Разработка набора данных для определения удобочитаемости текста

Студент: Пучкова Екатерина Михайловна

Научный руководитель: Головин Игорь Геннадьевич

Содержание

Определение

- Удобочитаемость(“читабельность” англ. readability) – свойство текстового материала, характеризующее легкость восприятия его человеком в процессе чтения.

Определение

- Индекс удобочитаемости – мера определения сложности восприятия текста читателем.

Индекс удобочитаемости может вычисляться на основе нескольких параметров: длины предложений, слов, удельного количества наиболее частотных слов

Адапатация для русского языка

- Индексы удобочитаемости были придуманы для английского языка
- Для подсчета удобочитаемости в русском языке нужны адаптированные коэффициенты
- Темой занимался Бегтин И.В. Числа можно найти в его программе, выложенной в открытом доступе для проекта “Plain Russian”

Формулы readability. Коэффициенты в русском языке

Название формулы	Каркас	Коэффициенты в русском языке
Тест Флэша-Кинкайда	$C_1\left(\frac{total\ words}{total\ sentences}\right) + C_2\left(\frac{total\ syllables}{total\ words}\right) - C_3$	$C_1 = 0{,}49$
Тест Колман-Лиау	$C_1 L - C_2 S - C_3$	
Тест SMOG	$C_1\sqrt{pollysyllables}\frac{C_2}{sentences}$	
Формула Дэйла-Чалл	$C_1\left(\frac{difficult\ words}{words}\right) + C_2\left(\frac{words}{sentences}\right)$	
Автоматизированный индекс удобочитаемости	$C_1\left(\frac{characters}{words}\right) + C_2\left(\frac{words}{sentences}\right)$	

Датасет

- Состоит из:
  - Текстов презентаций
  - Учебников и методических пособий
- Причины:
  - Представляют из себя размеченный набор данных

Как получался датасет

- Нужен был корпус из размеченных данных
- Производился поиск учебных материалов в интернет ресурсах
- Проблема получения текстовых файлов

Как получался датасет

- Нужен был корпус из размеченных данных
- Производился поиск учебных материалов в интернет ресурсах
- Проблема получения текстовых файлов

Как получался датасет. Проблема получения текстовых файлов

- Русские школьные учебники издаются ОАО “Просвещение”. Издательство контролирует нарушение авторских прав
- В открытом доступе представлены сканы оригиналов, распознавание которых является отдельной задачей
- Идея брать учебники из библиотеки Московской Электронной Школы. Скрапинг сайта.
- Двухфакторная аутентификация

Скрапинг. Двухфакторная аутентификация

- При попытке обратиться к сайту при помощи Requests, программа зависала.
- Базовая авторизация через Requests тоже не дала результатов, так как до нее просто не дошло дело из-за сокетов.
- Кроме этого обычный механизм авторизации не подходит при наличии двухфакторной аутентификации.

Как получался датасет. Проблема получения текстовых файлов. Решение.

- Получения pdf методических пособий и вузовских учебников – легко
- Получение презентация в формате pptx – легко
- Отдельный ноутбук для создания датасета

Как получался датасет. Проблема получения текстовых файлов

- Получения pdf методических пособий и вузовских учебников – легко
- Получение презентация в формате pptx – легко
- Ноутбук для создания датасета

Ноутбук для создания датасета. Получение текстов учебников

- Создание конвертера для перевода из pdf в txt на Питоне
- Библиотека PyPDF2
- Получение текстовых файлов и сохранение их память

Ноутбук для создания датасета. Получение текстов презентаций

- Создание конвертера для перевода из формата pptx в txt на Питоне
- Библиотека pptx
- Получение текстовых файлов и сохранение их память