Анализ неструктурированных данных

Домашняя работа 2

Извлечение именованных сущностей

Дедлайн: 20.12.2018

Результатом выполнения задания является ipython-ноутбук, в котором представлены все скрипты, реализующие проделанные эксперименты и отчет, описывающий каждый шаг и всю логику вычислений и все полученные результаты. Если использован Томита-парсер, все его файлы также необходимо представить. Отчёт также должен содержать краткое описание всех использованных моделей (с указанием на цитируемый источник описания модели, если такой имеется) и инструментов. Вся проделанная работа должна быть понятна из этого текста.

На усмотрение проверяющего остается штраф (т.е. снижение оценки) за неаккуратное оформление, копирование Википедии и любого другого ресурса без указания источника, списывание, неформальный стиль изложения и обилие стилистических ошибок.

Домашнее задание выполняется в группах по 1-3 человека.

Выполненное домашнее задание сдается через систему AnyTask. Инструкции по использованию системы AnyTask будут дополнительно опубликованы в телеграм-канале.

Это домашнее задание посвящено извлечению именованных сущностей на материале соревнования FactRuEval. Ссылка на обучающие тестовые данные будет отправлена в чат группы в Телеграм.

Корпус FactRuEval предобработан: тексты токенизированы, выполнен POS-тэггинг. Из исходного соревнования удалены все типы сущностей, кроме LOC, ORG, PER. Токены размечены в IOB-схеме. Ниже приведен список возможных методов извлечения именованных сущностей. За реализацию каждого метода можно получить 1 балл (если не указано иное количество баллов).

Все реализованные методы следует сравнить между собой по F-мере, accuracy. Кроме того, следует провести детальный анализ ошибок методов (в каких случаях какой метод ошибается и почему) и корректных ответов методов (в каких случаях методы дают верные ответы и почему? в каких случаях согласны между собой и почему?). Без такого сравнения методов между собой и детального анализа ошибок и корректных ответов работа оцениваться не будет.

Методы извлечения именованных сущностей

1. Правила на Томита-парсере
2. Использование библиотеки Natasha и правила на языке Yargy
3. Локальные классификаторы на основе любого ML алгоритма
4. Скрытые цепи Маркова (HMM)

5. Марковские модели максимальной энтропии (MEMM)
6. Условные случайные поля (CRF)
7. SoTA модель CNN-BiLSTM-CRF
8. Модель Senna
9. Модель ELMo-BiLSTM-CRF (2 балла)
10. Модель Transformer (2 балла)