Модель для определения национальности по именам и фамилиям

Иванова Дарья
Студент курса
Машинное обучение: фундаментальные инструменты и практики

Содержание

Постановка задачи

1

Постановка задачи

Задача:

Определить национальность по фамилии и имени

1. Сгенерировать* набор данных: имя, фамилия, национальная принадлежность
2. Провести анализ сгенерированных данных
3. Выбрать модель для классификации
4. Разделить данные на тренировочное и тестовое подмножества
5. Обучить модель на тренировочных данных
6. Рассчитать точность классификации для тестового подмножества

Набор данных:

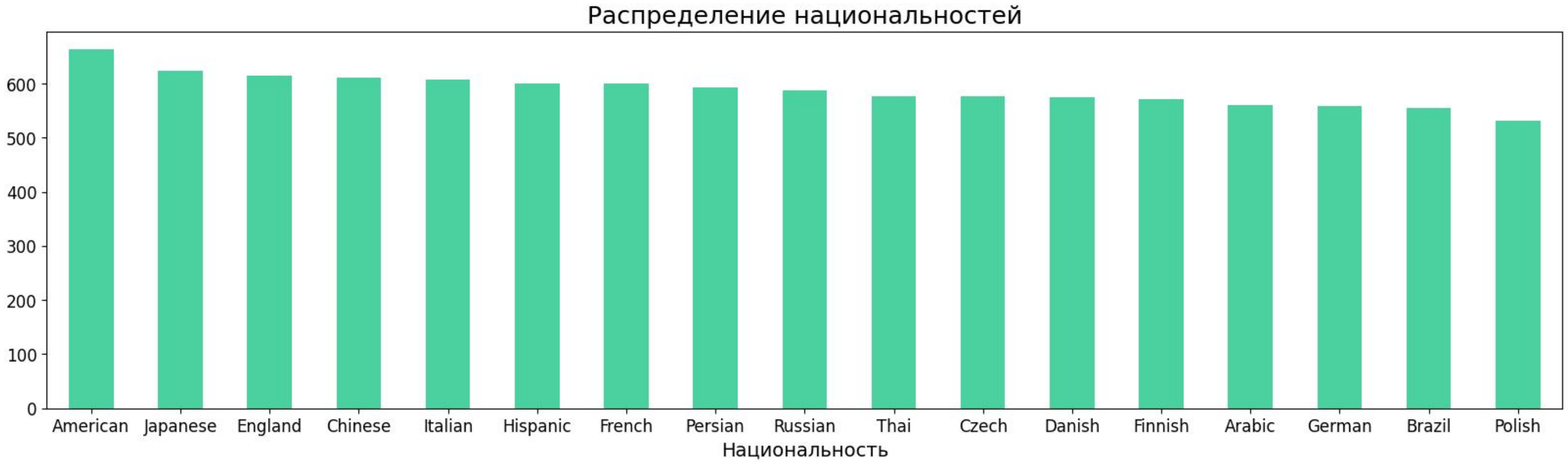
№	Имя	Фамилия	Национальность
0	Rhys	May	England
1	Arkhip	Bykov	Russian
2	Luke	Curl	American
3	Maya	Perry	England
4	Rudolph	Gorshkov	Russian
…	…	…	…
9999	Onni	Salo	Finnish

* https://www.fakenamegenerator.com/

Анализ исходных данных

2

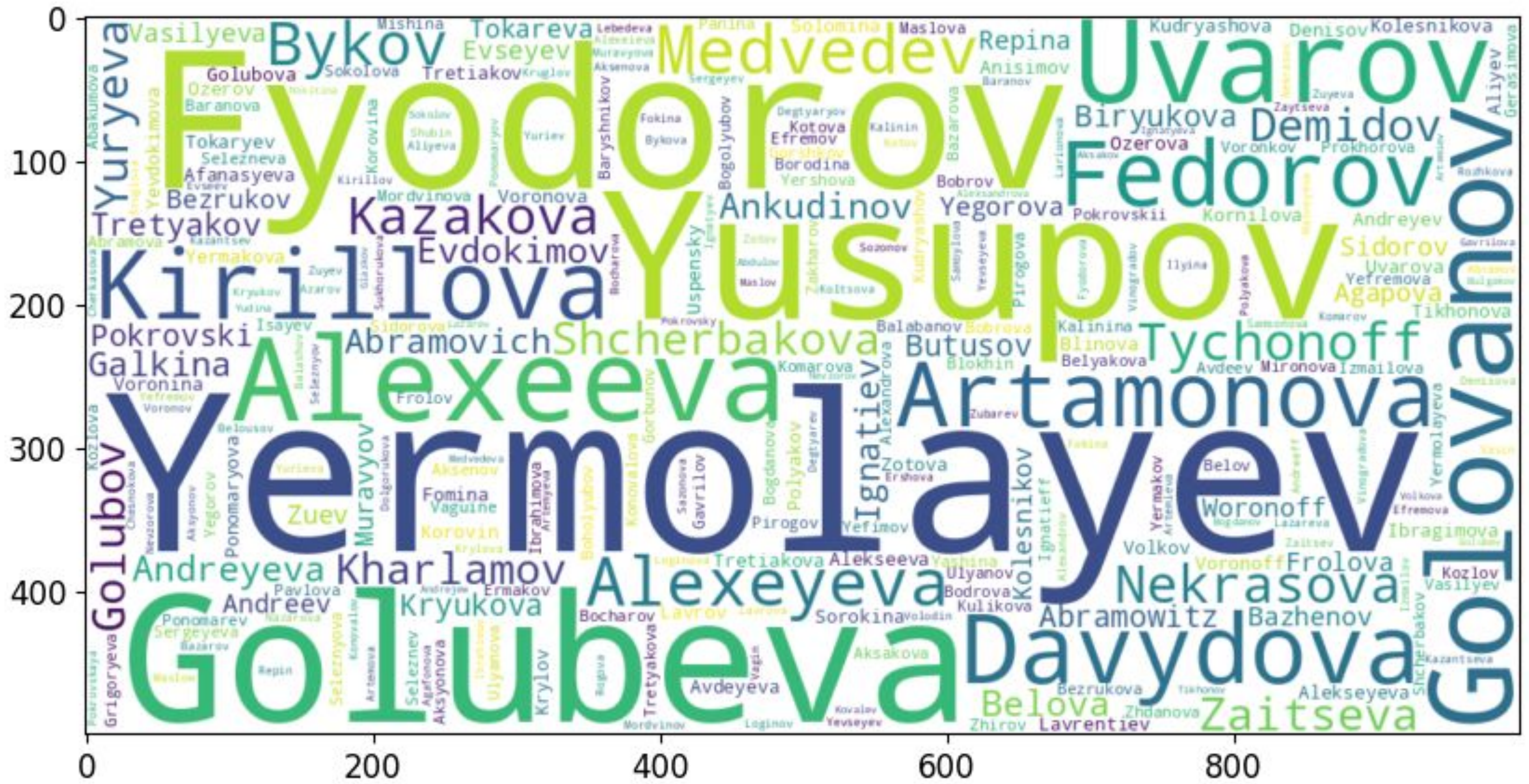
Анализ исходных данных



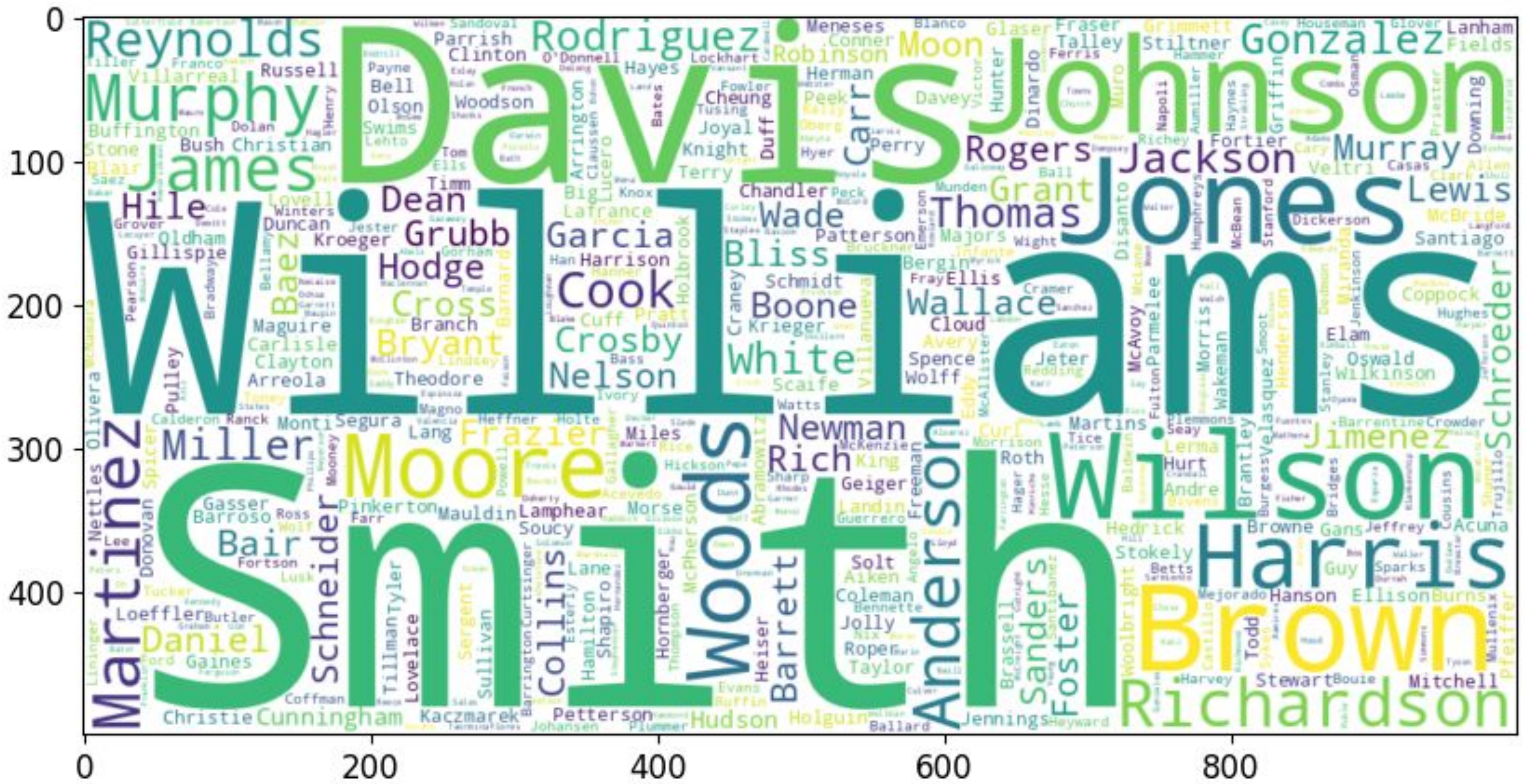
Набор данных содержит **10 000** строк, пропусков не обнаружено.

Всего **17** национальностей, которые распределены сбалансированно.

Облако слов



Облако слов для русских фамилий



Облако слов для американских фамилий

Токенизация слов

Для токенизации имен и фамилий будем использовать 2- и 3-граммы.

Ниже приведены наиболее часто встречающиеся в фамилиях 3-граммы для разных национальностей.

Топ-10 3-грамм по национальностям

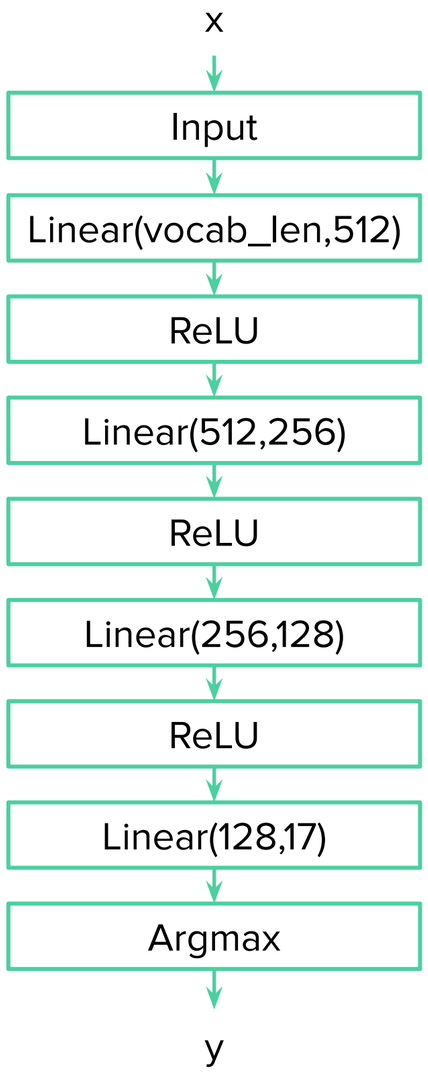
	Russian		American		Finnish		German		French		Hispanic	
N°	3-грамма	Частота	3-грамма	Частота	3-грамма	Частота	3-грамма	Частота	3-грамма	Частота	3-грамма	Частота
1	ova	202	son	58	nen	146	man	48	eau	60	rre	28
2	rov	82	ill	28	ine	70	Sch	43	ier	40	era	23
3	kov	81	lli	20	one	36	ann	42	ard	28	ado	22
4	eva	67	Wil	19	ala	32	ler	39	ois	26	arr	19
5	yev	65	ter	18	ain	31	ber	34	lle	26	ill	19
6	nov	65	ers	18	ari	28	ger	29	our	25	err	18
7	lov	54	arr	17	ane	27	sch	25	Cha	25	ero	18
8	oro	51	ton	15	kka	24	ner	25	ill	25	ara	18
9	ono	42	ing	15	ila	24	ter	23	aul	20	rez	18
10	mov	40	Smi	14	ola	20	ste	18	Mar	19	lla	17

Подход к решению задачи

3

Выбор модели для обучения

Полносвязная нейронная сеть



Разбиение данных:

80% - обучающее подмножество, 20% - тестовое подмножество

Параметры модели:

Функция потерь - CrossEntropyLoss (перекрестная энтропия)

Оптимизатор - Adam

Подход к решению задачи

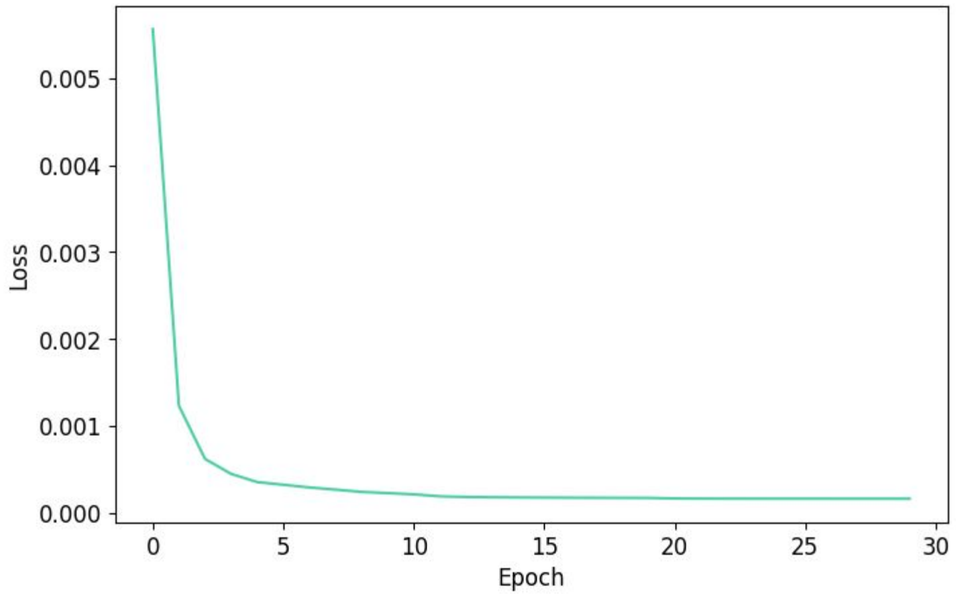
Для определения национальности сравним два подхода: использование модели, обученной только на наборе фамилий, и использование двух моделей с объединением полученных результатов классификации

Обучение модели

4

Обучение модели на наборе фамилий

Число эпох для обучения — 30, коэффициент скорости обучения — 0,005



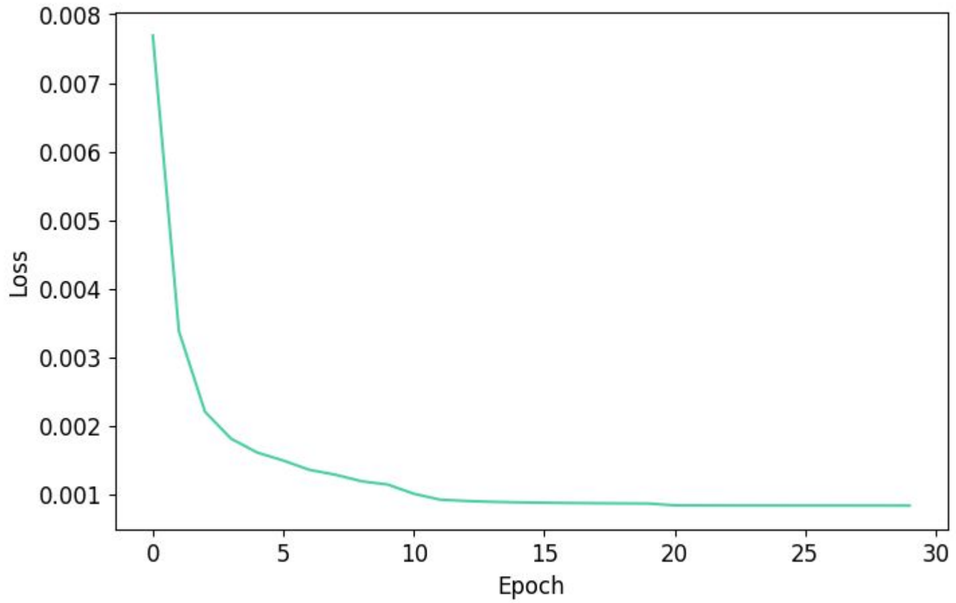
Зависимость функции потери от эпохи

89%

ТОЧНОСТЬ

Обучение модели на наборе имен

Число эпох для обучения — 30, коэффициент скорости обучения — 0,005



Зависимость функции потери от эпохи

74%

ТОЧНОСТЬ

Тестирование модели

5

Тестирование модели

Модель, обученная на наборе фамилий

Фамилия	№	Национальность	Вероятность
Bezrukov	1	Russian	100%
	2	Japanese	0%
	3	Hispanic	0%
	...	...	...
Komarova	1	Russian	100%
	2	Arabic	0%
	3	Japanese	0%
	...	...	...
Hauta-aho	1	Finnish	99,99%
	2	American	0,01%
	3	German	0%
	...	...	...
Chu	1	Chinese	99,97%
	2	American	0,02%
	3	Japanese	0%
	...	...	...

Модель, обученная на наборе имен

Фамилия	№	Национальность	Вероятность
Nicephorus	1	Russian	100%
	2	French	0%
	3	Hispanic	0%
	...	...	...
Arina	1	Japanese	99,52%
	2	Hispanic	0,31%
	3	Finnish	0,07%
	...	...	...
Anniina	1	Finnish	100%
	2	Czech	0%
	3	Japanese	0%
	...	...	...
Huan Yue	1	Chinese	100%
	2	Hispanic	0%
	3	Arabic	0%
	...	...	...

Полученные результаты

Полученные результаты

Модель, обученная на наборе фамилий и имен

Фамилия	Имя	Предсказанная национальность	Произведение вероятностей
Bezrukov	Nicephorus	Russian	100%
Komarova	Arina	Russian	0,03%
Hauta-aho	Anniina	Finnish	99,99%
Chu	Huan Yue	Chinese	99,97%

93%
ТОЧНОСТЬ

Рассмотренный способ определения национальности показывает лучший результат по сравнению с моделью, обученной только на наборе фамилий.

В дальнейшем планируется протестировать модель на русскоязычных данных, рассмотреть другие методы токенизации и типы моделей, позволяющие повысить точность классификации.

https://colab.research.google.com/drive/1AIY_Ym72z8hGoWDIUYxRkeIXCWvUltRz?usp=sharing

Спасибо за внимание

Иванова Дарья
Студент курса
Машинное обучение: фундаментальные инструменты и практики