Uppsala University

Large Datasets For Scientific Applications

1TD268

Assignment A2

Author:
Sudarsan Bhargavan

June 9, 2018

1 A - Working with the RDD API

1.1 Question A.1

1. Read the English transcripts with Spark and count the number of lines.

   DataSet: Bulgarian dataset(europarl-v7.bg-en.en)
   Number of Lines: 406934

2. Do the same with the other language (so that you have a separate lineage of RDDs for each).

   No of Lines mentioned in the repository document : 406934

   WordCount Script

```
//WordCount.sh
  #!/bin/bash

  lineCount_en=‘/usr/bin/wc --lines ./europarl-v7.bg-en.en‘
  lineCount_bg=‘/usr/bin/wc --lines ./europarl-v7.bg-en.bg‘

  # Print Only The Number of Lines

  output_en=‘echo $lineCount_en |/usr/bin/awk ’{print $1}’‘
  output_bg=‘echo $lineCount_bg |/usr/bin/awk ’{print $1}’‘

  echo "Line Count of English Dataset: $output_en"
  echo
  echo "Line Count of Bulgarian Dataset: $output_bg"
```

   Output

```
Line Count of English Dataset: 406934

Line Count of Bulgarian Dataset: 406934
```

3. Verify that the line counts are the same for the two languages.

   Verifying the line counts for both the languages yields: [1]

   Bulgarian

   ```
   lines_en = sparkC.textFile(”/home/ubuntu/DATA/europarl-v7.bg-en.en”)
   lines_en.count()
   ```

   406934

   English

   ```
   lines_bg = sparkC.textFile(”/home/ubuntu/DATA/europarl-v7.bg-en.bg”)
   lines_bg.count()
   ```

   406934

4. Count the number of partitions.

   Trying to find the number of partitions yields:

   ```
   lines_en.getNumPartitioins()
   ```

   2

1.2 Question A.2

1. Inspect 100 entries from your RDD to verify your pre-processing.

   Inspecting 100 entries from RDD

   Output:

   For The Output Please See The File

   inspectedData.txt

2. Verify that the line counts still match after the pre-processing.

   Inspecting line counts after pre-processing still yields the same results for both the languages
   406934

1.3 Question A.3

1. Use Spark to compute the 10 most frequently according words in the English language corpus. Repeat for the other language.

   English

   Frequent Words List For English : [('the', 698563), ('of', 362452), ('to', 326291), ('and', 293700), ('in', 222084), ('a', 162764), ('is', 157336), ('that', 155812), ('for', 119429), ('I', 108253)]

   Bulgarian

   See File Referenced Below
   frequentsBulgarian.txt

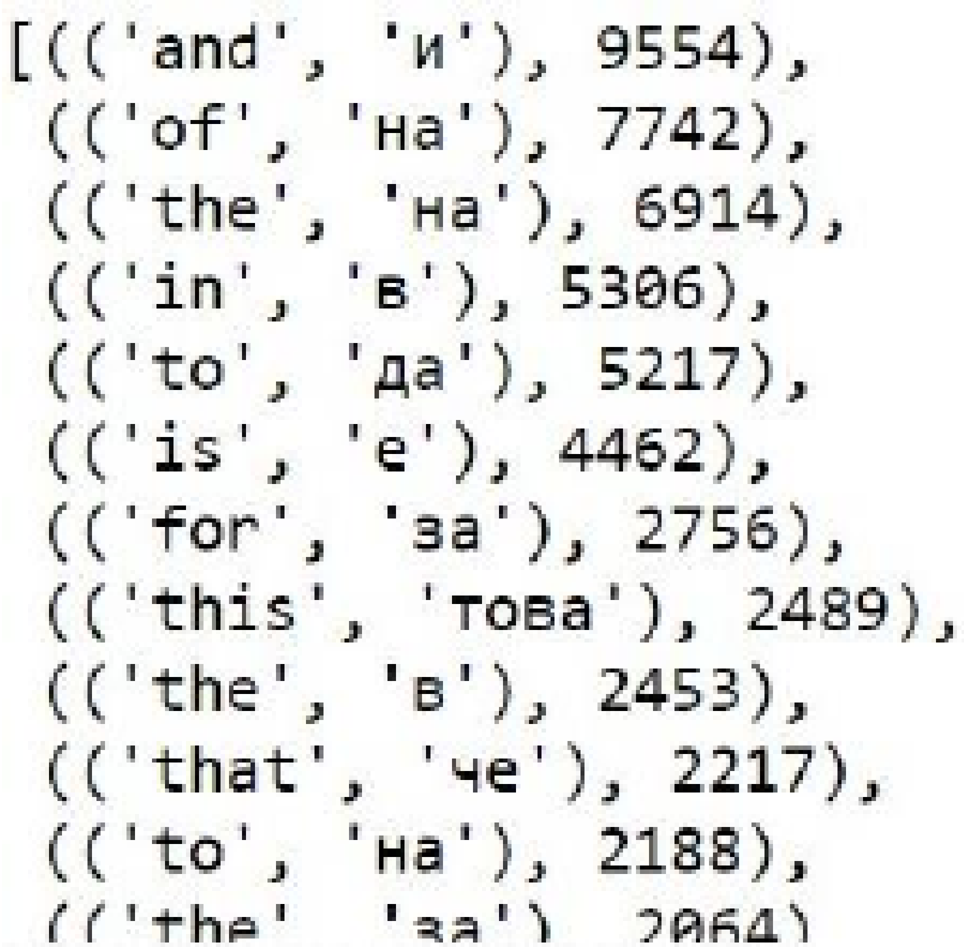
2. Verify that your results are reasonable.

   After Translation Bulgarian —— English
   It was found that many matched with the frequent English words.
   Please See Matched Frequent Words

1.4 Question A.4

1. Do your translations seem reasonable?

   While manually comparing with google translate, the translation seemed reasonable.



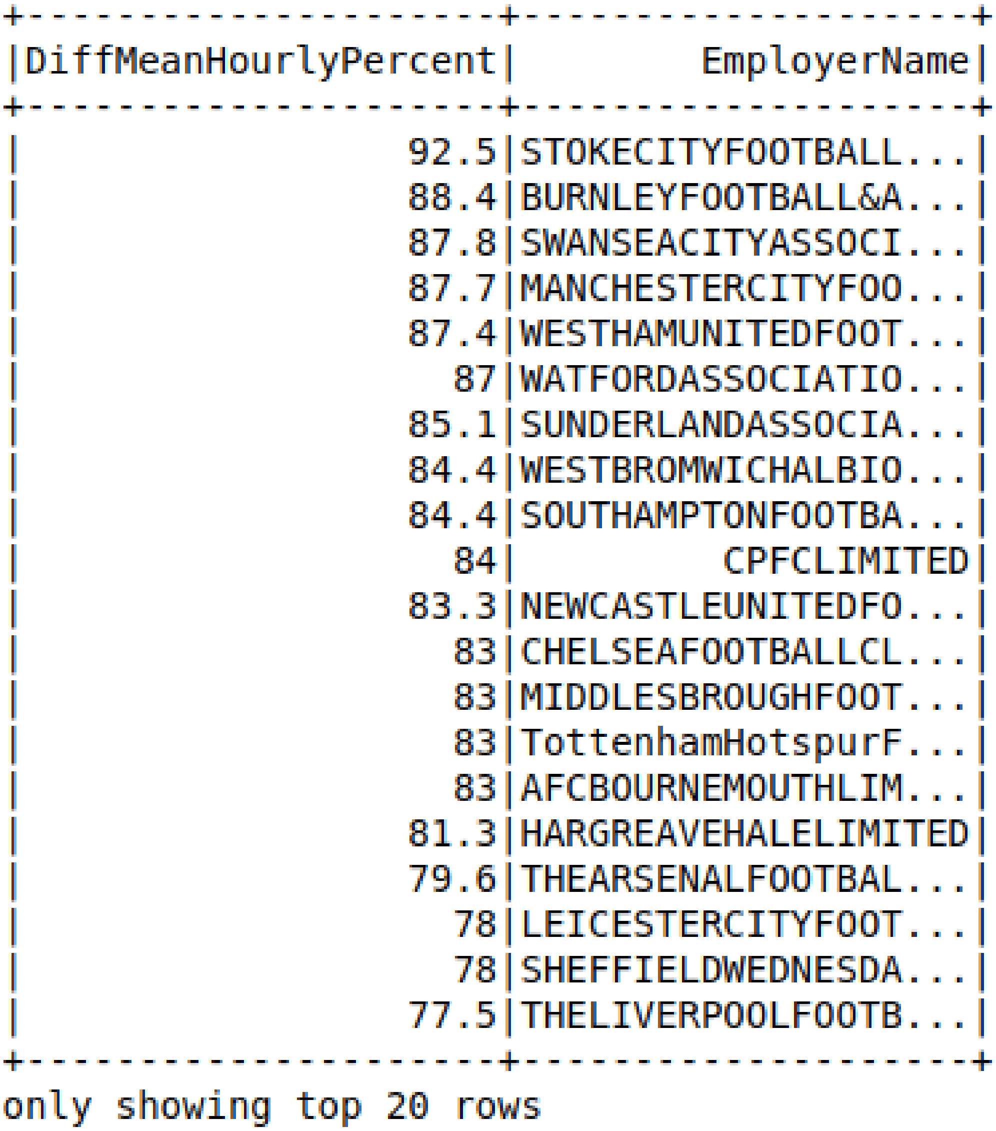
```
[(('and', 'и'), 9554),
 (('of', 'на'), 7742),
 (('the', 'на'), 6914),
 (('in', 'в'), 5306),
 (('to', 'да'), 5217),
 (('is', 'е'), 4462),
 (('for', 'за'), 2756),
 (('this', 'това'), 2489),
 (('the', 'в'), 2453),
 (('that', 'че'), 2217),
 (('to', 'на'), 2188),
 (('the', 'за'), 2064)
```

2 B - Working with DataFrames and SQL

2.1 Question B.1 - Analysis with DataFrames / SQL

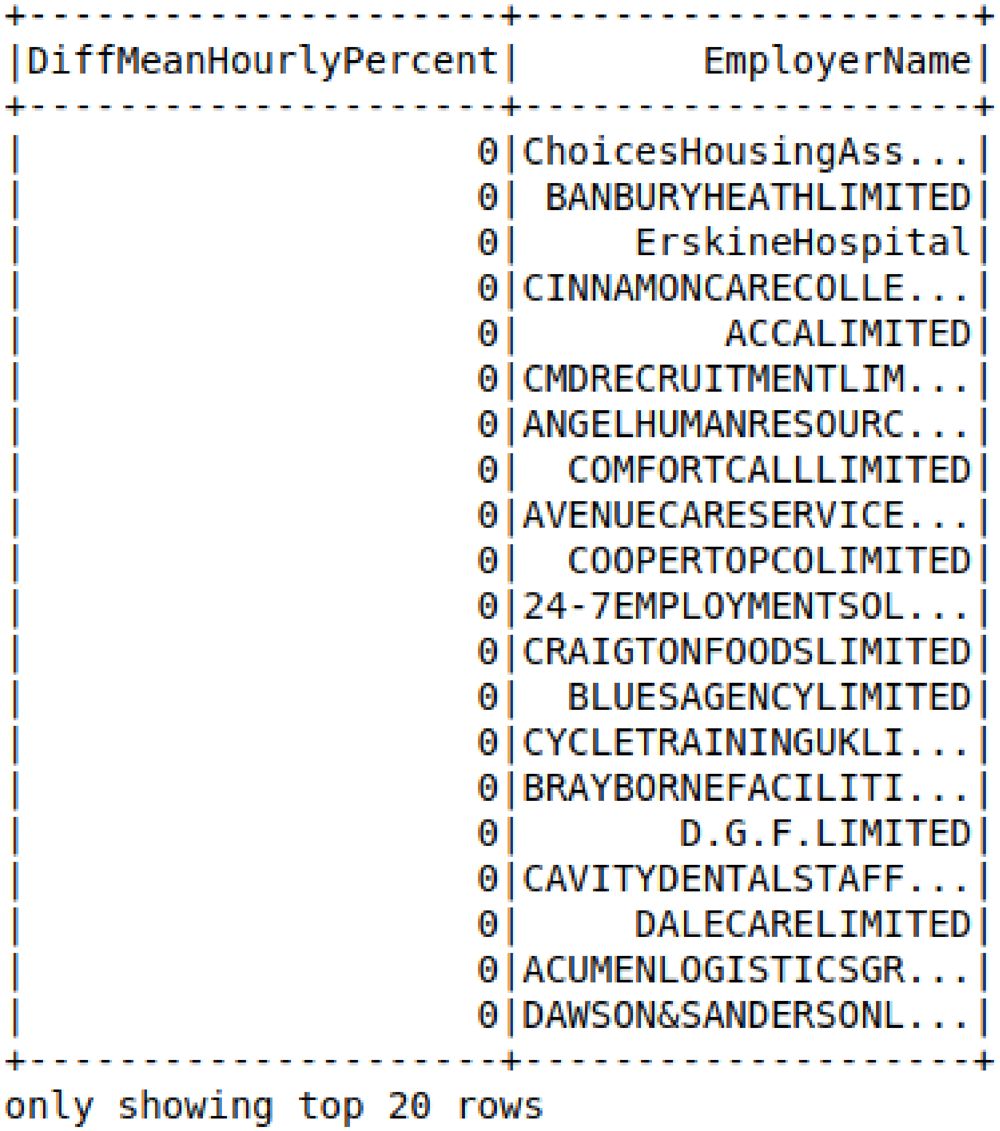
1. Which organization has the largest gender pay gap? Which the least?

   Largest Gender Pay Gap:



```
+---------------------+--------------------+
|DiffMeanHourlyPercent|        EmployerName|
+---------------------+--------------------+
|                 92.5|STOKECITYFOOTBALL...|
|                 88.4|BURNLEYFOOTBALL&A...|
|                 87.8|SWANSEACITYASSOCI...|
|                 87.7|MANCHESTERCITYFOO...|
|                 87.4|WESTHAMUNITEDFOOT...|
|                   87|WATFORDASSOCIATIO...|
|                 85.1|SUNDERLANDASSOCIA...|
|                 84.4|WESTBROMWICHALBIO...|
|                 84.4|SOUTHAMPTONFOOTBA...|
|                   84|         CPFCLIMITED|
|                 83.3|NEWCASTLEUNITEDFO...|
|                   83|CHELSEAFOOTBALLCL...|
|                   83|MIDDLESBROUGHFOOT...|
|                   83|TottenhamHotspurF...|
|                   83|AFCBOURNEMOUTHLIM...|
|                 81.3|HARGREAVEHALELIMITED|
|                 79.6|THEARSENALFOOTBAL...|
|                   78|LEICESTERCITYFOOT...|
|                   78|SHEFFIELDWEDNESDA...|
|                 77.5|THELIVERPOOLFOOTB...|
+---------------------+--------------------+
only showing top 20 rows
```

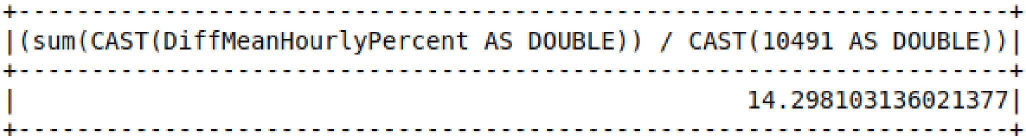
Least Gender Pay Gap:



```
+---------------------+--------------------+
|DiffMeanHourlyPercent|        EmployerName|
+---------------------+--------------------+
|                    0|ChoicesHousingAss...|
|                    0| BANBURYHEATHLIMITED|
|                    0|     ErskineHospital|
|                    0|CINNAMONCARECOLLE...|
|                    0|         ACCALIMITED|
|                    0|CMDRECRUITMENTLIM...|
|                    0|ANGELHUMANRESOURC...|
|                    0|  COMFORTCALLLIMITED|
|                    0|AVENUECARESERVICE...|
|                    0|  COOPERTOPCOLIMITED|
|                    0|24-7EMPLOYMENTSOL...|
|                    0|CRAIGTONFOODSLIMITED|
|                    0|  BLUESAGENCYLIMITED|
|                    0|CYCLETRAININGUKLI...|
|                    0|BRAYBORNEFACILITI...|
|                    0|       D.G.F.LIMITED|
|                    0|CAVITYDENTALSTAFF...|
|                    0|     DALECARELIMITED|
|                    0|ACUMENLOGISTICSGR...|
|                    0|DAWSON&SANDERSONL...|
+---------------------+--------------------+
only showing top 20 rows
```

2. What is the mean gender pay gap across all organization?

   Mean Gender Pay Gap:



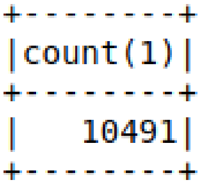
```
+--------------------------------------------------------------------+
|(sum(CAST(DiffMeanHourlyPercent AS DOUBLE)) / CAST(10491 AS DOUBLE))|
+--------------------------------------------------------------------+
|                                                  14.298103136021377|
+--------------------------------------------------------------------+
```

3. Export the results of B.1.2 to a CSV file. Inspect the output file to check it looks reasonable.

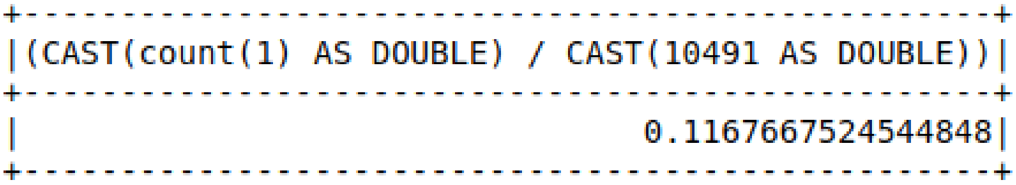
   Please See File
   csv.file

4. What proportion of organizations pay women more than men on average?

   Proportion of Organization That Pay Women More:



```
+--------+
|count(1)|
+--------+
|   10491|
+--------+
```



```
+--------------------------------------------------+
|(CAST(count(1) AS DOUBLE) / CAST(10491 AS DOUBLE))|
+--------------------------------------------------+
|                                0.1167667524544848|
+--------------------------------------------------+
```

2.2 Question B.2- Advanced DataFrames / SQL

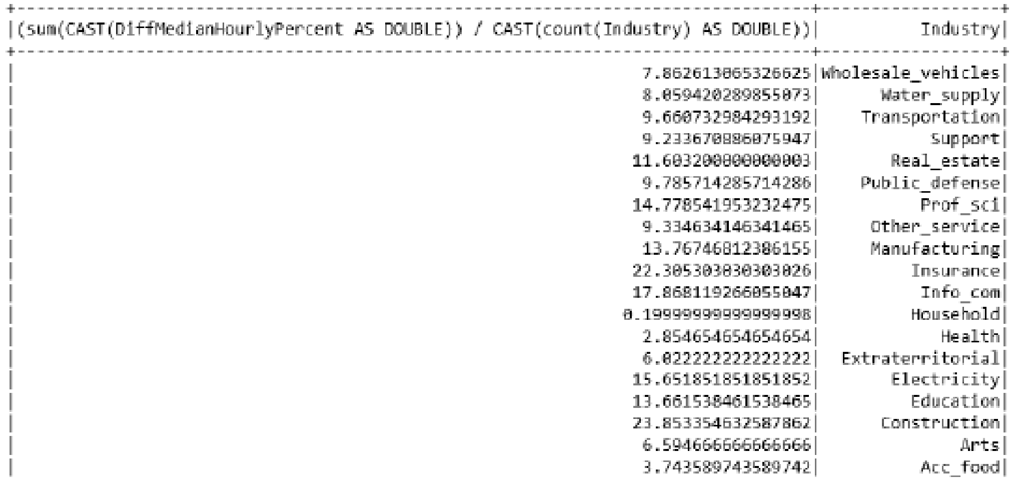
1. Create a new column for the industry sector (for each company) using the SIC code:

   The **broadcast** and **join** variables were used to modify the **Data Frame**. Also as per the instructions given the **sic_codes** with value **-1** has been ignored.

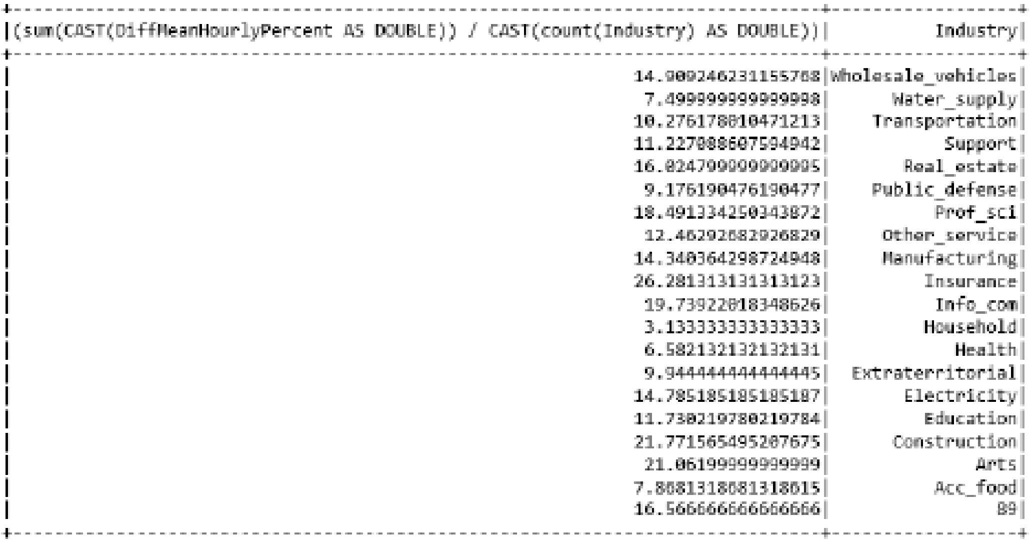
   The **broadcast** variable is used to maintain a read-only cached data of the variable. Data has been joined as per the required conditions, with help of the **join** command. [2]

2. Compute the mean gender pay gap per sector.

   Mean Gender Pay Gap:



```
+--------------------------------------------------------------------------------+------------------+
|(sum(CAST(DiffMedianHourlyPercent AS DOUBLE)) / CAST(count(Industry) AS DOUBLE))|          Industry|
+--------------------------------------------------------------------------------+------------------+
|                                                               7.862613065326625|Wholesale_vehicles|
|                                                               8.059420289855073|      Water_supply|
|                                                               9.660732984293192|    Transportation|
|                                                               9.233670886075947|           Support|
|                                                              11.603200000000003|       Real_estate|
|                                                               9.785714285714286|    Public_defense|
|                                                              14.778541953232475|          Prof_sci|
|                                                               9.334634146341465|     Other_service|
|                                                               13.76746812386155|     Manufacturing|
|                                                              22.305303030303026|         Insurance|
|                                                              17.868119266055047|          Info_com|
|                                                             0.19999999999999998|         Household|
|                                                               2.854654654654654|            Health|
|                                                               6.022222222222222|  Extraterritorial|
|                                                              15.651851851851852|       Electricity|
|                                                              13.661538461538465|         Education|
|                                                              23.853354632587862|      Construction|
|                                                               6.594666666666666|              Arts|
|                                                               3.743589743589742|          Acc_food|
```



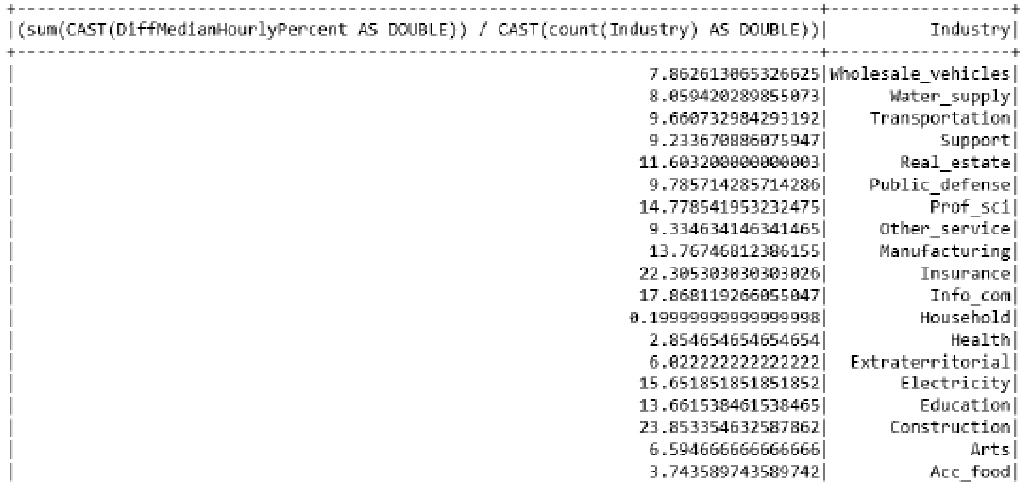
```
+------------------------------------------------------------------------------+------------------+
|(sum(CAST(DiffMeanHourlyPercent AS DOUBLE)) / CAST(count(Industry) AS DOUBLE))|          Industry|
+------------------------------------------------------------------------------+------------------+
|                                                            14.909246231155768|Wholesale_vehicles|
|                                                             7.499999999999998|      Water_supply|
|                                                            10.276178010471213|    Transportation|
|                                                            11.227088607594942|           Support|
|                                                            16.024799999999995|       Real_estate|
|                                                             9.176190476190477|    Public_defense|
|                                                            18.491334250343872|          Prof_sci|
|                                                             12.46292682926829|     Other_service|
|                                                            14.340364298724948|     Manufacturing|
|                                                            26.281313131313123|         Insurance|
|                                                             19.73922018348626|          Info_com|
|                                                             3.133333333333333|         Household|
|                                                             6.582132132132131|            Health|
|                                                             9.944444444444445|  Extraterritorial|
|                                                            14.785185185185187|       Electricity|
|                                                            11.730219780219784|         Education|
|                                                            21.771565495207675|      Construction|
|                                                             21.06199999999999|              Arts|
|                                                            7.8081318081318015|          Acc_food|
|                                                            16.566666666666666|                89|
+------------------------------------------------------------------------------+------------------+
```

3. How does gender pay equality compare per sector? Compute some additional statistics.

   Calculating the mean values yields the following information :

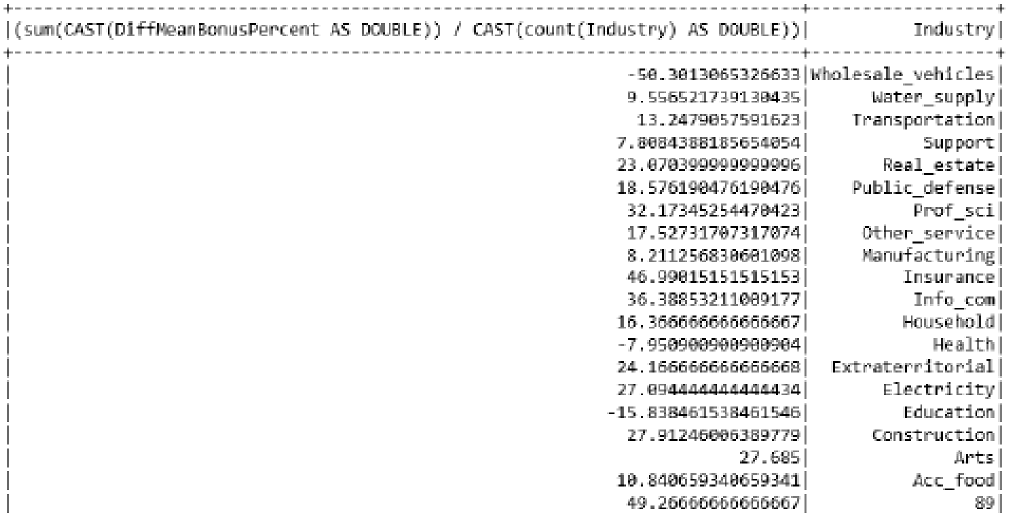
   In some cases **women** were paid more than **mean**, but in most cases it was the other way around.

   While calculating median mean per sector the gender pay equality was **netural**.



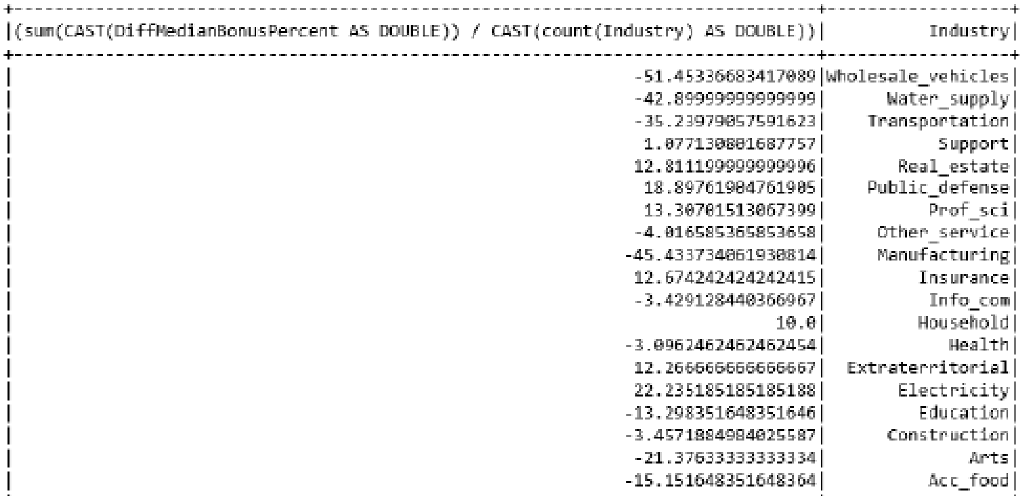
```
+--------------------------------------------------------------------------------+------------------+
|(sum(CAST(DiffMedianHourlyPercent AS DOUBLE)) / CAST(count(Industry) AS DOUBLE))|          Industry|
+--------------------------------------------------------------------------------+------------------+
|                                                               7.862613065326625|Wholesale_vehicles|
|                                                               8.059420289855073|      Water_supply|
|                                                               9.660732984293192|    Transportation|
|                                                               9.233670886075947|           Support|
|                                                              11.603200000000003|       Real_estate|
|                                                               9.785714285714286|    Public_defense|
|                                                              14.778541953232475|          Prof_sci|
|                                                               9.334634146341465|     Other_service|
|                                                               13.76746812386155|     Manufacturing|
|                                                              22.305303030303026|         Insurance|
|                                                              17.868119266055047|          Info_com|
|                                                             0.19999999999999998|         Household|
|                                                               2.854654654654654|            Health|
|                                                               6.022222222222222|  Extraterritorial|
|                                                              15.651851851851852|       Electricity|
|                                                              13.661538461538465|         Education|
|                                                              23.853354632587862|      Construction|
|                                                               6.594666666666666|              Arts|
|                                                               3.743589743589742|          Acc_food|
```

While calculating the mean per sector of **mean bonus pay**, it was found that **women** were paid more.



```
+-----------------------------------------------------------------------------+------------------+
|(sum(CAST(DiffMeanBonusPercent AS DOUBLE)) / CAST(count(Industry) AS DOUBLE))|          Industry|
+-----------------------------------------------------------------------------+------------------+
|                                                            -50.3013065326633|Wholesale_vehicles|
|                                                            9.556521739130435|      Water_supply|
|                                                             13.2479057591623|    Transportation|
|                                                           7.8084388185654054|           Support|
|                                                           23.070399999999996|       Real_estate|
|                                                           18.576190476190476|    Public_defense|
|                                                            32.17345254470423|          Prof_sci|
|                                                            17.52731707317074|     Other_service|
|                                                            8.211256830601098|     Manufacturing|
|                                                            46.99015151515153|         Insurance|
|                                                            36.38853211009177|          Info_com|
|                                                           16.366666666666667|         Household|
|                                                           -7.950900900900904|            Health|
|                                                           24.166666666666668|  Extraterritorial|
|                                                           27.094444444444434|       Electricity|
|                                                          -15.838461538461546|         Education|
|                                                            27.91246006389779|      Construction|
|                                                                       27.685|              Arts|
|                                                           10.840659340659341|          Acc_food|
|                                                            49.26666666666667|                89|
```

While calculating the median per sector of **mean bonus pay**, a lot of negative values were found, which according to the references provided, means that **women** were paid more. [3]



```
+-------------------------------------------------------------------------------+------------------+
|(sum(CAST(DiffMedianBonusPercent AS DOUBLE)) / CAST(count(Industry) AS DOUBLE))|          Industry|
+-------------------------------------------------------------------------------+------------------+
|                                                             -51.45336683417089|Wholesale_vehicles|
|                                                             -42.89999999999999|      Water_supply|
|                                                             -35.23979057591623|    Transportation|
|                                                              1.077130801687757|           Support|
|                                                             12.811199999999996|       Real_estate|
|                                                              18.89761904761905|    Public_defense|
|                                                              13.30701513067399|          Prof_sci|
|                                                             -4.016585365853658|     Other_service|
|                                                            -45.433734061930814|     Manufacturing|
|                                                             12.674242424242415|         Insurance|
|                                                             -3.429128440366967|          Info_com|
|                                                                           10.0|         Household|
|                                                            -3.0962462462462454|            Health|
|                                                             12.266666666666667|  Extraterritorial|
|                                                             22.235185185185188|       Electricity|
|                                                            -13.298351648351646|         Education|
|                                                            -3.4571884984025587|      Construction|
|                                                             -21.37633333333334|              Arts|
|                                                            -15.151648351648364|          Acc_food|
```

3 C - Spark Clusters and Deployment

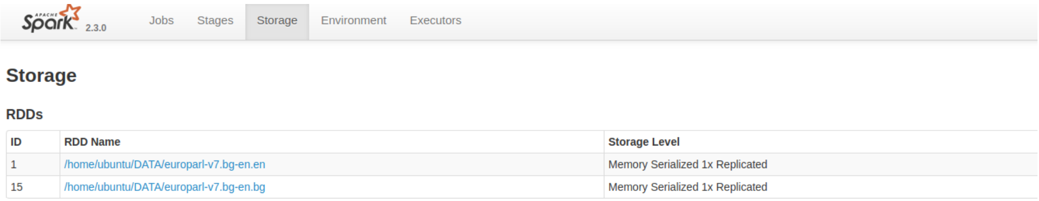
1. Modify a copy of your code from Section A, so that it runs on your cluster.

   Run Jobs - Cluster Mode
   In order to run a pyspark job in the cluster, the spark master url has to passed to the **SparkContext** method.

   ```
   //PySparkJobClusterMode.py
       #!/usr/bin/env python3
       import pyspark as pys
          sparkC = pys.SparkContext("spark://localhost:7077")
   ```

2. Run your code first without and then with .cache() - and look under the storage tab in the web GUI for your application. What do you notice? Explain briefly what's going on.

   Cache When the cache method is used, the **RDD** caches a copy of the imported data, for further operations.
   When we omit the cache method, then the **RDD** waits for an event to get triggered, after which it loads the data. [4]



Storage

RDDs

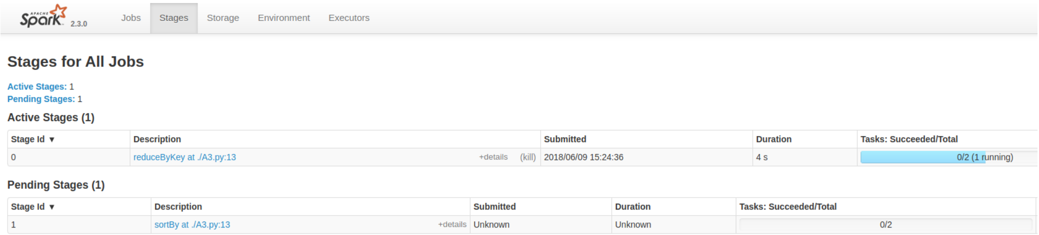
ID	RDD Name	Storage Level
1	/home/ubuntu/DATA/europarl-v7.bg-en.en	Memory Serialized 1x Replicated
15	/home/ubuntu/DATA/europarl-v7.bg-en.bg	Memory Serialized 1x Replicated

3. Use the Web GUI to explore your cluster and examine jobs, stages, and tasks. Create an example that requires a job with more than one stage. Explain, with reference to the Spark API methods you invoke in your code, why this is so.

   ### Multi-Stage Jobs

   A stage is a smaller set of tasks from a job. Stages can be parallelized if they are independent transformations are actions. [5]

   Here the **Task A3** has multiple-stages. But they cannot be parallelized because each stage is dependent on each other. **Stage Id: 0** represents the reduce operation. **Stage Id: 1** represents the sort operation.



Stages for All Jobs

Active Stages: 1
Pending Stages: 1

Active Stages (1)

Stage Id ▼	Description	Submitted	Duration	Tasks: Succeeded/Total
0	reduceByKey at ./A3.py:13 +details (kill)	2018/06/09 15:24:36	4 s	0/2 (1 running)

Pending Stages (1)

Stage Id ▼	Description	Submitted	Duration	Tasks: Succeeded/Total
1	sortBy at ./A3.py:13 +details	Unknown	Unknown	0/2

References

[1] statmt.org. (2018) European parliament proceedings parallel corpus 1996-2011. Accessed: 2018-06-09. [Online]. Available: http://www.statmt.org/europarl/

[2] learn4master.com. (2018) Pyspark broadcast variable example. Accessed: 2018-06-09. [Online]. Available: http://www.learn4master.com/big-data/spark/pyspark-broadcast-variable-example

[3] G. E. O. acas.org.uk, “Managing gender pay reporting,” Tech. Rep., 2017-December.

[4] D. Darabos. (2018) (why) do we need to call cache or persist on a rdd. Accessed: 2018-06-09. [Online]. Available: https://stackoverflow.com/questions/28981359/why-do-we-need-to-call-cache-or-persist-on-a-rdd

[5] javadba. (2018) How are stages split into tasks in spark? Accessed: 2018-06-09. [Online]. Available: https://stackoverflow.com/questions/37528047/how-are-stages-split-into-tasks-in-spark