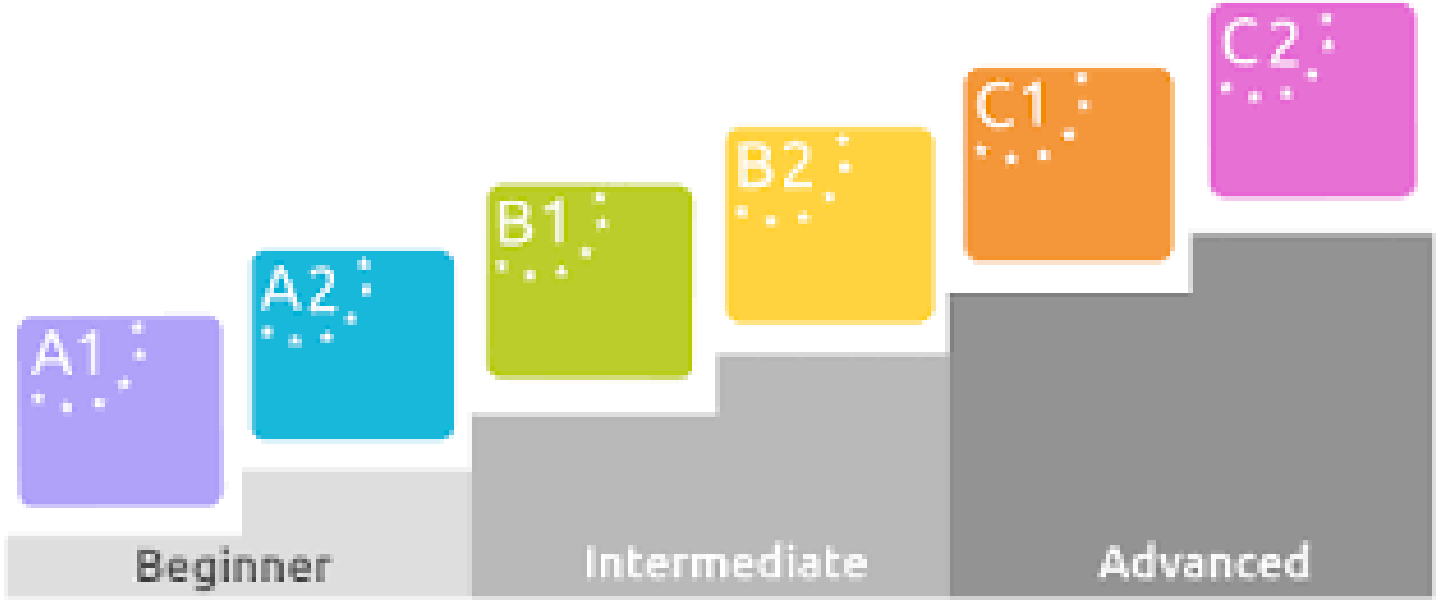
Автоматическое определение сложности русского текста как иностранного

Автоматически определять место текста на шкале из 6 уровней сложности



Сбор коллекции текстов

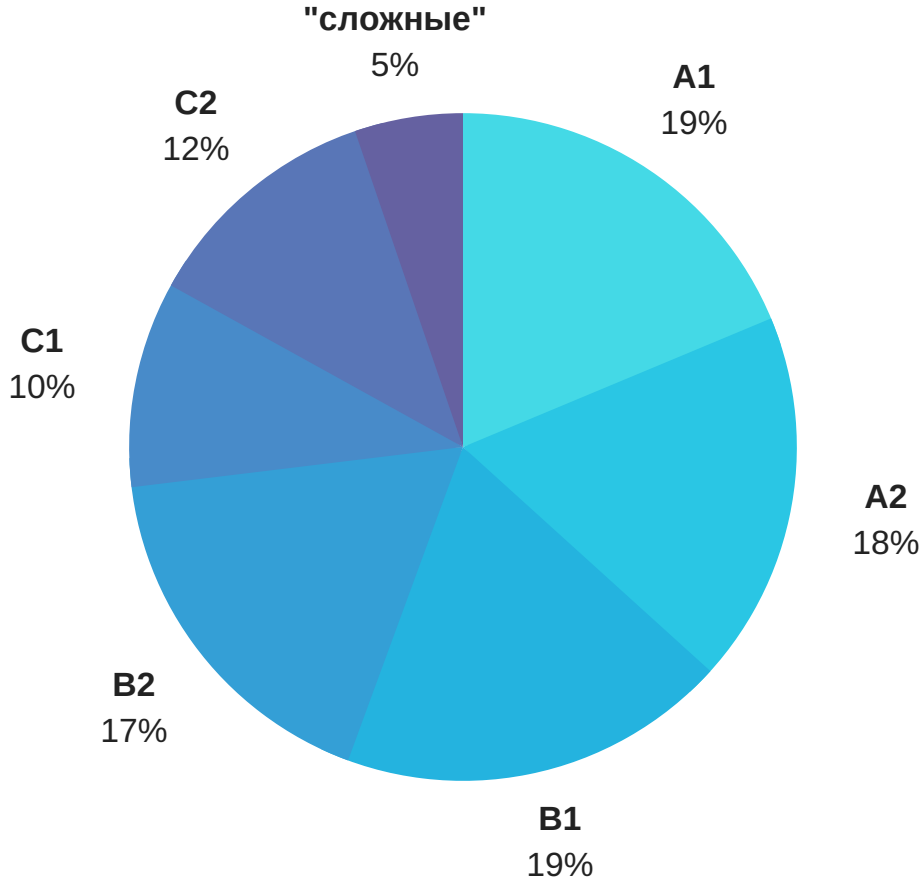
~ 600 текстов

Источники:

A1-C1 тексты из учебников и пособий по чтению для иностранных студентов

C2 тексты новостных сайтов, журналов различных тематик

C2+ законы, научные журналы



Признаки текста

Каждый текст характеризуется более 150 признаками

Лексические:
процент слов, входящих в лексические минимумы (A1, A2…)
процент абстрактных слов в тексте
процент слов, входящих в частотные списки (500, 1000, 2000…)

Грамматические:
процент слов в родительном падеже
среднее количество существительных на предложение
процент причастий в тексте

Традиционные метрики текста и формулы читабельности:
средняя длина предложения
количество знаков пунктуации на предложение
формула Флеша

Модель

Лучший результат показала
Ridge Regression на 44 признаках с наилучшей корреляцией.

Группы признаков	Explained variance score	Mean squared error
Грамматические признаки	0.5	1.69
Традиционные метрики текста и формулы читабельности	0.61	1.32
Лексические признаки	0.77	0.78
Все признаки	0.83	0.49
44 признака с наивысшей корреляцией (все группы)	0.84	0.46

Примеры работы модели

Источник текста	Уровень	Средняя длина предложения	Средняя длина слова	Процент слов, в лексическом минимуме B2	Процент слов в 33000 частотных слов	Процент абстрактной лексики
Народная сказка "Маша и медведи"	3.2	8.5	4.8	80%	96%	29%
Статья из блога про путешествия (ок. 1 тыс. слов)	3.9	12.17	5.1	82%	96%	60%
А.П.Чехов. "Общее образование"	4.1	11	4.8	78%	94%	44%
А.С.Пушкин. "Капитанская дочка" (отрывок ок. 3 тыс. слов)	5.1	11.4	4.9	75%	92%	46%
Типовой договор на аренду квартиры	5.5	9.4	6.3	63%	89%	77%
Л.Н. Толстой. "Анна Каренина" (отрывок ок. 3 тыс. слов)	5.8	22.9	5	79%	93%	48%
Правила пользования московским метрополитеном	6.5	10.2	6.8	67%	94%	66%
Алексей Навальный. Расследование "Он вам не Димон" (отрывок ок. 3 тыс. слов)	6.6	15.6	6.2	71%	90%	48%
В.Набоков. "Лолита" (отрывок ок. 3 тыс. слов)	6.9	23.4	5.5	71%	91%	54%

Проверка качества модели

70 студентов уровня B1

7 преподавателей

3 текста, оцененных моделью

A2 **B1** **B2**