

Публичные репозитории данных для машинного обучения на примере GitHub

















Способы доступа к данным

- Так как GitHub является публичным репозиторием то большинство пользователей выкладывают туда данные для публичного использования. Поэтому для доступа к этим данным нужно всего лишь зарегистрироваться на GitHub.com. Стоит отметить, что регистрация и использование репозитория совершенно бесплатные.



- Далее нужно найти подходящие нам данные. Для этого можно воспользоваться поиском по репозиторию с такими ключевыми словами, как “Data for machine learning” либо поиском в Google или других поисковиках с запросами, например: “data for machine learning GitHub”, “open data repository GitHub”
- Для конкретного примера я нашел репозиторий данных для машинного обучения пользователя под именем “[fivethirtyeight](#)”

- В его репозитории представлено большинство наборов данных на различные темы

Branch: master	New pull request	Create new file	Upload files	Find file	Clone or download
 ritchieking committed on GitHub Update README.md Latest commit a1ca215 17 days ago					
 airline-safety	Update README.md	2 years ago			
 alcohol-consumption	Revert "Update drinks.csv"	2 years ago			
 avengers	add avengers data	2 years ago			
 bad-drivers	add bad drivers data	2 years ago			
 bechdel	format email address	3 years ago			
 biopics	for race_known as unknown, make subject_race blank, not White	a year ago			
 births	fix README	6 months ago			
 bob-ross	cleaned up bob ross clustering script	3 years ago			
 buster-posey-mvp	add buster posey mvp scripts	a year ago			
 classic-rock	fixed entries with #REF! excel errors on two rows of classic-rock-raw...	2 years ago			
 college-majors	update college majors readme	2 years ago			
 comic-characters	Update README.md	2 years ago			
 comma-survey-data	clean comma survey data	3 years ago			
 congress-age	renamed congress_terms.csv to congress-terms.csv	3 years ago			
 cousin-marriage	Merge branch 'master' of https://github.com/fivethirtyeight/data	2 years ago			

Описание данных

- После того как вы выбрали интересующую вас тематику данных, вы можете посмотреть их описание, которое как правило находится в файле readme, а в .csv файле находятся непосредственно сами данные

fivethirtyeight / data

Watch

744

Star

4,533

Fork

1,593

Code

Issues11

Pull requests7

Projects0

Wiki

Pulse

Graphs

Branch: master

data / bad-drivers /

Create new file

Upload files

Find file

History

andrewflowers

add bad drivers data

Latest commit 5ec94e2 on Nov 1, 2014

..

README.md

add bad drivers data

2 years ago

bad-drivers.csv

add bad drivers data

2 years ago

- Так же как правило в описании данных есть ссылка на источник этих данных, в данном случае это article

Bad drivers

Data from the article [Dear Mona, Which State Has The Worst Drivers?](#)

Variable	Source
State	N/A
Number of drivers involved in fatal collisions per billion miles	National Highway Traffic Safety Administration, 2012
Percentage Of Drivers Involved In Fatal Collisions Who Were Speeding	National Highway Traffic Safety Administration, 2009
Percentage Of Drivers Involved In Fatal Collisions Who Were Alcohol-Impaired	National Highway Traffic Safety Administration, 2012
Percentage Of Drivers Involved In Fatal Collisions Who Were Not Distracted	National Highway Traffic Safety Administration, 2012
Percentage Of Drivers Involved In Fatal Collisions Who Had Not Been Involved In Any Previous Accidents	National Highway Traffic Safety Administration, 2012
Car Insurance Premiums (\$)	National Association of Insurance Commissioners, 2011
Losses incurred by insurance companies for collisions per insured driver (\$)	National Association of Insurance Commissioners, 2010

Получение набора данных

- Как я уже говорил GitHub является публичным и бесплатным репозиторием, поэтому для получения набора данных его можно “клонировать” себе в репозиторий, либо загрузить в виде .zip архива

The screenshot shows the GitHub repository page for `fivethirtyeight / data`. At the top, there are statistics: 621 commits, 3 branches, 0 releases, 25 contributors, and MIT license. Below this, there are tabs for Code, Issues (11), Pull requests (7), Projects (0), Wiki, Pulse, and Graphs. A dropdown menu is open for the 'Clone or download' button, showing options to 'Clone with SSH' (with a red circle around the button) and 'Download ZIP' (with a red circle around the button). The repository list shows files like `airline-safety`, `alcohol-consumption`, `avengers`, and `bad-drivers`.

После загрузки, данные можно импортировать, например в MS Excel, так как в большинстве случаев они представлены в виде csv

