Вопрос 1

Выполнен

Баллов: 1,00 из 1,00

Вы делаете задачу определения частей речи (POS-теггинга).

Вам дано предложение "мама мыла раму", токенизированное как $x_1 =$. Каждому токену x_{ij} может быть назначен один из нескольких классов $y_{ij} \in \{NOUN, VERB, PUNCT, \dots\}$.

Тогда задачу определения частей речи можно поставить как задачу оптимизации

$y_{i1}, y_{i2}, \dots, y_{in} = argmax_{y_{i1}, y_{i2}, \dots, y_{in}} P(y_{i1}, y_{i2}, \dots, y_{in} | x_{i1}, x_{i2}, \dots, x_{in})$Отметьте верные утверждения касательно распределения $P(y_{i1}, y_{i2}, \dots, y_{in} | x_{i1}, x_{i2}, \dots, x_{in})$.

Выберите один или несколько ответов:

- a. $P(y_{i1}, y_{i2}, \dots, y_{in} | x_{i1}, x_{i2}, \dots, x_{in}) = \prod_{j=1}^{n} P(y_{ij} | x_{i1}, x_{i2}, \dots, x_{in})$
- b. $P(y_{i1}, y_{i2}, \dots, y_{in} | x_{i1}, x_{i2}, \dots, x_{in}) = \prod_{j=1}^{n} P(y_{ij} | x_{i1}, x_{i2}, \dots, x_{in}, y_{i1}, \dots, y_{i(j-1)})$
- c. $P(y_{i1}, y_{i2}, \dots, y_{in} | x_{i1}, x_{i2}, \dots, x_{in}) = \prod_{j=1}^{n} P(y_{ij} | x_{ij})$

Ваш ответ верный.

Вопрос 2

Выполнен

Баллов: 1,00 из 1,00

Примените BIO кодирование к тексту "В Москве сегодня пасмурно, а в Комсомольске-на-Амуре солнечно", который токенизируется как "в москве сегодня пасмурно а в комсомольске на амуре солнечно".

Нужно разметить текст для одного типа сущности - локация.

Используйте заглавные латинские буквы B, I, O для обозначения меток.

Ответ запишите в одну строку, не используя разделители (всего 10 букв).

Ответ: OBOOOOBIIO

Ваш ответ верный.

Вопрос Инфо

Итак в этом видео мы поговорили о задаче распознавания плоской структуры коротких текстов. Её ещё называют "chunking" или "поверхностный разбор". Такая постановка задачи используется для извлечения именованных сущностей, определения частей речи и множества других прикладных задач. Мы выяснили, что задача "chunking" отличается от обычной классификации отсутствием независимости меток друг от друга — они теперь зависят друг от друга. А ещё мы поговорили о том, как готовить обучающую выборку, а именно — о том, как назначать золотые метки токенам. Мы рассмотрели три наиболее распространённые схемы кодирования. Для задач такого рода, в некотором смысле, есть золотой молоток — то есть общая архитектура, применяемая почти всегда: получить эмбеддинги, потом контекстуализировать, предсказать вероятности и сгладить их с помощью CRF. Её можно применять в любых задачах, если у вас достаточно данных. Если же данных меньше, то можно отбросить нейросети и оставить CRF. А ещё мы вкратце поговорили про CRF. Эта модель заслуживает гораздо большего внимания, но это не совсем вводная тема.

Вопрос 3

Выполнен

Баллов: 1,00 из 1,00

Отметьте задачи обработки текстов, которые могут быть решены методами анализа плоской структуры, то есть методами классификации токенов.

Выберите один или несколько ответов:

- [x] a. Извлечение именованных сущностей (named entity recognition)
- [x] b. POS-теггинг
- [ ] c. Полный синтаксический анализ (построение дерева зависимостей)
- [ ] d. Семантический анализ (semantic role labeling, построение семантических связей)
- [x] e. Сегментация текста (выделение заголовков, списка литературы)
- [ ] f. Извлечение отношений между сущностями

Ваш ответ верный.