СБЕР УНИВЕРСИТЕТ

Введение в Data Science

Программа «Перезапуск»

План занятия

1. Работа инженера, аналитика, исследователя
2. IDE, Colab, Kaggle
3. Linux и основы работы в консоли
4. Git

DS, DE и DA в общем и целом

Data Scientist

- Обучает модели
- Генерирует фичи
- Оптимизирует метрики качества

Data Engineer

- Формирует данные для DS и DA
- Выстраивает потоки данных, ETL –процессы (extract transform load)
- Отвечает для хранилища данных

Data Analyst

- Готовит разного рода аналитику
- Формирует отчеты
- Разрабатывает дашборды

DS, DE и DA инструменты

Data Scientist

- Jupyter Notebook + Python
- Rstudio + R
- SQL

Data Engineer

- Jupyter Notebook + Python
- HADOOP
- SPARK
- SQL
- Airflow

Data Analyst

- Jupyter Notebook + Python
- Excel
- BI
- Tableau

DS, DE и DA навыки

Data Scientist

- Python
- Статистика
- Знание ML Алгоритмов
- Написание SQL запросов
- Презентационные навыки

Data Engineer

- Написание SQL запросов, процедур, сложных джоинов
- Администрирование БД
- Миграции БД
- Python, Java, Scala

Data Analyst

- Python
- Написание SQL запросов
- Презентационные навыки
- Визуализация

Integrated Development Environment

Как и где писать код?

1. Vim
2. Nano
3. Atom
4. Sublime

Anaconda

1. https://www.anaconda.com/products/individual
2. Jupyter Notebook
3. Jupyter Lab и Jupyter Hub
4. Spyder

Pycharm & DataSpell

IDE для написания кода на python от компании JetBrains

- https://www.jetbrains.com/pycharm
- Синтаксический анализатор кода, дебагер
- Встроенный интерфейс для работы БД
- Удобная навигация в коде
- При желании можно работать только клавишами

VSCode

IDE полиглот от компании Microsoft

1. https://code.visualstudio.com/
2. Бесплатная и при этом не уступает PyCharm
3. С помощью плагинов установить необходимые опции, которых нет в базовой версии

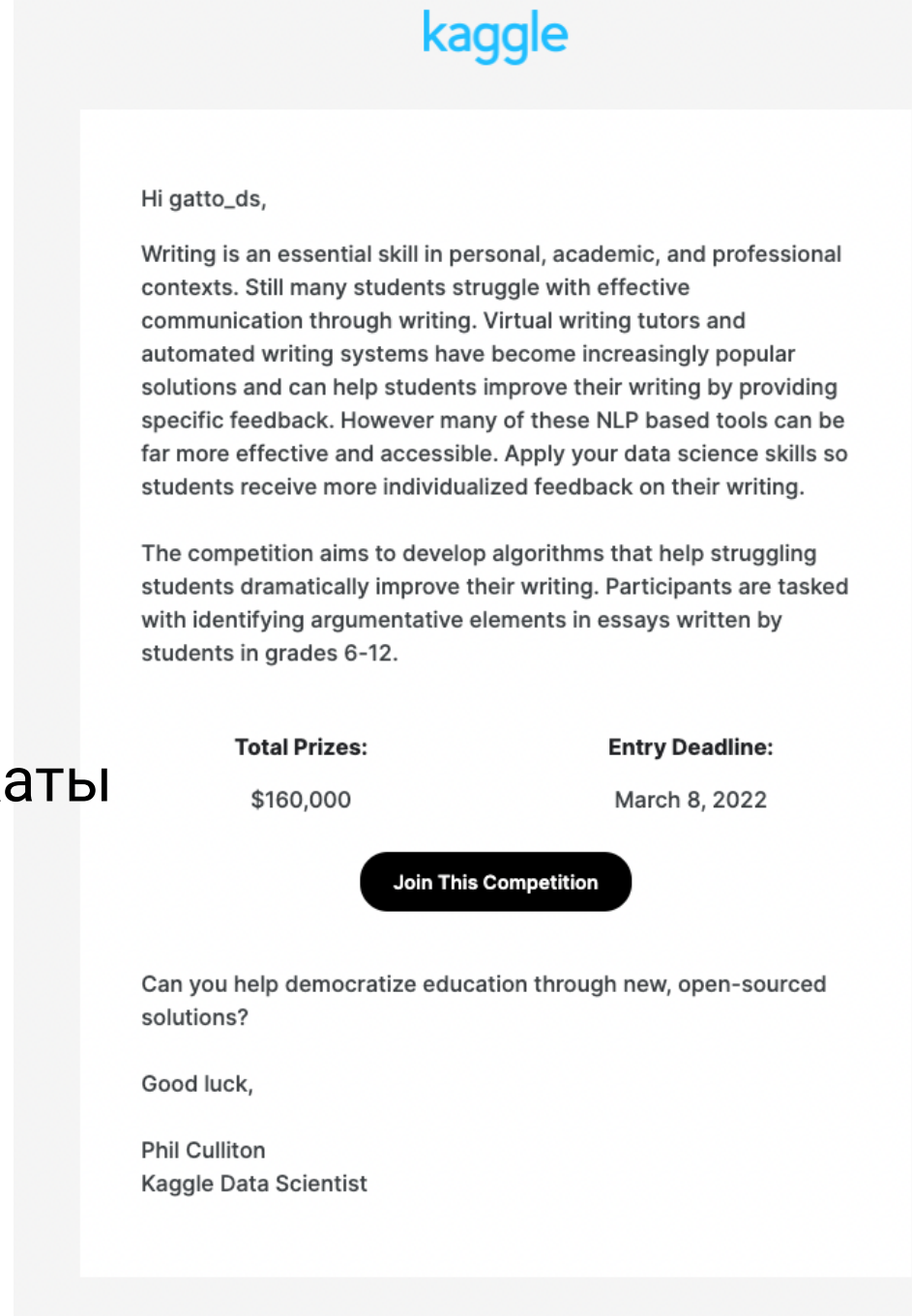
Colab

Облачная IDE для Data Science задач от компании Google

1. https://colab.research.google.com/
2. Можно работать на слабом компьютере используя мощности Colab
3. Можно производить вычисления с использованием GPU
4. Удобно делиться кодом

Kaggle

1. Соревнования и интересные решения
2. Различные датасеты
3. Облачная IDE для вычислений как в Colab
4. Небольшие бесплатные курсы и Сертификаты



Командная строка Linux

Функционал	Команда
Перейти на уровень выше	cd ..
Перейти на уровень ниже	cd some_directory
Перейти в домашнюю директорию	cd ~
Создать файл	touch filename
Создать файл в редакторе	vim filename
Посмотреть текущие процессы	ps
Посмотреть все процессы	ps aux
Найти подстроку в тексте	grep
Найти питоновский процесс	ps aux | grep python
Остановить процесс	kill pid процесса

Командная строка Linux – базовые команды

- Параметры передаются через дефис, например: -r
- Проверка загрузки памяти и процессов TOP
- Отправка запросов curl
- Скачивание файлов wget
- Очистка терминала clear

```
root@flightclubber-bot:~# curl http://www.sberbank.ru/
<html>
<head><title>301 Moved Permanently</title></head>
<body>
<center><h1>301 Moved Permanently</h1></center>
<hr><center>nginx</center>
</body>
</html>
root@flightclubber-bot:~#
```

```
root@flightclubber-bot:~# wget https://wordpress.org/latest.zip
--2021-03-23 18:31:51--  https://wordpress.org/latest.zip
Resolving wordpress.org (wordpress.org)... 198.143.164.252
Connecting to wordpress.org (wordpress.org)|198.143.164.252|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 16865632 (16M) [application/zip]
Saving to: 'latest.zip'

latest.zip                          100%[=============================================================================>]  16.08M  12.4MB/s    in 1.3s

2021-03-23 18:31:53 (12.4 MB/s) - 'latest.zip' saved [16865632/16865632]

root@flightclubber-bot:~#
```

Командная строка Linux – базовые команды

- Запуск нескольких команд
  - ; последовательно *sudo apt update ; sudo apt upgrade*
  - && последовательно если первая выполнилась успешно

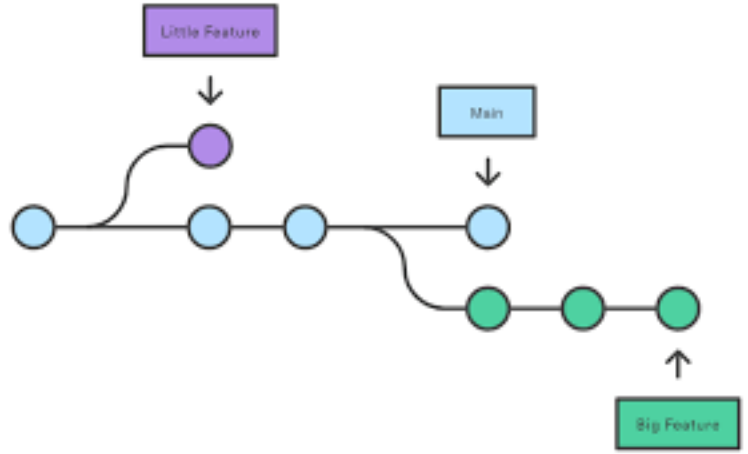
- Pipe | когда результаты выполнения команды передаются другой

  ls -la| tee test

```
root@flightclubber-bot:~# ps aux | grep python
root         661  0.0  1.2  29256 12256 ?        Ss   Mar13   0:00 /usr/bin/python3 /usr/bin/networkd-dispatcher --run-startup-triggers
root         740  0.0  1.2 108076 12908 ?        Ssl  Mar13   0:00 /usr/bin/python3 /usr/share/unattended-upgrades/unattended-upgrade-shutdown --wait
-for-signal
root       19167  0.0  5.5 186300 56252 ?        Sl   Mar13   1:58 python3 app.py
root       81770  0.0  0.0   8160   740 pts/0    S+   18:25   0:00 grep --color=auto python
root@flightclubber-bot:~#
```

Систем контроля версий GIT

1. Скачать git можно отсюда https://git-scm.com/
2. Где создать репозиторий:
   - Github.com
   - Bitbucket.org
   - Gitlab.com



- Основные команды:
  - git init
  - git add filename
  - git commit -m "first commit"
  - git branch -M main
  - git remote add origin https://github.com/user/repo.git
  - git push -u origin main

Домашнее задание

1. Попробовать запустить ipynb ноутбуки в разных IDE и понять, где больше нравится работать
2. Установить git, создать аккаунт в Github, создать тестовый проект и запушить код, создать отдельную ветку внести изменения, сделать pull request и merge

Рекомендуемые материалы

1. Канал PyCharm в youtube https://www.youtube.com/channel/UCak6beUTLlVmf0E4AmnQkmw
2. VSCode и его плагины для написания кода на python https://youtu.be/WkUBx3g2QfQ
3. Небольшой курс про Git https://youtube.com/playlist?list=PLDyvV36pndZFHXjXuwA_NywNrVQO0aQqb
4. Статьи + код https://paperswithcode.com/
5. Научные статьи по DS https://arxiv.org/

Ваши вопросы