

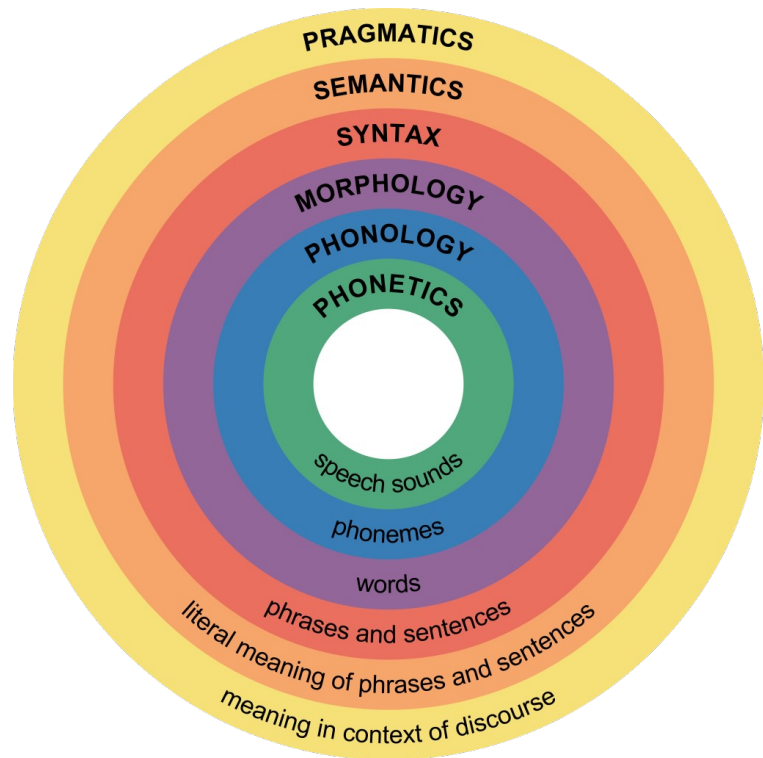


**NLP**

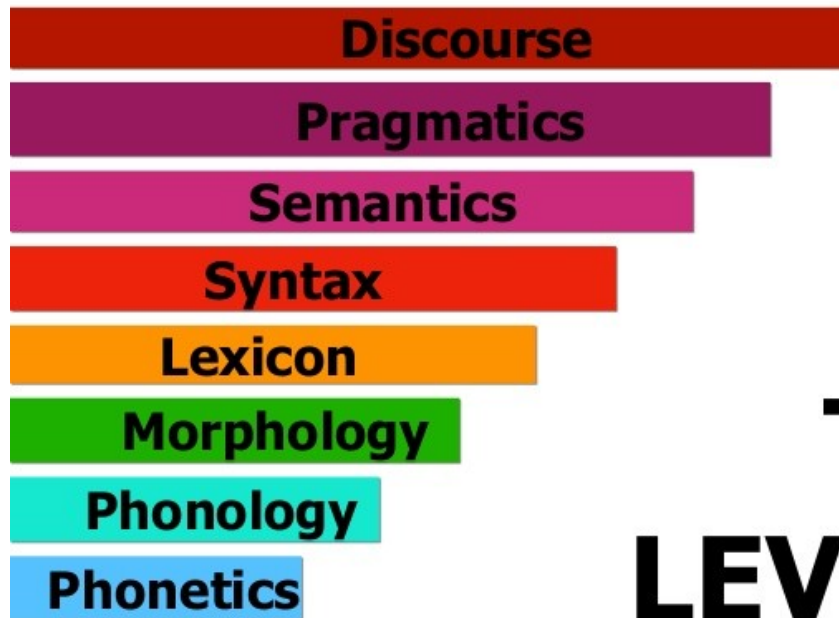
**Natural language  
processing**



# Немного лингвистики



## LANGUAGE



**THE  
LEVELS**

# Разобраться поподробнее поможет:

- Энциклопедия “Языкознание для детей” – название обманчиво.
- Реформатский. Введение в языкознание.

# NLP



**Raw text**

**Tokenization**

**Text\_cleaning**

**POS tagging**

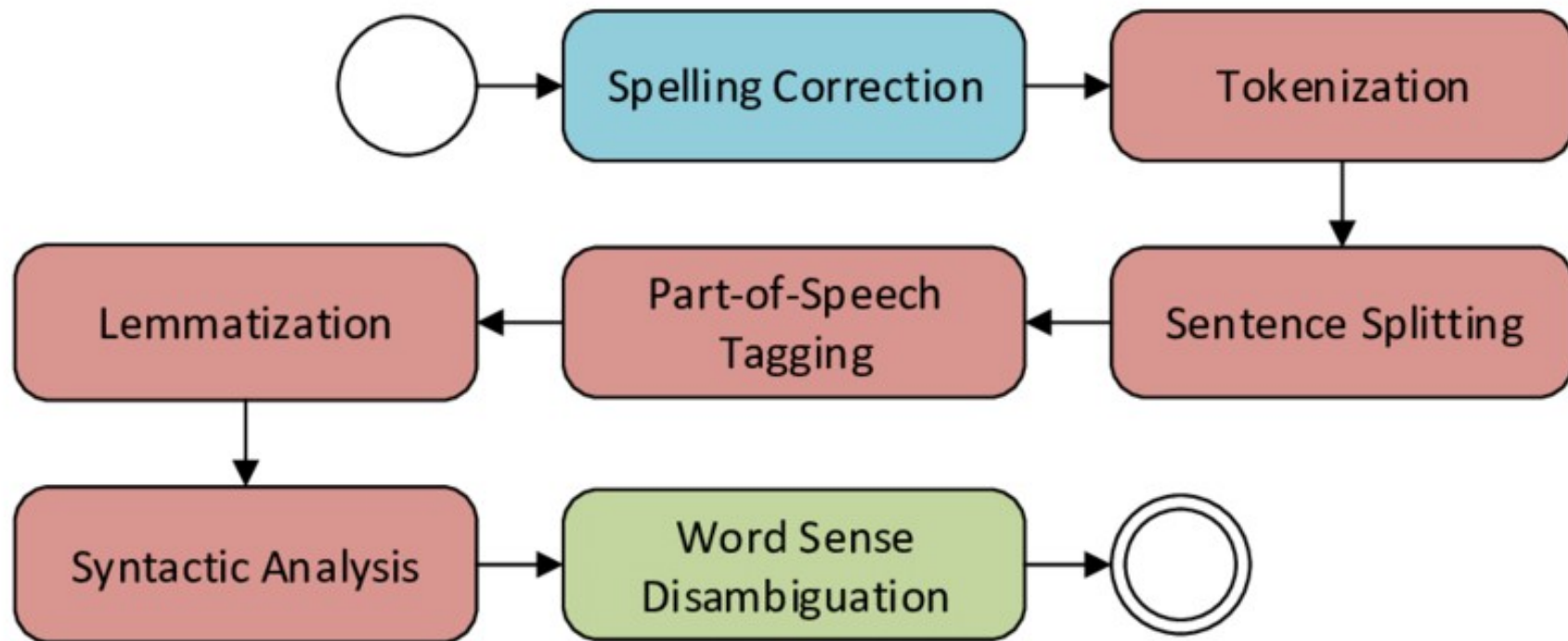
**Stopwords**

**Lemmetization**

**Cleaned text**

**ML Model**





# Морфологический анализ

Василий	Пупкин	приобрёл	контрольный	пакет	акций
<b>сущ.</b> им. падеж, ед. число, муж. род	<b>сущ.</b> им. падеж, ед. число, муж. род	<b>глагол</b> сов. вид прош. время ед. число муж. род	<b>прил.</b> им. падеж, ед. число, муж. род	<b>сущ.</b> им. падеж, ед. число, муж. род	<b>сущ.</b> род. падеж мн. число жен. род

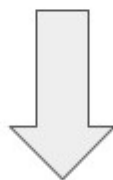


## Синтаксический анализ



# Распознавание именованных сущностей

Василий Пупкин приобрёл контрольный пакет акций ООО "Рога и копыта". Сделка состоялась в Москве в минувшую среду.



**Василий Пупкин** приобрёл контрольный пакет акций  
**ООО "Рога и копыта"**. Сделка состоялась в **Москве** в  
минувшую среду .

# Построение диалоговых систем (чат-ботов)

**Source text:** Василий Пупкин приобрёл контрольный пакет акций ООО "Рога и копыта". Сделка состоялась в Москве в минувшую среду.

**Question 1:** Кто приобрёл контрольный пакет акций?

**Answer 1:** *Василий Пупкин*

**Question 2:** А где состоялась сделка?

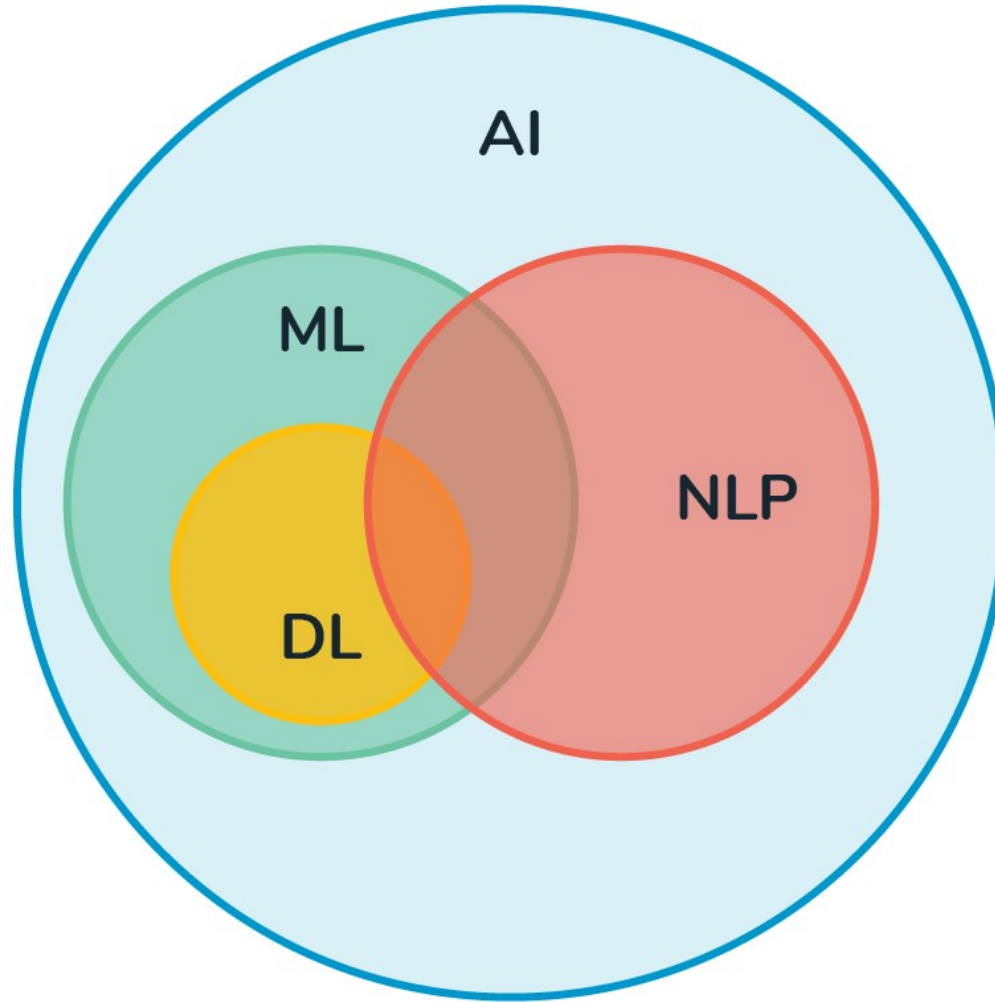
**Answer 2:** *в Москве*





# Тематическое моделирование

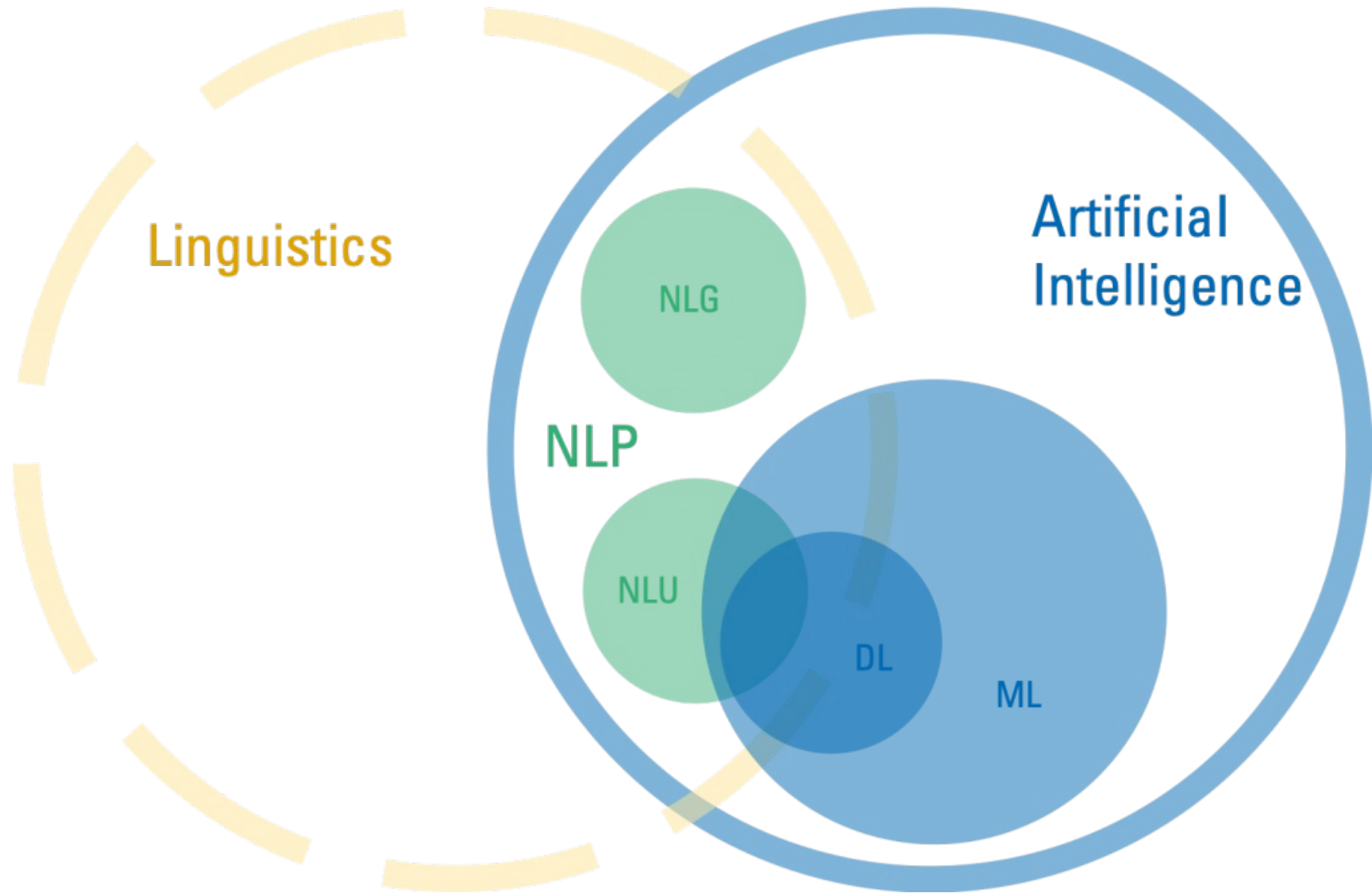
Экономика	Наука	...
<u>Ключевые слова:</u> акция, контрольный пакет, сделка ...	<u>Ключевые слова:</u> учёный, наблюдение, исследовать ...	
Василий Пупкин приобрёл контрольный пакет акций ООО "Рога и копыта". Сделка состоялась в Москве в минувшую среду.	Британские учёные выяснили, что кошки делают по ночам, пока их хозяева спят. Группа специалистов провела наблюдения за 50 котами и кошками из графства Саррей.	

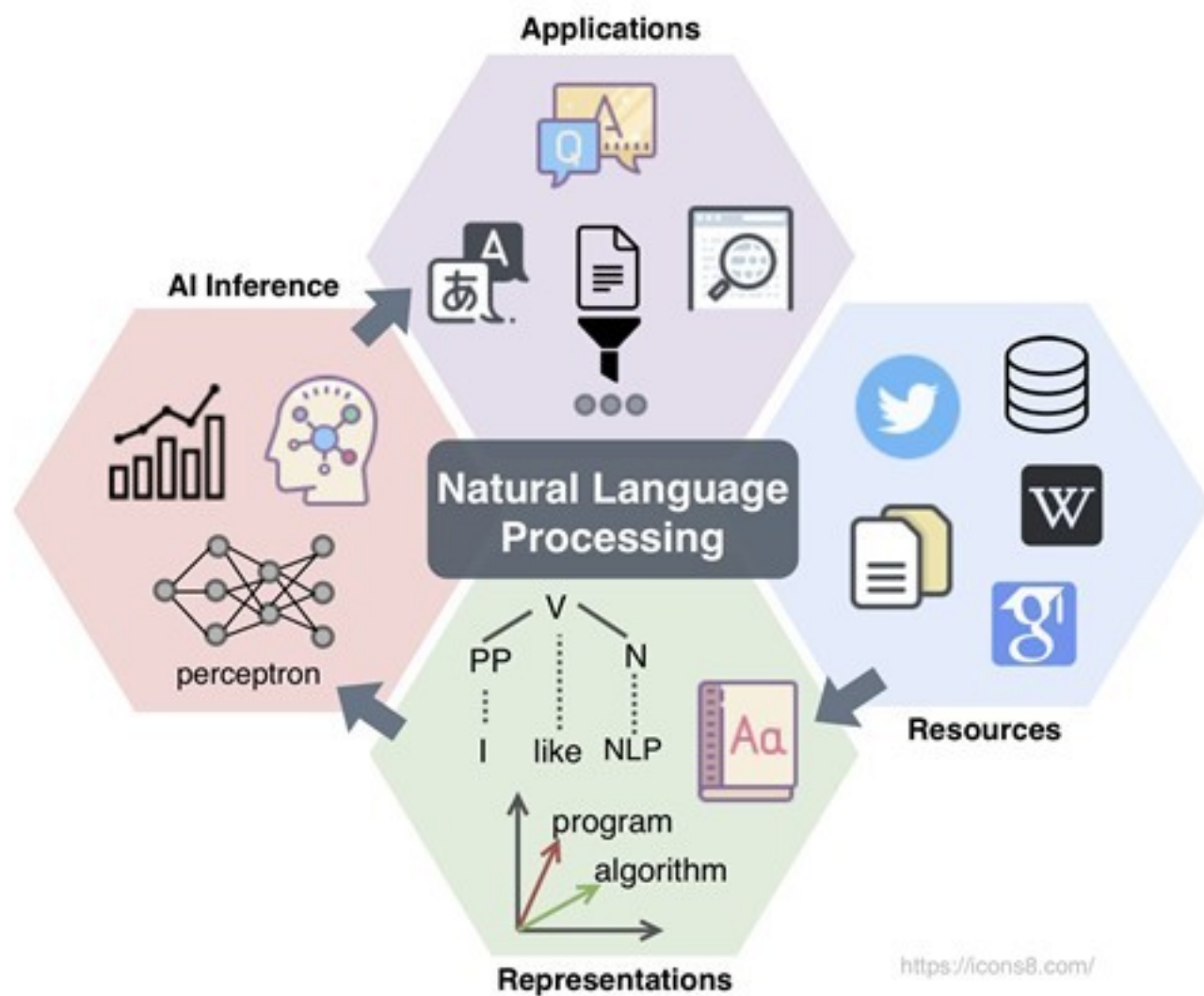
# Распознавание тональности высказывания

Василий Пупкин приобрёл контрольный пакет акций ООО "Рога и копыта".	нейтральная эмоция
Василий Пупкин - негодяй, ведь это был мой пакет акций!	негативная эмоция
Василий Пупкин - успешный бизнесмен, учредивший уникальный холдинг по заготовлению рогов и копыт	позитивная эмоция



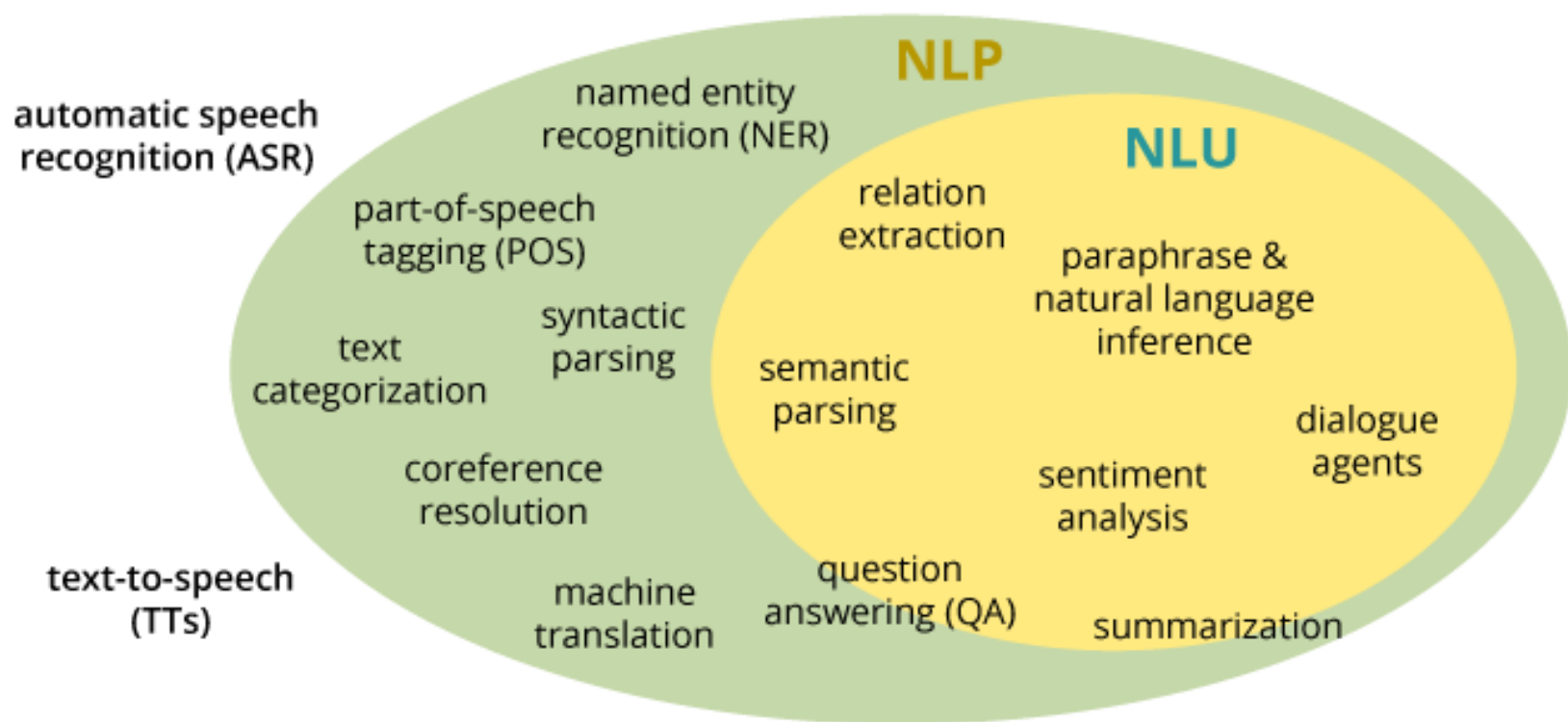
-  Artificial intelligence
-  Machine learning
-  Language Processing
-  Deep learning



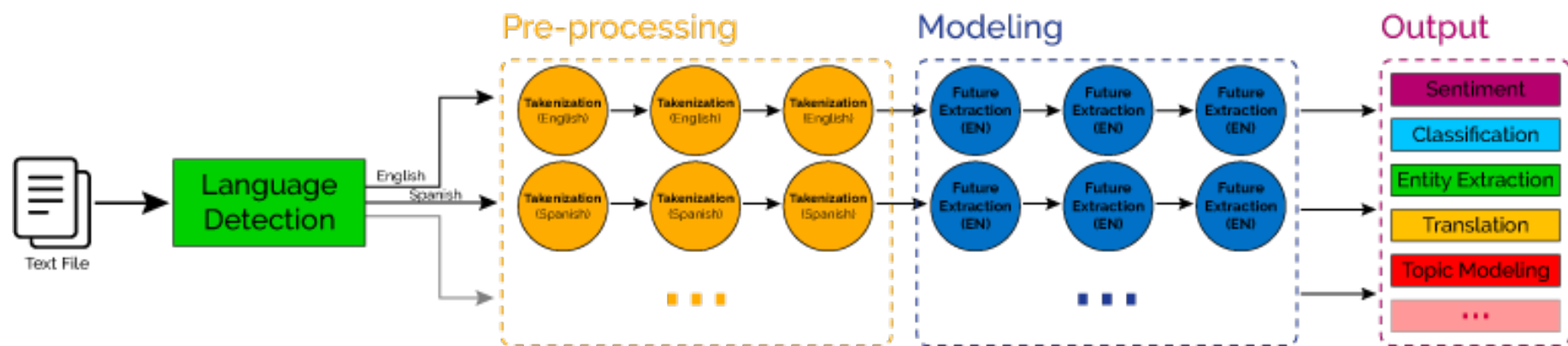




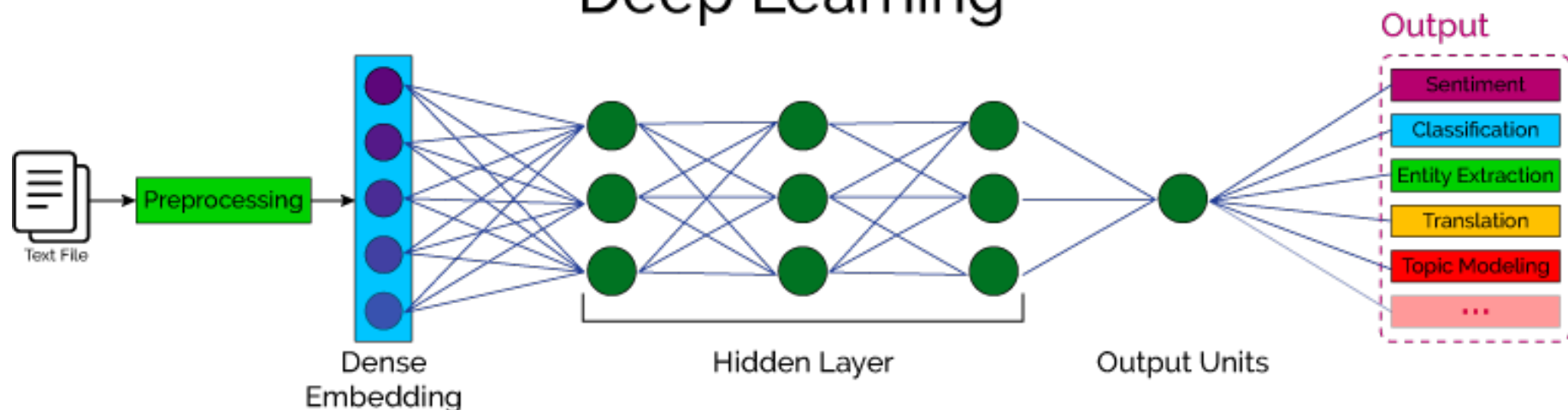
# Terminology: NLU vs. NLP vs. ASR



# Classical NLP

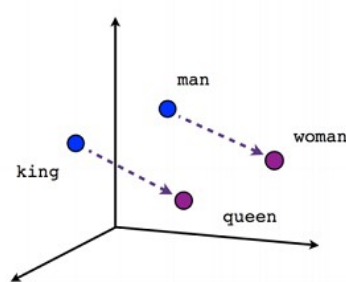


# Deep Learning

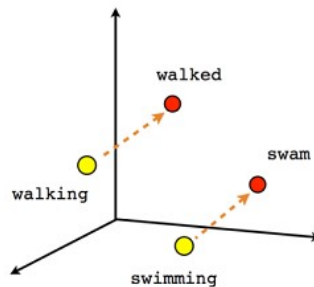


# Эмбеддинги – сила!

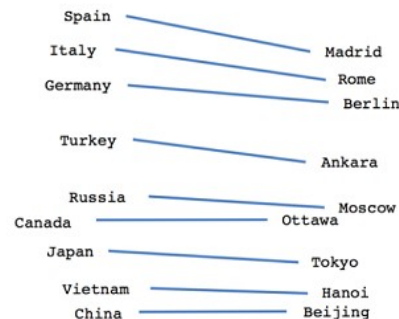
1. Контексто-независимые – Word2Vec, GloVe
2. Контексто-зависимые – ELMo, BERT
3. fasttext – может предсказать вектор для нового слова/несуществующего.



Male-Female



Verb tense



Country-Capital

Неплохой пост про BERT - <https://habr.com/ru/post/487358/>

# Как превратить текст в набор чисел?

## Вариант 1 - “мешок слов”

Словарь: 1) в; 2) купить; 3) магазин; 4) мама; 5) Маша; 6) мыло; 7) мыть; 8) рама

Признаки:  $f_1, \dots, f_8$ , частоты соответствующих слов словаря в текстах

Тексты:

	$f_1$	$f_2$	$f_3$	$f_4$	$f_5$	$f_6$	$f_7$	$f_8$
Мама мыла мылом Машу.	0	0	0	1	1	1	1	0
Мама мыла, мыла раму.	0	0	0	1	0	0	2	1
В магазине купила мама мыло.	1	1	1	1	0	1	0	0

# Как превратить текст в набор чисел?

## Вариант 2 - “цепочка слов-векторов”

Словарь: 1) в; 2) купить; 3) магазин; 4) мама; 5) Маша; 6) мыло; 7) мыть; 8) рама

Признаки:  $f_1, \dots, f_8$ , бинарный признак (есть слово – нет слова)

	$f_1$	$f_2$	$f_3$	$f_4$	$f_5$	$f_6$	$f_7$	$f_8$
Маша	0	0	0	0	1	0	0	0

	$f_1$	$f_2$	$f_3$	$f_4$	$f_5$	$f_6$	$f_7$	$f_8$
купить	0	1	0	0	0	0	0	0

# Word2Vec - слово в семантический вектор



2013 год

**Tomas Mikolov** и другие

*Efficient Estimation of Word  
Representations in Vector Space*

---

## Дистрибутивная гипотеза

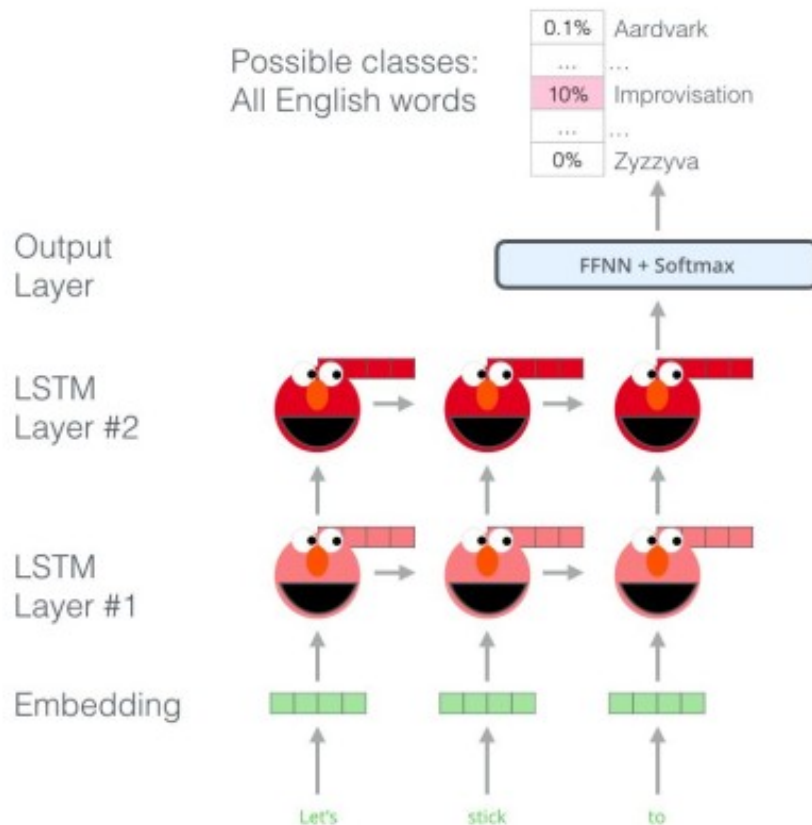
Слова, встречающиеся в схожих контекстах, имеют близкие значения!

Ученые провели **анализ** алгоритмов



Ученые провели **исследование** алгоритмов

# EMBEDDINGS FROM LANGUAGE MODELS



# BERT

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

- BERT: Pre-training of Deep Bidirectional Transformers for Language
- Understanding
- <https://arxiv.org/pdf/1810.04805>
- Реализация на Python:  
<https://github.com/google-research/bert>



Крупные крутые проекты

# Stanford NLP Group

- <https://nlp.stanford.edu/software/>



# IPavlov



- проект по компьютерной лингвистике

iPavlov.ai <http://ipavlov.ai/>

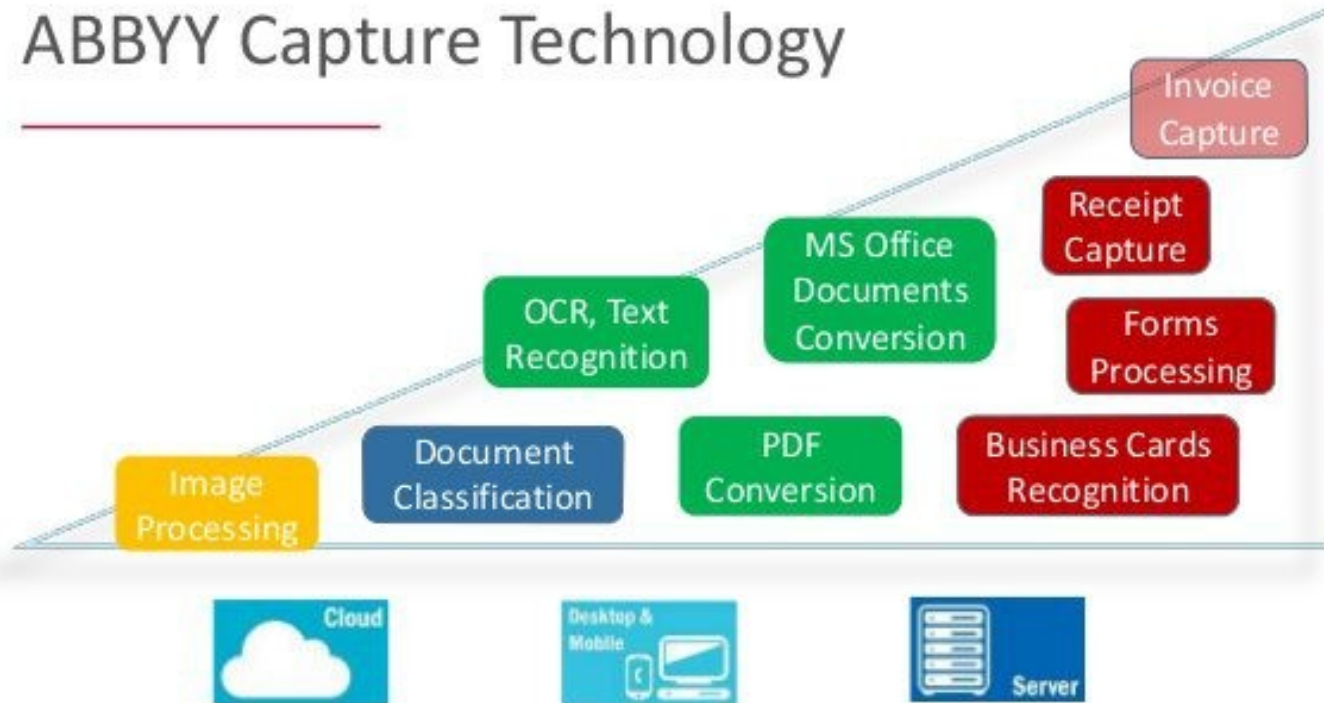
Демонстрация работы:

<http://demo.ipavlov.ai/>

# ABBYY

## ABBYY Capture Technology

---



# Яндекс

- Подробнее тут - <https://habr.com/ru/post/516098/>
- Томита-парсер - <https://yandex.ru/dev/tomita/?turbo=true>

# Библиотеки, которые могут МНОГОЕ...

- SpaCy
- gensim
- NLTK
- UDPipe

тут ещё много всего —

<https://towardsdatascience.com/5-heroic-tools-for-natural-language-processing-7f3c1f8fc9f0>

- <https://opensource.com/article/19/3/natural-language-processing-tools>
- <https://lionbridge.ai/articles/best-nlp-tools-libraries-and-services/>

# Почему для русского всё не так просто?

\*кроме отсутствия финансирования и не такого большого сообщества, как у английского языка

# ОМОНИМИЯ

Коса





# Лексическая и синтаксическая многозначность



“Он из Германии туманной привез  
учености плоды” Пушкин



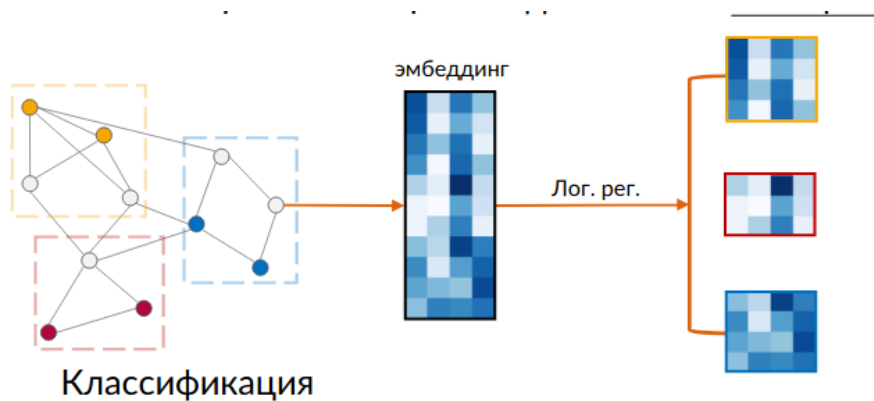
# НО! Выход есть!

Cross-lingual подходы:

- Обучаемся на межъязыковых моделях;
- Дообучаем на данных на русском языке;

# Графы и лингвистика

- Node2Vec models - <http://tsitsul.in/talks/datastart/>
- Семантические сети :)



# Где поискать статьи?

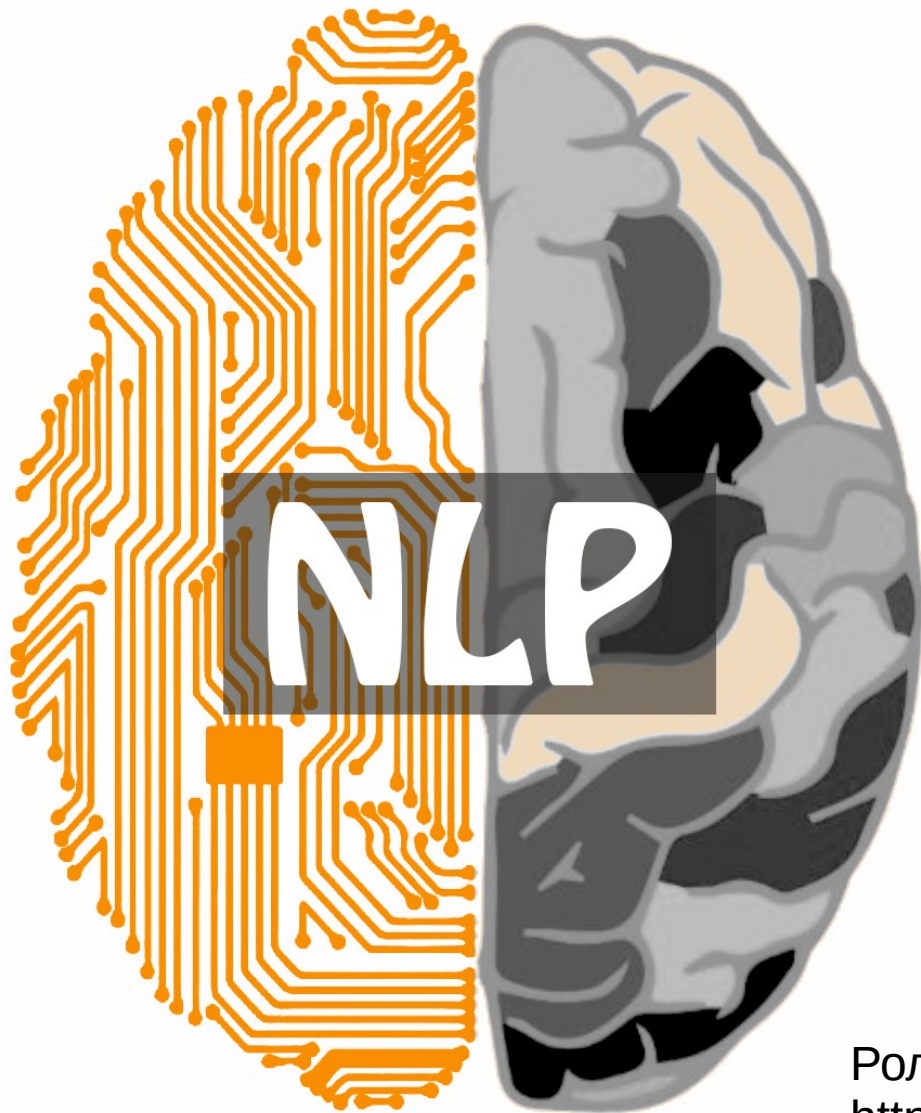
- [aclweb.org](http://aclweb.org) – статьи по компьютерной лингвистике.
- [arxiv.org](http://arxiv.org) – статьи по чему бы то ни было.
- Ежегодные соревнования FactRuEval, SentiRuEval, SemEval
- Конференция Диалог
- Прекрасный курс от Стендфорда  
<http://web.stanford.edu/class/cs224n/>
- Не забываем про Хабр:)

## Доклады про разметку данных из Яндекса

- В своём докладе Алексей Друца, руководитель отдела эффективности и развития в поисковом портале Яндекса, предлагает воспользоваться краудсорсингом. Именно так мы формируем большинство наших датасетов — для Поиска, Такси, Алисы и других проектов. Узнайте, что такое краудсорсинг и как начать им пользоваться. <https://youtu.be/sZJx0SYvRns>
- Кирилл Данилюк сейчас работает техническим менеджером в проекте беспилотных автомобилей Яндекса, а раньше был дата-саентистом в компании RnD Lab. Он и его команда применили подход, альтернативный краудсорсингу, и получили набор данных неплохого качества. Это называется полуавтоматическая разметка: часть работы по формированию датасета выполняется людьми вручную, а часть делает машина. Подробности — в докладе Кирилла. <https://youtu.be/nK2Ra0Xq0So>
- Перейдём к следующему этапу. Предположим, вы собрали подходящий датасет, но он получился неоднородным — отдельные примеры в нём оказывают слишком большое влияние на процесс машинного обучения. Эффект от этого может оказаться как позитивным, так и негативным. Борис Шарчилов, руководитель группы счастья авторов в Яндекс.Дзене, делится опытом поиска «влиятельных» фрагментов выборки. \*\*\* <https://youtu.be/X8DqoWQVU-M>
- Рассмотрим пример датасета, построенного на основе текстовых документов. В анализе данных задача, связанная с текстами, часто сводится к их аннотированию — нужно либо за счёт краудсорсинга, либо другими методами понять, о чём текст, и составить краткую аннотацию. Виктор Кантор, который раньше руководил разработкой машинного обучения в Яндекс.Такси, рассказывает, что важно в аннотировании текстов для ML-моделей. Это видео взято из онлайн-курса «Прикладные задачи анализа данных» Физтеха и Яндекса на Coursera: <https://ya.cc/7JXRf> , <https://ya.cc/7JXTU>

# Книги на почитать

- NLP in action - файл в чате в телеграм.
- Speech and Language Processing -  
<https://web.stanford.edu/~jurafsky/slp3/>
- Handbook of NLP



Остались вопросы?

- тг: @museling
- discord: AnastasiaMmm

Ролик на английском -

<https://www.youtube.com/watch?v=8S3qHHUKqYk>