          type| 9245|  56|
|       contact| 9241|  57|
|        memory| 9142|  58|
|          info| 8976|  59|
|      function| 8974|  60|
|           mar| 8940|  61|
|       address| 8918|  62|
|           git| 8846|  63|
|           job| 8561|  64|
|          2020| 8512|  65|
|        github| 8473|  66|
|        static| 8406|  67|
|       release| 8399|  68|
|        create| 8300|  69|
|          long| 8020|  70|
|          void| 7903|  71|
|        device| 7613|  72|
|          good| 7477|  73|
|       queries| 7462|  74|
|          pull| 7429|  75|
|           bug| 7423|  76|
|       package| 7218|  77|
|infrastructure| 7165|  78|
|         start| 7114|  79|
|     automated| 7061|  80|
|       respond| 7026|  81|
|            g1| 7014|  82|
|        failed| 6954|  83|
|         build| 6905|  84|
|       default| 6884|  85|
|        source| 6823|  86|
|           dec| 6725|  87|
|        python| 6623|  88|
|        update| 6579|  89|
|       current| 6456|  90|
| configuration| 6450|  91|
|           kvm| 6392|  92|
|         guest| 6359|  93|
|        kernel| 6252|  94|
|       command| 6179|  95|
|            00| 6143|  96|
|        number| 6094|  97|
|    additional| 6061|  98|
|      messages| 6006|  99|
|       provide| 5787| 100|
+--------------+-----+----+
only showing top 100 rows

>>> df2.count()
5022277
```