CSC 369 Final Project Report

US Traffic Accidents (2016-2019)

Contributors: Derek Kelley, Joulien Ivanov, Kevin Yoo, Tyler Davis

Q1

Q2 – Severity & Description

The natural language description of the accidents had 2 main formats: “Accident on…” and “Lane blocked due to…”. The program in Q2.scala calculates the average severity of accidents based on 1) the accident description format and 2) the API source.

First, 4 RDDs are created to hold:

1. All accidents
2. Accidents with the description of “Accident on…”
3. Accidents with the description of “Lane blocked due to…”
4. Everything else (2nd & 3rd RDDs subtracted from the 1st)

Then, I split each of the 2nd, 3rd, and 4th RDDs to contain accidents from the Bing API and accidents from the MapQuest API. Finally, I average the severity for each of the 6 RDDs I’ve created (“Accident on…” from Bing, “Accident on” from MapQuest, “Lane blocked…” from Bing, “Lane blocked…” from MapQuest, Everything else from Bing, and Everything else from MapQuest).

One set of experimental results was the average severity per description type **without** the API source distinction. Although the API source distinction doesn’t change the results too drastically (“Accident on…” records still had a lower average severity than “Lane blocked…” records), the source distinction was important to include in the analysis because the natural language descriptions could have entailed completely different results depending on the API they were generated from.

Hadoop was a necessity when performing these calculations because we had a lot of data in our dataset and the distributed capabilities of Hadoop and Spark made the program extremely efficient.

Q3

Q4