In [0]:
```
list_data = sc.parallelize(['hat', 'car', 'mysql', 'December', 'May', 'thunderstorm', 'red', 'blue', 'yellow',
                            'pink', 'green', 'silver', 'black', 'mary', 'joe', 'aqua', 'new', 'gray', 'orange'
                            , 'teal'])
```

In [0]:
```
list_data.count()
```

Out[2]: 20

In [0]:
```
month_data = sc.parallelize(['January', 'February', 'April', 'June', 'December', 'October', 'August', 'March',
                             'November', 'May', 'July', 'September'])
month_data.collect()
```

Out[3]: ['January',
 'February',
 'April',
 'June',
 'December',
 'October',
 'August',
 'March',
 'November',
 'May',
 'July',
 'September']

In [0]:
```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('SparkByExamples.com').getOrCreate()
month_data_y = month_data.filter(lambda x: "y" in x)
month_data_y.collect()
```

Out[4]: ['January', 'February', 'May', 'July']

In [0]:
```
path = "dbfs:/FileStore/shared_uploads/mishr212@umn.edu/wikipedia_public_domain.txt"
file_data = spark.read.text(path)
```

In [0]:
```
path = "dbfs:/FileStore/shared_uploads/mishr212@umn.edu/wikipedia_public_domain.txt"
file_data2 = sc.textFile(path).map(lambda x: x.split('\n'))
file_data2.collect()
```

Out[6]: [['These are excerpts taken from the "public domain" page of Wikipedia.'],
 [''],
 ['Definitions of the boundaries of the public domain in relation to copyright, or intellectual property more generally, regard the public domain as a negative space; that is, it consists of works that are no longer in copyright term or were never protected by copyright law.[18] According to James Boyle this definition underlines common usage of the term public domain and equates the public domain to public property and works in copyright to private property. However, the usage of the term public domain can be more granular, including for example uses of works in copyright permitted by copyright exceptions. Such a definition regards work in copyright as private property subject to fair-use rights and limitation on ownership.[1] A conceptual definition comes from Lange, who focused on what the public domain should be: "it should be a place of sanctuary for individual creative expression, a sanctuary conferring affirmative protection against the forces of private appropriation that threatened such expression".[18] Patterson and Lindberg described the public domain not as a "territory", but rather as a concept: "[T]here are certain materials — the air we breathe, sunlight, rain, space, life, creations, thoughts, feelings, ideas, words, numbers — not subject to private ownership. The materials that compose our cultural heritage must be free for all living to use no less than matter necessary for biological survival."[19] The term public domain may also be interchangeably used with other imprecise or undefined terms such as the public sphere or commons, including concepts such as the "commons of the mind", the "intellectual commons", and the "information commons"'],
 [''],
 ['A public-domain book is a book with no copyright, a book that was created without a license, or a book where its copyrights expired[20] or have been forfeited.[21]'],
 [''],
 ['In most countries the term of protection of copyright expires on the first day of January, 70 years after the death of the latest living author. The longest copyright term is in Mexico, which has life plus 100 years for all deaths since July 1928.'],
 [''],
 ['A notable exception is the United States, where every book and tale published before 1927 is in the public domain; American copyrights last for 95 years for books originally published between 1925 and 1978 if the copyright was properly registered and maintained.[22]'],
 [''],
 ['For example: the works of Jane Austen, Lewis Carroll, Machado de Assis, Olavo Bilac and Edgar Allan Poe are in the public domain worldwide as they all died over 100 years ago.'],
 [''],
 ['Project Gutenberg and the Internet Archive make tens of thousands of public domain books available online as ebooks.'],
 [''],
 ['Determination of whether a copyright has expired depends on an examination of the copyright in its source

```
country.'],
 [''],
 ['In the United States, determining whether a work has entered the public domain or is still under copyright
can be quite complex, primarily because copyright terms have been extended multiple times and in different wa
ys—shifting over the course of the 20th century from a fixed-term based on first publication, with a possible
renewal term, to a term extending to 50, then 70, years after the death of the author. The claim that "pre-19
27 works are in the public domain" is correct only for published works; unpublished works are under federal c
opyright for at least the life of the author plus 70 years.'],
 [''],
 ["In most other countries that are signatories to the Berne Convention, copyright term is based on the life
of the author, and extends to 50 or 70 years beyond the death of the author. (See List of countries' copyrigh
t lengths.)"],
 [''],
 ['Legal traditions differ on whether a work in the public domain can have its copyright restored. In the Eur
opean Union, the Copyright Duration Directive was applied retroactively, restoring and extending the terms of
copyright on material previously in the public domain. Term extensions by the US and Australia generally have
not removed works from the public domain, but rather delayed the addition of works to it. However, the United
States moved away from that tradition with the Uruguay Round Agreements Act, which removed from the public do
main many foreign-sourced works that had previously not been in copyright in the US for failure to comply wit
h US-based formalities requirements. Consequently, in the US, foreign-sourced works and US-sourced works are
now treated differently, with foreign-sourced works remaining under copyright regardless of compliance with f
ormalities, while domestically sourced works may be in the public domain if they failed to comply with then-e
xisting formalities requirements—a situation described as odd by some scholars, and unfair by some US-based r
ightsholders.[47]'],
 [''],
 ["The Reiss-Engelhorn-Museen, a German art museum, brought a suit against Wikimedia Commons in 2016 for phot
ographs uploaded to the database depicting pieces of art in the museum. The museum claimed that the photos we
re taken by their staff, and that photography within the museum by visitors was prohibited. Therefore, photos
taken by the museum, even of material that itself had fallen into the public domain, were protected by copyri
ght law and would need to be removed from the Wikimedia image repository. The court ruled that the photograph
s taken by the museum would be protected under the German Copyright Act, stating that since the photographer
needed to make practical decisions about the photograph that it was protected material. The Wikimedia volunte
er was ordered to remove the images from the site, as the museum's policy had been violated when the photos w
ere taken.[48]"],
 [''],
 ['A trademark registration may remain in force indefinitely, or expire without specific regard to its age. F
or a trademark registration to remain valid, the owner must continue to use it. In some circumstances, such a
s disuse, failure to assert trademark rights, or common usage by the public without regard for its intended u
se, it could become generic, and therefore part of the public domain.'],
 [''],
 ['Because trademarks are registered with governments, some countries or trademark registries may recognize a
mark, while others may have determined that it is generic and not allowable as a trademark in that registry.
```

```
For example, the drug acetylsalicylic acid (2-acetoxybenzoic acid) is better known as aspirin in the United S
tates—a generic term. In Canada, however, Aspirin, with an uppercase A, is still a trademark of the German co
mpany Bayer, while aspirin, with a lowercase "a", is not. Bayer lost the trademark in the United States, the
UK and France after World War I, as part of the Treaty of Versailles. So many copycat products entered the ma
rketplace during the war that it was deemed generic just three years later.[67]'],
 [''],
 ['Informal uses of trademarks are not covered by trademark protection. For example, Hormel, producer of the
canned meat product Spam, does not object to informal use of the word "spam" in reference to unsolicited comm
ercial email.[68] However, it has fought attempts by other companies to register names including the word \'s
pam\' as a trademark in relation to computer products, despite that Hormel\'s trademark is only registered in
reference to food products (a trademark claim is made within a particular field). Such defences have failed i
n the United Kingdom.[69]']]
```

In [0]:
```
file_data2.take(4)
```

```
Out[7]: [['These are excerpts taken from the "public domain" page of Wikipedia.'],
 [''],
 ['Definitions of the boundaries of the public domain in relation to copyright, or intellectual property more
generally, regard the public domain as a negative space; that is, it consists of works that are no longer in
copyright term or were never protected by copyright law.[18] According to James Boyle this definition underli
nes common usage of the term public domain and equates the public domain to public property and works in copy
right to private property. However, the usage of the term public domain can be more granular, including for e
xample uses of works in copyright permitted by copyright exceptions. Such a definition regards work in copyri
ght as private property subject to fair-use rights and limitation on ownership.[1] A conceptual definition co
mes from Lange, who focused on what the public domain should be: "it should be a place of sanctuary for indiv
idual creative expression, a sanctuary conferring affirmative protection against the forces of private approp
riation that threatened such expression".[18] Patterson and Lindberg described the public domain not as a "te
rritory", but rather as a concept: "[T]here are certain materials — the air we breathe, sunlight, rain, spac
e, life, creations, thoughts, feelings, ideas, words, numbers — not subject to private ownership. The materia
ls that compose our cultural heritage must be free for all living to use no less than matter necessary for bi
ological survival."[19] The term public domain may also be interchangeably used with other imprecise or undef
ined terms such as the public sphere or commons, including concepts such as the "commons of the mind", the "i
ntellectual commons", and the "information commons"'],
 ['']]
```

In [0]:
```
path = "dbfs:/FileStore/shared_uploads/mishr212@umn.edu/wikipedia_public_domain-1.txt"
file_data3 = sc.textFile(path).flatMap(lambda x: x.split(' '))
file_data3.collect()
```

```
Out[26]: ['These',
 'are',
 'excerpts',
 'taken',
 'from',
 'the',
 '"public',
 'domain"',
 'page',
 'of',
 'Wikipedia.',
 '',
 'Definitions',
 'of',
 'the',
 'boundaries',
 'of',
 'the',
 'public',
```

In [0]:
```
path = "dbfs:/FileStore/shared_uploads/mishr212@umn.edu/wikipedia_public_domain-1.txt"
file_data3 = sc.textFile(path).flatMap(lambda x: x.split(' '))
file_data3 = file_data3.map(lambda x: (x,1))
file_data3.collect()
```

```
Out[30]: [('These', 1),
 ('are', 1),
 ('excerpts', 1),
 ('taken', 1),
 ('from', 1),
 ('the', 1),
 ('"public', 1),
 ('domain"', 1),
 ('page', 1),
 ('of', 1),
 ('Wikipedia.', 1),
 ('', 1),
 ('Definitions', 1),
 ('of', 1),
 ('the', 1),
 ('boundaries', 1),
 ('of', 1),
 ('the', 1),
 ('public', 1),
```

In [0]:
```
file_data4 = file_data3.reduceByKey(lambda x,y: x + y)
file_data4.collect()
```

```
Out[29]: [('These', 1),
 ('are', 10),
 ('excerpts', 1),
 ('"public', 1),
 ('page', 1),
 ('of', 37),
 ('Wikipedia.', 1),
 ('', 14),
 ('public', 23),
 ('domain', 14),
 ('in', 27),
 ('relation', 2),
 ('copyright,', 2),
 ('more', 2),
 ('regard', 3),
 ('as', 15),
 ('negative', 1),
 ('is,', 1),
 ('no', 3),
```

In [0]:
```
file_data5 = file_data4.map(lambda y: y[1])
file_data4.top(5 , lambda y:  y[1])
```

```
Out[31]: [('the', 88), ('of', 37), ('in', 27), ('to', 27), ('a', 25)]
```

In [0]:
```
file_data3.count()
```

```
Out[35]: 1185
```

In [0]:
```
file_data2.sample(0, 0.3, 7).count()
```

```
Out[37]: 10
```

In [0]:
```
sc
```

SparkContext

Spark UI (/?o=2894624732106252#setting/sparkui/0211-235710-5qgr11on/driver-7287641975591693240)

Version

v3.1.2

Master

local[8]

AppName

Databricks Shell

In [0]:
```
file_data_o2 = file_data4.sortBy(lambda y: y[1], ascending= 0)
file_data_o2.collect()
```

```
Out[45]: [('the', 88),
 ('of', 37),
 ('in', 27),
 ('to', 27),
 ('a', 25),
 ('public', 23),
 ('copyright', 21),
 ('and', 21),
 ('that', 19),
 ('as', 15),
 ('', 14),
 ('domain', 14),
 ('is', 13),
 ('works', 13),
 ('by', 13),
 ('for', 12),
 ('or', 11),
 ('with', 11),
 ('trademark', 11),
```

In [0]:
```

```