

Город Актау

nFactorial Incubator 2023

**Задание отборочного тура**

**nFactorial Incubator 2023**

**Задача для апликанта:**

**- реализовать Bigram Language Model**

**- добавить инструкцию по запуску**

**Тема: «****Написание программы Bigram Language Model на любом из удобных языков программирования»**

Работа: \_\_\_\_\_Сейтджан Айзере Оразбекқызы\_\_\_\_\_

(ФИО) email: seitjan,aktau@gmail.com

Проверятющий: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(ФИО) email: admin@nfactorial.school

Актау, 2023

# **СОДЕРЖАНИЕ**

1. **Задача для апликанта……………………………………………………………………………..** [3](#_heading=h.gjdgxs)
2. **Процесс работы………………………………………………………………………………....,** [4](#_heading=h.4d34og8)
3. **Запуск…………………………………………………………………………………….............** 6
4. **Код……………..………………………………………………………………………………......** [7](#_heading=h.4d34og8)
5. **Как открыть код………………………………………………………………………………....** 8
6. **Онлайн материалы……………………………………………………………………………....** [8](#_heading=h.4d34og8)

# **ЗАДАЧА ДЛЯ АПЛИКАНТА**

**Задание отборочного тура**

Привет, аппликант nFactorial Incubator! Ваша задача сегодня - реализовать **Bigram Language Model**

**Общие заметки**

- Вы можете выбрать любой язык, на котором вы сможете продемонстрировать умение писать структурированный код

- Приложите инструкцию по запуску вашей программы и запускаемый файл (filename.py, filename.java и так далее)

- Можете использовать command line interface, веб страницу на локалкe, использовать библиотеки языков для демонстрации работы вашей программы

- Даже если вы не закончили проект полностью, не бойтесь отправлять задание на 50%

**Функционал**

**Биграмма** букв - это пара букв, которые стоят рядом в слове. Например, в слове "кот" биграммы это "^к", "ко", "от", и "т$" **(^ и $ — начало и конец слова.)**

Мы используем биграммы букв, чтобы лучше понимать, какие пары букв чаще всего встречаются в словах. Это может быть полезно, например, для предсказания следующего слова в предложении или определения языка, на котором написан текст.

Представьте что вам нужно придумать имя для вашего ребенка. У вас есть список уже существующих имен. Ваша задача: придумать **новое имя!**

Вам даны данные (dataset) в следующем файле:

[names.txt]

**Логика модели:**

- Прочитайте данные с файла в структуры данных удобных для высчитывания вероятностей

- Высчитайте вероятность всех существующих биграмм (строим выборку)

- Возьмите букву из выборки которое может придти как первая буква имени (рандомно)

- Продолжать тянуть следующую букву из выборки, таким образом генерируя имя. Это нужно делать пока вы не вытянули конец имени.

**Пользователь**

- Generate function - возможность создавать имя.

- получить таблицу визуализирующие вероятности биграмм

**Бонус**

- использование и оптимизирование с помощью библиотеки **pytorch**

- визуализация таблицы в картинку (подойдет любая библиотека)

**2x Бонус**

Создать нейронную сеть которая учится на выборке. Гугл в помощь!

Есть вопросы? Напишите нам на email: nfactorialincubator2022@gmail.com

**Deadline: 05.05.2023, 23:59**

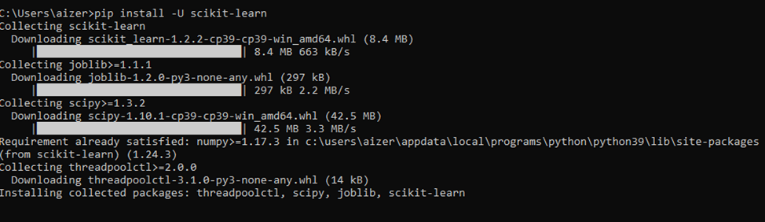
**Чтобы сдать работу:**

Заполните форму: [https://form.typeform.com/to/s0kO34KF]

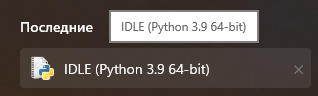
# **ПРОЦЕСС РАБОТЫ**

1. Изображение выглядит как текст

   Автоматически созданное описаниеЯ выбрала язык программирования – Python. Хотела С++, но в первом быстрее разобраться, а из-за экзаменов (МЭСКов) я решила не рисковать. Потом мне надо было скачать нужные libraries, поэтому открыв командную строку, я скачала pandas, -u scikit-learn, которые к сожалению не пригодились ☹



1. Я решила для таблицы использовать pygame, потому что есть опыт работы с данным модулем. А программу для написания и запуска, я выбрала IDLE

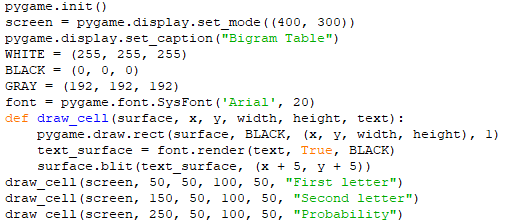
 Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

1. Я использовала модули Random и Pygame



1. Cоздала окно для таблицы



1. Тут я определила биграмы из файлы names.txt

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

1. Изображение выглядит как текст, снимок экрана, Шрифт, число

   Автоматически созданное описаниеПоследнее, я пыталась вставить в таблицу данные из биграма

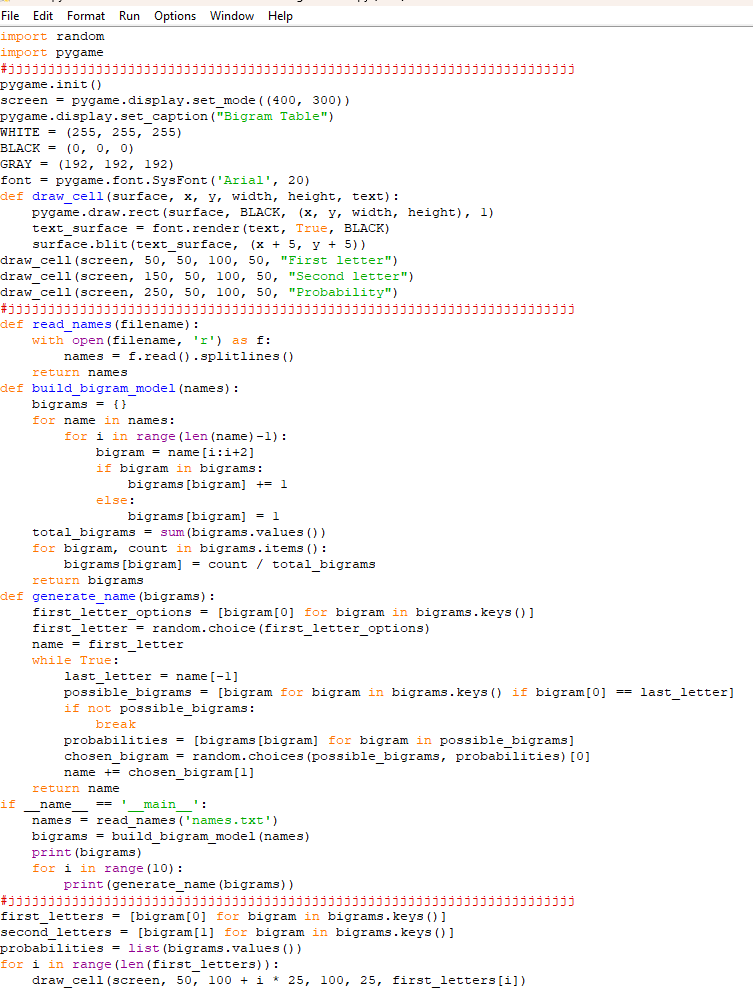
# **ЗАПУСК**

****Изображение выглядит как текст, электроника, компьютер, дисплей

Автоматически созданное описаниеМоя программа недоработана из-за недостатка времени от экзаменов, поэтому пока только такие результаты:

В желтом button находятся bigrams(снизу нее синие символы – это биграмы). Таблица не показывается к сожалению

# **КОД МОЕЙ ПРОГРАММЫ**



**Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание**

# **КАК ОТКРЫТЬ КОД МОЕЙ ПРОГРАММЫ**

Надо зайти на IDLE и нажать на File 🡪 Open… 🡪 выбрать мою папку с файлом .py и открыть

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт, дизайн

Автоматически созданное описание

# **ГУГЛЬ В ПОМОЩЬ!**

**Онлайн ресурсы**

1. <https://stackoverflow.com/questions/73893586/how-to-put-the-names-in-the-code-from-txt-document-for-it-to-read>
2. <https://towardsdatascience.com/text-analysis-basics-in-python-443282942ec5> - моя начальная идея
3. Ну я использовала много сайтов, так что мой код это франкенштейн