1.

txt = sc.textFile('file:///home/cloudera/testespark/python_conceito.txt')

a) Quantidade de linhas no documento: txt.count()
7

b) lines = txt.filter(lambda line: 'Python' in line.lower())
print(lines())

‘Python é uma linguagem de programação de alto nível,[5] interpretada, de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte.’
‘Foi lançada por Guido van Rossum em 1991.[1] Atualmente possui um modelo de desenvolvimento comunitário, aberto e gerenciado pela organização sem fins lucrativos Python Software Foundation.’
‘Python é uma linguagem de propósito geral de alto nível, multiparadigma, suporta o paradigma orientado a objetos, imperativo, funcional e procedural.’
‘O nome Python teve a sua origem no grupo humorístico britânico Monty Python,[8] criador do programa Monty Python's Flying Circus, embora muitas pessoas façam associação com o réptil do mesmo nome (em português, píton ou pitão).’
‘Fonte: https://pt.wikipedia.org/wiki/Python’

c) words = txt.flatMap(lambda line: line.split(" "))
words_values = words.map(lambda word: (word, 1))
wordCounts = words_values.reduceByKey(lambda a,b:a +b)

2.

a) filepath = file:///home/cloudera/testespark/vgsales_mod2.csv

lines = sc.textFile(filepath).map(lambda x: x.split(";"))
vgsales_mod2 = lines.toDF([“Id”, “Rank”, “Name”, “Platform”, “Year”, “Genre”, “Publisher”, “NA_Sales”, “EU_Sales”, “JP_Sales”, “Other_Sales”, “Global_Sales"])

vgsales_mod2.count()
16291

b) vgsales_mod2.head(2)

c) vgsales_sports = vgsales_mod2.where('Genre = Sports')
vgsales_sports.count()

d) vgsales_global_sales = vgsales_mod2.where('Global_Sales > 20')
vgsales_global_sales.take(3)

e) vgsales_year = vgsales_mod2.where('Year > 2010')
vgsales_year.take(10)