Публичные репозитории данных для машинного обучения на примере GitHub

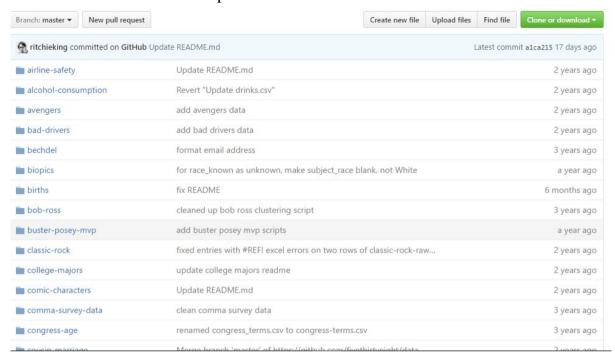
Способы доступа к данным

• 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.



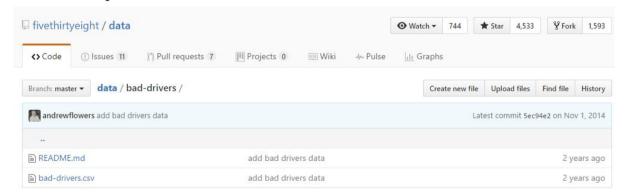
- Далее нужно найти подходящие нам данные. Для этого можно воспользоваться поиском по репозиторию с такими ключевыми словами, как "Data for machine learning" либо поиском в Google или других поисковиках с запросами, например: "data for machine learning GitHub", " open data repository GitHub"
- Для конкретного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

• В его репозитории представлено большинство наборов данных на различные темы



Описание данных

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

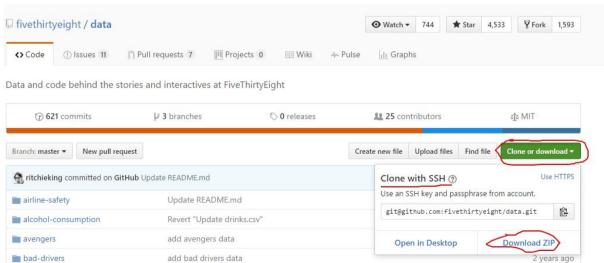


 Так же как правило в описании данных есть ссылка на источник этих данных, в данном случае это article



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



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