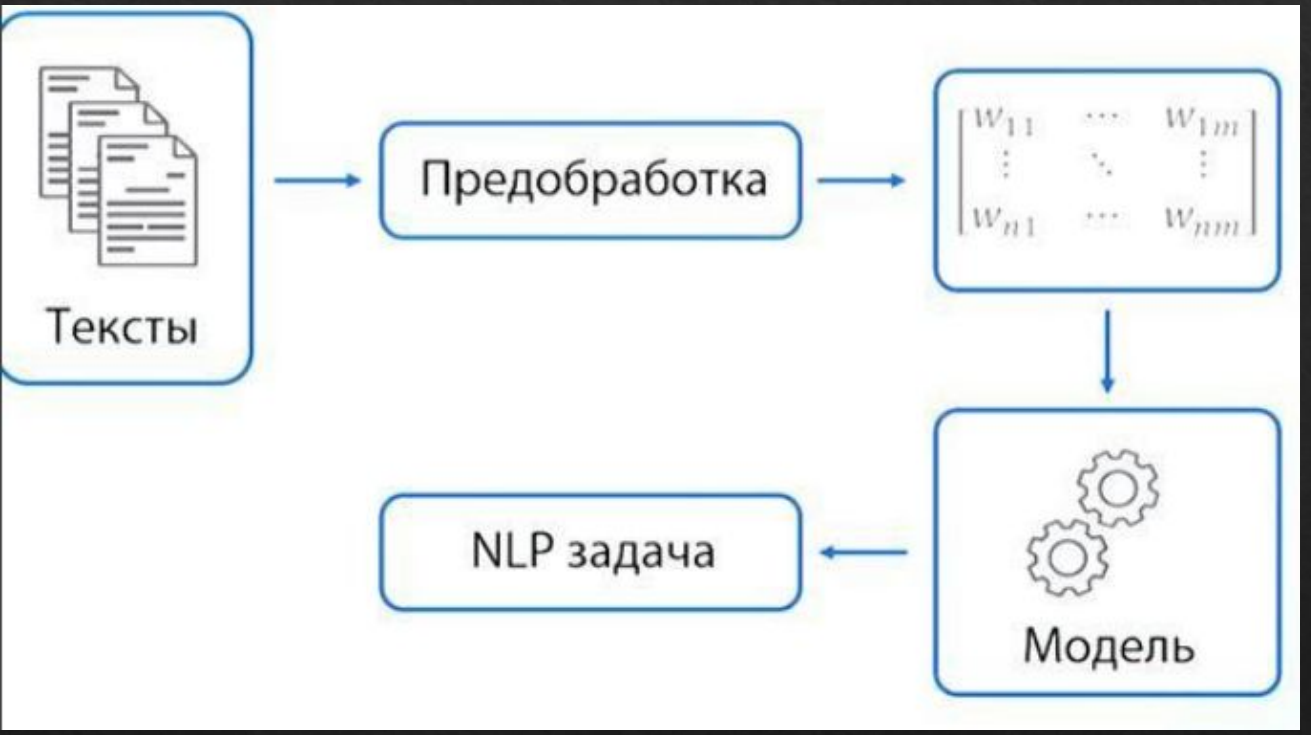
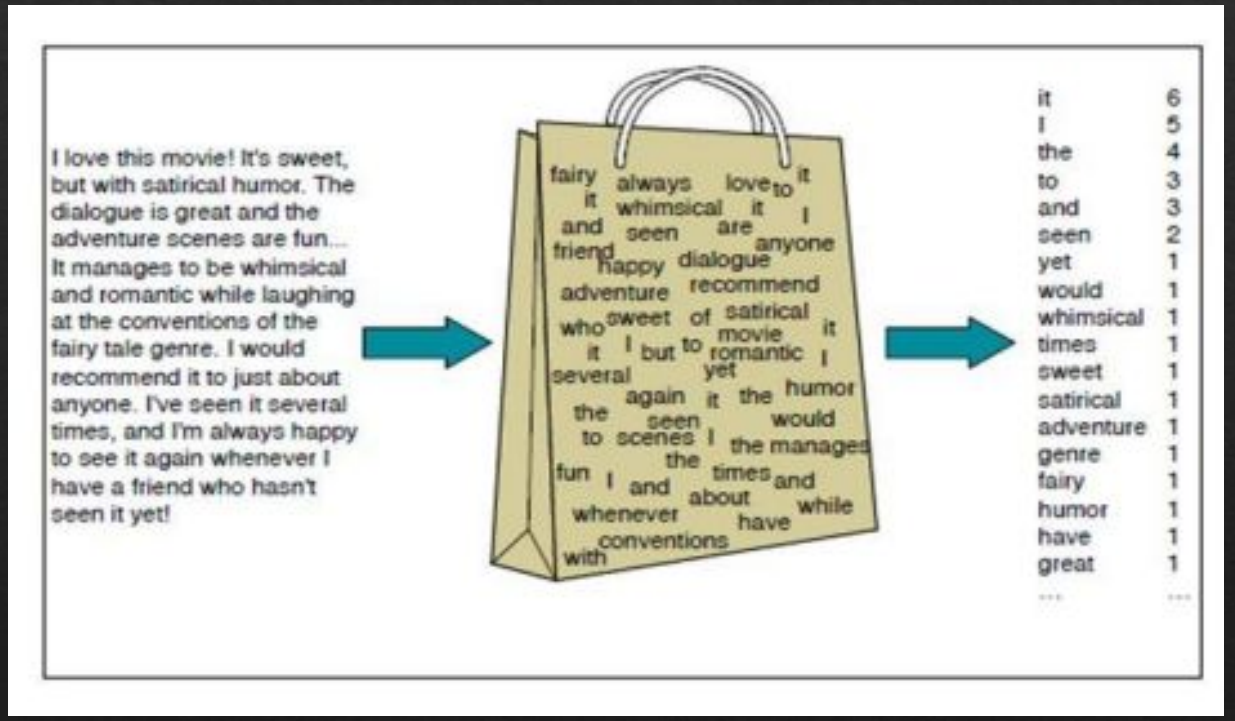
Исследование методологии морфологического анализа

Успанов Азамат

Предмет

Методология морфологического анализа



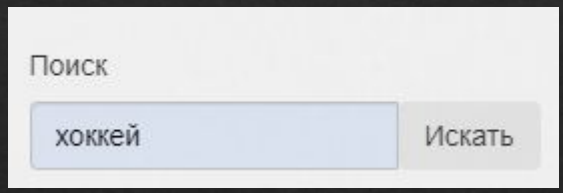


Объект

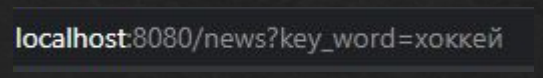
- Интеграция морфологической обработки текста в веб-сайт с поиском статей.

```
public String main(Model model,
                   @RequestParam(value = "page", defaultValue = "1") int page,
                   @RequestParam(value = "key_word", defaultValue = "", required = false) Strin


    PageRequest pageable = PageRequest.of( page: page - 1,  size: 12);

    Page<Post> postOptional;
    if (keyWord.equalsIgnoreCase( anotherString: "")) {
        postOptional = postService.findAllByOrderByDate(pageable);
        model.addAttribute( s: "isSearch",  o: false);
    } else {
        postOptional = postService.findByTitleLikeOrderByDate(keyWord, pageable);
        InputStream is = getClass().getResourceAsStream( name: "/models/en-lemmatizer.dict");
        lemmatizer = new DictionaryLemmatizer(is);
        is.close();
        List<String> listTOkens = Arrays.asList(simpleTokenizer.tokenize(keyWord));
        List<String> listPosTags = new ArrayList<>();
        listPosTags.add("returned\tVBD\treturn");
        String lemma = String.valueOf(lemmatizer.lemmatize(listTOkens,listPosTags));

        model.addAttribute( s: "keyWord", lemma);
        model.addAttribute( s: "isSearch",  o: true);
    }

    model.addAttribute( s: "posts", postOptional);
    model.addAttribute( s: "searchDto", new SearchForm( searchText: ""));
```

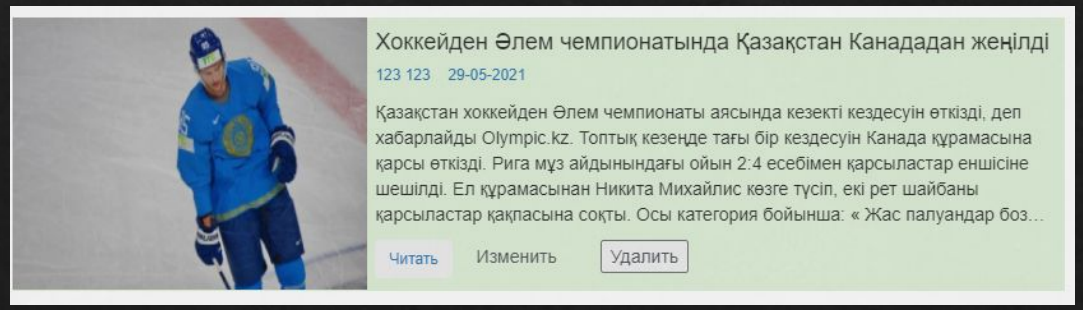
Search form



Key word



Search results



Технологии

IntelliJ IDEA

Apache OpenNLP

Java +

NAZARBAYEV
UNIVERSITY

Этапы предобработки ЕЯ

```
[ ]  from nltk.tokenize import sent_tokenize

     sents = sent_tokenize(text)
     print(len(sents))
     sents
```

```
10
['\nПродаётся LADA 4x4.',
 'ПТС 01.12.2018, куплена 20 января 19 года, 10 000 км пробега.',
 'Комплектация полная.',
 'Новая в салоне 750 000, отдам за 650 000.',
 'Возможен обмен на ВАЗ-2110 или ВАЗ 2109 с вашей доплатой.',
 'Краснодар, ул.',
 'Миклухо-Маклая, д.',
 '4/5, подъезд 1 \nТел.',
 '8(999)1234567, 8 903 987-65-43, +7 (351) 111 22 33 \nИ.И.',
 'Иванов (Иван Иванович)']
```

```
[ ]  # самая простая токенизация: разбиение по пробелам

     text = '''
     Продаётся LADA 4x4. ПТС 01.12.2018, куплена 20 января 19 года, 10 000 км пробега.
     Комплектация полная. Новая в салоне 750 000, отдам за 650 000.
     Возможен обмен на ВАЗ-2110 или ВАЗ 2109 с вашей доплатой.
     Краснодар, ул. Миклухо-Маклая, д. 4/5, подъезд 1
     Тел. 8(999)1234567, 8 903 987-65-43, +7 (351) 111 22 33
     И.И. Иванов (Иван Иванович)
     '''

     tokens = text.split()
     print(tokens)
     len(tokens)
```

```
['Продаётся', 'LADA', '4x4.', 'ПТС', '01.12.2018,', 'куплена', '20', 'января', '19',
56
```

```
[ ]  from nltk.corpus import stopwords

     # смотрим, какие языки есть
     stopwords.fileids()
```

```
['arabic',
 'azerbaijani',
 'danish',
 'dutch',
 'english',
 'finnish',
 'french',
 'german',
 'greek',
 'hungarian',
 'indonesian',
 'italian',
 'kazakh',
 'nepali',
 'norwegian',
 'portuguese']
```

```
[ ]  sw = stopwords.words('kazakh')
     print(sw)
```

```
['ах', 'ох', 'эх', 'ай', 'эй', 'ой', 'тағы', 'тағыда', 'әрине', 'жоқ'
```

```
[ ]  print([w if w not in sw else print(w) for w in clean_words])
```

```
['Продаётся', 'LADA', '4x4', '', 'ПТС', '01.12.2018', '', 'куплена',
```

Модели представления слова и текста

Изученный материал

1. Assembling the kazakh language corpus [O. Makhambetov, A. Makazhanov, Zh. Yessenbayev, B. Matkarimov, I. Sabyrgaliyev, A. Sharafudinov] 2013/10
2. Вычислительная Обработка Казахского Языка. Сборник научных трудов(Computational Processing of the Kazakh Language. Collection of scientific papers) [К.С. Дуйсебекова, Л.С. Копболсын] 2020
3. Preprocessing and Morphological Analysis in Text Mining [Krishna Kumar Mohbey, Sachin Tiwari] 2011
4. Development of the information system for the Kazakh language preprocessing [D. Akhmed-Zaki, M. Mansurova, G. Madiyeva, N. Kadyrbek & M. Kyrgyzbayeva] 2021