

The slide features decorative illustrations of various leaves in the corners. The top-left and top-right corners contain delicate, light green line-art leaves. The bottom-left and bottom-right corners feature more prominent, solid-colored leaves, including a large green monstera leaf and a bright yellow autumn leaf. The main title is centered in a large, black, serif font.

Русская силлаботоника и машинное обучение

Борис Орехов (НИУ ВШЭ, ИРЛИ РАН)



Содержание

01. Машинное обучение

Что такое машинное обучение? Можно ли использовать его как исследовательский инструмент?

02. Стих и проза

Чем стих отличается от прозы? Можно ли доказать определение?



03. Силлаботоника и ударение

Ударение позволяет различать размеры

04. Метр без ударений



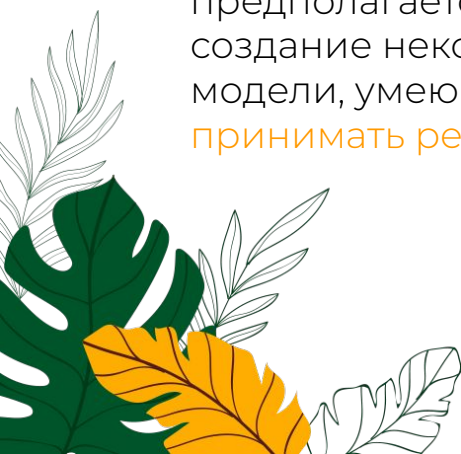
Можно ли различить метры без ударений?



Машинное обучение

Концепция,

в которой
предполагается
создание некоторой
модели, умеющей
принимать решения




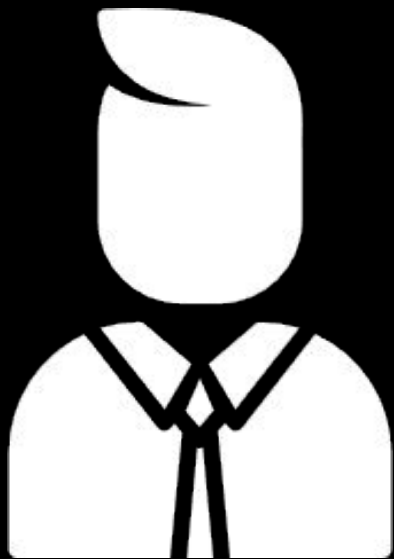
Но при этом

алгоритм сам
«обозревает»
материал (обучающие
данные), «обучается»

Статистически

выводит
закономерности,
производит
распознавание
паттернов
посредством
статистической
индукции





Почему машинное

Человек-эксперт (с некоторыми оговорками) не участвует в создании модели, принимающей решения



Стих и проза



Определения стиха



01

Михаил Гаспаров

- стих — это прежде всего речь, четко расчлененная на относительно короткие отрезки, **соотносимые и соизмеримые** между собой. Каждый из таких отрезков тоже называется “стихом” и на письме обычно выделяется в отдельную строку



02

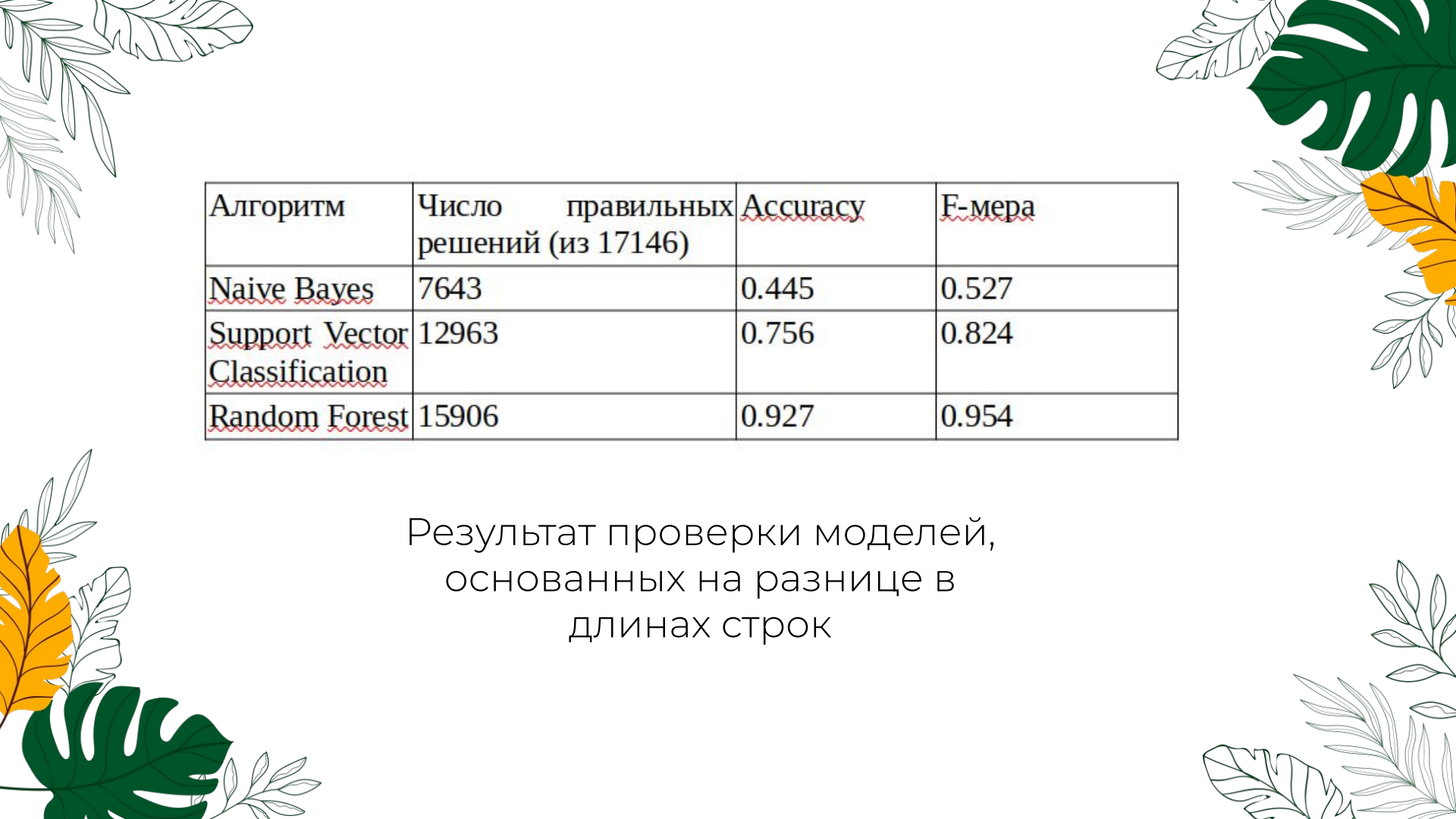
Максим Шапир

- стих — это система сквозных принудительных парадигматических членений, структурирующих дополнительное измерение текста



Обучающие данные

№ текста и строки	Длина строки n	Длина строка n-1	Длина строка n-2	Длина строка n-3	Длина строки n-4	Длина строка n-1	Длина строки n+2	Длина строки n+3	Длина строки n+4	Статус
256, 42	11	0	10	11	10	10	11	10	0	стих
256, 43	10	11	0	10	11	11	10	0	11	стих
256, 44	11	10	11	0	10	10	0	11	10	стих
256, 45	10	11	10	11	0	0	11	10	11	стих
1020, 12	792	593	635	515	111	0	5	167	816	проза

The slide features decorative illustrations of various leaves in the corners. Top-left: a cluster of light green, feathery leaves. Top-right: a large, dark green monstera leaf and a smaller yellow leaf. Bottom-left: a large dark green monstera leaf and a yellow leaf. Bottom-right: a cluster of light green, feathery leaves.

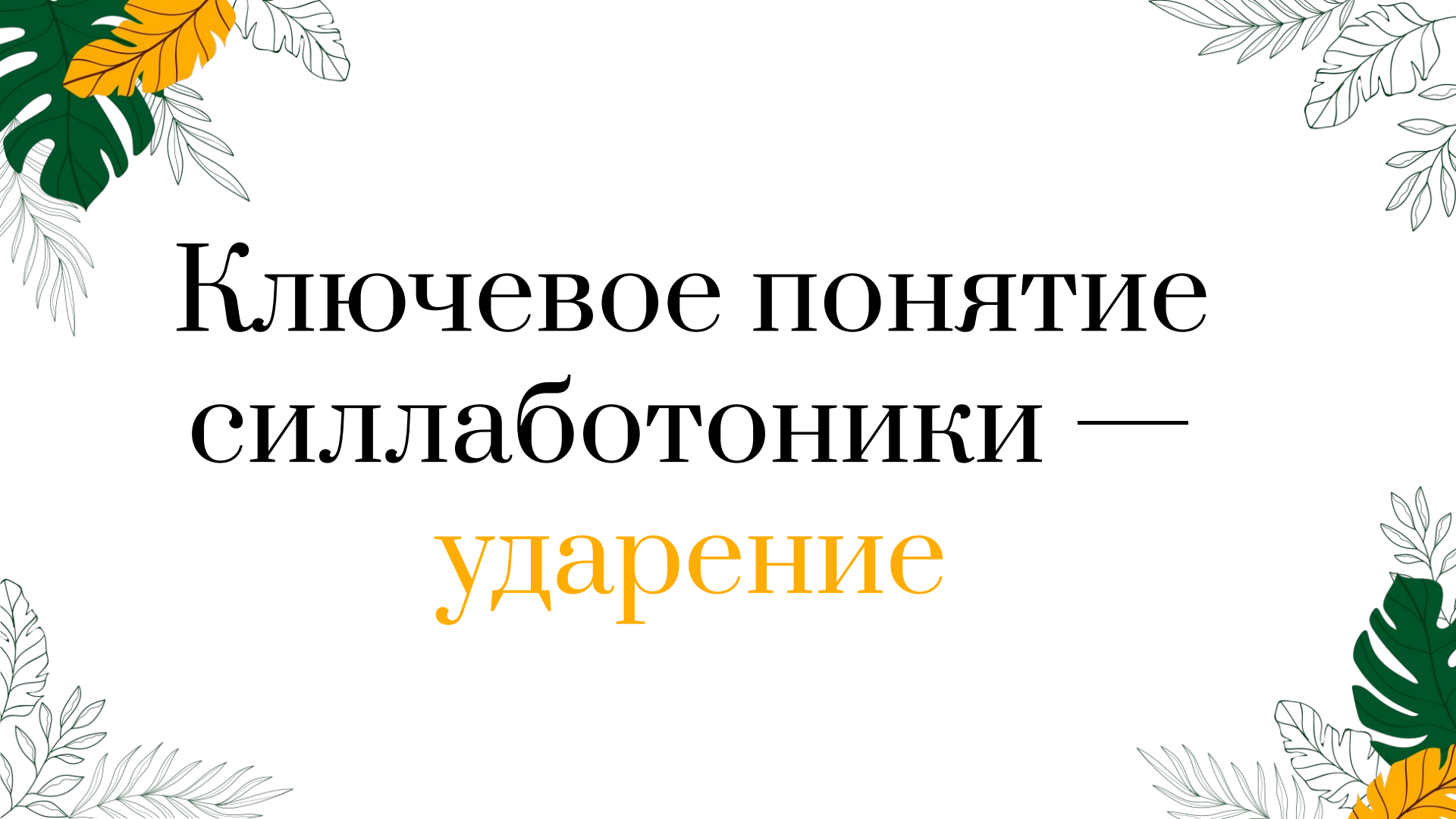
Алгоритм	Число правильных решений (из 17146)	<u>Accuracy</u>	<u>F-мера</u>
<u>Naive Bayes</u>	7643	0.445	0.527
<u>Support Vector Classification</u>	12963	0.756	0.824
<u>Random Forest</u>	15906	0.927	0.954

Результат проверки моделей,
основанных на разнице в
длинах строк



Ударение и метр



The image features decorative illustrations of various leaves in the corners. Top-left: A cluster of leaves including a large green monstera leaf, a smaller yellow leaf, and several white line-art leaves. Top-right: A branch with several white line-art leaves. Bottom-left: A branch with several white line-art leaves. Bottom-right: A cluster of leaves including a large green monstera leaf, a smaller yellow leaf, and several white line-art leaves.

Ключевое понятие силлаботоники — ударение



Нейросети умеют классифицировать текст

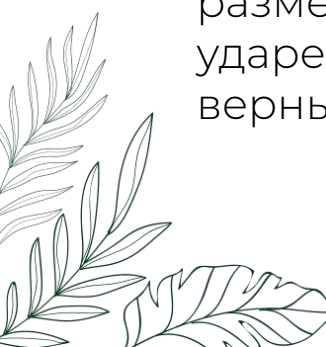

Сможет ли сеть правильно классифицировать строки по размерам, не имея данных об ударении?

01.

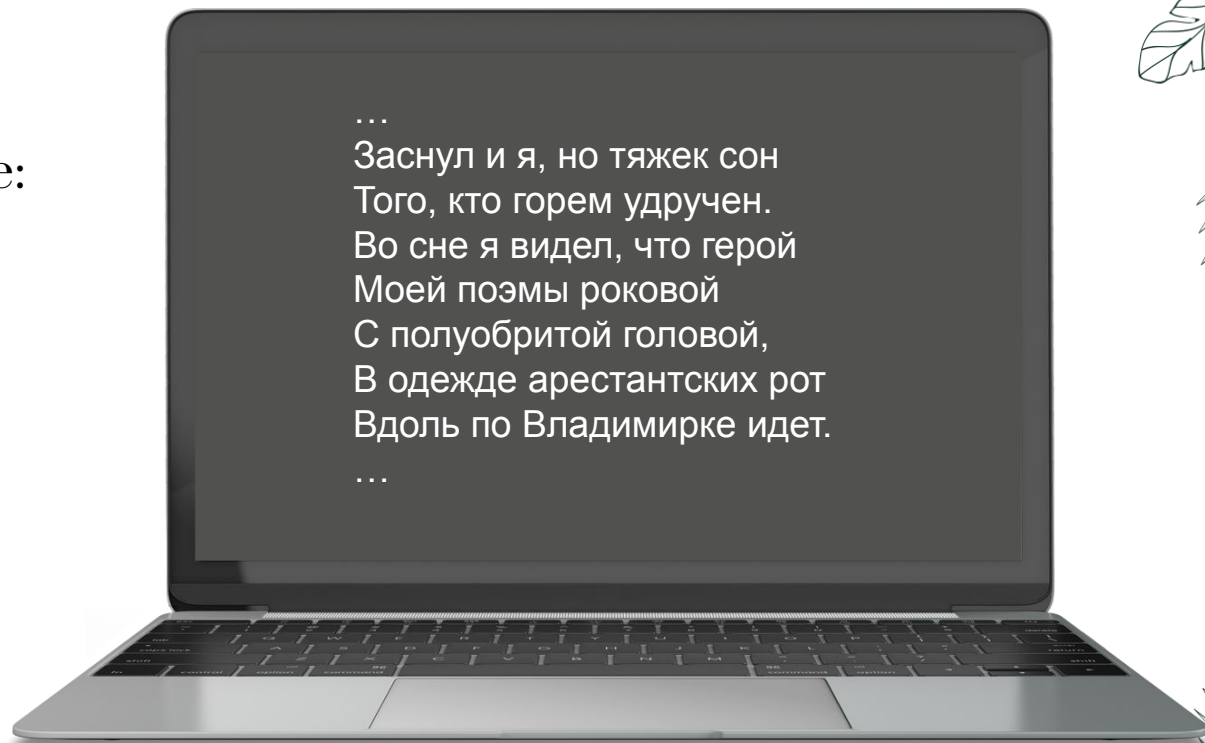
Если нет, то наше представление о размере и ударении было верным.

02.

Если да, то ударение не обязательно для определения размера.



Так выглядят
обучающие данные:



...

Заснул и я, но тяжек сон
Того, кто горем удручен.
Во сне я видел, что герой
Моей поэмы роковой
С полуобритой головой,
В одежде арестантских рот
Вдоль по Владимирке идет.

...



Многовариантная классификация

— тяжелая задача



Если не получится

то это ничего не будет значить



Поэтому

начнем с бинарной классификации



Отбор материала

01.

Сеть может
«понять», что на
размер
работает
количество
гласных

02.

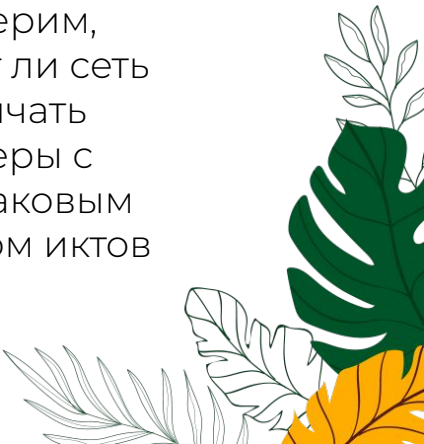
Поэтому
проверим,
умеет ли сеть
различать
размеры с
одинаковым
числом слогов

03.

Сеть может
понять, что
ударение — это
«слово» (группа
графем между
пробелами)

04.

Поэтому
проверим,
умеет ли сеть
различать
размеры с
одинаковым
числом иктов





6 СЛОГОВ

Я3м
Х3ж
~~Д2д~~
Аф2ж
Ан2м
~~Я2д~~



7 СЛОГОВ

Я3ж
Х4м
~~Д3м~~
~~Аф2д~~
Ан2ж
Х3д



8 СЛОГОВ

Я4м
Х4ж
~~Д3ж~~
Аф3м
~~Ан2д~~

Обучающие данные

Я4м
100 000 строк

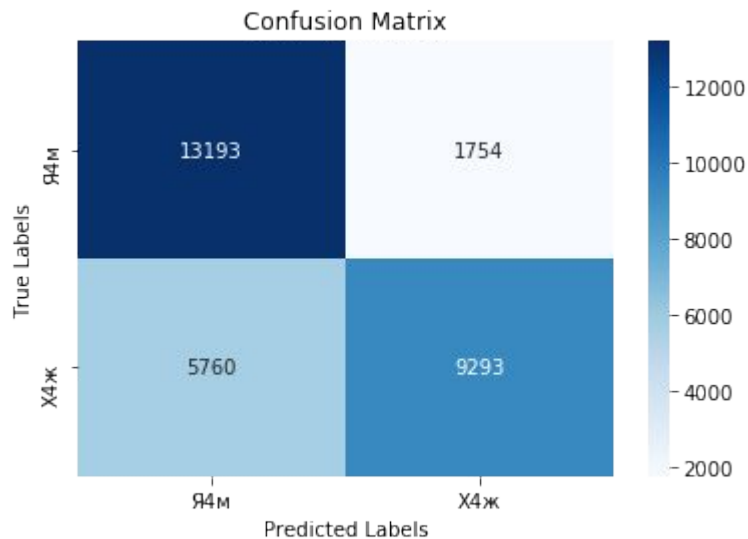
Х4ж
100 000 строк

"Я4м": "А сердцем больше
щедр и благ", "Поверхность
вод позолотит", "К ее
пленительным устам" ...
"Х4ж": "«Ничего... Жену
спровадил", "Нам лишь чудо
путь укажет", "О бананах
долгоплодых" ...

– Каждой строке соответствует
метка размера

– Модель обучается тому, что
такая последовательность
букв соответствует такому
размеру

– Нет данных об ударении!
Только буквы



Обучили модель

Теперь протестируем ее.
Данные для оценивания:

- 30 000 строк, которые сеть не видела в процессе обучения.



метр	точность	полнота	f1-мера	всего
Я4м	0.6961	0.8827	0.7783	14947
Х4ж	0.8412	0.6174	0.7121	0.7121

accuracy 0.7495

“

Это значит, что в
75 % случаев
размер
определяется
правильно.

”






Если бы речь шла о случайности, то значение было бы 50 %. В предельной формулировке это значит, что ударения **не нужны** для определения силлабо-тонического размера.



Результаты

Эксперимент показывает, что размер «звучит» и на других уровнях текста, **кроме просодического.**






01

- Модель ориентируется только на распределения букв (не слов!)

02

- Самыми частотными сочетаниями букв в тексте наряду со служебными словами являются морфологические форманты (суффикс+окончание)

03

- Форманты различают части речи
- 

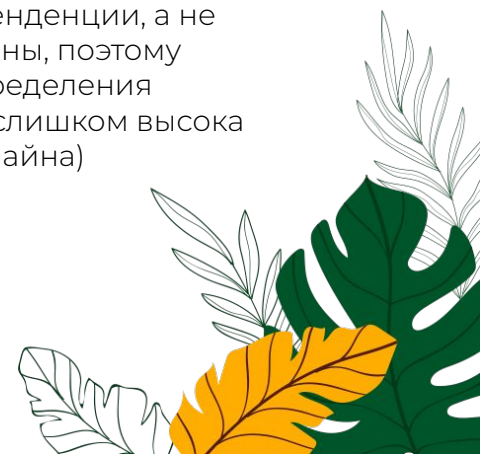
04

- Благодаря лингвистике стиха мы знаем, что распределение частей речи в строке не случайны

05

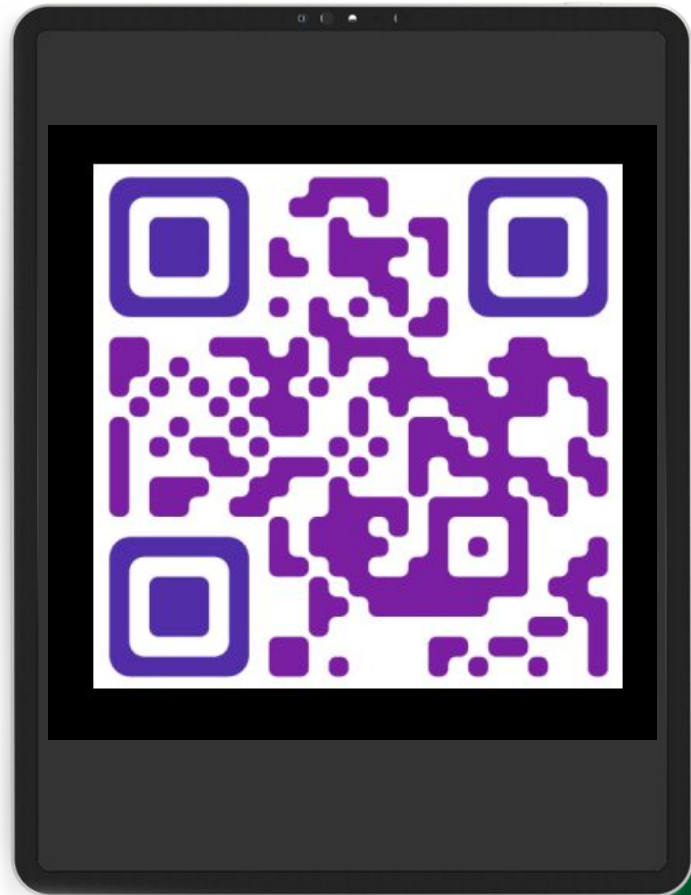
- По всей видимости, сеть выучивает тенденции в распределении морфологических форм в строке, характерном для конкретного размера

06

- Это только тенденции, а не строгие законы, поэтому точность определения размера не слишком высока (но и не случайна)
- 

nevmenandr.github.io

Статьи, сервисы, видео, подкасты



Подкаст «Лига Айвы»

Подкаст про университет, про атмосферу, про пространство,
про общение

@universitates_podcast

