МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ ИМЕНИ М.В.ЛОМОНОСОВА

ФАКУЛЬТЕТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ
И КИБЕРНЕТИКИ

КАФЕДРА АЛГОРИТМИЧЕСКИХ ЯЗЫКОВ

АВТОМАТИЧЕСКОЕ РАНЖИРОВАНИЕ РУССКОЯЗЫЧНЫХ ТЕКСТОВ ПО СЛОЖНОСТИ ВОСПРИЯТИЯ

КУРСОВАЯ РАБОТА

ВЫПОЛНИЛА: СТУДЕНТКА 325 ГРУППЫ
ПУЧКОВА Е.М.

НАУЧНЫЙ РУКОВОДИТЕЛЬ: К.Ф-М.Н.,
ГОЛОВИН И.Г.

МОСКВА 2023

ВВЕДЕНИЕ

В начале работы приводятся ключевые понятия, связанные с темой. На этом этапе должно сформироваться представление о том, что такое сложность, простота текста, понятность, краткость, доступность. Хотелось бы очертить точные рамки между этими определениями, чтобы картина была более точной.

Далее будет сформулирован ряд прикладных задач, в которых ключевой проблемой является определение сложности. На этом этапе станет понятным, что проблема данной работы - насущная задача, решение которой может принести много пользы.

После этого можно будет сформулировать основные подходы к определению сложности. Общепринятые решения будут чередоваться с моими собственными идеями, которые либо будут объединять существующие, предлагая гибридный подход, либо иметь некоторый элемент новаторства.

Хочется отметить, что это достаточно изведанная тема для английского языка, чего нельзя сказать о нашем родном русском. В английских текстах уже были подобраны нужные коэффициенты в индексах удобочитаемости, придуманы множество формул. К сожалению, просто взять и применить эти результаты на практике при работе с русским текстом просто так нельзя, ведь это отдельная структура с другой организацией. Однако взять за основу существующие идеи, изменить коэффициенты, исследовать возможность применения можно.

В данной работе исследуются три подхода - статистический, инженерный и определение сложности при помощи машинного обучения.

Инженерный подход основан на правилах, которые сформулировали эксперты. К примеру, если такая конструкция встречается в тексте, то его можно отнести к сложным, если такая комбинация лингвистических конструкций была использована, текст является простым.

Статистический подход - это построение вероятностной модели и получение результата при помощи данных статистики - частоты встречаемости лингвистических конструкций определённого вида, токенов определённого вида и т.д.

Если говорить о машинном обучении, то мы посмотрим на разные модели при решении данного вопроса. Далее выберем наиболее подходящую. Основной проблемой при использовании данного подхода является формирование датасета. Нужно получить экспертное мнение при определении сложности. Нужно, чтобы человек сказал, насколько трудно было воспринимать предложенный текст. Речь идёт о решении задачи

мультиклассовой классификации. Сложность будет оцениваться по шкале от 1 до 10. Задача - определить оптимальные значения весов при построении модели.

В конце работы будут сделаны выводы.

ПОСТАНОВКА ЗАДАЧ

1. Определить точные рамки между ключевыми определениями для более точного понимания проблемы.

2. Ввести новые подходы для решения проблемы автоматического ранжирования текстов.

3. Сформировать датасет, основанный на экспертной точке зрения.

4. Провести эксперименты по определению сложности, для выявления наиболее подходящего подхода.

ОБЗОР СУЩЕСТВУЮЩИХ РАБОТ ПО ВЫБРАННОЙ ТЕМЕ

РАБОТА ЛЯШЕВСКОЙ

Я ознакомилась с некоторыми работами.

Первая работа для обзора была взята у Ляшевской. Это был доклад не семинаре НУГ. Автор говорит о ряде презумпций.

1. Короткие предложения читать легче, чем длинные;

2. Длинные слова затрудняют чтение;

3. Читатель замедляется или «спотыкается», встречая низкочастотные и/ или незнакомые ему слова и т.п.

Благодаря данной работе стало понятно, что при измерении нам важно оценить сложность именно с точки зрения языка. Мы не должны брать в учёт сложность от визуального восприятия, графического оформления, шрифтового представления. Также нам нужно, чтобы читатель был абстрагирован от сложности, связанной с незнанием

темы текста. Здесь важно учесть сложность лексики, графических форм, строения предложений.

На оценку сложности оказывают влияние и субъективные факторы, такие как: языковой опыт, возраст носителя, мотивированность читателя.

В связи с разнообразием ситуаций, в которых встречаются в качестве объекта текст, а субъекта читатель, проводятся исследования в разных областях.

1. Оценка удобочитаемости упражнений и учебных текстов для иностранцев, изучающих язык как неродной.

2. Экспертиза школьных учебников, экзаменационных тестов и других материалов.

3. Оценка читабельности деловой документации;

В работе [6] также говорится о метриках сложности. В большинстве случаев метрика удобочитаемости представляет из себя формулу линейной регрессии, в которой значение представляет из себя категорию читателя либо же средний возраст. Коэффициенты подбираются таким образом, чтобы на заданной выборке тестов оценка наилучшим образом соответствовала оценкам, поставленным экспертами. Есть некоторые другие характеристики, которые привлекаются при оценке сложности текстов, их можно разделить на лексические, морфологические, синтаксические и дискурсивные.

Если мы говорим о лексических характеристиках, то обычно имеется в виду наличие в тексте слов, которые находятся в списке самых частотных.

Под морфологическими характеристиками понимается доля разных частеречных классов в тексте или же присутствие слов с определённой словообразовательной структурой.

Синтаксические факторы оценивают сложность синтаксической структуры предложений, в частности, среднюю долю подчинённых, сочинённых и т.п., причастных и деепричастных оборотов.

Дискурсивные характеристики учитывают количество диалоговых единиц на предложение, анафорических местоимений, то есть конструкций, которые требуют от себя понимание предшествующих единиц текста, когда задействуется краткосрочная память.

Автор подчёркивает, что линейная регрессия не является единственным readabilty. Есть более сложные зависимости. Это объясняется тем, что характеристики некоторым образом зависят друг от друга. Например, важность такого фактора, как длина слова,

может меняться. Она становится малорелевантной для взрослых образованных носителей языка.

ХАБР. АВТОР СТАТЬИ БЕГТИН ИВАН ВИКТОРОВИЧ

В начале статьи [3] было рассказано об указе, в котором президент США Барак Обама постановляет: «Наша система регулирования должна обеспечить, чтобы правила были доступны, согласованы, написаны простым языком, и легко понимаемы». После того, как узнаешь о существовании похожего указа сразу становится понятна насущность проблемы, затрагиваемой в работе.

В статье можно прочитать, что написанное простым(понятным)языком - это не расхожий термин и не оборот речи. Это сформулированные за десятилетия подход по переводу официальных текстов, документов, речей политиков, законов и всего что наполнено официальным смыслом в адаптированную форму.

В этой статье приводится термин «plain». Этот термин и означает, что язык является понятным.

В работе [3] приводятся точные формулы для расчёта сложности.

ТЕСТ ФЛЭША-КИНКАЙДА(FLESH-KINKAID READIBILITY TEST)

$$C_1(\frac{total\ words}{total\ sentences}) + C_2(\frac{total\ syllables}{total\ words}) + C_3$$

total words - всего слов

total sentences - всего предложений

total syllables - всего слогов

Результатом является число лет обучения, необходимых для понимания текста по американской градации образования.

Это лишь оценка слов и предложений, но никак не смысла. Можно написать полную бессмыслицу из определённого количества слов и предложений, которая не будет никому нужна.

ТЕСТ КОЛМАН-ЛИАУ(COLEMAN-LIAN READABILITY TEST)

В данном тесте используются не слоги, а буквы. Формула расчёта учитывает среднее число букв на слово и среднее число слов на предложение.

$$CLI = C_1 L + C_2 S + C_3$$

L - среднее число букв на 100 слов

S - среднее число предложений на 100 слов

ТЕСТ SMOG(SMOG GRADE)

Формула SMOG была разработана Harry McLaighlin в 1969 и опубликована в работе «Smog Grading - a New Readability Formula».

Основной идеей являлось то, что на сложность текста более всего влияют трудные слова, которые всегда являются словами со множеством слогов.

Итоговая формула учитывала число многосложный - слов с 3-мя и более слогами, и число предложений.

$$grade = C_1 \sqrt{number\ of\ polysyllables \times \frac{C_2}{number\ of\ sentences}} + C_3$$

ФОРМУЛА ДЭЙЛА-ЧАЛЛ(DALE-CHALE READABILITY FORMULA)

Эта формула была разработана в 1948 году Эдгаром Дэйлом и Джоан Чалл на основе списка из 763 слов, некоторые из которых не были понятны ученикам 4ого класса. Таким образом определились сложные слова. Позднее появилась обновлённая формула, которая учитывала уже 3000 узнаваемых слов.

Сама формула:

$$C_1(C_2 \frac{difficult\ words}{words}) + C_3(\frac{words}{sentences})$$

АВТОМАТИЗИРОВАННЫЙ ИНДЕКС УДОБОЧИТАЕМОСТИ(AUTOMATED READABILITY INDEX)

Эта формула была опубликована в 1967, и, как и формула Колеман-Лиау, была построена на оценке сложности текстов по числу букв. Это позволило использовать формулу в электрических печатных машинках для измерения сложности текстов в реальном времени.

$$C_1(\frac{characters}{words}) + C_2(\frac{words}{sentences}) + C_3$$

ПРИМЕНЕНИЕ НА ПРАКТИКЕ

В статье [3] указаны ситуации, в которых данная задача может иметь практическое применение.

- Управление социальной защиты США использует специальное ПО - StyleWriter, помогающее оценивать и упрощать тексты.
- Администрация штата Орегон проверяет и выверяет все публикуемые ими тексты до уровня 10 класса школа - Oregon Readability.
- В кодексе штата Вирджиния присутствуют требования по обязательному уровню удобочитаемости для всех договоров по страхованию жизни и несчастных случаев и проверка уровня удобочитаемости по формуле Flesch-Kinkaid.

ИССЛЕДОВАНИЕ АВТОРА

Автор брал тексты для внеклассного чтения, для каждого из них обычно есть пометка для какого класса они предназначены.

В работе [3] были собраны качественные метрики по каждому тексту: среднее число слогов на слово, среднее число слов на предложение, среднее число букв на слово и так далее. Также поскольку был взят размеченный набор данных, для каждого текста был проставлен уровень образования, который необходим для точного понимания смысла. Как пишет сам автор, далее нужно было найти три коэффициента для формулы(за основу брался Automated Readability). Подход был таким:

1. Для констант был подобран диапазон значений с шагом 0.0001.
2. Для каждой тройки констант рассчитывались метрики удобочитаемости по выбранной формуле.

3. Далее рассчитывалось отклонение от правильного значения по каждому тексту.

4. Отклонение по всем текстам пересчитывались и получалось среднее значение по массиву.

В результате из всех вариантов констант отбирались те, по которым средние отклонения минимальны. Вот каких результатов удалось достичь. Значения констант получились равными 6.26, 0.2805 и 31.04. Сложно объяснить, чем вызваны изменения в формуле. Можно предположить, что предложения в русском языке короче, среднее количество слов на одно предложение соответственно меньше, таким образом, нужно немного поднять константу, чтобы маленькое значение среднего всё равно вносило некоторый вклад в результат.

В выводе автор говорит о том, что полностью полагаться на индексы удобочитаемости нельзя, так как они могут давать иногда ошибочный или недостаточно точный результат. Поэтому, несмотря на широкое применение, возникает вопрос об их развитии.

СТАТЬЯ ОБ NLP. ОСНОВЫ: ИЗМЕРЕНИЕ ЯЗЫКОВОЙ СЛОЖНОСТИ ТЕКСТА.

Автор статьи [5] подчёркивает важность предварительной обработки текста. Сначала нужно токенизировать текст. Далее удалить знаки препинания, символы, цифры и преобразовать все слова в нижний регистр. Также нужно убрать стоп-слова. Это позволяет получить более значимые результаты при измерении лексической сложности текста. Стоп-слова имеют небольшое лексическое богатство и используются только для связывания слов в предложения.

В статье я прочитала про лексическое богатство. Это отношение количества словоформ к количеству словоупотреблений - так называемый TTR. Нужно понимать отличия между словоформой и словоупотреблением. Словоупотребление - это количество слов в нашем корпусе. Словоупотребления включают слова вне зависимости от их повторения. Общее количество словоформ отражает количество уникальных слов, найденных в корпусе.

Основная идея этой меры заключается в том, что если текст более сложный автор использует более разнообразный словарный запас, поэтому существует большее количество словоформ. В результате, чем выше TTR, тем выше лексическая сложность.

Хотя TTR является полезным показателем, следует помнить, что, к сожалению, на него может влиять длина текста. Чем длиннее текст, тем менее вероятно, что новое слово встретится в нём.

В конце статьи [5], можно прочитать, что разработка новых мер сложности текста - это постоянно развивающаяся область.

ВЫПОЛНЕННАЯ РАБОТА

КЛЮЧЕВЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

В ЧЁМ РАЗНИЦА?

ПРИКЛАДНЫЕ ЗАДАЧИ

ИНЖЕНЕРНЫЙ ПОДХОД К ОПРЕДЕЛЕНИЮ СЛОЖНОСТИ

СТАТИСТИЧЕСКИЙ ПОДХОД К ОПРЕДЕЛЕНИЮ СЛОЖНОСТИ

МАШИННОЕ ОБУЧЕНИЕ

ВЫВОДЫ

СПИСОК ЛИТЕРАТУРЫ

1. https://scikit-learn.ru/1-12-multiclass-and-multioutput-algorithms/
2. Учебные материала курса «Анализ Неструктурированных Данных» для 3 курса ВМК МГУ.
3. https://habr.com/ru/company/infoculture/blog/238875/
4. https://habr.com/ru/post/239511/
5. https://machinelearningmastery.ru/linguistic-complexity-measures-for-text-nlp-e4bf664bd660/
6. https://ling.hse.ru/data/2016/12/15/1111563794/Readability%20talk.pdf