Big Data Architecture Final Project: Analyzing CDC Mortality Data set with Apache Spark and Scala

SCHOOL OF ENGINEERING

Big Data Architecture SEIS - 736

Authors:
Ethan Wang

Instructor's name: Bradley Rubin

Why "Mortality"

I chose this project after listening to NPR's podcast regarding the efficacy of the "Lazaraus Drug.(Vedantam 2017)" In the podcast I heard harrowing stories as users would use powerful synthetic opioids like fentanyl and overdose. As they lost consciousness they would often be revived by friends and emergency staff with the use of a Nasal spray called Narcan. After this close call with death, they remain undeterred and continue to further their addiction. Sadly, the CDC has reported that seventy-two thousand Americans have passed away due to drug overdoes in 2017. More dead then the names on the Vietnam Memorial(US News 2017).

As such I wanted to use the skills, I learned in class to find ways to slice and interpret the data to gain a better understanding of the leadup to this growing epidemic.

The Data Source

This data source was located on Kaggle and was hosted by the CDC. It contains the record of every death in the country from 2005-2015 (Dane, 2017) . They include 10 .CSV and .JSON files and in total add up to about 4 GB of data. The JSON files contained metadata and descriptions for the ICD-10 medical classification codes for the millions of different injuries and states of deaths.

Data Pre-Processing required

I was initially unsure on how to read in multiple CSV files through RDD’s but was very pleased to see that the process went smoothly once I uploaded it into the cluster. My first challenge was finding a method to remove the header for all 10 CSV files and was successfully able to do so by the following code:

```
val header = mortality.first
val mortalityrdd = mortality.mapPartitions{x => x.filter(_ != header)}
```

This allowed me to point to the first row and turn it into the variable header. After which I created filter to go through the rest of the RDD to not include whatever matched that header.

Unfortunately there was another header I think nested somewhere else within the CSV that did not follow the same format so I created this filter to remove it.

```
val data = mortalityrdd.filter(s => !(s.contains("resident_status")))
```

Afterwards the header was removed I was able to go through most of the data without too much hindrance.

Bad Data Issues

There weren't too many bad data issues that I encountered as I went through the CSV files. I did have a little trouble trying to figure out what column I was working with at times and would utilize the following code to help guide me along.

```
mortality.map(x=>x.split(",")).map(x=>x(1)).first()
```

I found that naming the columns as follows to be very helpful to keeping me oriented with my code.

```
val current_data_year: Int = 16
val ic10code: Int = 23
val drug_code: Int = 24
val gender: Int = 5
val age: Int = 11
val race: Int = 74
val edu: Int = 2
```

Otherwise I did not encounter too many bad data issues other than that errant header.

Your Spark Algorithm and output

The goal for my spark program was to filter out all non-overdose deaths and to see change over time from 2010 to 2015. I utilized the ICD 10- Code and found the necessary codes: "420" ,"425","433","443","454" that were specifically related to drug overdoeses and utilized them in the below code to find only drug related deaths:

```
val drugdeaths=cleandata.map(x=>(x(current_data_year),x(drug_code)))

.filter{case(x,y)=>y.contains("420")||y.contains("425")||y.contains("433")||y.contains("443")||y.contains("454")}.map{case(x,y)=>(x,1)}

.reduceByKey(_+_,1)

.sortByKey(ascending = true).
```

This produced the following output:

```
(2005,30274)
(2006,34896)
(2007,36474)
(2008,37028)
(2009,37521)
(2010,38850)
(2011,41901)
(2012,42054)
(2013,44509)
(2014,47599)
(2015,53096)
```

Which shows a steady increase of drug deaths every year.

I then tried to find a way to narrow down the deaths and I was able to succeed by going through the specific death code: X40–X44, X60–X64, X85, and Y10–Y14. After going through all of these I found that the most numerous deaths were specifically related with X42 and X44.

X42: Accidental poisoning by and exposure to narcotics and psychodysleptics [hallucinogens], not elsewhere classified includes:

cannabis (derivatives)
cocaine
codeine
heroin
lysergide [LSD]
mescaline
methadone
morphine
opium (alkaloids)

My output for X42 deaths:

```
(2005,11091)
(2006,13354)
(2007,13078)
(2008,12827)
(2009,12502)
(2010,12317)
(2011,13261)
(2012,13976)
(2013,15299)
(2014,16879)
(2015,20010)
```

X44: Accidental poisoning by and exposure to other and unspecified drugs, medicaments and biological substances includes:

agents primarily acting on smooth and skeletal muscles and the respiratory system
anaesthetics (general)(local)
drugs affecting the:
cardiovascular system
gastrointestinal system
hormones and synthetic substitutes
systemic and haematological agents
systemic antibiotics and other anti-infectives
therapeutic gases
topical preparations
vaccines
water-balance agents and drugs affecting mineral and uric acid metabolism

My output for X44 deaths:

```
(2005,9677)
(2006,11373)
(2007,12804)
(2008,13569)
(2009,14375)
(2010,15589)
(2011,17584)
(2012,16774)
(2013,17296)
(2014,18632)
(2015,20178)
```

These two death codes were the most predominant causes of drug deaths in America. After I was able to do this, I switched back to include codes: "420" ,"425","433","443","454" looked utilize joins to make direct 5-year comparisons from 2010 and 2015 with gender, age, ethnicity and education.

My analysis for comparing deaths among gender groups over time:

```
Female drug overdose related deaths
(2005,11301)
(2006,12752)
(2007,13920)
(2008,14274)
(2009,14648)
(2010,15588)
(2011,16600)
(2012,16658)
```

```
(2013,17435)
(2014,18496)
(2015,19698)
```

```
Male drug overdose related deaths
(2005,18973)
(2006,22144)
(2007,22554)
(2008,22754)
(2009,22873)
(2010,23262)
(2011,25301)
(2012,25396)
(2013,27074)
(2014,29103)
(2015,33398)
```

```
My output on how much more Men in comparison to Women die in drug
overdoses.
(2005,7672)
(2006,9392)
(2007,8634)
(2008,8480)
(2009,8225)
(2010,7674)
(2011,8701)
(2012,8738)
(2013,9639)
(2014,10607)
(2015,13700)
```

I then examined in my analysis the age distribution of these deaths. I first compared 2015’s age distribution with the following code:

```
Age distribution of drug related deaths in 2015
(age code:
01: Under 1 year (includes not stated infant ages)
02:1 - 4 years
03:5 - 14 years
04:15 - 24 years
05:25 - 34 years
06:35 - 44 years
07:45 - 54 years
08:55 - 64 years
09:65 - 74 years
10:75 - 84 years
11:85 years and over
12:Age not stated)
```

My analysis for 2015 deaths in these age codes:

```
(01,29)
(02,61)
(03,68)
(04,4261)
(05,11937)
(06,11582)
(07,13105)
(08,9056)
(09,2073)
(10,581)
(11,333)
(12,10)
```

My analysis for 2010 deaths in these age codes:

```
(01,25)
(02,47)
(03,70)
(04,3596)
(05,7608)
(06,8595)
(07,11398)
(08,5573)
(09,1107)
(10,523)
(11,303)
(12,5)
```

Then as my output I compared the difference to 2010 to see the change in frequency in 5 years as my eventual output. You can see the result below:

```
(01,4)
(02,14)
(03,-2)
(04,665)
(05,4329)
(06,2987)
(07,1707)
(08,3483)
```

```
(09,966)
(10,58)
(11,30)
(12,5)
```

From this information we can see that the ages of 25-34 and ages 55-64 showed a large increase in drug overdoses compared to 10 years ago but the age group 45-54 being the largest group of total deaths.

The next attribute I investigated was the different ethnic groups that were impacted by drug deaths and compare them from 2010 to 2015. During my analysis I again compared the year 2010 of different ethnic drug overdoses with the year 2015's overdoses. In my analysis this was the output I looked at.

```
Drug deaths among ethnic groups on year 2010

(CDC race codes)
0:Other (Puerto Rico only)
1:White
2:Black
3:American Indian
4:Asian or Pacific Islander

(1,34795)
(2,3264)
(3,442)
(4,349)

Drug deaths among ethnic groups on year 2015

(1,46557)
(2,5318)
(3,606)
(4,615)
```

For my final code, I utilized a join and subtracted 2010's numbers from 2015. This gave me the following output:

```
Differences among ethnic groups on year 2015 and year 2010

(CDC race codes)
```

0:Other (Puerto Rico only)
1:White
2:Black
3:American Indian
4:Asian or Pacific Islander

(1,11762)
(2,2054)
(3,164)
(4,266)

The final group I looked at were education levels among drug overdoses. In my analysis I pulled the different groups from 2010 and from 2015.

Drug deaths among different education groups in 2010

(CDC education codes)
1:8th grade or less
2:9 - 12th grade, no diploma
3:high school graduate or GED completed
4:some college credit, but no degree
5:Associate degree
6:Bachelor’s degree
7:Master’s degree
8:Doctorate or professional degree
9:Unknown

(,10817)
(1,993)
(2,4536)
(3,12403)
(4,4702)
(5,1792)
(6,1940)
(7,531)
(8,213)
(9,923)

Drug deaths among different education groups in 2015

(,1459)
(1,1583)
(2,7702)
(3,23447)
(4,8340)

```
(5,3326)
(6,3409)
(7,882)
(8,368)
(9,2580)
```

For my final code I created a output comparing the differences between 2010 and 2015 as shown below:

Differences among education groups on year 2015 and year 2010

```
(,-9358)
(1,590)
(2,3166)
(3,11044)
(4,3638)
(5,1534)
(6,1469)
(7,351)
(8,155)
(9,1657)
```

Description of any other ecosystems or additional tools

I initially used R to see if I could process the information but quickly found that while R was able to read the CSV files, it had a very difficult time handing the sheer size of the data. Thus I mostly used R to double check if I was using the correct columns and the correct count of deaths.

Performance/Scale Characteristics

Generally, the performance of my final code only utilizes 2 executors and 2 executor cores. The code ran well and most certainly outperformed R's analytical capabilities. My final program was able to execute on the cluster with no problems.

What would you have done differently if you did this again?

I think a good opportunity for me in the future now that I have been able to utilize Apache Spark Ecosystem would be to spend more time on learning how I could use regular expressions and slice my own information within this data set. Or perhaps compare it with a different data set like Twitter. Another opportunity is to see if I could create profiles of the individuals that died to understand the common behaviors and characteristics that lead to drug overdose deaths.

I also would like to have found a way to pair the JSON metadata to all the codes in the CSV file. That would make the information much more presentable, but I am unsure on how it would impact performance.

Another opportunity for me would be to utilize more Scala Classes. I think by making classes I would have been able to simply my workflow processes in my analysis to find the information needed. Most likely this would have been the best method for me to pair the JSON metadata to the codes found in the CSV file.

Conclusions

Overdoes deaths seem to have seen a steady increase from 2005-2015. In particular we see increases of deaths in all ages, ethnicities, gender and education. With the largest bin of deaths among men, white’s, ages 45-54, with a high school/GED education.

Apache Spark’s performance and ability to pull and aggregate this information is quite impressive. Working on the RDD analysis was challenging for me but after learning how to apply filters for multiple individual columns I was able to pull the information fairly easily and find trends in my data.

References

(Life, Death And The Lazarus Drug: Confronting America's Opioid Crisis.(2017). Retrieved from https://www.npr.org/2018/10/26/661011560/life-death-and-the-lazarus-drug-confronting-americas-opioid-crisis

(Drug Overdose Deaths (2017). Retrieved from https://www.cdc.gov/drugoverdose/data/statedeaths.html

(3 New Names Added to Vietnam Veterans Memorial Wall(2017). Retrieved from

https://www.usnews.com/news/best-states/washington-dc/articles/2017-05-29/3-new-names-added-to-vietnam-veterans-memorial-wall