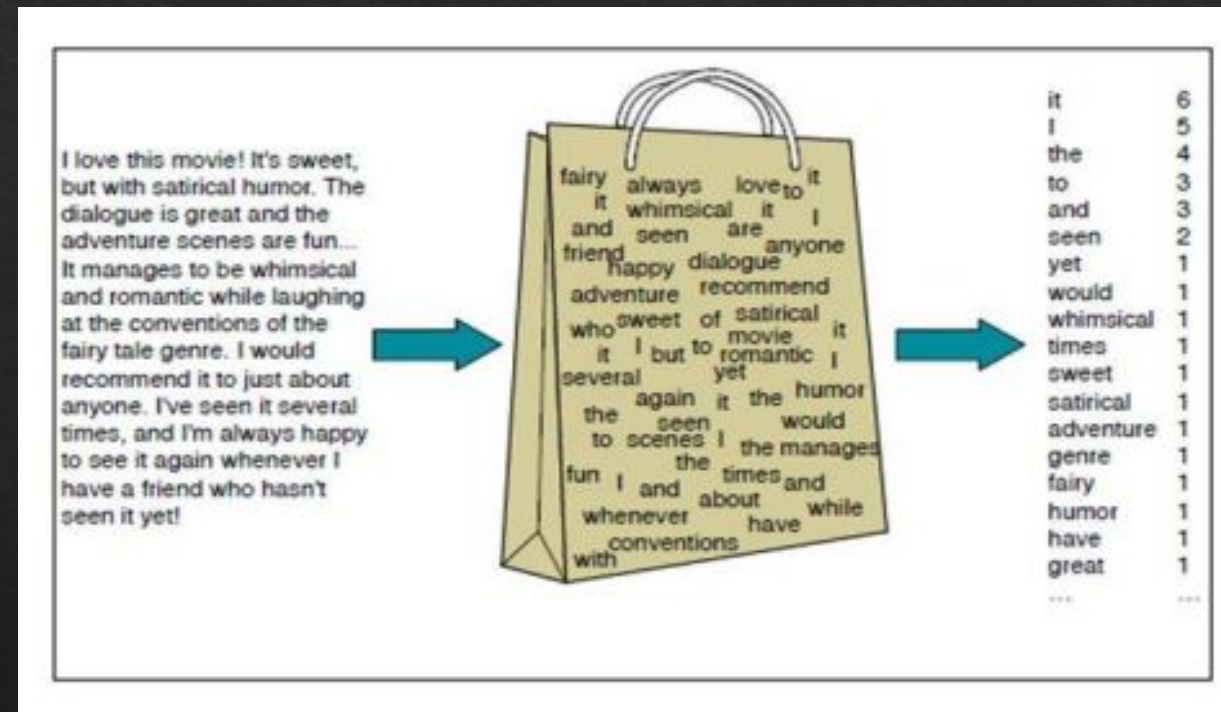
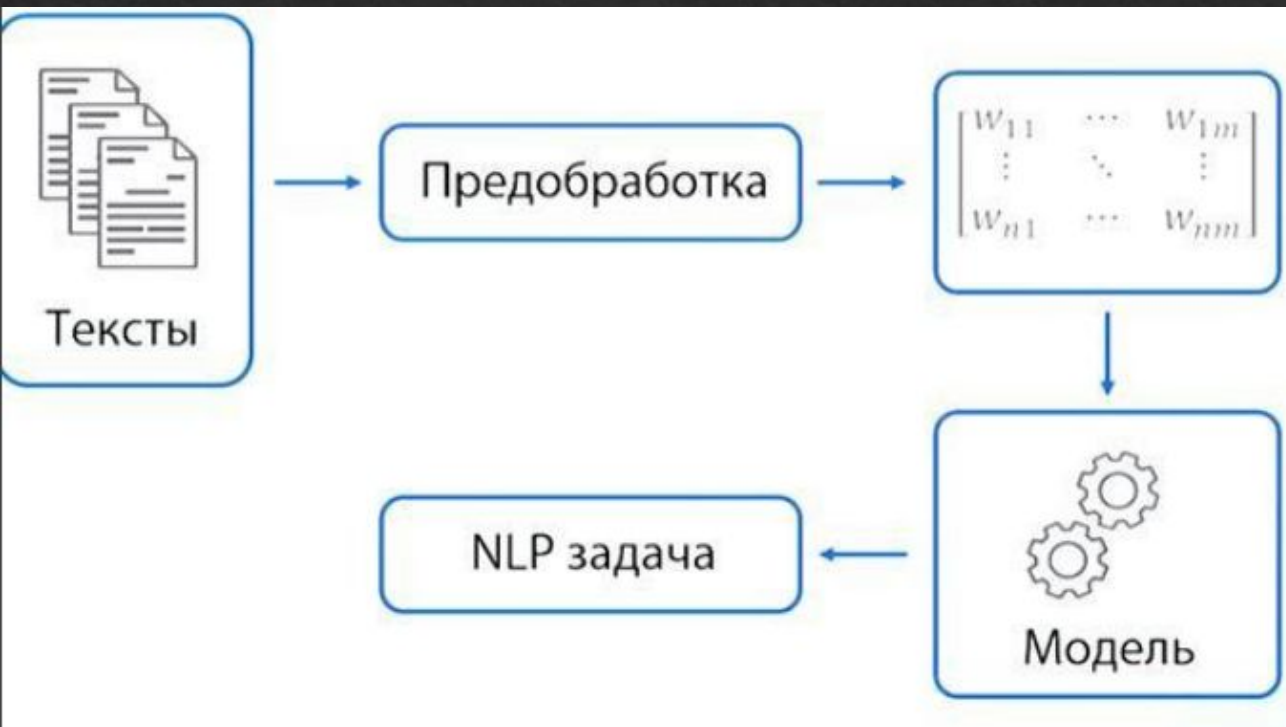


Исследование методологии морфологического анализа

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) String keyWord) {

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

    Page<Post> postOptional;
    if (keyWord.equalsIgnoreCase("")) {
        postOptional = postService.findAllByOrderByDate(pageable);
        model.addAttribute("isSearch", false);
    } else {
        postOptional = postService.findByTitleLikeOrderByDate(keyWord, pageable);
        InputStream is = getClass().getResourceAsStream("models/en-lemmatizer.dict");
        Lemmatizer 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("keyWord", lemma);
        model.addAttribute("isSearch", true);
    }

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


Search form

Поиск

Key word

localhost:8080/news?key_word=хоккей

Search results



Хоккейден Әлем чемпионатында Қазақстан Канададан жеңілді
123 123 29-05-2021

Қазақстан хоккейден Әлем чемпионаты аясында кезекті кездесуін өткізді, деп хабарлайды Олутрис.kz. Топтық кезеңде тағы бір кездесуін Канада құрамасына қарсы өткізді. Рига мұз айдынындағы ойын 2:4 есебімен қарсыластар еншісіне шешілді. Ел құрамасынан Никита Михайлис көзге түсіп, екі рет шайбаны қарсыластар қақпасына соқты. Осы категория бойынша: « Жас палуандар боз...

[Читать](#) [Изменить](#) [Удалить](#)

Технологии



IntelliJ IDEA



Apache OpenNLP

Java + 
spring



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)

['ax', 'ox', '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

“Матрица слово-текст строится по большому корпусу”

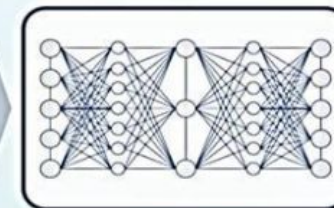
“матрица” “слово” “документ”

“большой” “корпус”

| | | | | |
|---|---|--|---|---|
| $\begin{bmatrix} 123 \\ 456 \\ 12 \\ \dots \\ 89 \end{bmatrix}$ | $\begin{bmatrix} 23 \\ 372 \\ 8 \\ \dots \\ 83 \end{bmatrix}$ | $\begin{bmatrix} 16 \\ 124 \\ 76 \\ \dots \\ 29 \end{bmatrix}$ | $\begin{bmatrix} 2 \\ 12 \\ 299 \\ \dots \\ 65 \end{bmatrix}$ | $\begin{bmatrix} 177 \\ 6 \\ 504 \\ \dots \\ 304 \end{bmatrix}$ |
|---|---|--|---|---|

Третье поколение: нейронная сеть производит представления слова и текста

В середине 2000-ных годов использование нейросетевых технологий стало де-факто индустриальным и академическим стандартом.



| |
|--|
| $\begin{bmatrix} 193 & \dots & 342 \\ \vdots & \ddots & \vdots \\ 858 & \dots & 276 \end{bmatrix}$ |
|--|

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