Artificial text detection

Настя Шабаева, Ира Лобанова

Мотивация

- Программа, распознающая сгенерированные тексты, необходима для защиты пользователей от получения фейкового и мошеннического контента.
- Такая программа также может быть полезна для разработчиков генераторов текста, поскольку она будет способствовать повышению их качества (имеет смысл делать генераторы такими, чтобы их текст программа посчитала естественным, подобрать оптимальную модель).

Настя

Ира

Обзор литературы, ведение trello, обсуждение и редактирование задач и результатов друг друга, презентация

Оценка бейслайна

Задача бинарной классификации (сгенерированный текст vs естественный текст)

Обработка данных

Задача мультиклассовой классификации (какая модель у сгенерированного текста)

Данные

- Датасеты, предоставленные организаторами соревнования.
- Для каждой подзадачи даны 3 файла: тестовые (test.csv) и обучающие данные (train.csv, val.csv).

Формат

- Тестовые: id, текст.
- Обучающие: id, текст, класс.

Baseline

- **Логистическая регрессия + tf-idf**

Логистическая регрессия - стандартный метод машинного обучения, использующийся в задачах классификации. tf-idf позволяет оптимально векторизовать текст с учетом значимости отдельных слов.

Метрики

- Стандартная метрика, используемая для задач классификации (и для оценки результатов соревнования) - **accuracy**. Помимо него мы планируем использовать **precision** и **recall**, поскольку они позволяют улучшить качество оценки для каждого из классов.

Наши планы

1. Посмотреть распределения классов в датасетах и в соответствии с этим скорректировать данные, если потребуется.
2. Сделать оценку качества бейзлайна.
3. Написать собственную нейросеть, решающую задачу классификации, попробовав разные архитектуры.
4. Сопоставить полученные результаты с бейзлайном и понять, насколько релевантно использование той или иной нейросети.
5. Если окажется совсем нерелевантно, попробуем разобраться с BERT.

Литература

1. Artificial Text Detection via Examining the Topology of Attention Maps* - статья авторов соревнования Dialog RuATD. Тема статьи пересекается с основной задачей соревнования. В статье авторы предлагают нестандартный для NLP метод "attention maps". Вряд ли мы будем его реализовывать, но интересно будет сравнить его эффективность с другими методами.
2. Automatic Detection of Machine Generated Text: A Critical Survey: в ней описаны разные модели генерации текстов, а также уже существующие детекторы, описаны возникающие проблемы. Это может быть полезно, так как мы можем столкнуться с такими трудностями и, кроме того, здесь описан метод, который мы будем использовать в качестве бейзлайна (tf-idf + логистическая регрессия).
3. Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover: авторы также рассматривают разные модели, а главное, разные методы (NN, логистическая регрессия, random forest, SVM) + есть хорошее описание того, в чем лингвистически отличаются сгенерированные тексты от естественных.

**подчеркнутые названия - это ссылки на статьи:)*