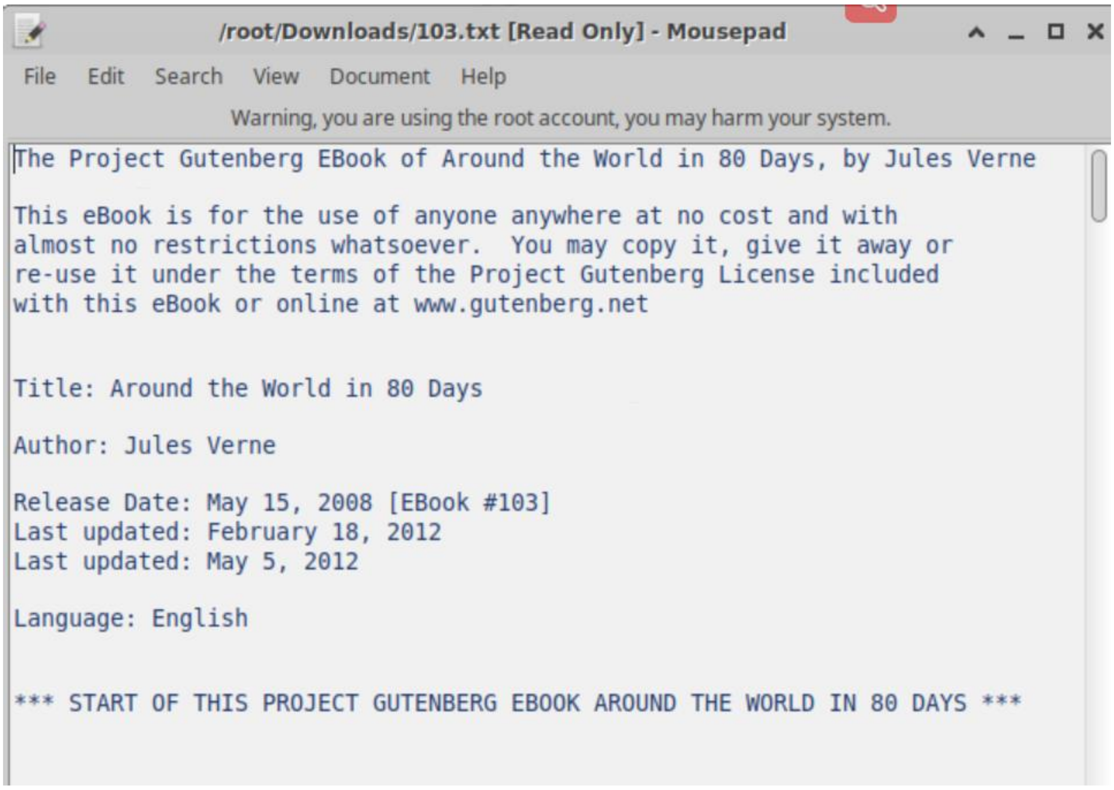
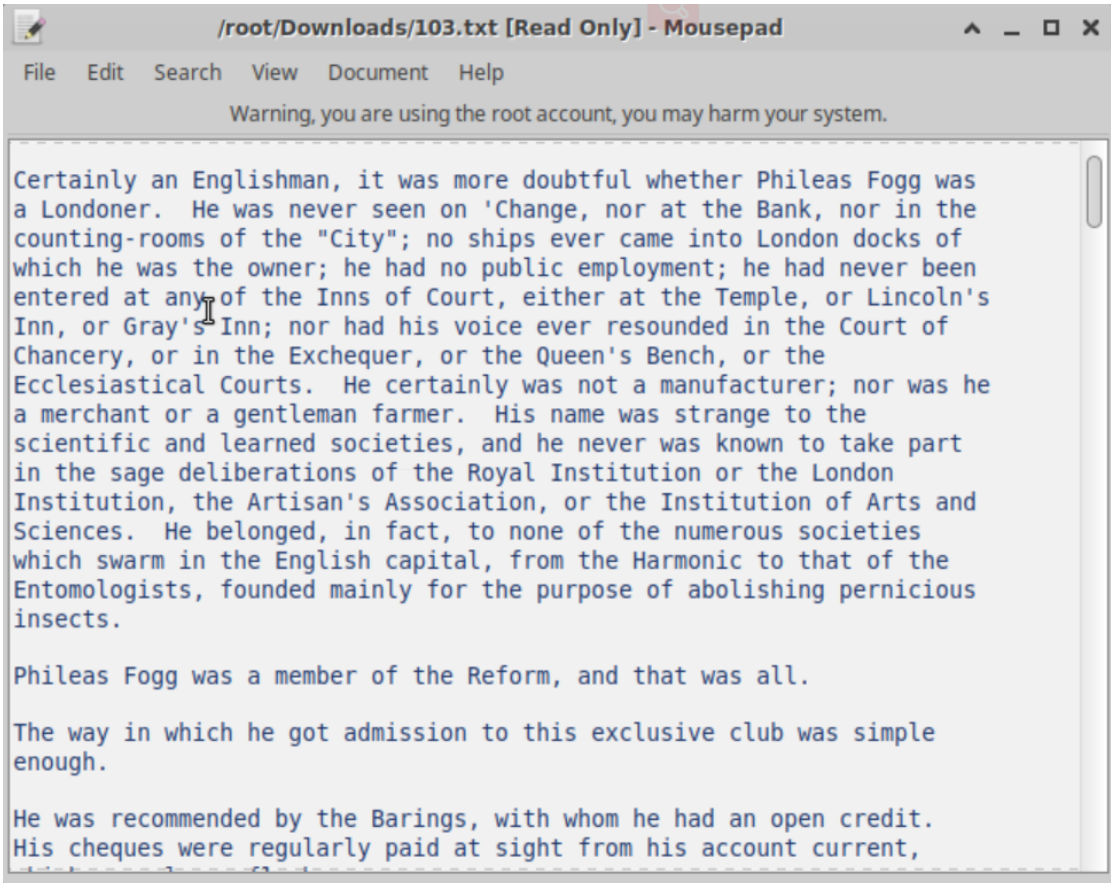
Intermediate



Warning, you are using the root account, you may harm your system.

```
The Project Gutenberg EBook of Around the World in 80 Days, by Jules Verne

This eBook is for the use of anyone anywhere at no cost and with
almost no restrictions whatsoever.  You may copy it, give it away or
re-use it under the terms of the Project Gutenberg License included
with this eBook or online at www.gutenberg.net


Title: Around the World in 80 Days

Author: Jules Verne

Release Date: May 15, 2008 [EBook #103]
Last updated: February 18, 2012
Last updated: May 5, 2012

Language: English


*** START OF THIS PROJECT GUTENBERG EBOOK AROUND THE WORLD IN 80 DAYS ***
```

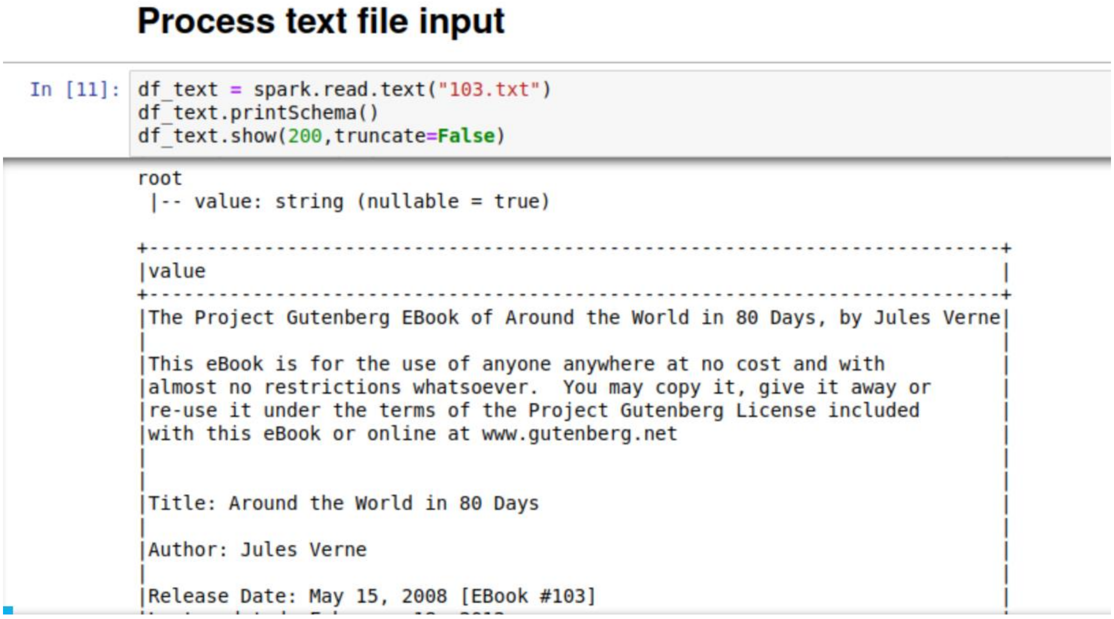
Find an English text book on Gutenberg. (Around the world in 80 days)



Warning, you are using the root account, you may harm your system.

```
Certainly an Englishman, it was more doubtful whether Phileas Fogg was
a Londoner.  He was never seen on 'Change, nor at the Bank, nor in the
counting-rooms of the "City"; no ships ever came into London docks of
which he was the owner; he had no public employment; he had never been
entered at any of the Inns of Court, either at the Temple, or Lincoln's
Inn, or Gray's Inn; nor had his voice ever resounded in the Court of
Chancery, or in the Exchequer, or the Queen's Bench, or the
Ecclesiastical Courts.  He certainly was not a manufacturer; nor was he
a merchant or a gentleman farmer.  His name was strange to the
scientific and learned societies, and he never was known to take part
in the sage deliberations of the Royal Institution or the London
Institution, the Artisan's Association, or the Institution of Arts and
Sciences.  He belonged, in fact, to none of the numerous societies
which swarm in the English capital, from the Harmonic to that of the
Entomologists, founded mainly for the purpose of abolishing pernicious
insects.

Phileas Fogg was a member of the Reform, and that was all.

The way in which he got admission to this exclusive club was simple
enough.

He was recommended by the Barings, with whom he had an open credit.
His cheques were regularly paid at sight from his account current,
```

Here is the partial content of the text file

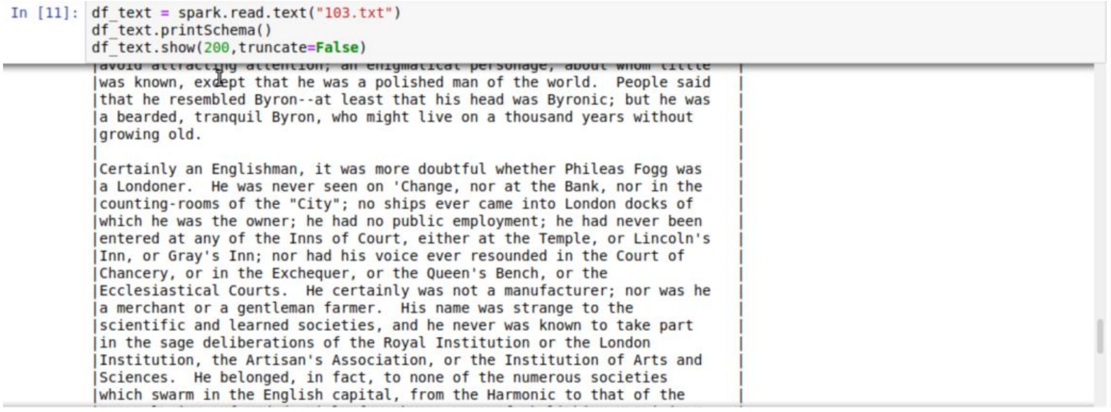


Process text file input

```
In [11]: df_text = spark.read.text("103.txt")
         df_text.printSchema()
         df_text.show(200,truncate=False)
```

```
root
 |-- value: string (nullable = true)

+-----------------------------------------------------------------------------+
|value                                                                        |
+-----------------------------------------------------------------------------+
|The Project Gutenberg EBook of Around the World in 80 Days, by Jules Verne|
|                                                                             |
|This eBook is for the use of anyone anywhere at no cost and with            |
|almost no restrictions whatsoever.  You may copy it, give it away or        |
|re-use it under the terms of the Project Gutenberg License included         |
|with this eBook or online at www.gutenberg.net                              |
|                                                                             |
|                                                                             |
|Title: Around the World in 80 Days                                          |
|                                                                             |
|Author: Jules Verne                                                         |
|                                                                             |
|Release Date: May 15, 2008 [EBook #103]                                     |
```

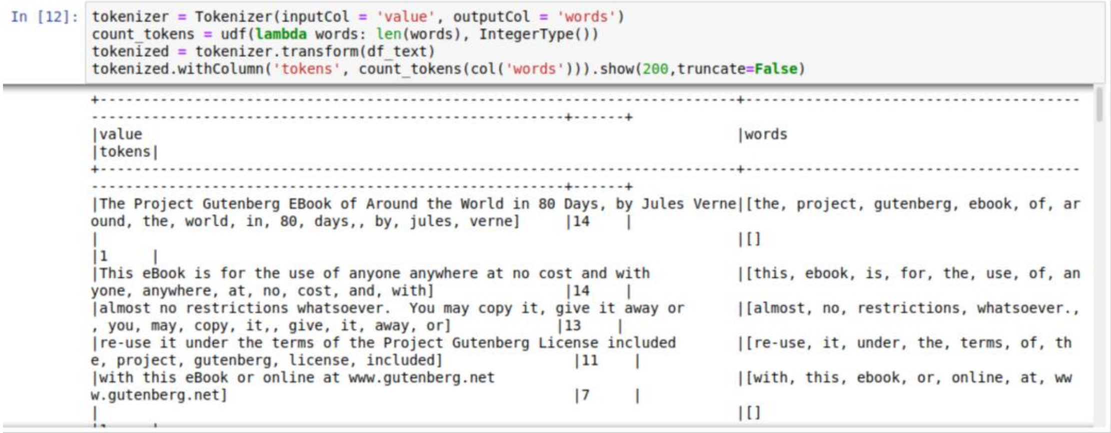
Input such text file as a dataframe in spark



```
In [11]: df_text = spark.read.text("103.txt")
         df_text.printSchema()
         df_text.show(200,truncate=False)
```

```
|avoid attracting attention; an enigmatical personage, about whom little     |
|was known, except that he was a polished man of the world.  People said     |
|that he resembled Byron--at least that his head was Byronic; but he was     |
|a bearded, tranquil Byron, who might live on a thousand years without       |
|growing old.                                                                 |
|                                                                             |
|Certainly an Englishman, it was more doubtful whether Phileas Fogg was      |
|a Londoner.  He was never seen on 'Change, nor at the Bank, nor in the      |
|counting-rooms of the "City"; no ships ever came into London docks of       |
|which he was the owner; he had no public employment; he had never been      |
|entered at any of the Inns of Court, either at the Temple, or Lincoln's     |
|Inn, or Gray's Inn; nor had his voice ever resounded in the Court of        |
|Chancery, or in the Exchequer, or the Queen's Bench, or the                 |
|Ecclesiastical Courts.  He certainly was not a manufacturer; nor was he     |
|a merchant or a gentleman farmer.  His name was strange to the              |
|scientific and learned societies, and he never was known to take part       |
|in the sage deliberations of the Royal Institution or the London            |
|Institution, the Artisan's Association, or the Institution of Arts and      |
|Sciences.  He belonged, in fact, to none of the numerous societies          |
|which swarm in the English capital, from the Harmonic to that of the        |
```

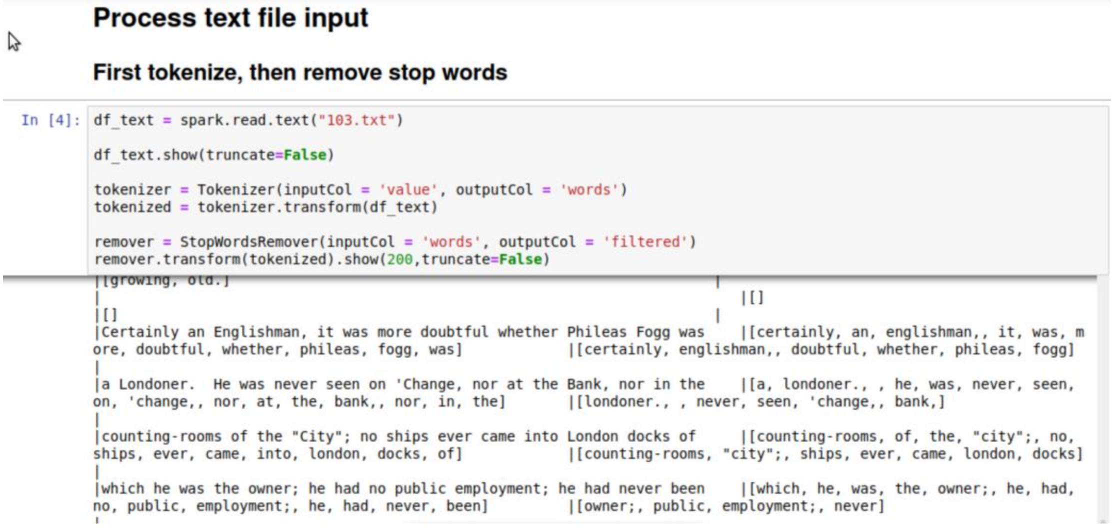
Here is the dataframe content.



```
In [12]: tokenizer = Tokenizer(inputCol = 'value', outputCol = 'words')
         count_tokens = udf(lambda words: len(words), IntegerType())
         tokenized = tokenizer.transform(df_text)
         tokenized.withColumn('tokens', count_tokens(col('words'))).show(200,truncate=False)
```

```
+--------------------------------------------------------------------------+-------------------------------------
--------------------------------------------------------+------+
|value                                                                     |words
|tokens|
+--------------------------------------------------------------------------+-------------------------------------
--------------------------------------------------------+------+
|The Project Gutenberg EBook of Around the World in 80 Days, by Jules Verne|[the, project, gutenberg, ebook, of, ar
ound, the, world, in, 80, days,, by, jules, verne]     |14    |
|                                                                          |[]
|1     |
|This eBook is for the use of anyone anywhere at no cost and with          |[this, ebook, is, for, the, use, of, an
yone, anywhere, at, no, cost, and, with]               |14    |
|almost no restrictions whatsoever.  You may copy it, give it away or      |[almost, no, restrictions, whatsoever.,
, you, may, copy, it,, give, it, away, or]            |13    |
|re-use it under the terms of the Project Gutenberg License included       |[re-use, it, under, the, terms, of, th
e, project, gutenberg, license, included]              |11    |
|with this eBook or online at www.gutenberg.net                            |[with, this, ebook, or, online, at, ww
w.gutenberg.net]                                       |7     |
|                                                                          |[]
```

First apply tokenization.



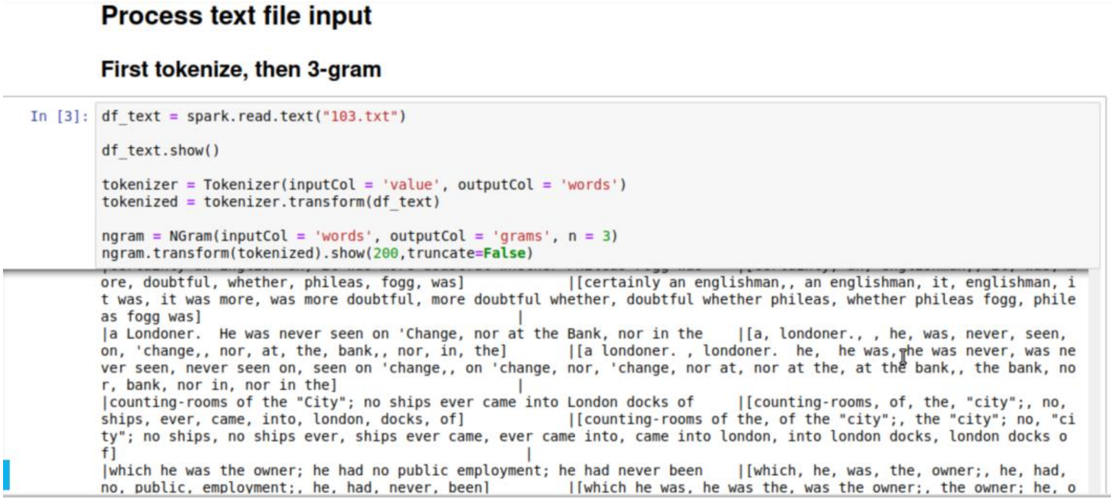
Process text file input

First tokenize, then remove stop words

```
In [4]: df_text = spark.read.text("103.txt")

        df_text.show(truncate=False)

        tokenizer = Tokenizer(inputCol = 'value', outputCol = 'words')
        tokenized = tokenizer.transform(df_text)

        remover = StopWordsRemover(inputCol = 'words', outputCol = 'filtered')
        remover.transform(tokenized).show(200,truncate=False)
```

```
|[growing, old.]                                                       |
|                                                                         |[]
|[]                                                                    |
|Certainly an Englishman, it was more doubtful whether Phileas Fogg was    |[certainly, an, englishman,, it, was, m
ore, doubtful, whether, phileas, fogg, was]           |[certainly, englishman,, doubtful, whether, phileas, fogg]
|
|a Londoner.  He was never seen on 'Change, nor at the Bank, nor in the    |[a, londoner., , he, was, never, seen,
on, 'change,, nor, at, the, bank,, nor, in, the]      |[londoner., , never, seen, 'change,, bank,]
|
|counting-rooms of the "City"; no ships ever came into London docks of     |[counting-rooms, of, the, "city";, no,
ships, ever, came, into, london, docks, of]           |[counting-rooms, "city";, ships, ever, came, london, docks]
|
|which he was the owner; he had no public employment; he had never been    |[which, he, was, the, owner;, he, had,
no, public, employment;, he, had, never, been]        |[owner;, public, employment;, never]
|
```

Then remove stop words by the built-in function



Process text file input

First tokenize, then 3-gram

```
In [3]: df_text = spark.read.text("103.txt")

        df_text.show()

        tokenizer = Tokenizer(inputCol = 'value', outputCol = 'words')
        tokenized = tokenizer.transform(df_text)

        ngram = NGram(inputCol = 'words', outputCol = 'grams', n = 3)
        ngram.transform(tokenized).show(200,truncate=False)
```

```
ore, doubtful, whether, phileas, fogg, was]           |[certainly an englishman,, an englishman, it, englishman, i
t was, it was more, was more doubtful, more doubtful whether, doubtful whether phileas, whether phileas fogg, phile
as fogg was]                                    |
|a Londoner.  He was never seen on 'Change, nor at the Bank, nor in the    |[a, londoner., , he, was, never, seen,
on, 'change,, nor, at, the, bank,, nor, in, the]      |[a londoner. , londoner.  he,  he was, he was never, was ne
ver seen, never seen on, seen on 'change,, on 'change, nor, 'change, nor at, nor at the, at the bank,, the bank, no
r, bank, nor in, nor in the]                    |
|counting-rooms of the "City"; no ships ever came into London docks of     |[counting-rooms, of, the, "city";, no,
ships, ever, came, into, london, docks, of]           |[counting-rooms of the, of the "city";, the "city"; no, "ci
ty"; no ships, no ships ever, ships ever came, ever came into, came into london, into london docks, london docks o
f]                                               |
|which he was the owner; he had no public employment; he had never been    |[which, he, was, the, owner;, he, had,
no, public, employment;, he, had, never, been]        |[which he was, he was the, was the owner;, the owner; he, o
```

Get 3-gram of the text file