Текстовая аналитика

Школа прикладного анализа данных Data-Diving

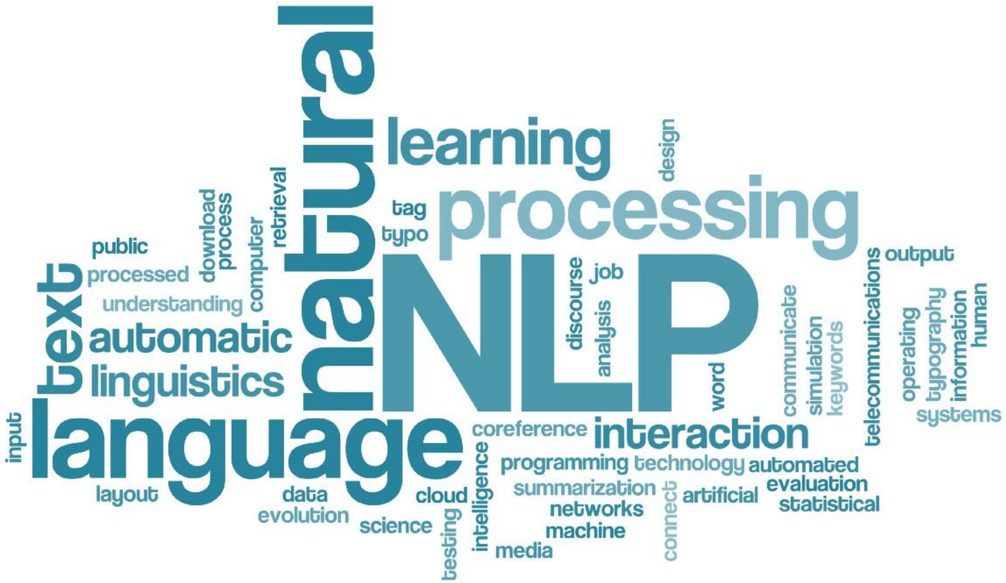
Басина Полина

аналитик Центра прикладного
анализа больших данных

Что такое обработка естественного языка (NLP)?

Обработка текстов на естественном языке
(Natural Language Processing, NLP)

общее направление искусственного интеллекта и математической лингвистики, направленное на изучение методов анализа и синтеза текстов на естественных языках

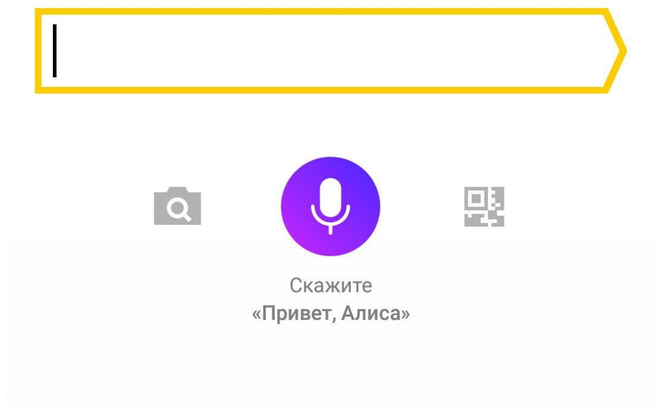


- Распознавание речи
- Анализ текста
- Синтез речи
- Генерирование текста

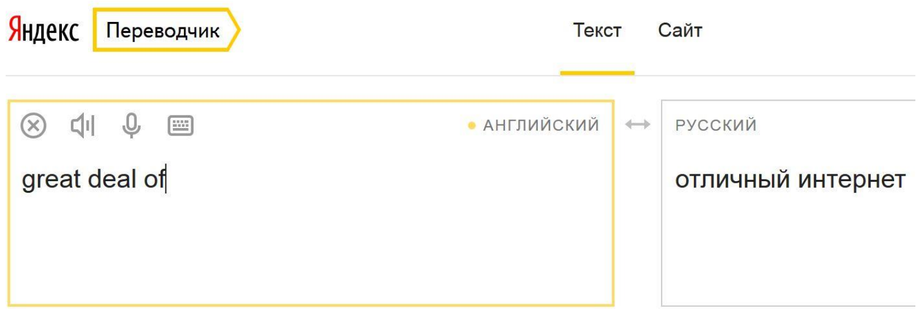
Встречались ли Вы в своей повседневной жизни с **примерами использования обработки естественного языка**?

Популярные задачи

Формирование ответов на вопросы (Question Answering)

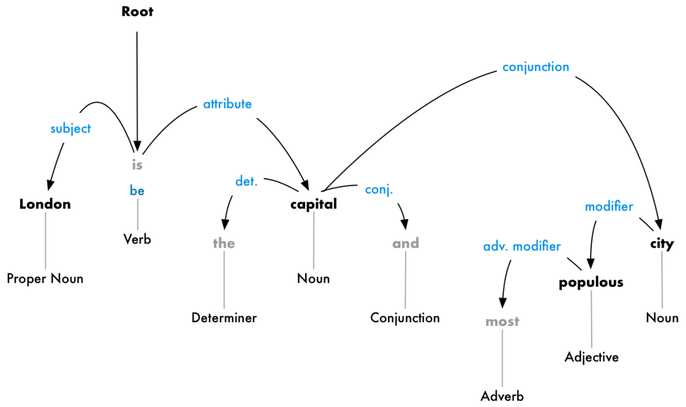


Машинный перевод



Анализ эмоциональной окраски высказываний

Частеречная разметка

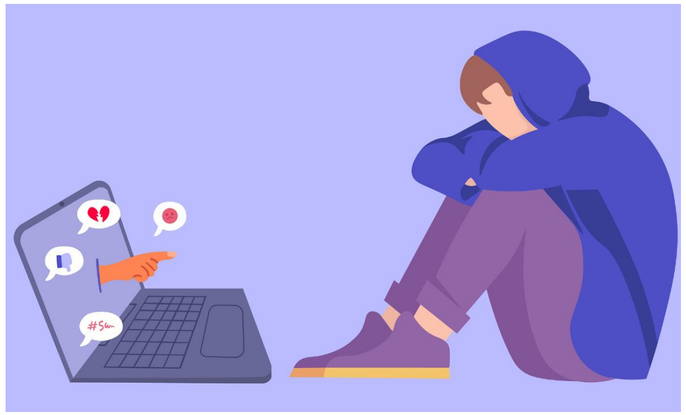


Источник:
https://m.rusvesna.su/news/1507621847
https://www.utwente.nl/nieuws/2017/4/520595/feedbacknow
https://fintraining.livejournal.com/778778.html?view=comments
https://www.pinterest.ru/pin/390476230193508045/

Популярные задачи

Определение фальшивых новостей

Определение токсичных высказываний

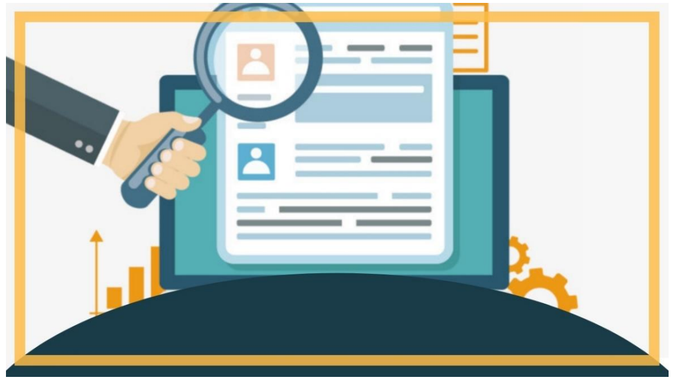


Классификация текстов по различным категориям

Источник:
https://ug.ru/ostatsya-v-zhivyh-chto-izvestno-pro-rasprostranenie-smertelnogo-kontenta-v-soczsetyah/
https://gazeta-hmrn.ru/index.php/pervaya-polosa/22-vlast/2361-nakazanie-za-lozh

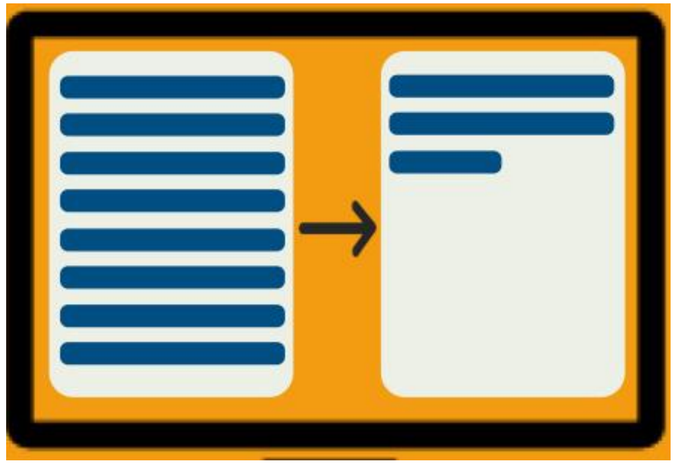
Популярные задачи

Поиск дубликатов



Извлечение фактов

Реферирование текста



Извлечение сущностей

Руководитель НИИ транспорта и дорожного хозяйства Михаил Блинкин утверждает , что может помочь Сергею Собянину решить транспортные проблемы столицы , в частности разгрузить МКАД . « Сама постановка задачи — вернуть МКАД исходное значение — мне чрезвычайно импонирует . Я все последние годы писал и говорил о том , что МКАД превратилась у нас в просёлочную дорогу , подъезд к магазинам , даже в городскую улицу межквартальную » , — цитирует РБК Блинкина .

Закончить разметку абзаца

СПАНЫ | УПОМИНАНИЯ

[НИИ] [НИИ транспорта и дорожного хозяйства]	Org
[Михаил] [Блинкин]	Person
[Сергею] [Собянину]	Person
[МКАД]	Location
[МКАД]	Location
[МКАД]	Location
[РБК]	Org
[Блинкина]	Person

Источник:
https://vk.com/@-168564273-vpesher-profi-vs-deletant
https://aviav.ru/chto-vy-ne-znali-ob-aviacii.html
https://www.pvsm.ru/data-mining/107327
https://www.pngegg.com/id

Основные термины

Токен – текстовые единицы
Символы, слова, словосочетания и предложения.

Документ – совокупность токенов, которые принадлежат одной смысловой единице
Предложение, комментарий или пост пользователя.

Словарь – совокупность всех токенов, встречающихся в корпусе текстов.

Корпус – генеральная совокупность всех документов.

Документ 1
«Сегодня мы едем на дачу»

Документ 2
«Мы поехали на речку»

Корпус

Токен – слово

Словарь: сегодня, мы, едем, на, дачу, поехали, речку

.

Поработаем самостоятельно…

Документ 1
«Мы занимаемся обработкой языка»

Документ 2
«Язык сложно анализировать»

Корпус

Токен – слово

Словарь - ?

Поработаем самостоятельно…

Документ 1
«Мы занимаемся обработкой языка»

Документ 2
«Язык сложно анализировать»

Токен – слово

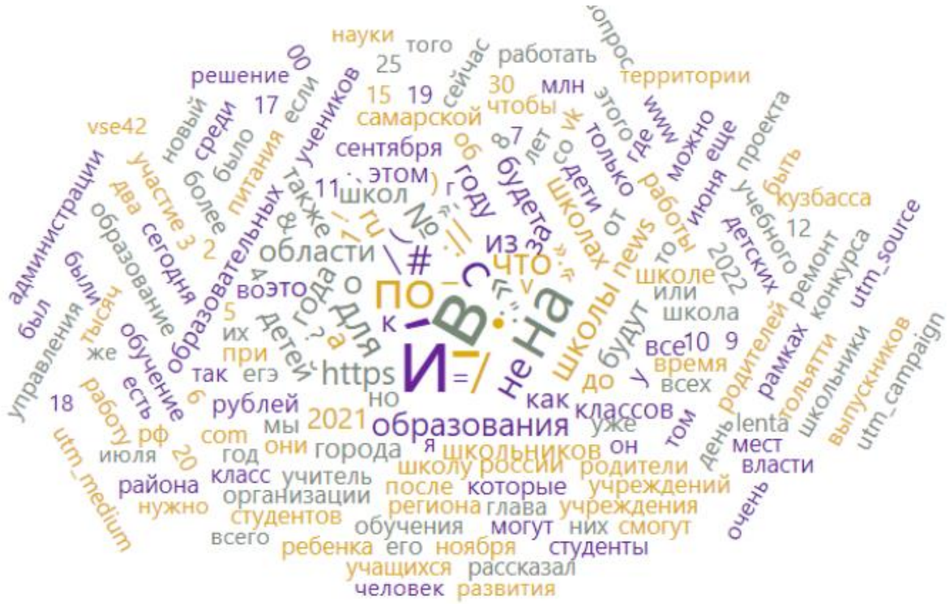
Корпус

Словарь:
Мы, занимаемся, обработкой, языка, язык, сложно, анализировать

Этапы текстового анализа данных

Шаг 2. Предварительная обработка

Токенизация → Приведение к нижнему регистру → Удаление знаков пунктуации → Нормализация лексики → Удаление стоп-слов → Фильтрация



?

Шаг 3. Векторизация

Прямое кодирование

- *вектор - слово*
- *признаки – токены из словаря*

Токен представлен бинарным вектором, который указывает наличие или отсутствие признака; единица ставится тому элементу, который соответствует номеру токена в словаре

Словарь: обработка, естественного, языка

обработка	1	0	0
естественного	0	1	0
языка	0	0	1

Попробуем сами…

Прямое кодирование

Словарь: мама, мыла, раму

мама			
мыла			
раму			

Попробуем сами…

Прямое кодирование

Словарь: мама, мыла, раму

мама	1	0	0
мыла	0	1	0
раму	0	0	1

Шаг 3. Векторизация

Мешок слов BoW

- *вектор всего документа*
- *признаки – токены из словаря*

Каждое число соответствует появлению или количеству появлений соответствующего слова в тексте

Словарь: обработка (1), естественный (2), язык (3), русский (4), английский (5), я (6), занимаюсь (7), нравиться (8)

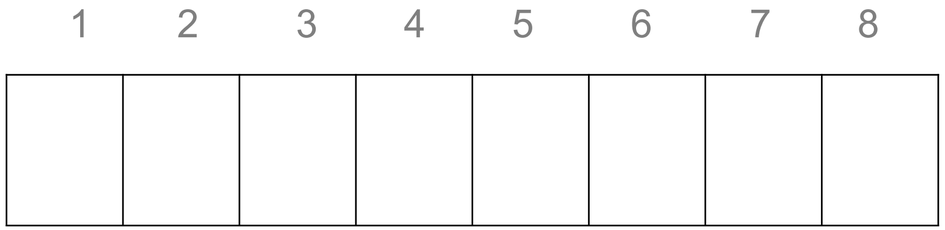
	1	2	3	4	5	6	7	8
Документ 1: Мне нравиться заниматься обработкой естественного языка	1	1	1	0	0	1	1	1
Документ 2: Мне нравиться заниматься обработкой русского языка	1	0	1	1	0	1	1	1
Документ 3: Мне нравиться заниматься обработкой естественного языка	1	0	0	1	1	1	1	1

Попробуем сами…

Мешок слов BoW

Словарь: обработка (1), образовательный(2), выбинар (3), русский (4), английский (5), я (6), смотреть(7), любить(8)

Документ 1:
Я люблю смотреть образовательные вебинары



Попробуем сами…

Мешок слов BoW

Словарь: обработка (1), образовательный(2), выбинар (3), русский (4), английский (5), я (6), смотреть(7), любить(8)

Документ 1:
Я люблю смотреть образовательные вебинары

1	2	3	4	5	6	7	8
0	1	1	0	0	1	1	1

Шаг 3. Векторизация

TF-TDF

- *вектор всего документа*
- *признаки – токены из словаря*

Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции

Словарь: обработка (1), естественный (2), язык (3), данные (4), фотографий(5), я (6), занимаюсь (7), нравиться (8)

Документ 1:
Мне нравиться заниматься обработкой естественного языка

0.1	0.92	0.92	0	0	0.1	0.1	0.1

Документ 2:
Мне нравиться заниматься обработкой фотографий

0.1	0	0	0	0.92	0.1	0.1	0.1

Документ 3:
Мне нравиться заниматься обработкой данных

0.1	0	0	0.92	0	0.1	0.1	0.1

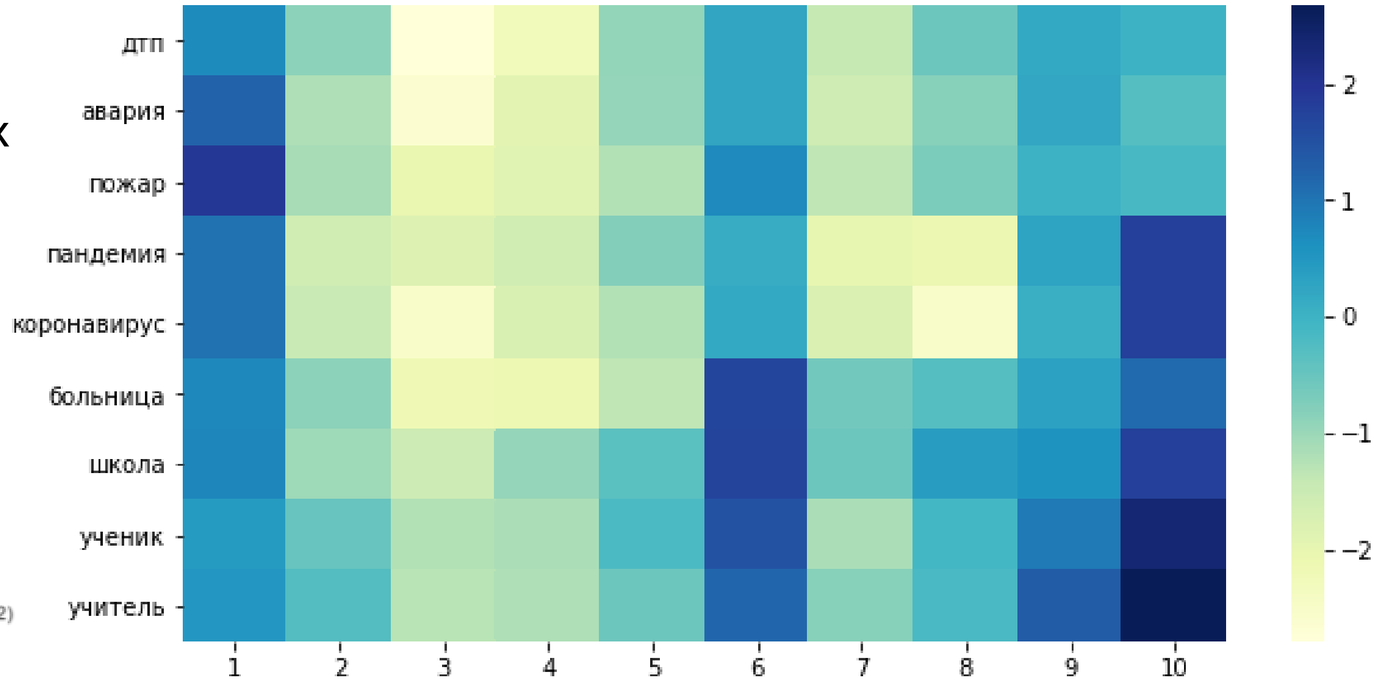
Шаг 3. Векторизация
Недостатки классических подходов

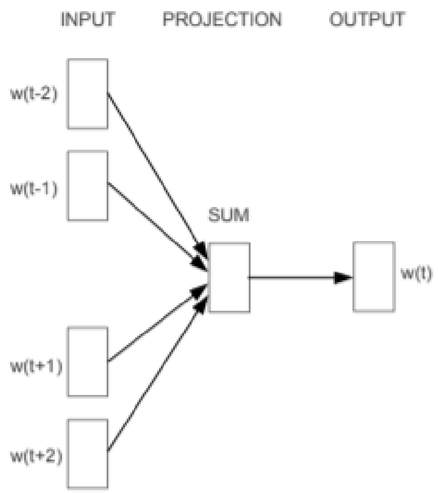
- отсутствие учета контекста и порядка слов
- высокая размерность получаемых векторов
- невозможно эффективно представить новые слова, отсутствующие в словаре

Шаг 3. Векторизация

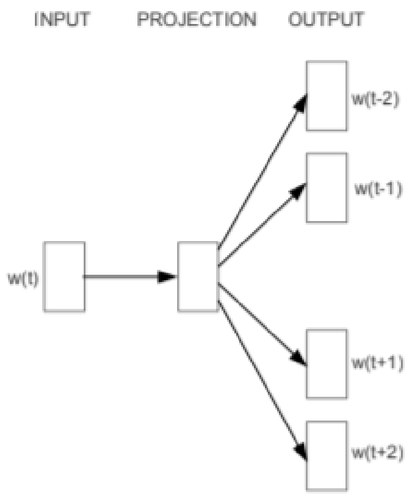
Статичные представления Word2vec, FastText

Статичные модели построены на принципах дистрибутивной семантики, основная идея которых заключается в том, что значение определяется употреблением, а семантика может быть получена из контекстов, в которых употребляется данное слово





CBOW



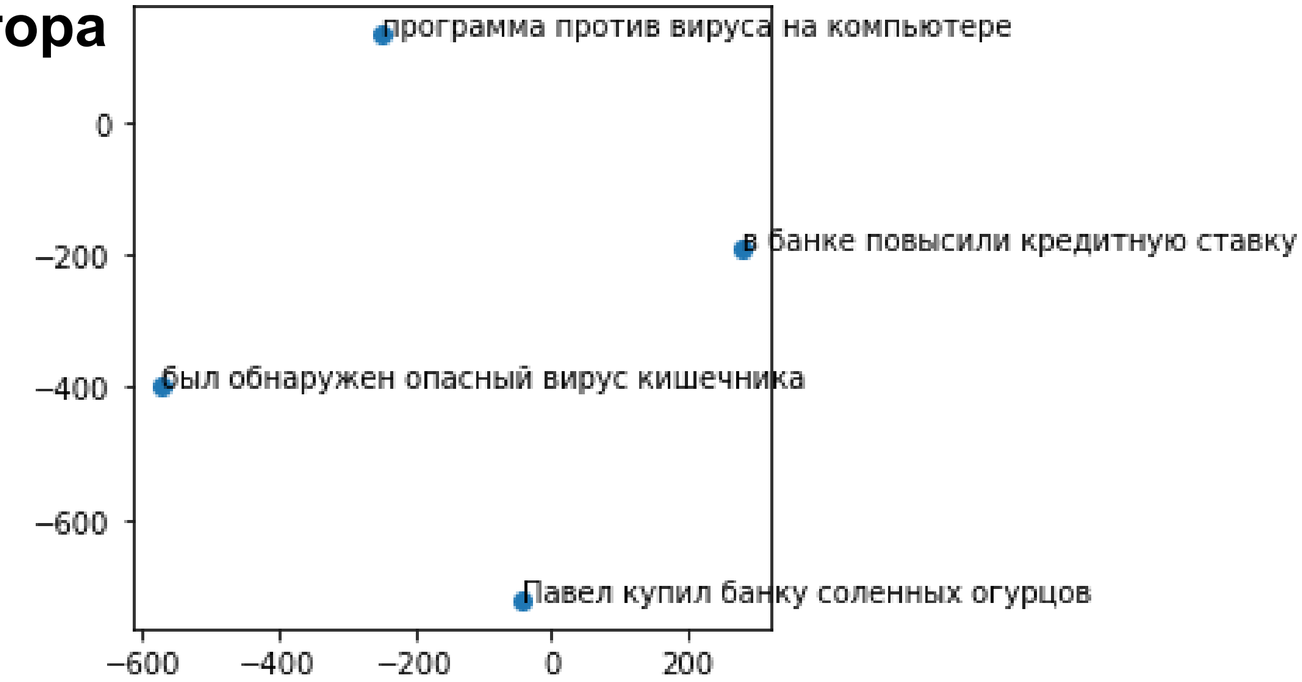
Skip-gram

- *вектор – слово*
- *признаки – токены из словаря*

Шаг 3. Векторизация

Контекстуализированные вектора ELMo, RuBERT

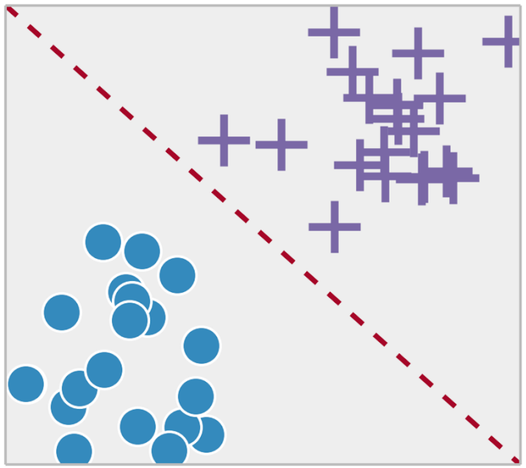
В разных контекстах одна и та же текстовая конструкция может означать различные понятия. В связи с этим, в последние годы получило развитие направление по формированию зависящих от контекста векторных представлений.



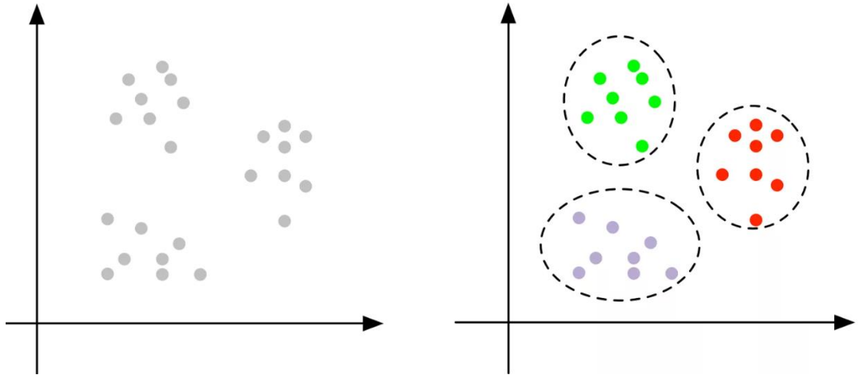
- *вектор – слово*
- *признаки – токены из словаря*

Шаг 4. Методы интеллектуального анализа данных

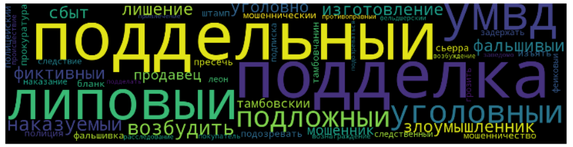
Классификация



Тематическое моделирование

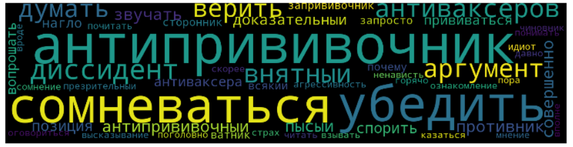


Topic 33

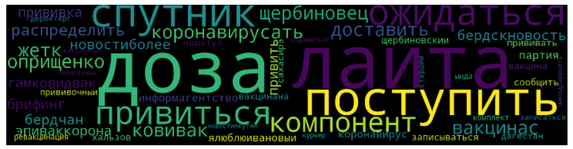


Topic 8

Topic 13



Topic 7

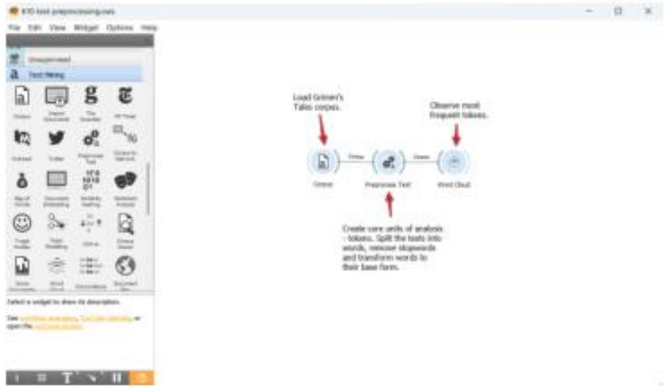


Источник:
https://medium.com/@meghanakantharaj/machine-learning-is-not-overwhelming-50cc785b12ab
https://videouroki.net/tests/test-1-4-chietviert.html

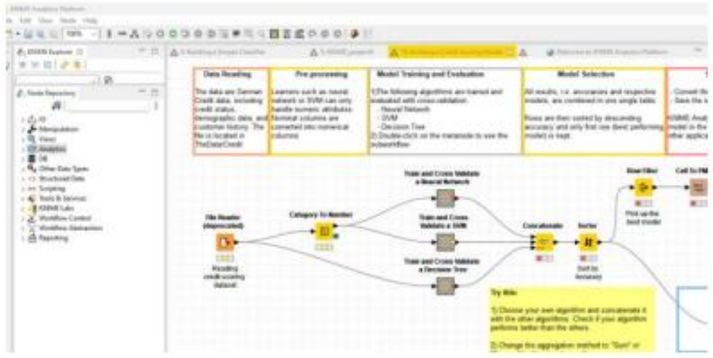
Инструменты

Python

Orange



KNIME



PolyAnalyst

Текстовая аналитика

Школа прикладного анализа данных Data-Diving

Басина Полина

аналитик Центра прикладного
анализа больших данных