Ezana N. Beyenne

MSDS 453, Section 57 2020

Assignment 1: **Focused Web Crawler – Autonomous Vehicle Safety Corpus**

Abstract

The goal of this assignment is to build a focused web crawler gathering as much information about the safety of autonomous vehicles to build a corpus, so that we can conduct further research to determine the general sentiment. Fully autonomous vehicles, like cars and trucks driving without human intervention, are at the top of many industries minds. There are at least forty-six corporations at varying stages of research that are engaged in some form of autonomous vehicles research, (Alessandro, 2020).

The hope is autonomous vehicle technology will usher in a new era where traffic fatalities and street congestion are reduced, the need for parking lots eliminated, and decreasing car ownership (Walker, 2020). This technology will become an inevitable part of our future, yet data, research, and development relating to its safety are still in the early stages and because of this the consumer sentiment is divided (Alessandro, 2020).

Introduction

I am fascinated by the artificial intelligence technology going into building autonomous vehicles, specifically the computer vision, natural language processing, and software engineering necessary to build a complex autonomous vehicle. The technology is expected to have a positive impact on issues ranging from the environment to road congestion. Autonomous vehicles are expected to reduce the instance of impaired driving due to either alcohol or drug impairment. Secondly, it will help reduce emissions and costs by finding the fastest routes to destinations thereby improving fuel efficiency. According to the Department of Transportation and the

National Highway Traffic Safety Administration, human error causes almost 94% of accidents on US roads (Alessandro, 2020).

I started out examining the technology, but as I conducted the research, one thing stood out. As fascinating as this technology is with its disruptive and revolutionary abilities, it will not win public support if safety is not at the forefront of this revolution. In a research conducted by the Governor Highway Safety Association's (GHSA) on "Autonomous Vehicles", the public remains skeptical about the safety of autonomous vehicles. Their surveys showed that only 33% of people would feel comfortable enough to take a ride in a highly automated vehicle and less than a quarter of the people surveyed would buy an autonomous vehicle if it became available.

Literature review

When an Uber-operated autonomous vehicle struck and killed a pedestrian crossing the street on March 18, 2018 in Tempe, Arizona, the safety of this technology came into forefront, since more of this technology is being conducted on public streets (Walker, 2020). Some advocates stated that the technology is immature, is introducing new problems to cities, and the focus should center on public transportation infrastructure development and encouraging people to bike and walk (Walker, 2020).

To ascertain how safe autonomous vehicles are and to develop sound policies regarding their deployment, a framework would be needed to test their safety. A proposed safety evaluation framework would be to deploy autonomous vehicles in real traffic situations and do a statistical comparison with cars being driven by humans (Kalra et al., 2016). Although it might seem like a viable testing framework, it might not be practical, since current traffic accidents are rare and not related to how many miles a car has traveled (Kalra et al., 2016). According to Rand Corporation's research, autonomous vehicles would have to be driven hundreds of millions of

miles, (sometimes hundreds of billions of miles), translating to decades, if not centuries to achieve these goals. Since this is not feasible, Rand Corporation recommends regulations adaptive in nature, so such regulations harness the benefits while mitigating the risks of this rapidly evolving technology (Kalra et al., 2016). There are currently thirty-seven states, along with the District of Columbia, that have created legislation or issued executive orders about autonomous vehicles.

Some of the dangers of autonomous vehicles include: 1) the inability of current infrastructure such as bad road designs, and 2) failure to respond to mistakes by autonomous vehicles that could lead to accidents. There are risks of accidents being caused by bugs in the software, or the cyber-security threats. Lastly, there is the risk of operator-negligence in relying on the autonomous vehicle technology to respond to unexpected circumstances, thereby leading to a slowed reaction time and an accident.

A way of testing and learning how to develop safety measure and evolving policies is to introduce autonomous technology in a phased approach. Hence, there are five autonomous levels with level 5 being full automation.

Table 1. Automation levels

Automation levels	**Autonomous Technology**	**Description**
1	Driver Assistance	Driver controls vehicle, but driver assist features are included.
2	Partial Automation	Has combined automated functions like acceleration and steering, but driver must remain actively engaged.
3	Conditional Automation	Driver is needed, but is not required to actively monitor the environment.
4	High Automation	Vehicle can perform all driving functions, under some conditions. Driver might control the vehicle.
5	Full Automation	Same as High Automation, except vehicle is performing under **all** driving conditions.

Methods

This assignment is to develop a focused web crawler to collect relevant high-quality documents found on web with regards to autonomous vehicle safety. The web crawler uses multi-threading to look through an initial search page, extract all the links from that search page related to autonomous vehicle safety. The web-crawler then proceeds to extract all the html files from those links and then saves those raw html files under the raw directory. The next step, is to extract the title and article information using customized CSS selectors related to the articles, remove script and style elements, remove leading and trailing spaces. I then break multi-headlines into a single line, and drop blank lines. I then saved each of those files into a corpus directory. The next step was to iterate through the corpus object list and save them as a JSON lines file, with each JSON line having an “ID”, “URL”, “TITLE” and “BODY”. The “ID” corresponded to the index of the corpus object in that corpus object list, while the “BODY” contains the text of each file and each of the text files contain more than 300 words. I extracted the first two characters of the domain name, and the first fifteen characters of the filename to create raw and corpus file names. The reasoning is that long file names of the URLs can cause problems on systems that have limited path lengths. I did not have that problem on my Windows 10 laptop.

Results

I am still in the corpus creation stage of my research, as I found that safety is a big concern that can sway the rate of adoption of this technology. Additionally, not all sites provided search pages with links to extract additional URLs, so I had to add a custom URLs to extract information from government and other sites that provided interesting perspectives on

autonomous vehicle safety. In total, I found seventeen interesting sites used to create a high-quality corpus.

Conclusions

Autonomous vehicle technology is not a matter of if it happens, but when. This technology would evolve faster if safety is adopted and assessed at all stages. There are a lot of articles suggesting that safety is at the primary concern from developers, to the consumers, to the regulators. This research assignment's first step is to gather as much information to decide on whether this technology is sufficiently advanced for a wide-scale adoption.

Works Cited

Lori, Alessandro. (2020). Are Self-Driving Cars Safe? Retrieved from https://www.verizonconnect.com/resources/article/are-self-driving-cars-safe/

Allssa, Walker. (2020). Are self-driving cars safe for our cities? Retrieved from https://www.curbed.com/2016/9/21/12991696/driverless-cars-safety-pros-cons

Autonomous Vehicles. Retrieved from https://www.ghsa.org/issues/autonomous-vehicles

Kalra, N., Paddock, Susan M. (2016). Driving to Safety- How many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability. Retrieved from https://www.rand.org/pubs/research_reports/RR1478.html

Allssa, Walker. (2018). It's time to delete Uber from our cities. Retrieved from https://www.curbed.com/transportation/2018/3/23/17153200/delete-uber-cities

Code Output

1. Starting the Web Scraping

```
In [1]: runfile('C:/MSDS453/A1_CorpusCreation/corpus-creation.py', wdir='C:/MSDS453/A1_CorpusCreation')

Starting Web Crawling...

Scraping URL: https://www.wired.com/search/?q=self%20driving%20cars%20safety&page=1&sort=score
Scraping URL: https://www.wired.com/story/news-rules-clear-way-self-driving-cars/
Scraping URL: https://www.wired.com/story/california-self-driving-cars-log-most-miles/
Scraping URL: https://www.wired.com/story/snow-ice-pose-vexing-obstacle-self-driving-cars/
Scraping URL: https://www.wired.com/story/why-are-parking-lots-so-tricky-for-self-driving-cars/
Scraping URL: https://www.wired.com/story/how-self-driving-car-makers-measure-progress/
Scraping URL: https://www.wired.com/story/regulating-self-driving-cars-no-one/
Scraping URL: https://www.wired.com/story/safety-board-faults-tesla-regulators-fatal-2018-crash/
Scraping URL: https://www.wired.com/story/teaching-self-driving-cars-watch-unpredictable-humans/
Scraping URL: https://www.wired.com/story/gms-cruise-rolls-back-target-self-driving-cars/
Scraping URL: https://www.wired.com/story/uber-self-driving-crash-volvo-polestar-1-roundup/
Scraping URL: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety
Scraping URL: https://www.ghsa.org/issues/autonomous-vehicles
Scraping URL: https://www.vox.com/recode/2019/5/17/18564501/self-driving-car-morals-safety-tesla-waymo
Scraping URL: https://www.curbed.com/2016/9/21/12991696/driverless-cars-safety-pros-cons
Scraping URL: https://www.curbed.com/transportation/2018/3/20/17142090/uber-fatal-crash-driverless-pedestrian-safety
Scraping URL: https://www.curbed.com/transportation/2018/3/23/17153200/delete-uber-cities
Scraping URL: https://www.rand.org/pubs/research_reports/RR1478.html
```

2. The list of the current items in the directory

```
A1_CorpusCreation/
    autonomous_vehicles_safety_corpus.jl
    corpus-creation.py
    corpus_dto.py
    threaded_web_crawler.py
    corpus/
        cu-delete-uber-cit.txt
        cu-driverless-cars.txt
        cu-uber-fatal-cras.txt
        gh-autonomous-vehi.txt
        nh-automated-vehic.txt
        ra-RR1478.txt
        vo-self-driving-ca.txt
        wi-california-self.txt
        wi-gms-cruise-roll.txt
        wi-how-self-drivin.txt
        wi-news-rules-clea.txt
        wi-regulating-self.txt
        wi-safety-board-fa.txt
        wi-snow-ice-pose-v.txt
        wi-teaching-self-d.txt
        wi-uber-self-drivi.txt
        wi-why-are-parking.txt
    raw_data/
        cu-delete-uber-cit.html
        cu-driverless-cars.html
        cu-uber-fatal-cras.html
        gh-autonomous-vehi.html
        nh-automated-vehic.html
        ra-RR1478.html
        vo-self-driving-ca.html
        wi-california-self.html
        wi-gms-cruise-roll.html
        wi-how-self-drivin.html
        wi-news-rules-clea.html
        wi-regulating-self.html
        wi-safety-board-fa.html
        wi-snow-ice-pose-v.html
        wi-teaching-self-d.html
        wi-uber-self-drivi.html
        wi-why-are-parking.html
    __pycache__/
        corpus_dto.cpython-37.pyc
        threaded_web_crawler.cpython-37.pyc

End of Web Crawling....
```