

---

# МЕТОДЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

ГАЗАНОВА Н.Ш.

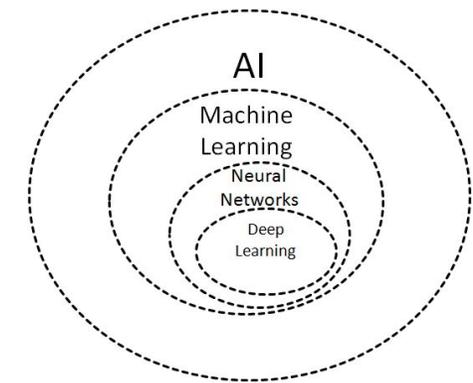
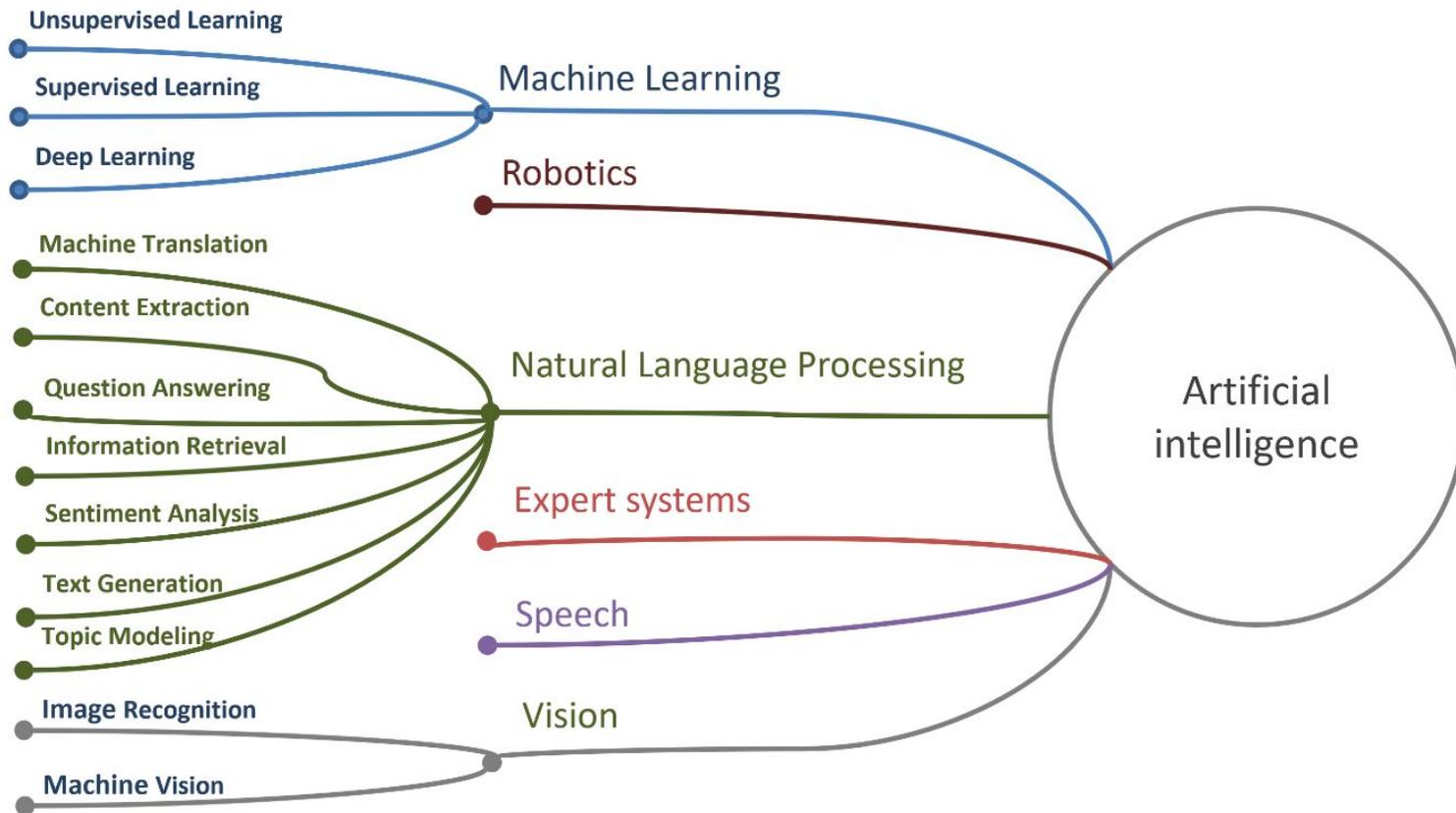




# 1. НАПРАВЛЕНИЯ



# ПОДРАЗДЕЛЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА





# ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

- a. Синтаксис
- b. Поиск
- c. Семантика
- d. Векторная модель и машинное обучение

# СИНТАКСИС

- Формальные грамматики, грамматики Хомского
- Прагматика, семантика и синтаксис
- Применение синтаксиса для токенизации
- Применения синтаксиса для задач семантики



# СИНТАКСИС-ИНСТРУМЕНТЫ

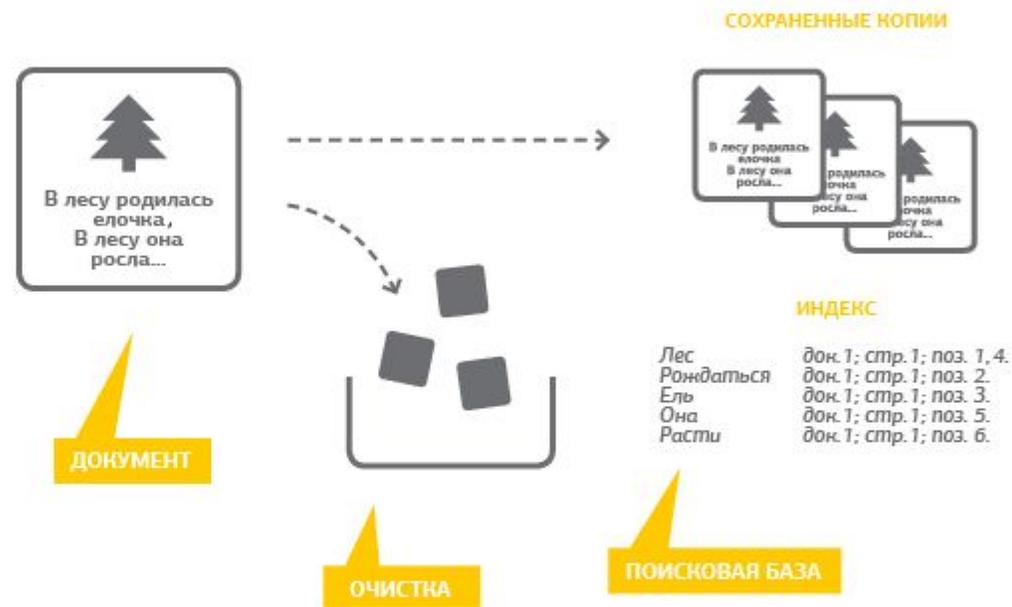
- <https://yandex.ru/dev/tomita/> — парсер для русского
- <https://github.com/natasha/natasha> — NLP-библиотека для русского
- <https://github.com/kmike/pymorphy2> — склонения и падежи для русского и украинского
- <https://deeppavlov.ai/> — анализ, ответы на вопросы, общение
- <https://github.com/nlpub/pymystem3> — стеммер для русского

Стемминг - отсечение от слова окончаний и суффиксов, чтобы оставшаяся часть, называемая *stem*, была одинаковой для всех грамматических форм слова.

<http://snowball.tartarus.org/algorithms/russian/stemmer.html> - алгоритм

# ПОИСК

- Предобработка текста
- Построение поискового индекса
- Выполнение запроса
- Закон Ципфа и его влияние на селективность и масштабирование



Чтобы понять, что именно ищет пользователь, поисковая система проводит тщательный лингвистический анализ запроса. Сначала определяется язык, на котором был сформулирован запрос. Например, в Яндексе индикатором языка является алфавит, используемый в запросе, а также характерные особенности сочетания букв, присущие различным языковым группам.

Далее проводится работа по трактовке морфологии. Поисковая система различает не только слова из запроса во всех их морфологических формах, но и синонимы, однако при ранжировании предпочтение отдается точному вхождению.

Также поисковым системам приходится разграничивать омонимы (слова с одинаковым написанием, но разным значением).

Например, одно и то же слово может быть истолковано и как глагол, и как существительное.

Определиться с наиболее вероятным списком форм помогает статистика совместной встречаемости слов и грамматических признаков. Для сбора статистики Яндекс использует [национальный корпус русского языка](#) и свои собственные корпуса, в которых собрано огромное количество текстов.

# ПОИСК

- Предобработка текста
- Построение поискового индекса
- Выполнение запроса
- Закон Ципфа и его влияние на селективность и масштабирование



## 1—100

Первые сто слов покрывают **37 %** всех текстов.

- |          |               |
|----------|---------------|
| 1. и     | 26. что       |
| 2. в     | 27. весь      |
| 3. не    | 28. год       |
| 4. на    | 29. от        |
| 5. я     | 30. так       |
| 6. быть  | 31. о         |
| 7. он    | 32. для       |
| 8. с     | 33. ты        |
| 9. что   | 34. же        |
| 10. а    | 35. все       |
| 11. по   | 36. тот       |
| 12. это  | 37. мочь      |
| 13. она  | 38. <u>вы</u> |
| 14. этот | 39. человек   |
| 15. к    | 40. такой     |
| 16. но   | 41. его       |
| 17. они  | 42. сказать   |
| 18. мы   | 43. только    |
| 19. как  | 44. или       |
| 20. из   | 45. ещё       |

# СЕМАНТИКА

## Дистрибутивная гипотеза и избыточность языка

На небе только и разговоров , что о море и о \_\_\_\_\_. Там говорят о том , как чертовски здорово наблюдать за огромным огненным шаром , как он тает в волнах . И еле видимый свет , словно от свечи , горит где - то в глубине ..

## TF-IDF и терм-документная матрица

Терм-документная матрица представляет собой математическую матрицу, описывающую частоту терминов, которые встречаются в коллекции документов. В терм-документной матрице строки соответствуют документам в коллекции, а столбцы соответствуют терминам. Существуют различные схемы для определения значения каждого элемента матрицы. Одной из таких является схема TF-IDF. Они полезны в области обработки естественного языка, особенно в методах латентно-семантического анализа.

При создании базы данных терминов, используемых в наборе документов, матрица терминов формируется как матрица инцидентности, строки которой соответствуют документам, а элементы строк - наличию соответствующих терминов в этих документах.

# СЕМАНТИКА

- Векторная модель : концепты , ортогональность и метрика  
Vector Space Model (VSM) – это математическая модель представления текстов, в которой каждому документу сопоставлен вектор, выражающий его смысл. Такое представление позволяет легко сравнивать слова, искать похожие, проводить классификацию, кластеризацию и многое другое. Но обо всём по порядку.
- Метод главных компонент для понижения размерности и выделения ортогональных концептов  
Один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации. Изобретён Карлом Пирсоном в 1901 году. Применяется во многих областях, в том числе, в эконометрике, биоинформатике, обработке изображений, для сжатия данных, в общественных науках.



## 2. ВЕКТОРНЫЕ МОДЕЛИ И МАШИННОЕ ОБУЧЕНИЕ



# ЗАДАЧА

Метод главных компонент рассматривает текст как мешок слов . Для коротких текстов это работает хорошо , но для длинных текстов это уже не так . Кроме того , разница между “А убил В” и “В убил А” будет потеряна .

Методы \*2vec рассматривают слово в маленьком контексте , что привносит элемент порядка в обучение .

Ваша задача построить поисковый движок на базе doc2vec . \* Пример обучения модели doc2vec по [ссылке](#)

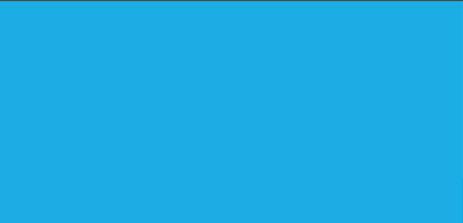
# В УГОДУ СКОРОСТИ

Натренированные векторные представления:

- EN: [English word vectors · fastText](#)
- RU: [natasha/navec](#); [RusVectōrēs: модели](#)



## 3. GOOGLE COLAB NOTEBOOK



# ТЕСТ

Почему использование Jupyter Notebooks и Google Colab для упаковки кода и текстов сейчас наиболее распространена при обмене решениями в индустрии. Отметьте все верные утверждения:

1. Позволяет использовать документ с кодом без переключения окон
2. Можно просматривать в браузере
3. Позволяет легко получить скомпилированный бинарный файл
4. Можно исполнять без установки ПО на собственный компьютер

# ТЕСТ

Почему использование Jupyter Notebooks и Google Colab для упаковки кода и текстов сейчас наиболее распространена при обмене решениями в индустрии. Отметьте все верные утверждения:

- 1. Позволяет использовать документ с кодом без переключения окон**
- 2. Можно просматривать в браузере**
3. Позволяет легко получить скомпилированный бинарный файл
- 4. Можно исполнять без установки ПО на собственный компьютер**



**СПАСИБО ЗА ВНИМАНИЕ!**