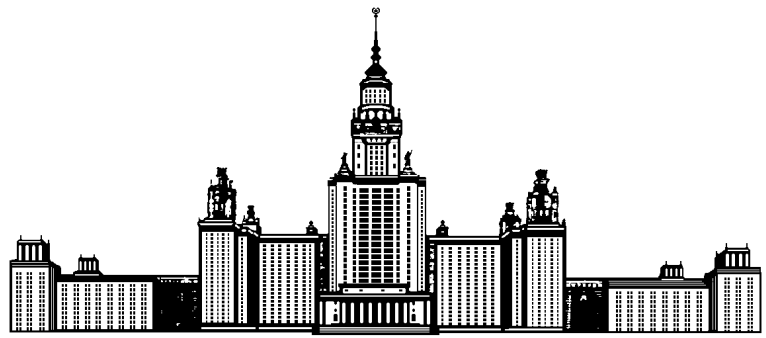
Ранжирование текстовых документов на основе оценок когнитивной сложности текста

Выполнил: Еремеев Максим Алексеевич, 417 гр.
Научный руководитель: д.ф.-м.н. Воронцов Константин Вячеславович
(ММП - ВМК - МГУ им. М.В. Ломоносова)



Москва • 11 июня 2020

Индексы удобочитаемости текста

Индексы удобочитаемости — меры сложности восприятия текста

- Наиболее распространенные:
  — *индекс Флеша*,
  — *Automated Readability index (ARI)*,
  — *SMOG-index*
- Вычисляются на основе простых параметров:
  — длины предложений и слов
  — среднего количества букв в слове
  — среднего количества слов в предложении
- содержат большое количество констант, которые подбираются отдельно для каждого языка

$$Flesh(d) = 206.835 - 1.015 \times \frac{w}{s} - 84.6 \times \frac{syl}{w}$$

Flesh, R. How To Test Readability. 1951.
Senter, R.J. and Smith, E.A. Automated Readability Index. 1967.

Ранжирование документов в порядке чтения

Reading Order — порядок документов от общих и простых к сложным и узкоспециализированным.

Применения:

- Сортировка выдачи разведочного информационного поиска
- Персонализация образовательных траекторий

Подход, основанный на тематическом моделировании:

- Оценка общности текста — энтропия тематического вектора
- Оценка близости документов — косинусное расстояние между тематическими эмбеддингами
- В цепочку попадают близкие документы в порядке убывания общности

Georgia Koutrika, Lei Liu, Steven Simske. Generating Reading Orders over Document Collections. 2015.

Психофизиологическая нагрузка декодирования речи

Что такое «тяжёлый текст» и чем он отличается от лёгкого?

Пример. Частота буквы р в русском языке 0.04, но здесь 0.17:

Гипотезы из нейрофизиологии и психофизиологии речи:

- декодирование каждого элемента языка вызывает нагрузку определенной зоны коры головного мозга
- зоне требуется *время рефрактерности* для восстановления
- в ходе эволюции языка устанавливаются существенно неравномерные распределения частот (закон Ципфа)
- в ходе освоения языка для высокочастотных элементов устанавливаются более короткие периоды рефрактерности

А.А.Биркин. Код речи. СПб.: Гиппократ, 2007.

Идея комплексного оценивания сложности текста

Основные положения предлагаемого подхода:

- *уровни языка*: фонетический, морфологический, лексический, синтаксический, дискурсивный
- на уровне h текст представляется в виде последовательности *токенов* алфавита A_h
- *сложность текста* на уровне h — это доля токенов, имеющих аномально высокую нагрузку
- нагрузка токена *аномально высокая*, если она превышает 95%-ю квантиль его нагрузки в референтном корпусе
- *референтный корпус* — тексты, которые можно считать простыми для выбранной читательской аудитории

M.Eremeev, K.Vorontsov. Lexical quantile-based text complexity measure. RANLP-2019.

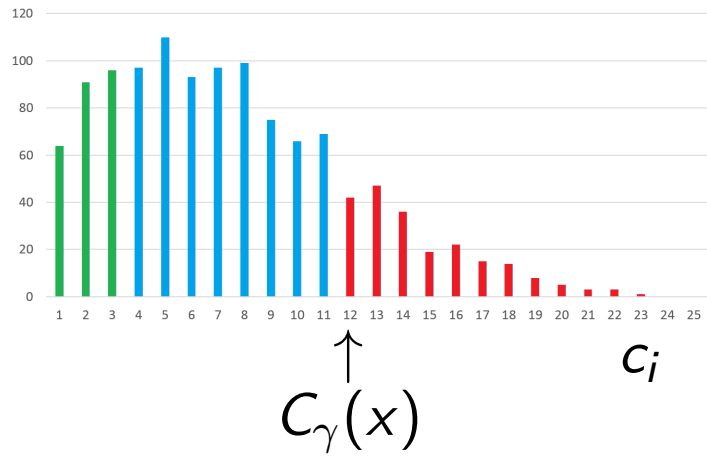
Квантильный подход к оцениванию сложности текста

$x_1, \ldots, x_n$ — последовательность токенов из A_h в тексте d;
c_i — *нагрузка* (оценка сложности) токена x_i
w_i — *вес нагрузки* токена x_i

Оценка сложности текста — суммарный вес сложных токенов:

$$W(d) = \sum_{i=1}^{n} w_i \left[c_i > C_\gamma(x_i)\right]$$

$C_\gamma(x)$ — γ-квантиль распределения сложности токена x в *референтном корпусе* несложных текстов



M.Eremeev, K.Vorontsov. Lexical quantile-based text complexity measure. RANLP-2019.

Два типа оценок токенов — частотные и сложностные

Частотная оценка нагрузки токена

- для каждого токена $a \in A_h$ строится эмпирическое распределение значений его нагрузок $\{c_i : x_i = a\}$ по референтному корпусу несложных текстов
- $c_i = f(r_i)$ — убывающая функция расстояния r_i от токена x_i до его предыдущего вхождения
- примеры: частоты букв, n-грамм, слов

Сложностная оценка нагрузки токена

- одноэлементный алфавит $A_h = \{a\}$, строится эмпирическое распределение всех нагрузок $\{c_i\}$ по референтному корпусу несложных текстов
- c_i — числовая характеристика сложности элемента x_i
- примеры: длина слова, предложения, синтаксической связи

Фонетический уровень: токены — это буквы

r_i — расстояние от токена x_i до его предыдущего вхождения:

$$\ldots \quad \boxed{x_{i-r_i} = a} \quad \underbrace{x_{i-r_i+1} \quad x_{i-r_i+2} \quad \ldots \quad x_{i-2} \quad x_{i-1} \quad \boxed{x_i = a}}_{r_i} \quad \ldots$$

Если предыдущего вхождения нет, доопределяем r_i «через хвост» (тогда сумма r_i по всем вхождениям $x_i = a$ равна длине текста n):

токен	г	е	р	о	й	н	а	ш	е	г	о	в	р	е	м	е	н	и
r_i исходное	–	–	–	–	–	–	–	–	7	9	7	–	10	5	–	2	11	–
r_i доопред.	9	4	8	11	18	7	18	18	7	9	7	18	10	5	18	2	11	18

Пример определения c_i как убывающей функции от r_i:

$$c_i = 1/r_i$$

А.А.Биркин. Природа речи. М.: Ликбез, 2009.

Морфологический уровень: токены — это буквенные n-граммы

Варианты определения токенов:

- слоги,
- буквенные n-граммы,
- буквенные n-граммы, не сохраняющие порядок букв,
- морфемы (приставки, корни, суффиксы, окончания)

Пример.

1-граммы	г	е	р	о	й	н	а	ш	е	г	о	в	р	е	м	е	н	и

n-граммы, сохраняющие порядок букв

2-граммы	ге	ер	ро	ой		на	аш	ше	ег	го		вр	ре	ем	ме	ен	ни	
3-граммы	гер	ero	рой			наш	аше	шег	его			вре	рем	еме	мен	ени		

n-граммы, не сохраняющие порядок букв

2-граммы	ге	ер	ор	йо		ан	аш	еш	ге	го		вр	ер	ем	ем	ен	ин	
3-граммы	гер	еор	йор			анш	аеш	геш	гео			вер	емр	еем	емн	еин		

Лексический уровень: токены — это слова

Варианты определения токенов для частотных оценок:

- слова после лемматизации,
- словные n-граммы,
- словные n-граммы, не сохраняющие порядок слов,

Варианты сложностных оценок нагрузки:

- c_i — длина слова
- $c_i = \frac{1}{\text{count}(x_i)}$ — редкость слова в референтном корпусе, где $\text{count}(x_i)$ — количество вхождений слова

Синтаксический уровень

Используется синтаксический парсер UDPipe. Токенами являются предложения.

Варианты определения токенов для частотных оценок:

- грамматические структуры — *синтагмы*,
  в которых отброшены слова и оставлены теги частей речи и/или членов предложения

Варианты сложностных оценок нагрузки:

- c_i — максимальная длина синтаксической связи в предложении
- c_i — суммарная длина синтаксических связей
- c_i — средняя длина синтаксических связей

Дискурсивный уровень

На дискурсивном уровне оценивается осмысленность текста, его связность и последовательность.

Варианты сложностных оценок нагрузки:

- c_i — число слов в предложении
- c_i — число логических связок в предложении (и, или, значит, который, чтобы, . . . около 150 выражений)

Обучаемая линейная модель когнитивной сложности текста

Пусть $W_k(d), k = 1, \ldots, K$ — различные оценки сложности.

Линейная агрегированная оценка сложности с параметрами α_k:

$$W(d, \alpha) = \sum_{k=1}^{K} \alpha_k W_k(d), \quad \alpha_k \geqslant 0.$$

Данные экспертного сравнения пар документов:
$d \prec d'$ — документ d' сложнее документа d.

Критерий обучения агрегированной оценки:

$$\sum_{d \prec d'} \mathscr{L}\big(\underbrace{W(d', \alpha) - W(d, \alpha)}_{\text{pair-wise margin}}\big) \to \min_{\alpha},$$

где $\mathscr{L}(M)$ — гладкая невозрастающая функция отступа M.

Сбор асессорских оценок

Для оценивания качества была сгенерирована выборка пар статей русскоязычной Википедии из категорий математики, физики, химии, информатики.

Асессорам предлагалось выбрать из двух статей ту, которая потребовала больше усилий для её понимания и содержала больше незнакомых терминов, либо указать, что статьи примерно равны по сложности, либо что они совершенно из разных областей.

Accuracy — доля пар, на которых и модель, и асессоры выбрали одну и ту же статью как сложную.

Эксперимент 1. Отдельные модели

Класс моделей	Модель	Accuracy
Индексы Удобочитаемости	Automated Readability Index	50.5%
	Flesch-Kincaid Grade	44.7%
	Flesch Reading Ease	50.7%
	Coleman-Liau	**52.1%**
Фонетические	Частотная (буквы)	62.5%
Морфологические	Частотная (с учетом порядка букв)	70.9%
	Частотная (без учета порядка букв)	73.1%
Лексические	Сложностная (длина слова)	42.4%
	Частотная	**75.0%**
	Сложностная	71.2%
Синтаксические	Сложностная (длина синтаксической связи)	62.0%
	Частотная (синтгамы)	64.2%
Дискурсивные	Сложностная (слова-связки)	62.5%

Эксперимент 2. Агрегированная модель

Модель	Функция отступа	**Accuracy**
Coleman-Liau	-	52.1%
Морфологическая (без учета порядка)	-	73.1%
Лексическая (частотная)	-	75.0%
Синтаксическая (синтгамы)	-	64.2%
Дискурсивная	-	62.5%
Агрегированная модель	Отрицательная квадр. ошибка	**88.1%**
Агрегированная модель	Отрицательная сигмоида	84.6%
Агрегированная модель	Отрицательная абс. ошибка	85.1%

- Агрегирование моделей со всех уровней дает прирост качества
- Все квантильные модели превосходят индексы удобочитаемости
- Частотная лексическая модель показывает лучшее качество среди квантильных моделей

Веб-ресурс TextComplexity.Net

Интерактивная веб-система подсвечивает сложных токены на каждом уровне в тексте при его модификации с целью упрощения. Ресурс использует оригинальную библиотеку с открытым кодом *Cognitive-Complexity*.

Выберите тип токенов:

Буквы | Слоги | Слова | Предложения

Рассчитать

Введите текст:

на дворе трава на траве дрова, не руби дрова на траве двора

Результат:

на дво ре тра ва на тра ве дро ва , не руби дро ва на тра ве двора

Результаты, выносимые на защиту

- *Предложен* квантильный подход к оцениванию сложности текстов, применимый единообразно на всех уровнях языка.
- *Предложена* методика верификации оценок сложности.
- *Предложен* обучаемый метод агрегирования оценок сложности со всех уровней.
- *Разработан* веб-ресурс TextComplexity.net, позволяющий определять сложность произвольного текста.
- *Разработана* библиотека для оценивания сложности текста с открытым кодом Cognitive-Complexity.
- *Показана* применимость оценок сложности для ранжирования текстовых документов.
- *В экспериментах показана* высокая точность квантильных оценок.