BDA _Assignment 1

gedit mapper.py

```
#!/usr/bin/python
"""mapper.py"""

import sys
for line in sys.stdin:
        line = line.strip()
        words = line.split()
        for word in words:
                print '%s\t%s' % (word, 1)
```

gedit 1234.txt

Business intelligence (**BI**) consists of strategies, methodologies, and technologies used by enterprises for data analysis and management of business information.[1] Common functions of BI technologies include reporting, online analytical processing, analytics, dashboard development, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics, and prescriptive analytics.

BI tools can handle large amounts of structured and sometimes unstructured data to help organizations identify, develop, and otherwise create new strategic business opportunities. They aim to allow for the easy interpretation of these big data. Identifying new opportunities and implementing an effective strategy based on insights is assumed to potentially provide businesses with a competitive market advantage and long-term stability, and help them take strategic decisions.[2]

Business intelligence can be used by enterprises to support a wide range of business decisions ranging from operational to strategic. Basic operating decisions include product positioning or pricing. Strategic business decisions involve priorities, goals, and directions at the broadest level. In all cases, BI is believed to be most effective when it combines data derived from the market in which a company operates (external data) with data from company sources internal to the business such as financial and operations data (internal data). When combined, external and internal data can provide a complete picture which, in effect, creates an "intelligence" that cannot be derived from any singular set of data

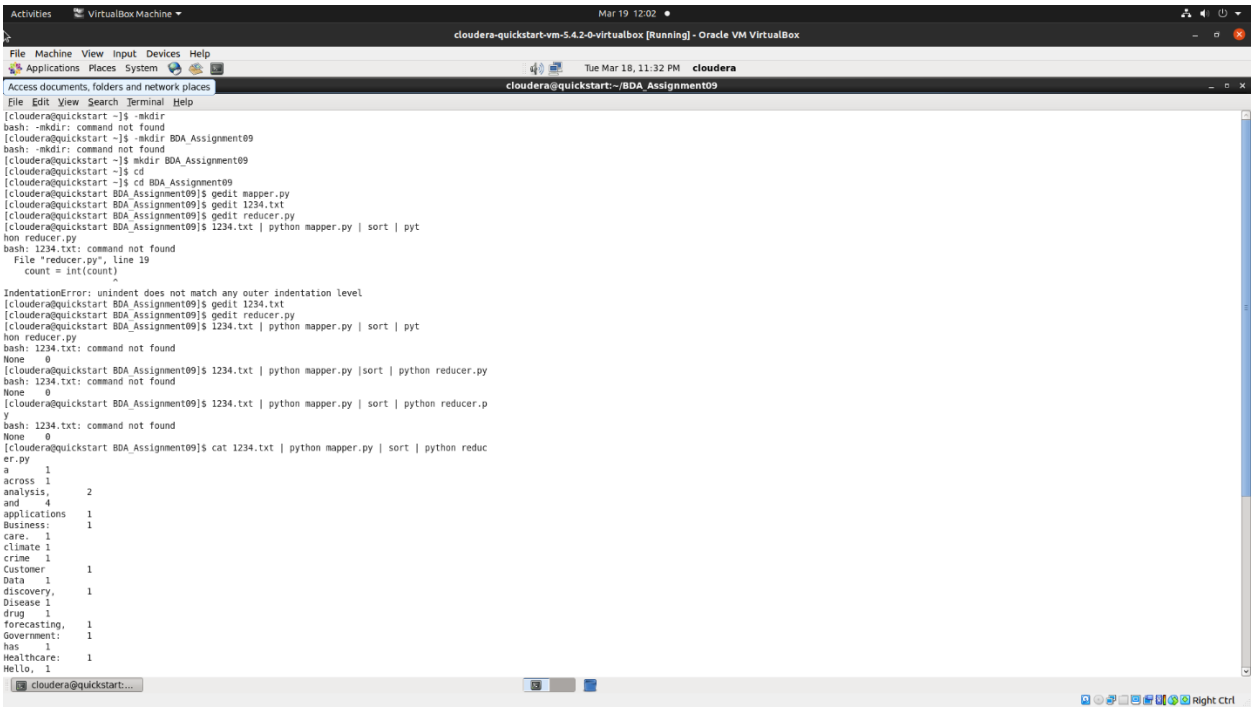
gedit reducer.py

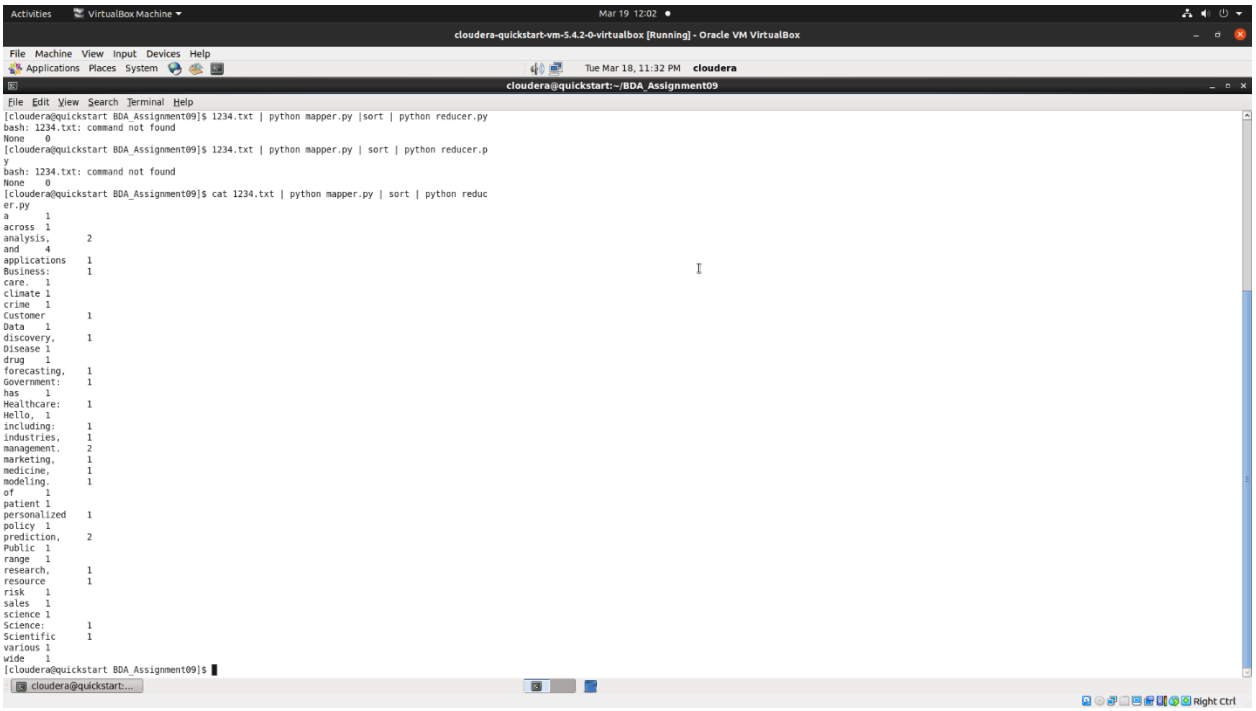
```
#!/usr/bin/env python
"""reducer.py"""

import sys

current_word = None
current_count = 0
word = None

# input comes from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()

    # parse the input we got from mapper.py
    word, count = line.split('\t', 1)

    # convert count (currently a string) to int
    try:
        count = int(count)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue
```

```
    # this IF-switch only works because Hadoop sorts map output
    # by key (here: word) before it is passed to the reducer
    if current_word == word:
        current_count += count
    else:
        if current_word:
            # write result to STDOUT
            print '%s\t%s' % (current_word, current_count)
        current_count = count
        current_word = word

# do not forget to output the last word if needed!
if current_word == word:
    print '%s\t%s' % (current_word, current_count)
```

```
cat 1234.txt | python mapper.py | sort | python reducer.py
```

final output


Activities
VirtualBox Machine
Mar 19 12:02
cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Tue Mar 18, 11:32 PM cloudera
Access documents, folders and network places
cloudera@quickstart:~/BDA_Assignment09
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ -mkdir
bash: -mkdir: command not found
[cloudera@quickstart ~]$ -mkdir BDA_Assignment09
bash: -mkdir: command not found
[cloudera@quickstart ~]$ mkdir BDA_Assignment09
[cloudera@quickstart ~]$ cd
[cloudera@quickstart ~]$ cd BDA_Assignment09
[cloudera@quickstart BDA_Assignment09]$ gedit mapper.py
[cloudera@quickstart BDA_Assignment09]$ gedit 1234.txt
[cloudera@quickstart BDA_Assignment09]$ gedit reducer.py
[cloudera@quickstart BDA_Assignment09]$ 1234.txt | python mapper.py | sort | pyt
hon reducer.py
bash: 1234.txt: command not found
File "reducer.py", line 19
count = int(count)
^
IndentationError: unindent does not match any outer indentation level
[cloudera@quickstart BDA_Assignment09]$ gedit 1234.txt
[cloudera@quickstart BDA_Assignment09]$ gedit reducer.py
[cloudera@quickstart BDA_Assignment09]$ 1234.txt | python mapper.py | sort | pyt
hon reducer.py
bash: 1234.txt: command not found
None 0
[cloudera@quickstart BDA_Assignment09]$ 1234.txt | python mapper.py |sort | python reducer.py
bash: 1234.txt: command not found
None 0
[cloudera@quickstart BDA_Assignment09]$ 1234.txt | python mapper.py | sort | python reducer.p
y
bash: 1234.txt: command not found
None 0
[cloudera@quickstart BDA_Assignment09]$ cat 1234.txt | python mapper.py | sort | python reduc
er.py
a 1
across 1
analysis, 2
and 4
applications 1
Business: 1
care. 1
climate 1
crime 1
Customer 1
Data 1
discovery, 1
Disease 1
drug 1
forecasting, 1
Government: 1
has 1
Healthcare: 1
Hello, 1
cloudera@quickstart:...
Right Ctrl


Activities
VirtualBox Machine
Mar 19 12:02
cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Tue Mar 18, 11:32 PM cloudera
cloudera@quickstart:~/BDA_Assignment09
File Edit View Search Terminal Help
[cloudera@quickstart BDA_Assignment09]$ 1234.txt | python mapper.py |sort | python reducer.py
bash: 1234.txt: command not found
None 0
[cloudera@quickstart BDA_Assignment09]$ 1234.txt | python mapper.py | sort | python reducer.p
y
bash: 1234.txt: command not found
None 0
[cloudera@quickstart BDA_Assignment09]$ cat 1234.txt | python mapper.py | sort | python reduc
er.py
a 1
across 1
analysis, 2
and 4
applications 1
Business: 1
care. 1
climate 1
crime 1
Customer 1
Data 1
discovery, 1
Disease 1
drug 1
forecasting, 1
Government: 1
has 1
Healthcare: 1
Hello, 1
including: 1
industries, 1
management. 2
marketing, 1
medicine, 1
modeling. 1
of 1
patient 1
personalized 1
policy 1
prediction, 2
Public 1
range 1
research, 1
resource 1
risk 1
sales 1
science 1
Science: 1
Scientific 1
various 1
wide 1
[cloudera@quickstart BDA_Assignment09]$
cloudera@quickstart:...
Right Ctrl