

In [36]:

```
import re
re.sub('[^a-zA-Z0-9]+', '', 'ci--ao')
```

Out[36]:

```
'ciao'
```

In [37]:

```
file = sc.textFile('da-vinci-notebook.txt')
```

In [38]:

```
file.map(lambda line: re.sub('[^a-zA-Z0-9 ]+', '' ,line).split()).take(4)
```

Out[38]:

```
[[u'The',
  u'Project',
  u'Gutenberg',
  u'EBook',
  u'of',
  u'The',
  u'Notebooks',
  u'of',
  u'Leonardo',
  u'Da',
  u'Vinci',
  u'Complete'],
[u'by', u'Leonardo', u'Da', u'Vinci'],
[u'3', u'in', u'our', u'series', u'by', u'Leonardo', u'Da', u'Vi
nci'],
[]]
```

In [40]:

```
file.flatMap(lambda line: re.sub('[^a-zA-Z0-9 ]+', '' ,line).split()).take(4)
```

Out[40]:

```
[u'The', u'Project', u'Gutenberg', u'EBook']
```

In [14]:

```
file.flatMap(lambda line: re.sub('[^a-zA-Z0-9 ]+', ' ',line).split()).map(lamb
```

Out[14]:

```
[(u'The', 1),  
(u'Project', 1),  
(u'Gutenberg', 1),  
(u'EBook', 1),  
(u'of', 1),  
(u'The', 1),  
(u'Notebooks', 1),  
(u'of', 1)]
```

In [41]:

```
(file.flatMap(lambda line: re.sub('[^a-zA-Z0-9 ]+', ' ',line).split())  
    .map(lambda w: (w, 1))  
    .reduceByKey(lambda a, b: a+b)  
    .take(8)  
)
```

Out[41]:

```
[(u'Dagh', 3),  
(u'yellow', 40),  
(u'four', 106),  
(u'travaux', 1),  
(u'verses', 2),  
(u'looking', 27),  
(u'granting', 1),  
(u'SocksThe', 1)]
```

In [42]:

```
(file.flatMap(lambda line: re.sub('[^a-zA-Z0-9 ]+', ' ',line).split())  
    .map(lambda w: (w, 1))  
    .reduceByKey(lambda a, b: a+b)  
    .sortBy(keyfunc = lambda e: -e[1])  
    .take(8)  
)
```

Out[42]:

```
[(u'the', 20670),  
(u'of', 10457),  
(u'and', 7713),  
(u'in', 5316),  
(u'to', 5110),  
(u'is', 4217),  
(u'a', 4157),  
(u'it', 2808)]
```

In [43]:

```
stop_words = re.sub('\r', '', open('stop-word-list.txt').read()).split('\n')
```

In [44]:

```
stop_words[:10]
```

Out[44]:

```
['a',  
'about',  
'above',  
'across',  
'after',  
'afterwards',  
'again',  
'against',  
'all',  
'almost']
```

In [45]:

```
(file.flatMap(lambda line: re.sub('[^a-zA-Z0-9 ]+', ' ', line).split())  
  .filter(lambda w: w.lower() not in stop_words)  
  .take(8)  
)
```

Out[45]:

```
[u'Project',  
 u'Gutenberg',  
 u'EBook',  
 u'Notebooks',  
 u'Leonardo',  
 u'Da',  
 u'Vinci',  
 u'Complete']
```

In [46]:

```
(file.flatMap(lambda line: re.sub('[^a-zA-Z0-9 ]+', ' ',line).split())
  .filter(lambda w: w.lower() not in stop_words)
  .map(lambda w: (w, 1))
  .reduceByKey(lambda a, b: a+b)
  .sortBy(keyfunc = lambda e: -e[1])
  .take(8)
)
```

Out[46]:

```
[(u'Footnote', 849),
 (u'light', 758),
 (u'eye', 571),
 (u'shadow', 463),
 (u'body', 439),
 (u'Leonardo', 422),
 (u'seen', 409),
 (u'Pl', 399)]
```

In []: