# Sports NER

Светлана Акопян
Илья Ершов
Кирилл Коновалов

Руководитель: Виталий Горабачев

# Что такое NER и где это применяется

Распознавание именованных сущностей = named entity recognition = NER

Поиск в тексте специальных объектов: имен, адресов, чисел, дат, и т.п.


*Figure 1:* An example of NER application on an example text

Может быть полезно везде, где нужно автоматически понимать смысл текста:

- информационный поиск
- чат-боты и диалоговые системы
- любые задачи машинного обучения с текстовыми признаками

# Способы выделения NER

- Rule-based:
  - dictionaries
  - regex
- Machine learning:
  - Unsupervised learning approaches:
    - Clustering
  - Feature-based Supervised learning approaches:
    - Classification + word-level features/document or corpus features
- Deep learning techniques

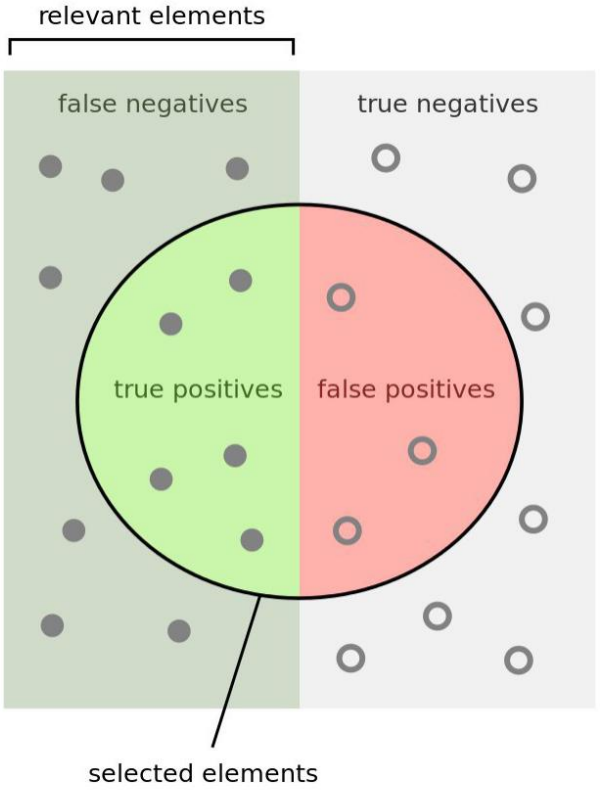
# Метрики качества

$$Precision = \frac{\text{число верно выделенных сущностей}}{\text{число всех выделенных сущностей}} \quad (1)$$

$$Recall = \frac{\text{число верно выделенных сущностей}}{\text{число сущностей в корпусе}} \quad (2)$$

$$F1 = \frac{2PR}{P + R} \quad (3)$$

# Метрики качества




How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

# Цель

Протестить разные методы распознавания именованных сущностей (в идеале state-of-the-art), а именно:

- Коэффициент
- Тип ставки
- Букмекер
- Дата и время матча
- Лига
- Название команды 1
- Название команды 2

# Задачи

- ресерч
- работа с данными и их анализ
- выбор методов, которые будем тестить
- применение выбранных методов
- оценка результатов
- написание статьи

# Данные

Датасет с прогнозами спортивных мероприятий

Всего прогнозов - 870 362

Уникальных прогнозов - 628 231

Прогнозов на английском - 89 657

Средняя длина прогноза - 52

Медианная длина прогноза - 42

Поля:

- Текст прогноза
- Коэффициент
- Тип ставки
- Букмекер
- Дата и время матча
- Лига
- Название команды 1
- Название команды 2

# Данные

| description | url | video_url | odd | tip_name | author_name | bookmaker_name | source_name | match_date_time | league_name_in_russian | team_of_season_1_name_in_russian | t |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Болгария. София принимает Витошу. Ну здесь вро... | https://vprognoze.ru/forecast/newbie/fcufootba... | NaN | 4.8 | NaN | dobruijin | 1xСтавка | Vprognoze.ru | 2019-02-25 15:30:00 | Первая лига | Септември София | |

```
df1['description'].unique()
```

```
array(['Болгария. София принимает Витошу. Ну здесь вроде обе команды выписали себе путевки в группу понижения и
       'Тридцатого марта ожидается внушительная порция матчей регулярного немецкого первенства. Нашим аналитикам
       'Встречаются лидеры второй лиги и одна из лучших линий обороны с одной из самых мощных линий атаки Бундес
       ...,
       'ЮКД играет на выезде через чур плохо, команда забила всего 15 голов в 22 встречах, в товарищеской игре т
       'Црвена Звезда - "Янг Бойз". По кадровому потенциалу команды примерно равны. В 56.95 млн. € и 66.45 млн.
       'Япония — Тринидад и Тобаго: прогноз и ставка на матч 5 июня 2019 года. Товарищеский матч начнется в 13:3
      dtype=object)
```

# Какие методы будем тестить?

https://paperswithcode.com/task/named-entity-recognition-ner

- векторизация и просто близость по порогу к словам из словаря
- logistic regression
- k-Nearest Neighbor + CRF
- bi-LSTM+CRF
- Random Forest + CRF
- BERT и его аналоги

**Бейзлайн** - прямой мэтч слова из словаря

# Сложности?

- нетипичные сущности
- большой выбор алгоритмов

# Задачи на ближайшее будущее

- продолжить ресерч области
- попробовать применить первые методы (словарные, нечеткие поиски с помощью расстояние Дамерау-Левенштейна и аналогов)