In [1]:
```
from pyspark import SparkContext
import os

stop_words = ['they', 'she', 'he', 'it', 'the', 'as', 'is', 'and']
sc = SparkContext("local", "app")
```

```
21/12/02 02:18:27 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
```

In [2]:
```
# get all files in the data folder
file_arr = []
def construct_rdd(root_dir, file_arr):
    for fname in os.listdir(root_dir):
        f = os.path.join(root_dir, fname)
        if os.path.isfile(f):
            file_arr.append(f)
        else:
            construct_rdd(f, file_arr)

    return file_arr
```

In [3]:
```
home_dir = '/notebooks/data'
construct_rdd(home_dir, file_arr)
```

Out[3]:
```
['/notebooks/data/Tolstoy/anna_karenhina.txt',
 '/notebooks/data/Tolstoy/war_and_peace.txt',
 '/notebooks/data/test.txt',
 '/notebooks/data/Hugo/Miserables.txt',
 '/notebooks/data/Hugo/NotreDame_De_Paris.txt',
 '/notebooks/data/shakespeare/histories/kingrichardiii',
 '/notebooks/data/shakespeare/histories/kingrichardii',
 '/notebooks/data/shakespeare/histories/kinghenryv',
 '/notebooks/data/shakespeare/histories/kinghenryviii',
 '/notebooks/data/shakespeare/histories/kingjohn',
 '/notebooks/data/shakespeare/histories/2kinghenryiv',
 '/notebooks/data/shakespeare/histories/3kinghenryvi',
 '/notebooks/data/shakespeare/histories/1kinghenryiv',
 '/notebooks/data/shakespeare/histories/1kinghenryvi',
 '/notebooks/data/shakespeare/histories/2kinghenryvi',
 '/notebooks/data/shakespeare/poetry/sonnets',
 '/notebooks/data/shakespeare/poetry/rapeoflucrece',
 '/notebooks/data/shakespeare/poetry/various',
 '/notebooks/data/shakespeare/poetry/venusandadonis',
 '/notebooks/data/shakespeare/poetry/loverscomplaint',
 '/notebooks/data/shakespeare/tragedies/juliuscaesar',
 '/notebooks/data/shakespeare/tragedies/titusandronicus',
 '/notebooks/data/shakespeare/tragedies/antonyandcleopatra',
 '/notebooks/data/shakespeare/tragedies/timonofathens',
 '/notebooks/data/shakespeare/tragedies/hamlet',
 '/notebooks/data/shakespeare/tragedies/kinglear',
 '/notebooks/data/shakespeare/tragedies/othello',
 '/notebooks/data/shakespeare/tragedies/macbeth',
 '/notebooks/data/shakespeare/tragedies/coriolanus',
 '/notebooks/data/shakespeare/tragedies/romeoandjuliet',
 '/notebooks/data/shakespeare/comedies/troilusandcressida',
 '/notebooks/data/shakespeare/comedies/periclesprinceoftyre',
 '/notebooks/data/shakespeare/comedies/twogentlemenofverona',
 '/notebooks/data/shakespeare/comedies/muchadoaboutnothing',
 '/notebooks/data/shakespeare/comedies/comedyoferrors',
 '/notebooks/data/shakespeare/comedies/tamingoftheshrew',
 '/notebooks/data/shakespeare/comedies/asyoulikeit',
 '/notebooks/data/shakespeare/comedies/merchantofvenice',
 '/notebooks/data/shakespeare/comedies/tempest',
 '/notebooks/data/shakespeare/comedies/twelfthnight',
 '/notebooks/data/shakespeare/comedies/cymbeline',
 '/notebooks/data/shakespeare/comedies/allswellthatendswell',
 '/notebooks/data/shakespeare/comedies/loveslabourslost',
 '/notebooks/data/shakespeare/comedies/winterstale',
 '/notebooks/data/shakespeare/comedies/midsummersnightsdream',
 '/notebooks/data/shakespeare/comedies/merrywivesofwindsor',
 '/notebooks/data/shakespeare/comedies/measureforemeasure',
 '/notebooks/data/shakespeare/glossary',
 '/notebooks/data/shakespeare/README']
```

In [11]:
```
# create RDD in format ((word, file), count)
rdd = None
for f_path in file_arr:
    file = sc.wholeTextFiles(f_path)

    f_rdd = file.flatMap(lambda x: [(x[0], word) for word in x[1].lower().split()]) \
            .filter(lambda x: x[1] not in stop_words) \
            .map(lambda x: ((x[1],x[0]), 1)).reduceByKey(lambda a,b: a+b)

    if rdd:
        rdd = rdd.union(f_rdd)
```

```
    else:
        rdd = f_rdd
```

In [12]:

```
# create output RDD in format (word, [(file, count)])
output_rdd = rdd.map(lambda x: (x[0][0], (x[0][1], x[1]))).groupByKey().mapValues(list)
output_rdd.take(10)
```

Out[12]:

```
[('project',
  [('file:/notebooks/data/Tolstoy/anna_karenhina.txt', 84),
   ('file:/notebooks/data/Tolstoy/war_and_peace.txt', 91),
   ('file:/notebooks/data/Hugo/Miserables.txt', 86),
   ('file:/notebooks/data/Hugo/NotreDame_De_Paris.txt', 85),
   ('file:/notebooks/data/shakespeare/histories/2kinghenryiv', 1),
   ('file:/notebooks/data/shakespeare/tragedies/antonyandcleopatra', 1),
   ('file:/notebooks/data/shakespeare/tragedies/hamlet', 1),
   ('file:/notebooks/data/shakespeare/comedies/muchadoaboutnothing', 1),
   ('file:/notebooks/data/shakespeare/comedies/tempest', 3),
   ('file:/notebooks/data/shakespeare/comedies/allswellthatendswell', 1),
   ('file:/notebooks/data/shakespeare/comedies/winterstale', 1),
   ('file:/notebooks/data/shakespeare/glossary', 1)]),
 ('gutenberg',
  [('file:/notebooks/data/Tolstoy/anna_karenhina.txt', 24),
   ('file:/notebooks/data/Tolstoy/war_and_peace.txt', 24),
   ('file:/notebooks/data/Hugo/Miserables.txt', 24),
   ('file:/notebooks/data/Hugo/NotreDame_De_Paris.txt', 24)]),
 ('ebook',
  [('file:/notebooks/data/Tolstoy/anna_karenhina.txt', 9),
   ('file:/notebooks/data/Tolstoy/war_and_peace.txt', 9),
   ('file:/notebooks/data/Hugo/Miserables.txt', 9),
   ('file:/notebooks/data/Hugo/NotreDame_De_Paris.txt', 9)]),
 ('of',
  [('file:/notebooks/data/Tolstoy/anna_karenhina.txt', 8641),
   ('file:/notebooks/data/Tolstoy/war_and_peace.txt', 14902),
   ('file:/notebooks/data/test.txt', 5),
   ('file:/notebooks/data/Hugo/Miserables.txt', 19865),
   ('file:/notebooks/data/Hugo/NotreDame_De_Paris.txt', 7333),
   ('file:/notebooks/data/shakespeare/histories/kingrichardiii', 742),
   ('file:/notebooks/data/shakespeare/histories/kingrichardii', 732),
   ('file:/notebooks/data/shakespeare/histories/kinghenryv', 736),
   ('file:/notebooks/data/shakespeare/histories/kinghenryviii', 535),
   ('file:/notebooks/data/shakespeare/histories/kingjohn', 561),
   ('file:/notebooks/data/shakespeare/histories/2kinghenryiv', 628),
   ('file:/notebooks/data/shakespeare/histories/3kinghenryvi', 433),
   ('file:/notebooks/data/shakespeare/histories/1kinghenryiv', 697),
   ('file:/notebooks/data/shakespeare/histories/1kinghenryvi', 659),
   ('file:/notebooks/data/shakespeare/histories/2kinghenryvi', 542),
   ('file:/notebooks/data/shakespeare/poetry/sonnets', 372),
   ('file:/notebooks/data/shakespeare/poetry/rapeoflucrece', 264),
   ('file:/notebooks/data/shakespeare/poetry/various', 42),
   ('file:/notebooks/data/shakespeare/poetry/venusandadonis', 126),
   ('file:/notebooks/data/shakespeare/poetry/loverscomplaint', 72),
   ('file:/notebooks/data/shakespeare/tragedies/juliuscaesar', 364),
   ('file:/notebooks/data/shakespeare/tragedies/titusandronicus', 322),
   ('file:/notebooks/data/shakespeare/tragedies/antonyandcleopatra', 447),
   ('file:/notebooks/data/shakespeare/tragedies/timonofathens', 342),
   ('file:/notebooks/data/shakespeare/tragedies/hamlet', 666),
   ('file:/notebooks/data/shakespeare/tragedies/kinglear', 475),
   ('file:/notebooks/data/shakespeare/tragedies/othello', 473),
   ('file:/notebooks/data/shakespeare/tragedies/macbeth', 346),
   ('file:/notebooks/data/shakespeare/tragedies/coriolanus', 476),
   ('file:/notebooks/data/shakespeare/tragedies/romeoandjuliet', 392),
   ('file:/notebooks/data/shakespeare/comedies/troilusandcressida', 503),
   ('file:/notebooks/data/shakespeare/comedies/periclesprinceoftyre', 356),
   ('file:/notebooks/data/shakespeare/comedies/twogentlemenofverona', 252),
   ('file:/notebooks/data/shakespeare/comedies/muchadoaboutnothing', 358),
   ('file:/notebooks/data/shakespeare/comedies/comedyoferrors', 618),
   ('file:/notebooks/data/shakespeare/comedies/tamingoftheshrew', 260),
   ('file:/notebooks/data/shakespeare/comedies/asyoulikeit', 421),
   ('file:/notebooks/data/shakespeare/comedies/merchantofvenice', 475),
   ('file:/notebooks/data/shakespeare/comedies/tempest', 294),
   ('file:/notebooks/data/shakespeare/comedies/twelfthnight', 407),
   ('file:/notebooks/data/shakespeare/comedies/cymbeline', 520),
   ('file:/notebooks/data/shakespeare/comedies/allswellthatendswell', 455),
   ('file:/notebooks/data/shakespeare/comedies/loveslabourslost', 440),
   ('file:/notebooks/data/shakespeare/comedies/winterstale', 474),
   ('file:/notebooks/data/shakespeare/comedies/midsummersnightsdream', 270),
   ('file:/notebooks/data/shakespeare/comedies/merrywivesofwindsor', 418),
   ('file:/notebooks/data/shakespeare/comedies/measureforemeasure', 381),
   ('file:/notebooks/data/shakespeare/glossary', 258),
   ('file:/notebooks/data/shakespeare/README', 9)]),
 ('anna',
  [('file:/notebooks/data/Tolstoy/anna_karenhina.txt', 443),
   ('file:/notebooks/data/Tolstoy/war_and_peace.txt', 290),
   ('file:/notebooks/data/shakespeare/comedies/tamingoftheshrew', 1)]),
 ('karenina,', [('file:/notebooks/data/Tolstoy/anna_karenhina.txt', 13)]),
 ('by',
  [('file:/notebooks/data/Tolstoy/anna_karenhina.txt', 1075),
   ('file:/notebooks/data/Tolstoy/war_and_peace.txt', 2354),
   ('file:/notebooks/data/test.txt', 3),
   ('file:/notebooks/data/Hugo/Miserables.txt', 2323),
```

```
('file:/notebooks/data/Hugo/NotreDame_De_Paris.txt', 777),
('file:/notebooks/data/shakespeare/histories/kingrichardiii', 164),
('file:/notebooks/data/shakespeare/histories/kingrichardii', 91),
('file:/notebooks/data/shakespeare/histories/kinghenryv', 96),
('file:/notebooks/data/shakespeare/histories/kinghenryviii', 122),
('file:/notebooks/data/shakespeare/histories/kingjohn', 97),
('file:/notebooks/data/shakespeare/histories/2kinghenryiv', 109),
('file:/notebooks/data/shakespeare/histories/3kinghenryvi', 108),
('file:/notebooks/data/shakespeare/histories/1kinghenryiv', 112),
('file:/notebooks/data/shakespeare/histories/1kinghenryvi', 96),
('file:/notebooks/data/shakespeare/histories/2kinghenryvi', 110),
('file:/notebooks/data/shakespeare/poetry/sonnets', 90),
('file:/notebooks/data/shakespeare/poetry/rapeoflucrece', 90),
('file:/notebooks/data/shakespeare/poetry/various', 11),
('file:/notebooks/data/shakespeare/poetry/venusandadonis', 46),
('file:/notebooks/data/shakespeare/poetry/loverscomplaint', 16),
('file:/notebooks/data/shakespeare/tragedies/juliuscaesar', 96),
('file:/notebooks/data/shakespeare/tragedies/titusandronicus', 80),
('file:/notebooks/data/shakespeare/tragedies/antonyandcleopatra', 101),
('file:/notebooks/data/shakespeare/tragedies/timonofathens', 59),
('file:/notebooks/data/shakespeare/tragedies/hamlet', 114),
('file:/notebooks/data/shakespeare/tragedies/kinglear', 88),
('file:/notebooks/data/shakespeare/tragedies/othello', 106),
('file:/notebooks/data/shakespeare/tragedies/macbeth', 47),
('file:/notebooks/data/shakespeare/tragedies/coriolanus', 106),
('file:/notebooks/data/shakespeare/tragedies/romeoandjuliet', 108),
('file:/notebooks/data/shakespeare/comedies/troilusandcressida', 100),
('file:/notebooks/data/shakespeare/comedies/periclesprinceoftyre', 96),
('file:/notebooks/data/shakespeare/comedies/twogentlemenofverona', 79),
('file:/notebooks/data/shakespeare/comedies/muchadoaboutnothing', 93),
 ('file:/notebooks/data/shakespeare/comedies/comedyoferrors', 83),
 ('file:/notebooks/data/shakespeare/comedies/tamingoftheshrew', 67),
 ('file:/notebooks/data/shakespeare/comedies/asyoulikeit', 83),
 ('file:/notebooks/data/shakespeare/comedies/merchantofvenice', 101),
 ('file:/notebooks/data/shakespeare/comedies/tempest', 73),
 ('file:/notebooks/data/shakespeare/comedies/twelfthnight', 83),
 ('file:/notebooks/data/shakespeare/comedies/cymbeline', 121),
 ('file:/notebooks/data/shakespeare/comedies/allswellthatendswell', 92),
 ('file:/notebooks/data/shakespeare/comedies/loveslabourslost', 114),
 ('file:/notebooks/data/shakespeare/comedies/winterstale', 118),
 ('file:/notebooks/data/shakespeare/comedies/midsummersnightsdream', 68),
 ('file:/notebooks/data/shakespeare/comedies/merrywivesofwindsor', 97),
 ('file:/notebooks/data/shakespeare/comedies/measureforemeasure', 112),
 ('file:/notebooks/data/shakespeare/glossary', 42)]),
('leo',
 [('file:/notebooks/data/Tolstoy/anna_karenhina.txt', 4),
  ('file:/notebooks/data/Tolstoy/war_and_peace.txt', 4),
  ('file:/notebooks/data/Hugo/Miserables.txt', 2)]),
('tolstoy',
 [('file:/notebooks/data/Tolstoy/anna_karenhina.txt', 4),
  ('file:/notebooks/data/Tolstoy/war_and_peace.txt', 4)]),
('this',
 [('file:/notebooks/data/Tolstoy/anna_karenhina.txt', 1170),
  ('file:/notebooks/data/Tolstoy/war_and_peace.txt', 1835),
  ('file:/notebooks/data/test.txt', 1),
  ('file:/notebooks/data/Hugo/Miserables.txt', 3658),
  ('file:/notebooks/data/Hugo/NotreDame_De_Paris.txt', 933),
  ('file:/notebooks/data/shakespeare/histories/kingrichardiii', 175),
  ('file:/notebooks/data/shakespeare/histories/kingrichardii', 160),
  ('file:/notebooks/data/shakespeare/histories/kinghenryv', 142),
  ('file:/notebooks/data/shakespeare/histories/kinghenryviii', 228),
  ('file:/notebooks/data/shakespeare/histories/kingjohn', 260),
  ('file:/notebooks/data/shakespeare/histories/2kinghenryiv', 155),
('file:/notebooks/data/shakespeare/histories/3kinghenryvi', 175),
('file:/notebooks/data/shakespeare/histories/1kinghenryiv', 171),
('file:/notebooks/data/shakespeare/histories/1kinghenryvi', 172),
('file:/notebooks/data/shakespeare/histories/2kinghenryvi', 166),
('file:/notebooks/data/shakespeare/poetry/sonnets', 91),
('file:/notebooks/data/shakespeare/poetry/rapeoflucrece', 107),
('file:/notebooks/data/shakespeare/poetry/various', 11),
('file:/notebooks/data/shakespeare/poetry/venusandadonis', 40),
('file:/notebooks/data/shakespeare/poetry/loverscomplaint', 7),
('file:/notebooks/data/shakespeare/tragedies/juliuscaesar', 141),
('file:/notebooks/data/shakespeare/tragedies/titusandronicus', 192),
('file:/notebooks/data/shakespeare/tragedies/antonyandcleopatra', 139),
('file:/notebooks/data/shakespeare/tragedies/timonofathens', 105),
('file:/notebooks/data/shakespeare/tragedies/hamlet', 249),
('file:/notebooks/data/shakespeare/tragedies/kinglear', 206),
('file:/notebooks/data/shakespeare/tragedies/othello', 183),
('file:/notebooks/data/shakespeare/tragedies/macbeth', 94),
('file:/notebooks/data/shakespeare/tragedies/coriolanus', 145),
('file:/notebooks/data/shakespeare/tragedies/romeoandjuliet', 190),
('file:/notebooks/data/shakespeare/comedies/troilusandcressida', 144),
('file:/notebooks/data/shakespeare/comedies/periclesprinceoftyre', 133),
('file:/notebooks/data/shakespeare/comedies/twogentlemenofverona', 99),
('file:/notebooks/data/shakespeare/comedies/muchadoaboutnothing', 130),
('file:/notebooks/data/shakespeare/comedies/comedyoferrors', 103),
('file:/notebooks/data/shakespeare/comedies/tamingoftheshrew', 122),
('file:/notebooks/data/shakespeare/comedies/asyoulikeit', 144),
('file:/notebooks/data/shakespeare/comedies/merchantofvenice', 123),
('file:/notebooks/data/shakespeare/comedies/tempest', 170),
('file:/notebooks/data/shakespeare/comedies/twelfthnight', 154),
('file:/notebooks/data/shakespeare/comedies/cymbeline', 198),
```

```
('file:/notebooks/data/shakespeare/comedies/allswellthatendswell', 155),
('file:/notebooks/data/shakespeare/comedies/loveslabourslost', 134),
('file:/notebooks/data/shakespeare/comedies/winterstale', 174),
('file:/notebooks/data/shakespeare/comedies/midsummersnightsdream', 142),
('file:/notebooks/data/shakespeare/comedies/merrywivesofwindsor', 110),
('file:/notebooks/data/shakespeare/comedies/measureforemeasure', 177),
('file:/notebooks/data/shakespeare/glossary', 2),
('file:/notebooks/data/shakespeare/README', 1)])]
```