**NER**

***Задача***

Одна из сложных задач в анализе текста - распознавание именованных сущностей или Named Entity Recognition (NER). В рамках соревнования требуется по тексту письма предсказать какие из его частей относятся к определенным категориям: имя магазина, номер заказа, имя товара, артикулы товаров, итоговая сумма, типы товаров.

***Данные***

Участникам будет выдан набор писем с разметкой, а именно: исходное eml-письмо, распаршенный дефолтным парсером текст, разметка для этого текста. Разметка состоит в сопоставлении каждого токена текста одной из категорий:

1. Продавцы товаров (магазины, сайты и т. п.) - метка «SHOP»
2. Заказы и их номера - метка «ORDER»,
3. Товары, которые покупаются (заказываются, бронируются и т. п.) пользователем в письме - "PRODUCT"
4. Артикулы (номера, коды, шифры и т. п.) этих товаров - "ARTICUL"
5. **Итоговая сумма** заказа в данном письме - "PRICE". Под итоговой суммой подразумевается общая сумма заказа, а не его отдельные части.
6. Типы товаров, которые, в принципе, продает магазин (в том числе, например, те товары, которые магазин рекламирует покупателю, но которые не включены в заказ) - "PRODUCT\_TYPE"
7. OUT - все прочие токены.

\*К метке первого токена каждой сущности всегда добавляется префикс "B-" (даже если сущность состоит из одного токена).

Пример разметки из одного из писем:

Магазин B-SHOP

Lamoda SHOP

. OUT

 Мужская B-PRODUCT\_TYPE

 одежда PRODUCT\_TYPE

 Женская B-PRODUCT\_TYPE

 одежда PRODUCT\_TYPE

Письма были получены путем регистрации заказов популярных магазинов на несколько тестовых ящиков, а затем размечены асессорами.

Участникам будет предоставлен код токенайзера и базовый пример для обучения на такой выборке для снижения порога входа в задачу.

***Evaluation***

Метрика качества предсказаний - средняя [F1-мера](https://www.evernote.com/OutboundRedirect.action?dest=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FF1_score) по токенам.

Под конец соревнования будет выдан тест-сет, на котором будет необходимо прогнать свою модель и сформировать submission файл. Чтобы исключить возможность влияния ручной разметки на результаты, топовые решения, претендующие на призы, должны будут подтвердить свои результаты, продемонстрировав жюри работу evaluation скрипта и его результата.

***Формат сабмита***

В качестве решения нужно предоставить файл, аналогичный train выборке, с размеченными категориями для каждого токена. Пример:

Вас OUT

приветствует OUT

замечательный OUT

интернет B-SHOP

- SHOP

магазин SHOP

канцелярских SHOP

товаров SHOP

" SHOP

Ромашка SHOP

" SHOP

, OUT

лучший OUT

магазин OUT

по OUT

выбору OUT

покупателей OUT

Мы предоставим документ-инструкцию для разметки, которая использовалась асессорами для составления этого датасета, в которой подробно изложены все аспекты и формат данных.

Пример кода скриптов, подсчитывающих результирующий скор по каждой задаче, будет предоставлен участникам.