Проект по курсу «Автоматическая обработка текстов»

Проект по курсу заключается в участии в одном из соревнований SemEval 2023. Проект выполняется в группах до 4-х человек. В ходе работы над проектом можно поучаствовать в одной или в нескольких дорожках выбранного соревнования.

SemEval 2023: https://semeval.github.io/SemEval2023/tasks

Примечание: студенты при желании могут выбрать и иной проект, но с обязательным согласованием и утверждением у преподавателя, ведущего курс.

1. Контрольная точка 1 (7 ноября)

К первой контрольной точке должен быть подготовлен отчет на 2-3 страницы. Отчет должен быть сдан одним из участников группы. Другие участники группы не обязаны сдавать отчет повторно, но подписать отчет должны все участники.

Регистрация в codalab и получение доступа к данным соревнования может занять некоторое время (которое зависит от организаторов соревнования), поэтому необходимо пройти регистрацию заранее.

Отчет должен содержать краткое, но разумное описание работы. Рекомендуемая структура отчета:

[2 балла] Краткое описание задачи (переводить все описание соревнования не нужно) и формальную постановку задачи. Приведите описание входных и выходных данных, опишите входные данные и структуру целевого признака (является ли он бинарной меткой класса, последовательностью и т.д.). Опишите как измеряется качество решения задачи. Проведите предварительный анализ данных, например, проверьте, сбалансированы ли классы, есть ли в данных опечатки или примеры некорректной разметки, можно ли часть задачи решить простыми правилами и т.д.

[2 балла] Краткий обзор литературы (научных статей и / или блогов), в которых описано решение похожей задачи. Не стоит ссылаться на тривиальные источники, например, статьи о word2vec или BERT, постарайтесь найти что-то более специфичное для вашей задачи. Проанализируйте 2-3 источника.

[2 балла] Опишите базовую архитектуру модели, которую вы собираетесь использовать, и объясните, почему вы будете использовать именно ее.

Опишите, как вы собираетесь улучшать эту модель в будущем.

[4 балла] Результаты, которые вы получили с помощью базовой модели:

a. Скриншот вашего submission в codalab

b. Если лидерборд соревнования открыт, то скриншот лидерборда

Если организаторы соревнования предоставили baseline, то разработанное вами решение должно быть лучше этого baseline. Если организаторы не предоставили baseline, объясните почему ваши результаты соответствуют ожидаемому baseline.

2. Контрольная точка 2 (17 декабря)

[1 балл] Опишите на чем вы остановились во время КТ1 и какие улучшения модели планировали сделать.

[4 балла] Опишите детально более продвинутую архитектуру модели, которую вы собираетесь использовать, и объясните почему вы будете использовать именно ее. Почему она лучше базовой модели? Не забудьте про ablation study: некоторые компоненты модели могут и снижать качество, проверьте, не происходит ли это в вашем случае.

[4 балла] Результаты, которые вы получили с помощью продвинутой модели:

a. Скриншот вашего submission в codalab, на котором видно, что ваша модель лучше общего baseline и лучше вашего собственного baseline

b. Если лидерборд соревнования открыт, то скриншот лидерборда

[1 балл] Подведите итоги. Каких результатов получилось достичь? Какие улучшения позволили повысить качество вашей модели? Получилось ли у вас превзойти остальных участников соревнования?