КОРПУС ФАНФИКОВ

Алексеева Анастасия, Карпова Полина

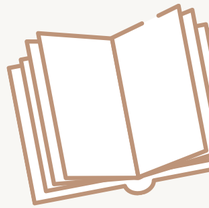
Фанфикшен

- Фанфики — это художественные произведения, написанные писателями-любителями (чаще всего фанатами определенной книги/фильма/т.п.), основанные на уже существующих произведениях.
- «Книга фанфиков» (также «Фикбук») — некоммерческий русскоязычный архив фанфикшена, а также оригинальной прозы, поэзии и публицистики, размещаемых пользователями на безвозмездной основе.
(с) Википедия

Как собирался корпус

- 40 фанфиков
- Длина до одной страницы
- Spacy

Сделали один лист tokens, по которому осуществлялся поиск. Получившаяся длина - 13917 токенов. В tokens для каждого токена лежал словарь с ключами: 'id_ff', 'id_sent', 'token', 'lemma', 'POS'.
Поиск осуществлялся линейно с кучей ифов.



```
def parse_inquiry(q):
  q = q.split(' ')
  new_q = []
  for i in q:
    new_q.append(i.split('+'))
  return new_q
```

```
def search(q, tokens):
  pos_tags = ['ADJ', 'ADP', 'ADV', 'AUX', 'CCONJ', 'DET', 'INTJ', 'NOUN', 'NUM', 'PART',

  k = len(q)
  for j in range(k):
    if not q[j][0].startswith('"') and q[j][0].upper() not in pos_tags:
      q[j][0] = lemmatize(q[j][0].lower())
  sss = []
  ids = []
  for i in range(len(tokens) - k + 1):
    sents = []
    s_searched = ''
    for j in range(k):

      if q[j][0].startswith('"'):
        s = q[j][0][1:(len(q[j][0]) - 1)].lower()
        if tokens[i + j]['token'].lower() == s:
          if len(q[j]) == 2 and q[j][len(q[j]) - 1].upper() == tokens[i + j]['POS']:
            sents.append([tokens[i + j]['id_ff'], tokens[i + j]['id_sent']])
            s_searched += tokens[i + j]['token'] + ' '
          elif len(q[j]) == 1:
            sents.append([tokens[i + j]['id_ff'], tokens[i + j]['id_sent']])
            s_searched += tokens[i + j]['token'] + ' '
```

```
    if len(q[j]) == 1:
      if q[j][0].upper() == tokens[i + j]['POS']:
        sents.append([tokens[i + j]['id_ff'], tokens[i + j]['id_sent']])
        s_searched += tokens[i + j]['token'] + ' '
      elif q[j][0].upper() not in pos_tags:
        s = q[j][0]
        if tokens[i + j]['lemma'].lower() == s:
          sents.append([tokens[i + j]['id_ff'], tokens[i + j]['id_sent']])
          s_searched += tokens[i + j]['token'] + ' '
    else:
      s = q[j][0].lower()
      if tokens[i + j]['lemma'].lower() == s and q[j][1].upper() == tokens[i + j]['POS']:
        sents.append([tokens[i + j]['id_ff'], tokens[i + j]['id_sent']])
        s_searched += tokens[i + j]['token'] + ' '

plus = 0
if len(sents) != 0:
  for m in range(len(sents) - 1):
    if sents[m] == sents[m + 1]:
      plus += 1
  if plus == len(q) - 1:
    ids.append(sents[0])
    sss.append(s_searched[:(len(s_searched) - 1)])
```

КТО ЧТО ДЕЛАЛ

- Сборка корпуса, разметка и функция поиска — Алексеева Ася

- Сайт, соединение кода, загрузка на сервер — Карпова Полина