Московский Авиационный Институт
(Национальный исследовательский институт)
Кафедра 319 «Системы интеллектуального мониторинга»

«Информатика: проблемы, методы, технологии» (IPMT-2021)

ПРИМЕНЕНИЕ АЛГОРИТМИЧЕСКИХ МЕТОДОВ И МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

к.т.н., доцент кафедры 319 Полицына Е.В.
к.т.н., доцент кафедры 319 Полицын С.А.
ст. преподаватель кафедры 319 Зеленова М.В.

ВАЖНОСТЬ NLP

- Люди выражают свои мысли, как правило, на естественном языке

ВАЖНОСТЬ NLP

- Люди создают множество документов на естественном языке
- Справочные ресурсы
- Литература
- Новости
- Управление знаниями и задачами в проектах
- Научные статьи
- Документация
- Отчетность
- Социальные сети
- Мессенджеры
- Электронная почта









ПОЛЕЗНОСТЬ NLP

1. Информационный **поиск**
2. **Сокращение** текста (рефераты и аннотации)
3. **Перевод** текста с одного языка на другой
4. **Классификация** и **кластеризация**
5. Выделение и поиск **ключевых** элементов
6. Определение **эмоциональной окраски** текста
7. **Формирование** рекомендаций
8. **Контекстная реклама**
9. Работа с текстовой **документацией**
10. **Определение тематики**
11. **Фильтрация по определенным признакам**

СИЛА NLP

- Поиск
- Перевод
- Анализ тональности
- NER (распознавание именованных сущностей)
- Чат-боты
- Голосовые ассистенты

ПРОБЛЕМЫ NLP

- Естественный язык - **сложная система**, изначально не отличающаяся однозначностью толкования.
- **Значение** единицы языка часто зависит от **контекста употребления**, слова в различных контекстах могут приобретать различные значения.
- **Выразительные средства**, использование синонимов, антонимов и т.д. для уточнения или усиления смысла еще более усложняют автоматический анализ текста.
- **Неоднозначность** проявляется на всех уровнях анализа языка, что затрудняет автоматизированный анализ текстов.

МНОГОЗНАЧНОСТЬ

- Морфологическая: «мой», «три», «село», «мыло»
- Фонетическая: «скрип колеса», «скрипка-лиса»
- Лексическая: «рожа»
- Синтаксическая: «мужу изменять нельзя», «советую ему помочь

ТОКЕНИЗАЦИЯ: УРОВНИ АБСТРАКЦИИ

- Токенизация - разбиение текста на токены (смысловые единицы), нужные для решения **конкретной задачи**.

- Токенами могут быть:
  - буквы
  - слова
  - N-граммы
  - предложения
  - параграфы
  - документы

ТОКЕНИЗАЦИЯ: РАЗДЕЛЕНИЕ НА СЛОВА

- Слово - *минимальный* фрагмент текста, имеющий смысловую значимость
- Нахождение границ слов - важная задача.

```
String fullText = "Съешь еще этих мягких французских булок да выпей чаю";
String[] words = fullText.split("\\s");
```

ТОКЕНИЗАЦИЯ: ОСОБЕННОСТИ ЯЗЫКА

- Lebensversicherungsangestellter - «сотрудник страховой компании»

- 今天有趣的會議– «сегодня интересная конференция»

ТОКЕНИЗАЦИЯ: РАЗДЕЛЕНИЕ НА ПРЕДЛОЖЕНИЯ

- Все просто: разделение по знакам препинания (.?!)

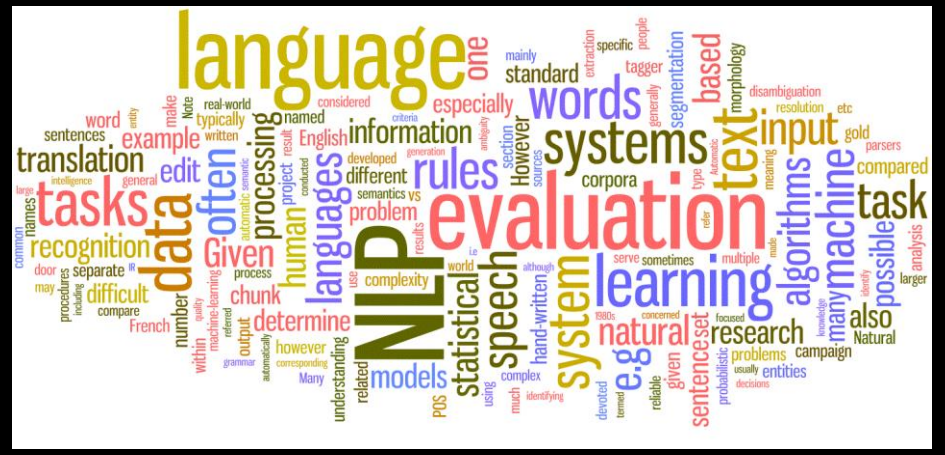
- *«Масла для бензиновых и дизельных двигателей с увеличенным сервисным интервалом, включая дизельные двигатели с сажевым фильтром и без дополнительных присадок в топливе. Альтернатива – VW 505.01, VW 506.00, VW 506.01. Исключение двигатели R5 TDI (2,5 л) и V10 TDI (5 л), требующие только VW 506.01.»*

ТОКЕНИЗАЦИЯ: ПУНКТУАЦИЯ

- Finland’s capital → Finland? Finlands? Finland’s ?
- what’re, I’m, isn’t
- L’ensemble → L? L’? Le?
- 9 a.m, i.e.
- Hewlett-Packard, Немирович-Данченко
- Сан-Франциско (San Francisco), Лос-Анджелес (Los Angeles), Нью-Васюки
- Анализ твитов - эмоджи (;), :( )

ЭВРИСТИКИ

- Предложение должно содержать слова
- Предложение должно начинаться с заглавной буквы
- Сокращения (из списка) требуют «особого внимания»
- г., тыс., млн., ул., км., …
- Отдельные большие буквы: А.Б. Иванов
- …
- …
- …



УРОВНИ/ЭТАПЫ АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТА

- **Графематический**
- **Морфологический**
- **Синтаксический**
- **Семантико-синтаксический**
- Семантический
- *Прагматический*

ИНСТРУМЕНТЫ NLP

TAWT

Unstructured Information Management Architecture
An Apache Project
TM

TEXTERRA

GATE general architecture for text engineering

Natasha

Summly

NLTK

ABBYY

wavii

Яндекс MyStem

Морфологический анализатор pymorphy2

LingPipe

ИНСТРУМЕНТЫ МОРФОЛОГИЧЕСКОГО АНАЛИЗА

GATE general architecture for text engineering

LingPipe

RussianPOSTagger
Status: Alpha Brought to you by: andrew_k

Unstructured Information Management Architecture
An Apache Project

Russian Morphology for Lucene

Яндекс MyStem

ABBYY

Pullenti

ИНСТРУМЕНТЫ РЕФЕРИРОВАНИЯ

TextAnalyst

- Языки: **русский**
- Методы: **статистический**

VisualWorld

- Языки: **русский**
- Методы: **статистический**

Text Compactor

- Языки: английский
- Методы: **статистический**, позиционный

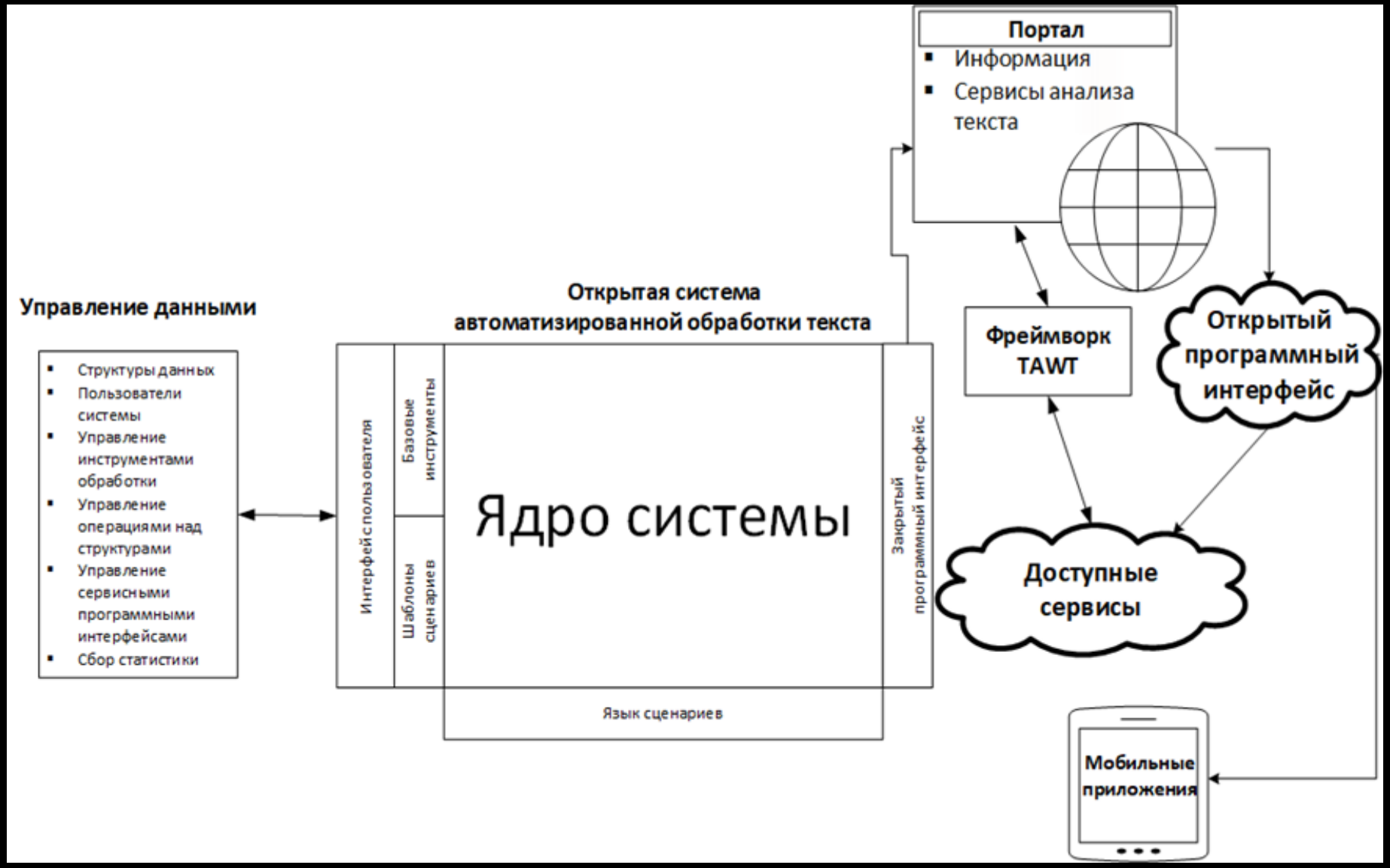
SWESUM

- Языки: английский, французский, немецкий, испанский, шведский, итальянский, норвежский, датский, греческий
- Методы: **статистический**, позиционный, индикаторный

ИНСТРУМЕНТЫ ВЫДЕЛЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ

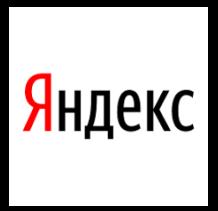
ИНСТРУМЕНТЫ ВЫДЕЛЕНИЯ КЛЮЧЕВЫХ СЛОВ

HTTP://TEXTANALYSIS.RU



ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ

- Автоматизация в системах, которых работают с большими объемами текстовых данных (поисковые, новостные, рекомендательные системы)
- Системы, построенные на применении автоматического анализа текста (системы антиплагиата, спам-фильтры)
- Проверка орфографии, чат-боты, голосовые помощники



ПРОБЛЕМЫ?

- Естественный язык – логически стройная система, но правила могут быть сложными и не все известны -> исследования в области лингвистики
- Сложность, многозначность, слабая формализуемость естественного языка -> исследования в области компьютерной лингвистики
- Сложность проектирования и реализации алгоритмов компьютерной лингвистики –> исследования в области лингвистики и компьютерной лингвистики
- Высокие требования к знанию языка и лингвистики, а также технических областей: теория алгоритмов, теория графов, математический анализ, математическая статистика, программирование и т.д.

ИЛИ ML….

Машинное обучение (англ. machine learning, ML):

- обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться;
- класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач.

Для построения таких методов используются средства математической статистики, численных методов, математического анализа, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными в цифровой форме.

ML – ЭТО ПРОСТО

```
import pickle

from sklearn.svm import SVC
import numpy as np
```

learn.txt

1,1,1,1
1,1,0,1
6,7,10,1
0,3,1,0
0,0,0,0
0,1,1,0
2,0,1,0

```
# Открыть файл с вектором признаков и преобразовать в датасет
rawData = open("learn.txt")
dataset = np.loadtxt(rawData, delimiter=",")

# Обучить модель для выбранного алгоритма ML
svmClassifier = SVC()
svmClassifier.fit(dataset[:, :-1], dataset[:, -1])

# Сохранить обученную модель
filename = 'model.dat'
pickle.dump(svmClassifier, open(filename, 'wb'))

# Загрузить ранее обученную модель
loadedSvmClassifier = pickle.load(open(filename, 'rb'))

# Классифицировать входные данные
testValues = np.array([[0, 0, 1]], float)
svmPredict = loadedSvmClassifier.predict(testValues)
```

ПРЕДОБРАБОТКА ТЕКСТА

- **Токенизация** – разбиение длинных участков текста на более мелкие (абзацы, предложения, слова).
- **Нормализация** – приведение текста к единообразному виду (единый регистр слов, отсутствие знаков пунктуации, расшифрованные сокращения, словесное написание чисел и т.д.).
- **Стеммизация** – приведение слова к его корню путем устранения придатков (суффикса, приставки, окончания).
- **Лемматизация** – приведение слова к начальной форме слова (инфинитив для глагола, именительный падеж единственного числа — для существительных и прилагательных).
- **Очистка** – удаление стоп-слов, которые не несут смысловой нагрузки (артикли, междометья, союзы, предлоги и т.д.).

ВЕКТОРИЗАЦИЯ ТЕКСТА

- **«Сумка слов» (bag of words)** – вектор признаков соответствует полному словарю текстовой выборки, для каждого слова считается количество вхождений в текст и это число подставляется на соответствующую позицию в векторе.
- **N-граммы —** комбинации из N последовательных терминов для упрощения распознавания текстового содержание, модель определяет и сохраняет смежные последовательности слов в тексте.
- **TF-IDF** – для учета соотношения частоты термина и частоты документа, в котором он встречается
- **Word2Vec** — набор моделей для анализа естественных языков на основе дистрибутивной семантики и векторного представления слов.

ПРИМЕНЕНИЕ ДЛЯ РЕШЕНИЯ ЗАДАЧ NLP

Классификация:

- Определение тональности
- Определение спама
- Классификация по предметным областям или другим признакам

Кластеризация:

- Группировка новостных статей

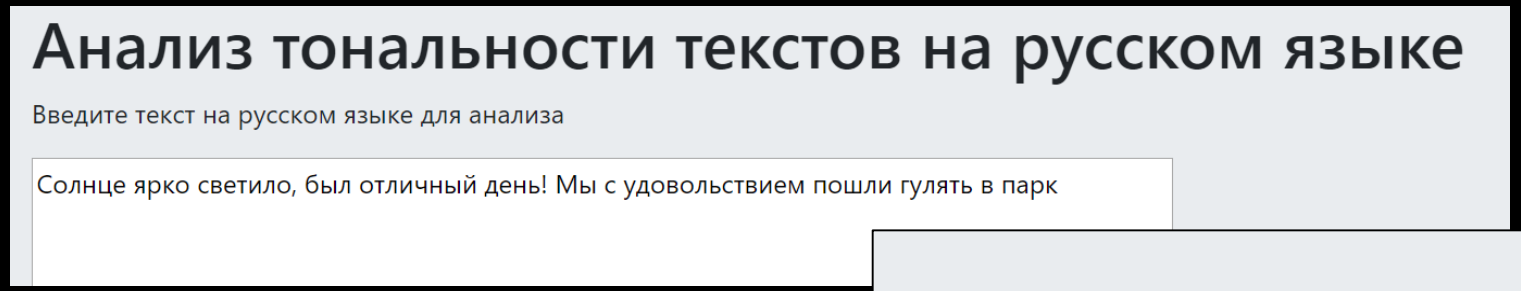
- и др.

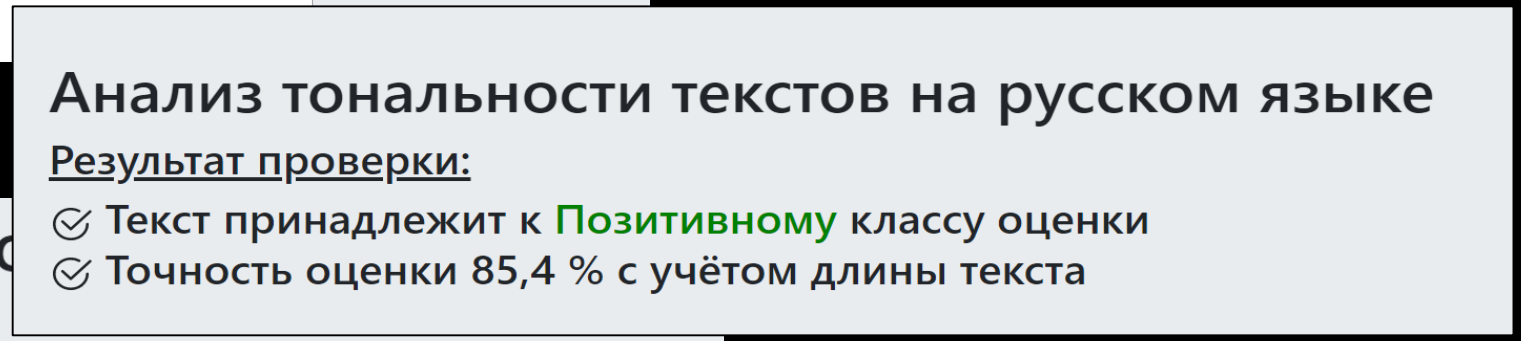
ОПРЕДЕЛЕНИЕ ТОНАЛЬНОСТИ ТЕКСТА

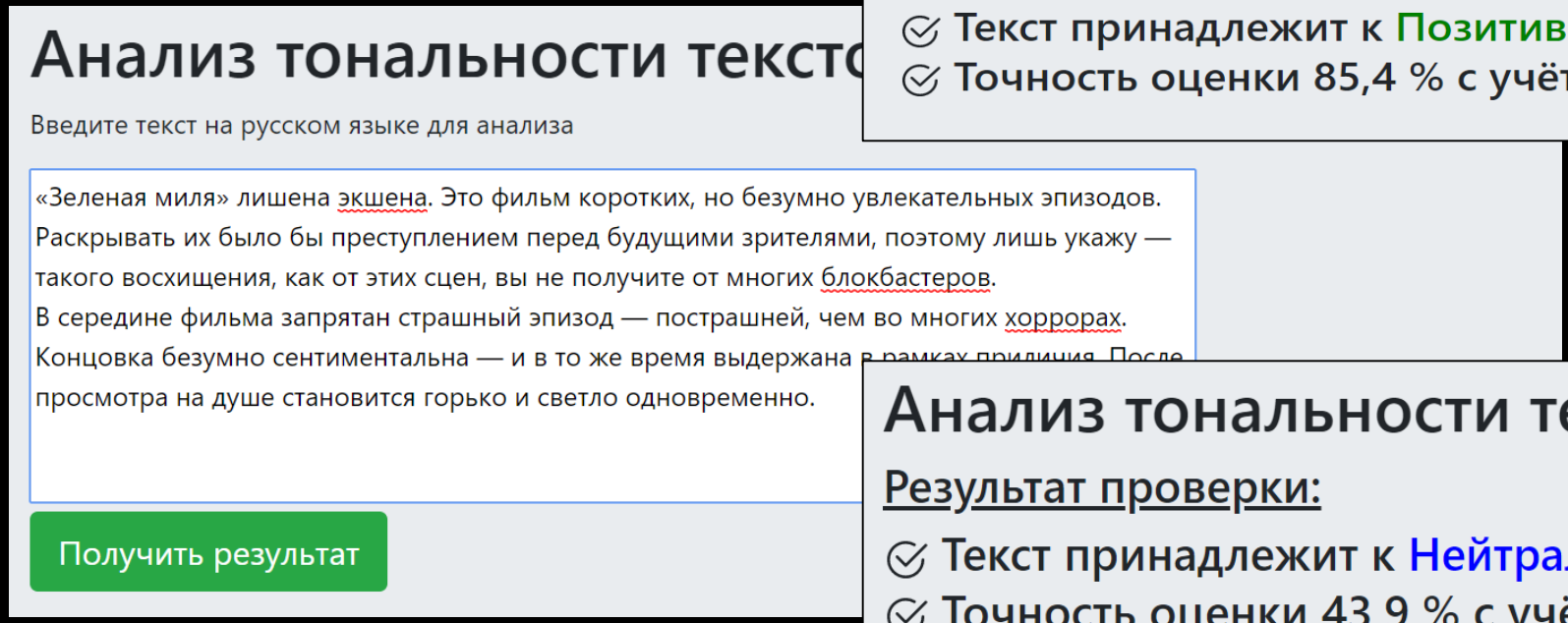
Текстовая выборка	Ансамбль 1 (ЛГ+Р)	Ансамбль 2 (К+ЛГ+Р)	Ансамбль 3 (Р)
Тексты длиной менее 280 символов	**90,8%**	90,2%	33,9%
Смешанная выборка (тексты длиной более и менее 280 символов)	82,1%	**85,4%**	38,7%
Тексты длиной более 280 символов	34,7%	43,9%	**47,1%**

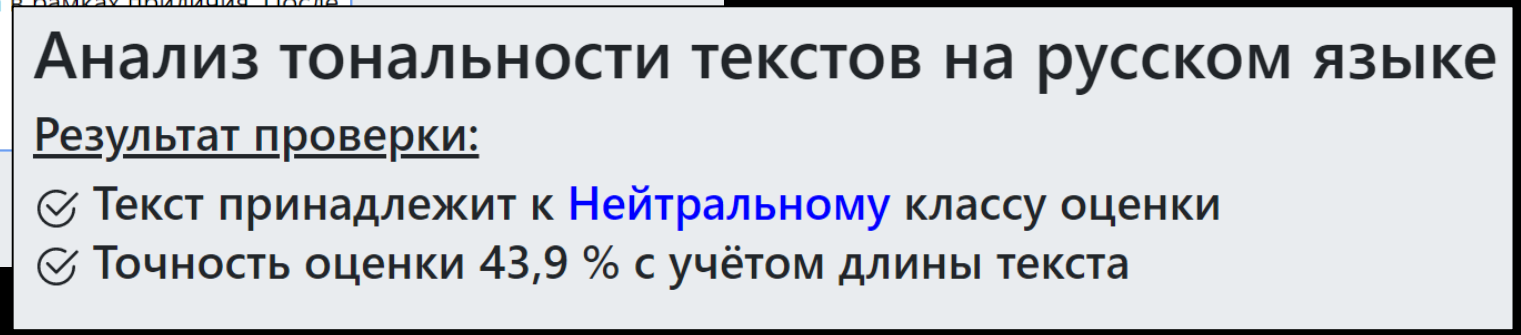
- **К** – метод *k* ближайших соседей
- **ЛГ** – метод логистической регрессии
- **Р** – метод «Random Forest»

ОПРЕДЕЛЕНИЕ ТОНАЛЬНОСТИ ТЕКСТА









КЛАССИФИКАЦИЯ ТЕКСТОВ ПО ПРЕДМЕТНЫМ ОБЛАСТЯМ

Название классификатора	Точность топ N (N = 3)
Наивный байесовский классификатор	0,8912
Классификатор на основе дерева принятия решений	0,9865
Классификатор k ближайших соседей	0,9549
Классификатор на основе метода опорных векторов	0,9366
Классификатор на основе логистической регрессия	0,9127
Стекинг	0,8915
Беггинг (на основе метода опорных векторов)	0,9317
Бустинг (на основе метода опорных векторов)	1
Нейронная сеть	0,9201

МОРФОЛОГИЧЕСКИЙ АНАЛИЗ

«Старику хотелось важных, серьезных мыслей…»

(А.П. Чехов. «Печенег»)

важный: (Часть речи - **18, прил.)** важных : морф. характеристики - **4272**

старик: (Часть речи - **17, сущ.)** старику : морф. характеристики - 230

хотелось: (Часть речи - **20**, **глаг. форма**) хотелось : морф. характеристики - 670764

серьёзный: (Часть речи - **18, прил.)** серьёзных : морф. характеристики - **4272**

мысль: (Часть речи - **17, сущ.)** мыслей : морф. характеристики - 187

ГРУППИРОВКА СЛОВ С ОДИНАКОВЫМИ МОРФОЛОГИЧЕСКИМИ ХАРАКТЕРИСТИКАМИ

«Старику хотелось важных, серьезных мыслей…» (*А.П. Чехов*)

Предложение - 1: морф. характеристики - **4272**
важный: 18: **важных** : морф. характеристики - **4272**
серьёзный: 18: **серьёзных** : морф. характеристики - **4272**

«Как я мелок, ничтожен… Я жалок, нищ духом…» (*И.А. Гончаров*)

Предложение - 1: морфологические характеристики - **4132**

ничтожен: 19:

ничтожен: морф. характеристики - **4132**

мелок: 19:

мелок: морф. характеристики - **4132**

Предложение - 2: морфологические характеристики - **4132**

жалок: 19:

жалок: морф. характеристики - **4132**

нищ: 19:

нищ: морф. характеристики - **4132**

ОПРЕДЕЛЕНИЕ СЛОВ СО СХОЖИМИ МОРФОЛОГИЧЕСКИМИ ХАРАКТЕРИСТИКАМИ

```
String wordInSentence = "стол";
JMorfSdk jMorfSdk = JMorfSdkFactory.loadFullLibrary();
List<String> initialForms = jMorfSdk.getStringInitialForm(wordInSentence);
List<Long> morphCharacteristics = jMorfSdk.getMorphologyCharacteristics(wordInSentence);
List<Byte> typeOfSpeeches = jMorfSdk.getTypeOfSpeeches(wordInSentence);
```

```
public class Word {
    private String word;
    private String initialForm;
    private byte typeOfSpeech;
    private long morphCharacteristics;
}
```

```
// Маска для поиска слов с одинаковыми
// указанными морфологическими характеристиками
final long NOUN_MASK = MorfologyParameters.Numbers.IDENTIFIER |
                       MorfologyParameters.Gender.IDENTIFIER |
                       MorfologyParameters.Case.IDENTIFIER |
                       MorfologyParameters.Animacy.IDENTIFIER;
```

ГРУППИРОВКА СЛОВ СО СХОЖИМИ МОРФОЛОГИЧЕСКИМИ ХАРАКТЕРИСТИКАМИ

«Белый снег летел за окном.»

Совпадение по роду, числу и падежу:

```
long NOUN_MASK = MorfologyParameters.Numbers.IDENTIFIER |
                 MorfologyParameters.Gender.IDENTIFIER |
                 MorfologyParameters.Case.IDENTIFIER;
```

Предложение – 1:

снег : 17:
снег : морф. характеристики – 103

белый : 18:
белый : морф. характеристики - 4196

МОРФОЛОГИЧЕСКАЯ ВЕКТОРИЗАЦИЯ ТЕКСТА

На сотни вёрст, на сотни миль,
На сотни километров
Лежала соль, шумел ковыль,
Чернели рощи **кедров**.
(А. Ахматова "С самолёта")

на	1 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
сотни	1 7 0 0 2 0 0 0 0 0 0 3 0 2 0 2 0
вёрст	1 7 0 0 2 0 0 0 0 0 0 3 0 2 0 3 0
на	1 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
сотни	1 7 0 0 2 0 0 0 0 0 0 3 0 2 0 2 0
миль	1 7 0 0 2 0 0 0 0 0 0 3 0 2 0 3 0
на	1 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
сотни	1 7 0 0 2 0 0 0 0 0 0 3 0 2 0 2 0
километров	1 7 0 0 2 0 0 0 0 0 0 3 0 1 0 3 0
лежала	2 0 0 0 0 0 0 0 3 3 0 0 2 2 2 2 0
соль	1 7 0 0 1 0 0 0 0 0 0 3 0 3 0 2 0
шумел	2 0 0 0 0 0 0 0 3 3 0 0 2 1 2 2 0
ковыль	1 7 0 0 1 0 0 0 0 0 0 3 0 1 0 2 0
чернели	2 0 0 0 0 0 0 0 3 3 0 0 2 0 2 3 0
рощи	1 7 0 0 2 0 0 0 0 0 0 3 0 2 0 2 0
кедров	1 7 0 0 2 0 0 0 0 0 0 3 0 1 0 3 0

РЕЗУЛЬТАТЫ КЛАСТЕРИЗАЦИИ

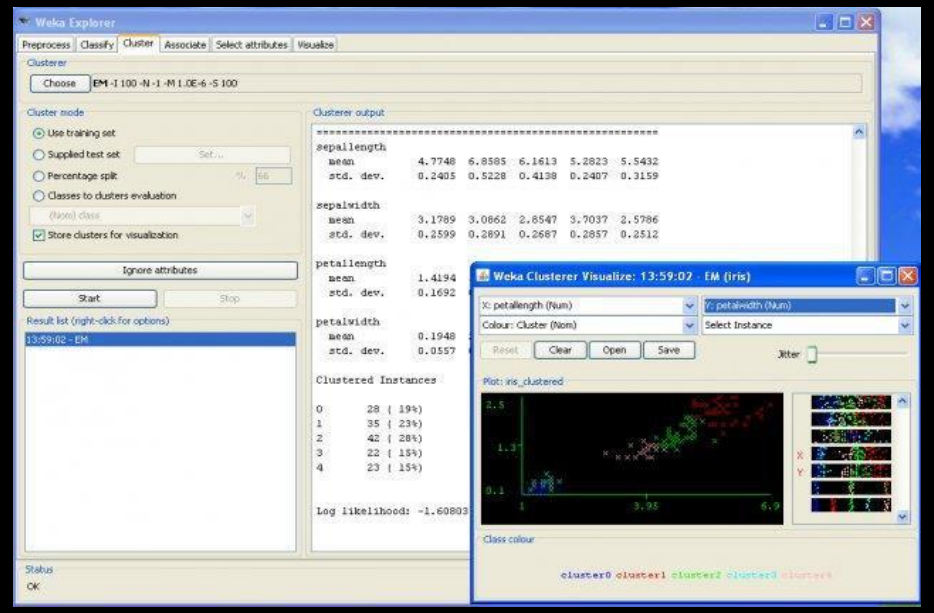
На сотни верст, на сотни миль,
На сотни километров
Лежала соль, шумел ковыль,
Чернели рощи кедров.
(А. Ахматова "С самолёта")

1 2 3 4 5 <- номера кластеров

3 0 0 0 0 | на
0 **3** 0 0 0 | **сотни**
0 **1** 0 0 0 | **вёрст**
0 **1** 0 0 0 | **миль**
0 **1** 0 0 0 | **километров**
0 0 1 0 0 | лежала
0 0 0 1 0 | соль
0 0 1 0 0 | шумел
0 0 0 0 1 | ковыль
0 0 1 0 0 | чернели
0 **1** 0 0 0 | **рощи**
0 **1** 0 0 0 | **кедров**

Сколько раз пытался я ускорить
Время, что несло меня вперед.
Подхлестнуть, вспугнуть его, пришпорить,
Чтобы слышать, как оно идет.
(С.Я. Маршак)

1 0 0 0 0 0 0 | сколько
1 0 0 0 0 0 0 | раз
0 1 0 0 0 0 0 | пытался
0 0 1 0 0 0 0 | я
0 0 0 1 0 0 0 | **ускорить**
0 0 0 0 1 0 0 | время
1 0 0 0 0 0 0 | что
0 1 0 0 0 0 0 | несло
0 0 1 0 0 0 0 | меня
0 0 0 **1** 0 0 0 | **подхлестнуть**
0 0 0 **1** 0 0 0 | **вспугнуть**
0 0 0 0 0 1 0 | его
0 0 0 **1** 0 0 0 | **пришпорить**
1 0 0 0 0 0 0 | чтобы
0 0 0 **1** 0 0 0 | **слышать**
1 0 0 0 0 0 0 | как



Weka 3.9.4

Метод кластеризации:
HierarchicalClusterer
(Иерархический классификатор) с настройкой измерения расстояния **EuclideanDistance**

РЕЗУЛЬТАТЫ КЛАСТЕРИЗАЦИИ

Мой новый друг снисходительно улыбнулся. Но как же я удивился, когда мой **строгий судья** вдруг просиял.

(А. де Сент-Экзюпери)

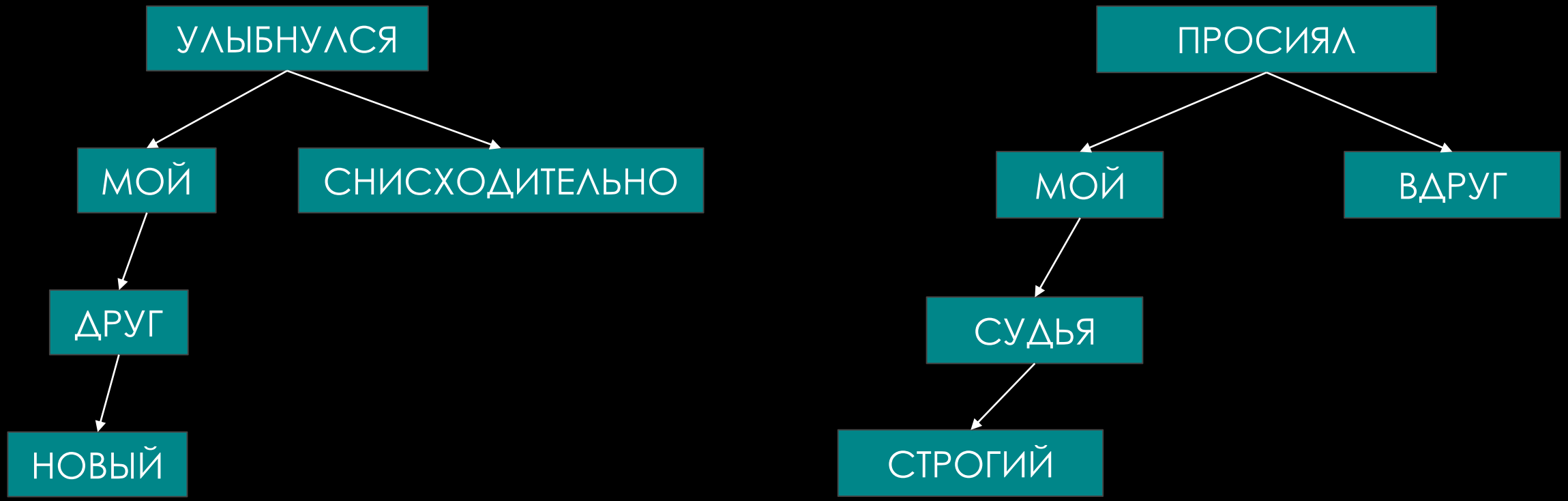
2 0 0 0 | **мой**
1 0 0 0 | **новый**
1 0 0 0 | **друг**
0 1 0 0 | снисходительно
0 0 1 0 | улыбнулся
0 1 0 0 | но
0 1 0 0 | как
0 1 0 0 | же
0 0 0 1 | я
0 0 1 0 | удивился
0 1 0 0 | когда
1 0 0 0 | **строгий**
1 0 0 0 | **судья**
0 1 0 0 | вдруг
0 0 1 0 | просиял

Белый снег летел за окном.

1 0 0 | **белый**
1 0 0 | **снег**
0 1 0 | летел
0 0 1 | за
1 0 0 | окном

ГРУППИРОВКА СИНТАКСИЧЕСКИХ СТРУКТУР

Прилагательное + существительное + предлог + глагол + наречие



ГРУППИРОВКА СИНТАКСИЧЕСКИХ СТРУКТУР

Мой	**новый**	**друг**	**снисходительно**	**улыбнулся**
18, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 2, 0	18, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 2, 0	17, 0, 0, 1, 0, 0, 0, 0, 0, 0, 2, 0, 1, 0, 2, 0	9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	20, 0, 0, 0, 0, 0, 0, 3, 2, 0, 0, 2, 1, 2, 2, 0
Мой	**строгий**	**судья**	**вдруг**	**просиял**
18, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 2, 0	18, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 2, 0	17, 0, 0, 1, 0, 0, 0, 0, 0, 0, 2, 0, 2, 0, 2, 0	9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	20, 0, 0, 0, 0, 0, 0, 3, 2, 0, 0, 2, 1, 2, 2, 0
Новый	**друг**	**снисходительно**	**улыбнулся**	
18, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 2, 0	17, 0, 0, 1, 0, 0, 0, 0, 0, 0, 2, 0, 1, 0, 2, 0	9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	20, 0, 0, 0, 0, 0, 0, 3, 2, 0, 0, 2, 1, 2, 2, 0	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
	Новый	**друг**	**снисходительно**	**улыбнулся**
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	18, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 2, 0	17, 0, 0, 1, 0, 0, 0, 0, 0, 0, 2, 0, 1, 0, 2, 0	9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	20, 0, 0, 0, 0, 0, 0, 3, 2, 0, 0, 2, 1, 2, 2, 0

РЕЗУЛЬТАТЫ ПРИМЕНЕНИЯ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ИССЛЕДОВАНИЙ В NLP

- Необходимость создания и реализации **лингвистических способов векторизации текста**, ориентированных на проводимое исследование.
- Результаты часто требуют **дополнительного анализа** и не могут напрямую использоваться на практике.
- При расширении набора учитываемых характеристик существенно **возрастает объем выборки** для кластеризации или классификации.
- Для сокращения размерности выборки эффективнее использовать **методы машинного обучения в сочетании с алгоритмическими методами.**
- **Результаты кластеризации** дают возможность получения исходных данных для создания выборки для **обучения классификатора**.

ПРОБЛЕМЫ ПРИМЕНЕНИЯ ML В NLP

- Необходимость применения алгоритмических методов для векторизации текстов.
- Сложность определения **набора признаков** для решения конкретных задач, особенно исследовательских.
- Трудоемкость подготовки **обучающей выборки**.
- Очень сильная **зависимость** от обучающей выборки текстов.
- **Высокие требования к ресурсам**: чем больше текстов и/или длина вектора признаков, тем объемнее модель.

- **Если это не решение типовой задачи на заданном наборе текстов, то в большинстве случаев нужны знания лингвистики.**

ВЫВОДЫ

- **Много текстов** с хорошо прослеживающейся **закономерностью** для решения частной задачи => применение методов ML:
  - Простая и быстрая реализация
  - Хорошая точность для частной задачи
  - Не нужно углубляться в алгоритмы лингвистики
- **Входные данные могут быть сильно разными**, нужно учитывать нюансы и особенности языковых явлений в текстах:
  - Алгоритмы, основанные на знании языка
  - Способы векторизации текстов, основанные на знании языка
- Сложность подготовки обучающей выборки текстов:
  - Алгоритмические методы

ПРИМЕНЕНИЕ АЛГОРИТМИЧЕСКИХ МЕТОДОВ И МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

СПАСИБО ЗА ВНИМАНИЕ!

к.т.н., доцент кафедры 319 Полицына Е.В.

к.т.н., доцент кафедры 319 Полицын С.А.

ст. преподаватель кафедры 319 Зеленова М.В.

tasystem@yandex.ru
http://textanalysis.ru/