Задание 1

1.

0. Неограниченные

1. Контекстно-зависимые

2. Контекстно-свободные (любые языки программирования, например, Pascal, Python, Java)

3. Регулярные (регулярные выражения)

2.

Бейзлайн – самое простое решение задачи (Bag-of-words для категоризации текста)

Пайплайн – этапы обработки текста (удаление стоп-слов, токенизация, лемматизация, морфологическая резметка)

SOTA (State of the art) – наилучшее решение задачи, показывающее наивысшие оценки по принятыми метрикам, в определенный момент времени (BERT, GPT-3 – 4)

3. Проверки

4. Языковая модель – вероятностное распределение слов в тексте (Пр.: униграммная, n-граммная модель (биграммная, триграммная), модель-трансформер, нейронные)

5. При кластеризации объекты разделяются на группы, чтобы в одной группе были объекты сходные между собой, но отличающихся от объектов, представленных в других группах. При классификации объект определяется как относящийся к определенному классу на основе его свойств или характеристик, эти классы заранее известны.

Задание 2.

Лемматизация:

Вероятно может возникнуть такая ситуация, что парсер не лемматизирует все слова, а что-то упустит. Тогда следует учитывать процентное соотношение предсказанных парсером ответов и правильных ответов. Это покажет, насколько в принципе распознаются слова, которые нужно лемматизировать. y\_pred/ y\_true

Теперь следует уточнить, какой процент лемм из выявленных верен. Обозначим верно выделенные парсером леммы как y\_pred(tl). Получим отношение y\_pred(tl)/ y\_pred

Теперь нужно объединить эти два выражения в одно. Возможно можно было бы использовать среднее арифметическое значение, но тогда может получится по-моему слишком большое значение. Можно попробовать использовать среднее гармоническое значение, которое больше учитывает малые значения и нередко используется в подобных задачах с соотношениями. И тогда полученная формула будет выглядеть следующим образом:

2/ (y\_true/ y\_pred + y\_pred/ y\_pred(tl))

Правда такая формула будет работать только если количество правильных ответов не равно количеству предсказанных парсером ответов. Если все же эти значения равны, то остается только вторая часть формулы: y\_pred(tl)/ y\_pred

Приписывание частей речи:

Для оценки качества приписывания частей речи, вероятно, можно использовать такую же формулу. Обозначим количество верно приписанных парсером частей речи как y\_pred(tps). Получаем формулу:

2/ (y\_true/ y\_pred + y\_pred/ y\_pred(tps)) или

y\_pred(tps)/ y\_true (эту формулу также можно использовать, если не учитывать, у скольких слов в целом была определена часть речи, а брать в расчет только верные варианты, автоматически маркируя все остальное как неверные)

Морфологические характеристики:

Для оценки качества приписывания морфологических характеристик нужно сначала отдельно вычислить вероятность для каждой грамматической категории.

Например, обозначим предсказанные парсером ответы касательно 1-ой категории как y\_pred1, правильные ответы y\_pred(1t), 2-ой - y\_pred2 и y\_pred(2t) соответственно, 3-ей y\_pred3 и y\_pred(3t), и далее в зависимости от количества.

Тогда получится 3 выражения y\_pred(1t)/ y\_true, y\_pred(2t)/ y\_true и y\_pred(3t)/ y\_true. Значения можно будет выразить в процентах, а потом найти среднее значение, т.е. формула будет такой:

(y\_pred(1t)/ y\_true + y\_pred(2t)/ y\_true + y\_pred(3t)/ y\_true)/ 3

Однако опять же формула работает, только в том случае когда количество правильных ответов равно количеству предсказанных парсером ответов для каждой категории. Если же у некоторых слов не удалось определить ту или иную категорию, то можно написать такую формулу:

(2/ (y\_true/ y\_pred1 + y\_pred1/ y\_pred(1t)) + 2/ (y\_true/ y\_pred2 + y\_pred2/ y\_pred(2t)) + 2/ (y\_true/ y\_pred3 + y\_pred3/ y\_pred(3t)))/ 3

Задание 3.

Категоризация текста

Предполагает распределение текстов по категориям. Обычные тексты на естественном языке не могут быть сразу категорированы, поэтому сначала должна быть проведена предварительная обработка, после которой в большинстве случаев текст будет представлен как вектор. Для обучения моделей часто используются нейронные сети, обучены они могут быть как с учителем, так и без учителя.

К бейзлайну обычно относят такие сравнительно простые модели, как модель логистической регрессии, основанная на TF-IDF, а также модели, основанные на теореме Байеса.

В данный момент в качестве SOTA принята модель-трансформер BERT (и ее вариант RoBERTa). Однако, данная область постоянно развивается, все чаще появляются новые модели. Например, модель PCRA-GP, согласно исследованиям, достигает SOTA и даже работает на 0,36% лучше, чем RoBERTa.

Категоризация текста может использоваться в разных областях. Может быть использована для категоризации электронных писем, отзывов, что облегчает и делает более эффективной работу, например, государственных учреждений. Это дает возможность оперативно реагировать на запросы граждан, при этом более экономно расходуя время и ресурсы. Используется при анализе общественного мнения, что важно для компаний, занимающихся производством товаров или предоставлениям услуг. Используется категоризация постов или сообщений в соцсетях для того, чтобы избежать нежелательного контента. К задачам категоризации также относится организация по темам различных статей, публикаций или документов, что помогает в значительной степени облегчить поиск. Активно используется для различения электронных писем, содержащих спам, и писем с полезной информаций, например, предложениями о сотрудничестве и т.п. Категоризация текстов может использоваться даже в медицине для анализа симптомов и самочувствия пациентов.

Сейчас исследование направлены на разработку еще более эффективных моделей для категоризации текста. В частности, идет работа над усовершенствованием уже существующих моделей и улучшением моделей для работы с разными языками.