

Подготовка к контрольной

Mystem

Яндекс

КОМПАНИЯ

БЛОГ

ВАКАНСИИ

РАЗРАБОТ

Все технологии /

MyStem

Программа MyStem производит морфологический анализ текста на русском языке. Она умеет строить гипотетические разборы для слов, не входящих в словарь. Первую версию программы написали [Илья Сегалович](#) и Виталий Титов.

Все вопросы, замечания и предложения отправляйте на mystem@yandex-team.ru.

XML

```
▼<html>
  ▼<body>
    ▼<se>
      ▼<w>
        <ana gr="APRO=дат,мн" lex="весь" />
        <ana gr="APRO=пр,ед,муж" lex="весь" />
        <ana gr="APRO=пр,ед,сред" lex="весь" />
        <ana gr="APRO=твор,ед,муж" lex="весь" />
        <ana gr="APRO=твор,ед,сред" lex="весь" />
        <ana gr="SPRO,мн=дат" lex="все" />
        <ana gr="SPRO,ед,сред,неод=пр" lex="все" />
        <ana gr="SPRO,ед,сред,неод=твор" lex="все" />
        Всем
      </w>
    ▼<w>
      <ana gr="S,муж,неод=вин,ед" lex="привет" />
      <ana gr="S,муж,неод=им,ед" lex="привет" />
      привет
    </w>
  </se>
  ▼<se>
    ▼<w>
      <ana gr="SPRO,ед,1-л=вин" lex="я" />
      <ana gr="SPRO,ед,1-л=род" lex="я" />
      <ana gr="S,фам,муж,од=вин,ед" lex="мень" />
      <ana gr="S,фам,муж,од=род,ед" lex="мень" />
      Меня
    </w>
  ▼<w>
    <ana gr="V,несов,пе=непрош,мн,изъяв,3-л" lex="звать" />
    зовут
  </w>
```

Данные для семинара



<https://vk.cc/9gx8tU>

Задача № 1

Открыть XML-файл и посчитать число строк внутри **первого** тега <se>, то есть между строкам <se> и </se>, открыть другой файл и записать туда число подсчитанных строк.

Задача №1. Решение

```
import re

with open('mystem.xml', encoding='utf-8') as file:
    data = file.read()

first_sent = re.search(
    '\s*<se>\n(?:.*?)\s*</se>',
    data,
    flags=re.DOTALL)

lines = first_sent.group(1).splitlines()
count_lines = len(lines)

with open('1.ans.txt', 'w', encoding='utf-8') as file:
    file.write(str(count_lines))
```

Задача №2

Создать словарь, в котором ключами являются строка с результатом морфологического разбора слова (то есть значения атрибута `gr`), а значениями — количество их вхождений в файле. Распечатать пары ключ-значение из словаря в открытый для записи файл таким образом, чтобы каждая пара располагалась на одной строке и была разделена символом “|”.

Задача №2. Решение

```
import re
from collections import Counter

with open('mystem.xml', encoding='utf-8') as file:
    data = file.read()

grs = re.findall('<ana gr="([^\"]+)".*>', data)

cnt = Counter(grs)

with open('2.ans.txt', 'w', encoding='utf-8') as file:
    for key, value in cnt.most_common(None):
        file.write(key + '|' + str(value) + '\n')
```


Задача №3

С помощью регулярных выражений выбрать из файла все словоформы, то есть теги <ana> и распределить заключенные в них леммы (lex) в разные файлы в зависимости от части речи.

<ana gr="S,сокр=вин,ед" lex="в"/> - является существительным, значит нужно записать “в” в файл с существительными

Задача №3. Решение

```
import re

with open('mystem.xml', encoding='utf-8') as file:
    data = file.read()

regex = '\s+<ana gr="([A-Z=]+)(?:,[^"]+)?" lex="([^"]+)" />'

grs = re.findall(regex, data)

pos_map = {}


# Распределим леммы по нужным частям речи
for pos, lex in grs:
    if pos not in pos_map:
        pos_map[pos] = []
    pos_map[pos].append(lex)

for pos, lemmas in pos_map.items():
    with open(pos + '.txt', 'a', encoding='utf-8') as file:
        file.write('\n'.join(lemmas) + '\n')
```

Задача №4

Преобразуйте содержимое корпуса в формат csv. Запишите результат в новый файл следующим образом: на одной строке должны находиться лемма, разбор, словоформа, разделённые точкой с запятой. Пунктуацию и служебную информацию можно удалить.

```
<w>  
  <ana gr="A=вин,ед,полн,муж,неод" lex="прошлый"/>  
  <ana gr="A=им,ед,полн,муж" lex="прошлый"/>  
  прошлый  
</w>
```



прошлый;A=вин,ед,полн,муж,неод;прошлый
прошлый;A=им,ед,полн,муж;прошлый

Задача №4. Решение

```
import re
```

```
word_regex = '\s+<w>\n(.+?)\s+</w>'
```

```
ana_regex = '<ana gr="([^\"]+)" lex="([^\"]+)" />'
```

```
with open('mystem.xml', encoding='utf-8') as file:  
    data = file.read()
```

```
with open('4.ans.txt', 'w', encoding='utf-8') as output:  
    for word in re.findall(word_regex, data, flags=re.DOTALL):  
        lines = re.sub('\t', ' ', word).splitlines()  
        anas = lines[:-1] # все, кроме последней  
        original = lines[-1]  
        for ana in anas:  
            match = re.search(ana_regex, ana)  
            gr = match.group(1)  
            lex = match.group(2)  
            output.write(';'.join([lex, gr, original]) + '\n')
```