

Извлечение структурированной информации из текстов

Евгений Борисов

Извлечение информации из текстов

Цель: преобразовать текст в структурированные данные

в течение 20 рабочих дней с даты заключения Контракта

```
{'begin_event_action': 'с даты заключения',  
 'begin_event_point': 'Контракта',  
 'begin_timeshift_unit': 'рабочих дней',  
 'begin_timeshift_value': '20'}
```

Оказание услуг осуществляется с даты заключения контракта в течение 10 (десяти) дней, но не позднее 31.12.2021г.

```
{'begin_event_action': 'с даты заключения',  
 'begin_event_point': 'контракта',  
 'begin_timeshift_unit': 'дней',  
 'begin_timeshift_value': '10'}  
{'end_abs': '31.12.2021г.'}
```

Извлечение информации из текстов

NER - Named Entity Recognition

- простое совпадение слов
- шаблоны RegExp, Rule-based
- грамматические шаблоны
- модели ML (классификатор контекста)

Kofi Atta Annan is a Ghanaian diplomat who served as the seventh Secretary General of the United Nations from January 1, 1997, to January 1, 2007, serving two five-year terms. Annan was the co-recipient of the Nobel Peace Prize in October 2001.

Kofi Annan was born on April 8, 1938, to Victoria and Henry Reginald Annan in Kumasi, Ghana. He is a twin, an occurrence that is regarded as special in Ghanaian culture. Efua Atta, his twin sister, shares the same middle name, which means 'twin'. As with most Akan names, his first name indicates the day of the week he was born: 'Kofi' denotes a boy born on a Friday. The name Annan can indicate that a child was the fourth in the family, but in his family it was simply a name which Annan inherited from his parents.

In 1962, Annan started working as a Budget Officer for the World Health Organization, an agency of the United Nations. From 1974 to 1976, he was the Director of Tourism in Ghana. Annan then returned to work for the United Nations as an Assistant Secretary General in three consecutive positions.

Person
Location
Organization
Date
Nationality
Title

Извлечение информации из текстов

NER: простое совпадение слов

```
from IPython.display import display, Markdown

keyword = 'plane'
sentence = 'The fastest plane in the World'
i = sentence.find(keyword)
if i > 0:
    j = i + len(keyword)
    display( Markdown( sentence[:i]+'__'+sentence[i:j]+'__'+sentence[j:] ] ) )
else:
    display('not found')
```

The fastest **plane** in the World

Извлечение информации из текстов

NER: шаблоны RegExr

```
import re

sentence = '''В это время Владимир Телескопов
             действительно сидит в закутке у буфетчицы Симы,
             волевой вдовы.'''

rule = r'[А-Я][а-я]+' # слово начинается с заглавной буквы
re.findall(rule, sentence)

['Владимир', 'Телескопов', 'Симы']
```

Извлечение информации из текстов

NER: модели ML (классификатор контекста)

Карл Фридрих Иероним фон Мюнхгаузен родился в Боденвердере

B-PER I-PER I-PER I-PER E-PER OUT OUT S-LOC

Разметка текста BIOES

B – (beginning) – первый token в сущности

I – (inside) – слово находится в середине

E – (ending) последний token сущности

S – (single). сущность состоит из одного слова

модели ML – строим классификатор

собираем размеченный датасет

[[<контекст>, слово] , метка слова]

обучаем классификатор размечать слова по контексту

Извлечение информации из текстов

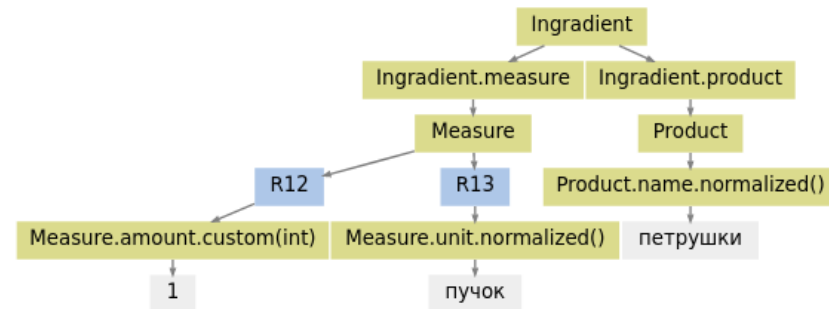
NER: грамматические шаблоны

строим грамматический шаблон с учётом морфологии

```
Ingradient -> Ingradient.product R2 Ingradient.measure | :  
Ingradient.product -> Product  
R2 -> e | in_(...)  
Ingradient.measure -> Measure
```

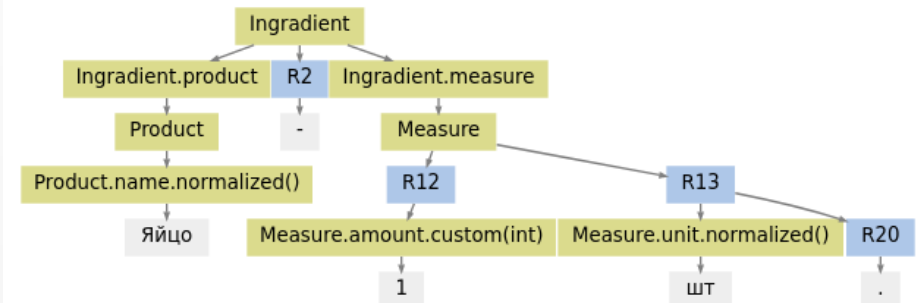
'1 пучок петрушки',
'● Молоко или вода - 2 стакана',
'20 г миндаль',
'Яйцо - 1 шт.',
'400 г варёного сгущённого молока',
'3 ст.л. красного винного уксуса',

1 пучок петрушки



```
Ingradient(  
  measure=Measure(  
    amount=1,  
    unit='пучок'  
  ),  
  product=Product(  
    name='петрушка',  
    modifiers=None  
  )  
)
```

Яйцо - 1 шт.



```
Ingradient(  
  measure=Measure(  
    amount=1,  
    unit='шт'  
  ),  
  product=Product(  
    name='яйцо',  
    modifiers=None  
  )  
)
```

Извлечение информации из текстов

Литература

git clone https://github.com/mechanoid5/ml_nlp.git

Блог компании ABBYY
NLP. Основы. Техники. Саморазвитие. Часть 2: NER
<https://habr.com/ru/company/abbyy/blog/449514/>

Sergey Kamov
Как найти что-то в тексте.
<https://habr.com/ru/post/530878/>

Александр Мазалов
Сравниваем работу open source Python-библиотек для NER.
<https://habr.com/ru/post/502366/>

Александр Кукушкин
Yargy парсер. Извлечение структурированной информации из текстов на русском языке.
PyData Moscow 2018
<https://www.youtube.com/watch?v=NQxzx0qYgK8>