Университет ИТМО, факультет программной инженерии и компьютерной техники

Двухнедельная отчётная работа по «Информатике»: аннотация к статье

Дата прошедшей лекции: 24.10.2023 Номер прошедшей лекции: №4 Дата сдачи: 08.11.2023

Выполнил(а) Молчанов Фёдор Денисович , № группы *Р3113* , оценка

Фамилия И.О. студента не заполнять

|  |  |  |
| --- | --- | --- |
| **Название статьи/главы книги/видеолекции**  Проект Natasha. Набор качественных открытых инструментов для обработки естественного русского языка (NLP). | | |
| **ФИО автора статьи (или e-mail)**  Александр Кукушкин | **Дата публикации**  **(не старше 2020 года)**  «24» августа 2020 г. | **Размер статьи**  **(от 400 слов)**  ~5000 |
| **Прямая полная ссылка на источник или сокращённая ссылка (bit.ly, tr.im и т.п.)**  https://habr.com/ru/articles/516098/ | | |
| **Теги, ключевые слова или словосочетания**  Обработка русского языка, Python, Машинное обучение, Искусственный интеллект | | |
| **Перечень фактов, упомянутых в статье (минимум три пункта)**   1. Библиотеку изучают на образовательных программах ВШЭ, МФТИ и МГУ. 2. Проект Natasha разделён на 9 репозиториев, каждый из которых выполняет определённую роль. 3. Сложность парсинга русского языка заключается в том числе и в том, что если мы разделим текст регулярным выражением по точкам [.?!]\s+, то такое решение даст 76 ошибок на 1000 предложений, из за Сокращений, инициалов и прочего. | | |
| **Позитивные следствия и/или достоинства описанной в статье технологии (минимум три пункта)**   1. Natasha использует результаты синтаксического разбора, учитывает связи между словами, нормализует именованные сущности. 2. Скорость и качество Razdel(библиотека Natasha) сопоставимы или выше, чем у других открытых решений для русского языка. 3. Natasha использует результаты синтаксического разбора, учитывает связи между словами, нормализует именованные сущности. | | |
| **Негативные следствия и/или недостатки описанной в статье технологии (минимум три пункта)**   1. В отличие от других библиотек-комбайнов, например, SpaСy, в Natasha пользователю потребуется самому инициализировать компоненты для анализа морфологии и синтаксиса и т.д. 2. Naeva(еще одна из библиотек Natasha) сравнивает 12 открытых решений задачи NER для русского языка. Некоторые из них так тяжело запустить и настроить, что мало людей ими пользуется. 3. Razdel обучен на аккуратно написанных текстах с правильной пунктуацей из датасетов. Из за этого он будет очень плохо работать на текстах соц. Сетей. | | |
| **Ваши замечания, пожелания преподавателю *или* анекдот о программистах[[1]](#footnote-1)**  Выбрал данную статью, так как она связана с парсингом(4 лаба) + это практически единственный отечественный проект такого масштабного уровня (остальные проекты, выполняющие схожие задачи, в основном закрытые и/или не такие эффективные) | | |

1. Наличие этой графы не влияет на оценку [↑](#footnote-ref-1)