1 - subset

5 pts

Download a small subset of the data (100 rows is plenty) to your personal computer, and examine it using any software you like. Briefly describe this subset of the data by picking out a couple rows that look interesting to you.

1. How many columns are there?
   61
2. Do the data values in each column seem to match the column definitions?
   Yes they do, although I am getting a length of 61 when imported to python rather than 62 as seen in the headers file.
3. What character delimits the records?
   Tab or \t
4. What is the CAMEO event code, what event does this correspond to, and what is the Goldstein score?
   The CAMEO event code describes the action that actor1 performed to actor2, this could be yield, investigate, reject, etc. The Goldstein score is a number between -10 and +10 that shows the potential impact that event will have with negative showing bad and positive showing good.
5. Are the URL's to the news articles still live, and do they match the CAMEO event code?
   I was only able to find 1 link that was potentially no longer live, but it was a Pakistani site so it may have just taken too long to load.
6. Does the Goldstein score appear to be doing what it was designed to do?

Yes

2 - histogram

10 pts

Create a histogram of the Goldstein scores for all of 2018, using the integers as bin endpoints for the histogram. It's possible to do this in less than 10 minutes using a single shell pipeline on a t2 micro instance with 1 vCPU, 1 GiB memory, and 8 GiB storage.

1. How long does your program take to run?
   4m28.212s
2. Explain in detail what each command in the pipeline does and how they work together.
   Time - records the time of the following commands

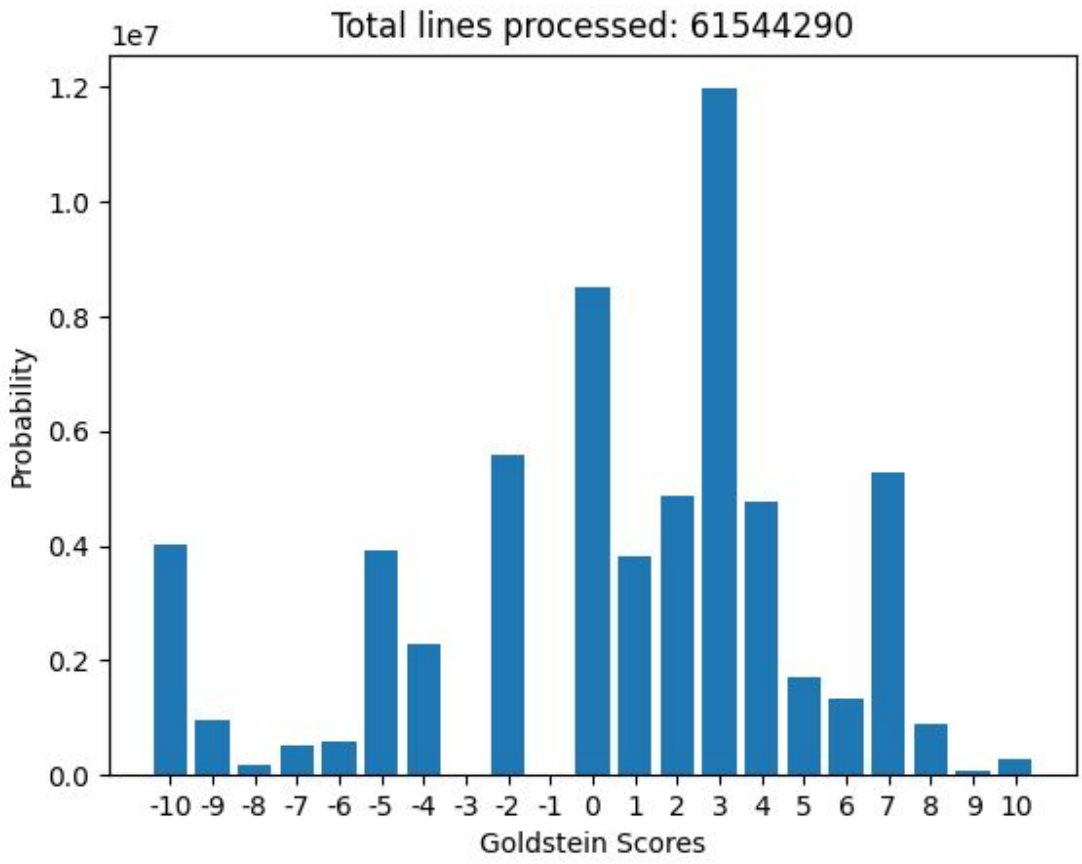
Aws s3 cp s3://stat196k-data-examples/2018.csv.gz - --no-sign-request - This pulls down the data from AWS s3

Gunzip - This unzips the zipped data

Cut --fields 31 - This goes through the tab delimited data and selects the 32nd column of each row

> goldsteinScores.txt - this writes the data out to a text file

3. Plot and interpret the histogram. You'll probably want to download the summary statistics (around 20 numbers) to your personal computer to plot the histogram. Do you notice anything strange?
   I noticed that -1 and -3 seem to have 0 values available. Perhaps these were taken out for our homework assignment as it seems unlikely that 61,544,290 data points left out 2 of 20 possible values.



4. Exactly how many events (rows) are in this data?
   61,544,290

3 - performance

5 pts

Print and interpret the output of `top` while your program is running.

1. What are the bottlenecks?
   Gzip seems to be taking up the most CPU power then cut then aws. However, aws takes up the most memory. Gzip also has the longest time out of all of these programs.
2. Run and time your program on an EC2 instance with more vCPU's and a faster network and show the results of `top` once more. Is the program faster on the more expensive instance?
   Yes but minimally considered that t3.large has a lot more resources, now it is 2m26.927s
3. Are you benefitting from pipeline parallelism?
   Yes
4. What's the bottleneck now?
   Network, because no process is using 100% of it's CPU
5. Compare and comment on the financial cost of using a more expensive instance versus the t2.micro. Is it worth it?
   No, the cost increase is much greater than the performance increase here.