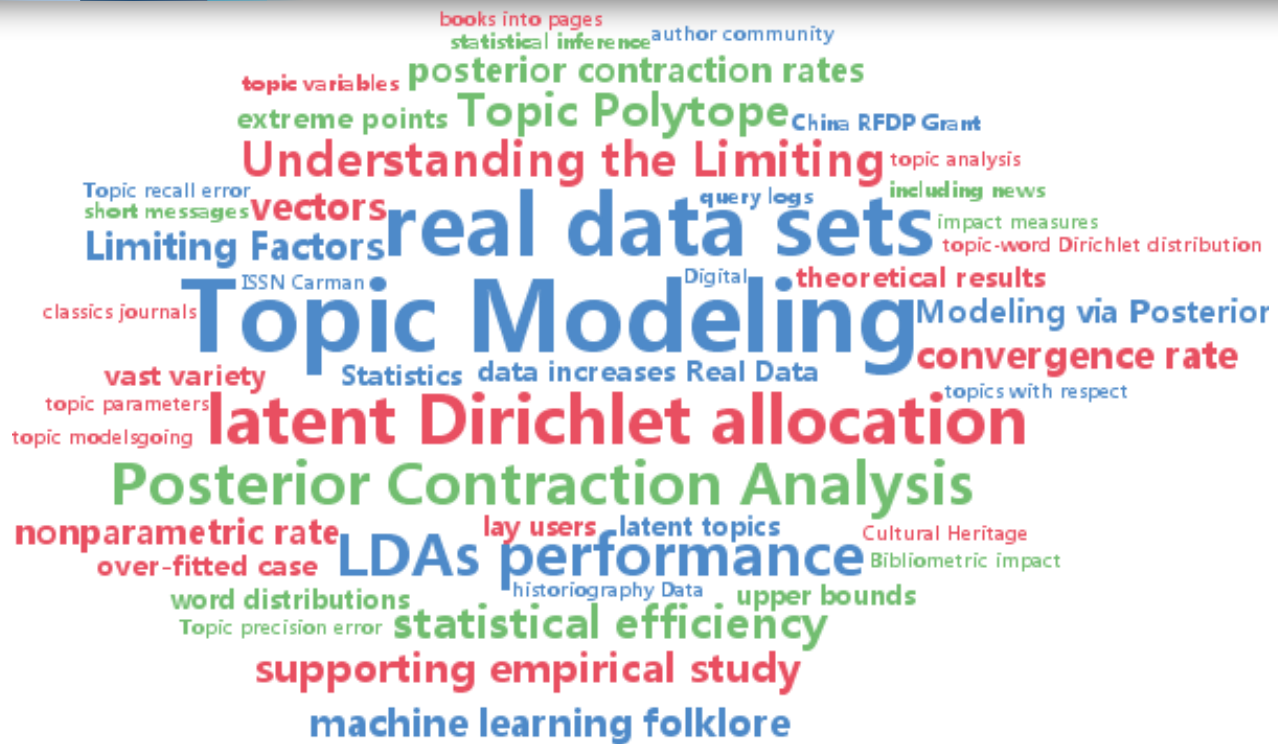




03/02/22



# Благодарень:

Елене Большаковой (МГУ, ВМК)



Татьяне Ландо (Яндекс, ШАД)



Венцеслав Йорданов  
(блогер)





# Семантический анализ -

— этап в последовательности действий алгоритма автоматического понимания текстов, заключающийся в выделении семантических отношений, формировании семантического представления текстов. Один из возможных вариантов представления семантического представления — структура, состоящая из «текстовых фактов».



## Natural Language Processing – (автоматическая) обработка естественного языка

Конечная цель – научить машину  
полноценно понимать обычный  
человеческий текст.



# АКТУАЛЬНОСТЬ

Рост объема текстовой информации, особенно в сети Интернет: человек не в состоянии охватить ее за приемлемое время. Нужны программы извлечения и преобразования информации в форму, удобную для дальнейшей обработки

Возможные приложения:

- мониторинг новостных лент
- составление дайджестов, рефератов, досье
- сбор данных для анализа экономической, производственной и др. деятельности

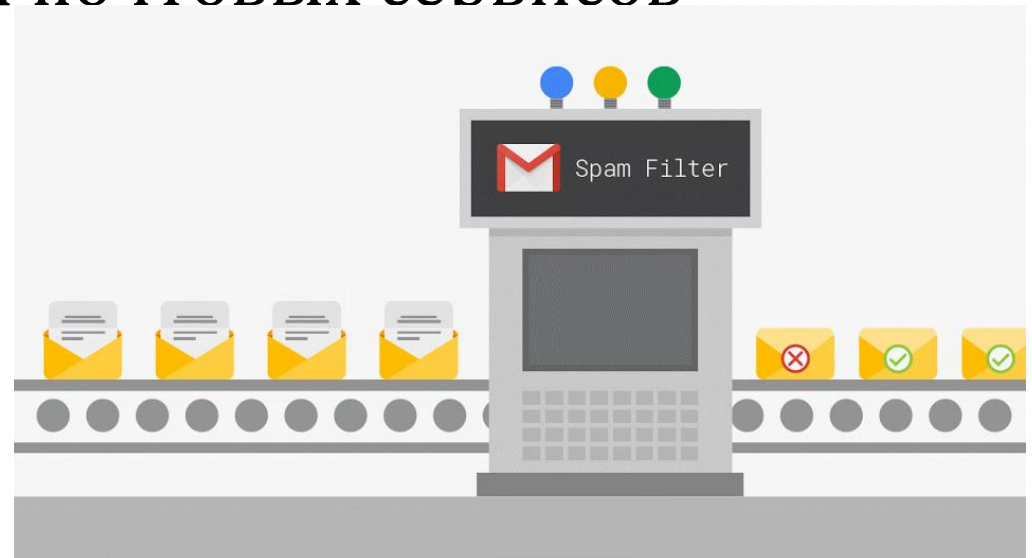


## Natural Language Processing: применения

- Текстовый поиск
- Извлечение фактов
- Диалоговые системы и Question Answering
- Синтез и распознавание речи
- Оценка тональности отзывов
- Кластеризация и классификация текстов.

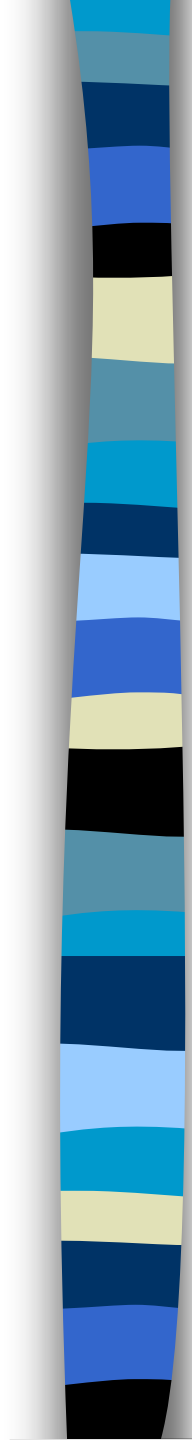
# Natural Language Processing: примеры

- Голосовые помощники типа Siri, Cortana и Алисы
- Спам фильтры почтовых сервисов
- Боты



What can I help you with?





## Извлечение структурированной информации из неструктурированного текста (Text Mining)

- Named Entity Recognition (NER) – извлечение именованных сущностей/объектов;
- Co-reference resolution – разрешение кореференции;
- Information Extraction (IE) – извлечение фактов.





## ВИДЫ ИЗВЛЕКАЕМОЙ ИНФОРМАЦИИ

- Именованные сущности (Named Entities, NE) – значимые объекты : персоны, названия фирм, белков, марки товаров, геогр. названия и т.п.
- Атрибуты объектов: для персоны – должность, место работы, телефон, подразделение
- Отношения между объектами: быть частью, быть владельцем
- Факты/события: прошла встреча, выдан кредит

А также:

- Термины ПО и их связи, ключевые слова текста
- Отзывы и мнения о товарах, услугах, кино и пр.



## ВИДЫ ИЗВЛЕКАЕМОЙ ИНФОРМАЦИИ

- Грейс Патриша Келли (12.11.1929 – 14.09.1982)  
– американская актриса, с 1956 года – супруга князя Монако Ренье III, 10-я княгиня Монако, мать ныне правящего князя Альбера II.
- Объекты (имен.сущности): ФИО, род занятий +даты
- Отношения:
  - супружество: Грейс Патриша Келли, Ренье III
  - быть матерью: Грейс Патриша Келли, Альбер II
- Факты и события:
  - замужество (1956, Грейс П. Келли, Ренье III)
  - правящий князь (Монако, Альбер II)

Можно ли извлечь : Ренье III – отец Альбера II ???



# ИМЕНОВАННЫЕ СУЩНОСТИ

Изначально именованные сущности – это:

- Имена персоналий: И. Сечин, Ben White
- Географические названия: р. Ока, гор. Москва
- Названия компаний/организаций: РЖД, ОАО «Уют»

Сейчас также выделяют:

- Даты и временные отрезки: 02.03.1913, 2 p.m.
- Номера телефонов: +7(123)456-78-90
- Адреса: 3-ая улица Строителей д. 25, кв.12
- Марки товаров: Nokia, Apple, Land Rover
- Обозначения денежных единиц: руб., \$, GBP
- Ссылки на литературу: [2], [Иванов, 1995]
- Гены, белки, хим. вещества:  $\text{H}_2\text{N}-\text{CH}(\text{R})-\text{COOH}$  11



# ИМЕНОВАННЫЕ СУЩНОСТИ:

## СЛОЖНОСТИ ИЗВЛЕЧЕНИЯ

Большое число разных сущностей/объектов,  
постоянно появляются новые

- Множество различных способов именования одной и той же сущности: ВВП, В.В.Путин

William H. Gates, Bill Gates, владелец Microsoft, BG

- Нередко требуется установление кореференции имен (тождества обозначаемых объектов – референтов) ГАИ, ГИБДД – это один референт или разные?

- В зависимости от контекста имен. сущность может относиться к разным видам (категориям): Лена, ВВП  
В России прошли ... – географический объект  
Россия отказалась от ... – страна



# ИМЕНОВАННЫЕ СУЩНОСТИ:

## СЛОЖНОСТИ ИЗВЛЕЧЕНИЯ

- Атрибуты конкретных объектов  
квартира (продажа/покупка): адрес, этаж, метраж, количество комнат, лифт, газ, ...
- Отношения (связи) конкретных объектов

Виды отношений:

- Общие: часть-целое, причина-следствие)
- Зависящие от ПО текста: работать\_в, быть\_владельцем, вступать\_в\_реакцию

При извлечении учитываются типичные конструкции описания атрибутов и отношений

Сложность: отношения непостоянны

# ИМЕНОВАННЫЕ СУЩНОСТИ:

## СЛОЖНОСТИ ИЗВЛЕЧЕНИЯ

При извлечении факта/события информация структурируется в виде семантического фрейма: (набора параметров-атрибутов события)

Примеры:

Яндекс купил Кинопоиск за 80\$ млн. в октябре 2013 г.

Фрейм покупки:

атрибуты: Сумма Покупатель Объект Продавец  
80 млн.\$ Яндекс Кинопоиск ?

Премьер-министр Казахстана Бакытжан Сагинтаев в апреле 2017 посетил офис Microsoft в Сан-Франциско

Фрейм делового визита:

Визитер	Принимающая сторона	Дата
Бакытжан Сагинтаев	офис Microsoft	04.2017 <sub>14</sub>



## Кореференция –

это попытка связать несколько разных отсылок в тексте к одному реальному объекту.

Одним из примеров кореференции является анафора – отсылка к объекту при помощи специальных указателей.

Второй пример кореференции – это синонимия. Она может быть выражена по-разному:

Транслитерация: Yandex – Яндекс.

- Аббревиация: ВТБ – Внешторгбанк – Банк Внешней Торговли.
- Синонимы: больница – госпиталь.
- Словообразование: Москва – московский.
- Графические: авто кредит – автокредит.



# Виды предварительного анализа текста

1. *Графематический*
2. *Лексический*
3. *Морфологический*
4. *Синтаксический*
5. *Семантический анализ*





# Виды предварительного анализа текста

*Текст делится на абзацы, предложения, слова. Затем слова нормализуются – выделяется их начальная форма. Далее проводится полный или частичный синтаксический разбор, определяются зависимости и связи между словами в предложениях.*

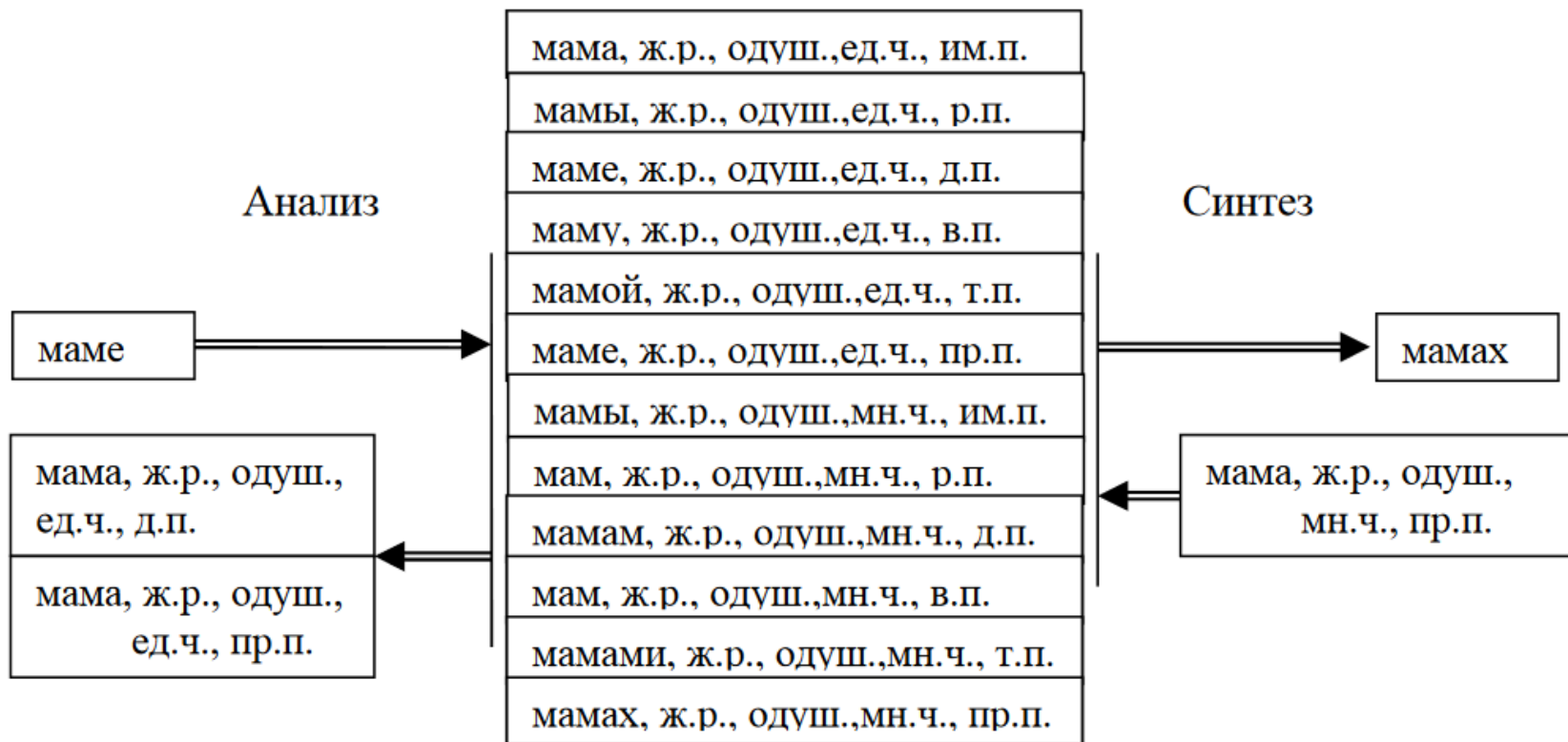
*На первый взгляд кажется, что разбить текст на предложения не составляет никакого труда. Нужно просто ориентироваться на знаки препинания, маркирующие конец предложения. Но работает этот метод далеко не всегда. Ведь, например, точка может обозначать и сокращение, использоваться в дробных числах или URL. Любой знак препинания может использоваться в названиях компаний или сервисов. Например, Yahoo! Или Яндекс.Маркет.*



## Задачей морфологического анализа

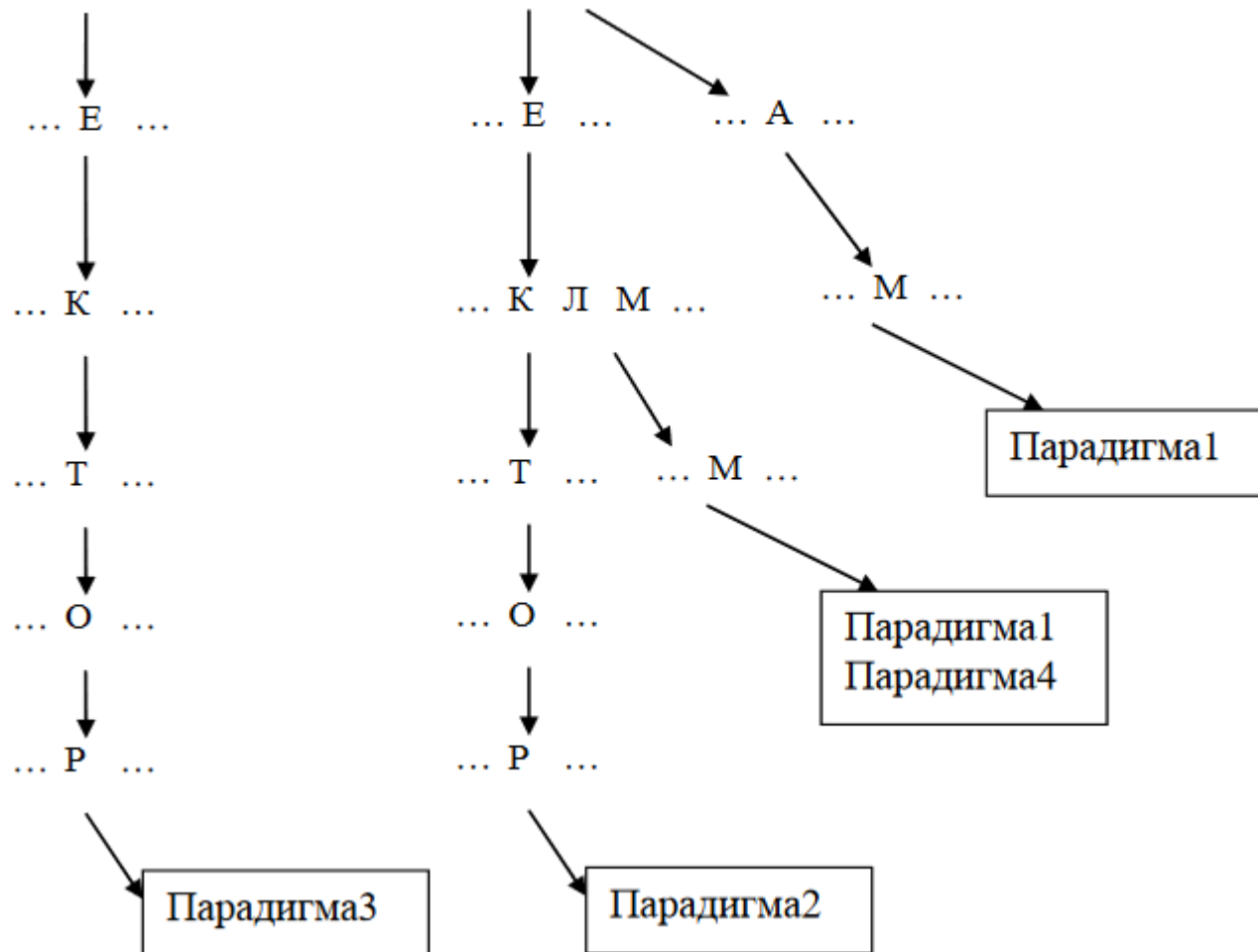
является определение по словоформе нормальной формы, от которой была образована данная словоформа, и набора параметров, приписанных к данной словоформе. При этом может оказаться, что одной словоформе может быть сопоставлено несколько таких пар.

# Пример морфологического анализа и синтеза

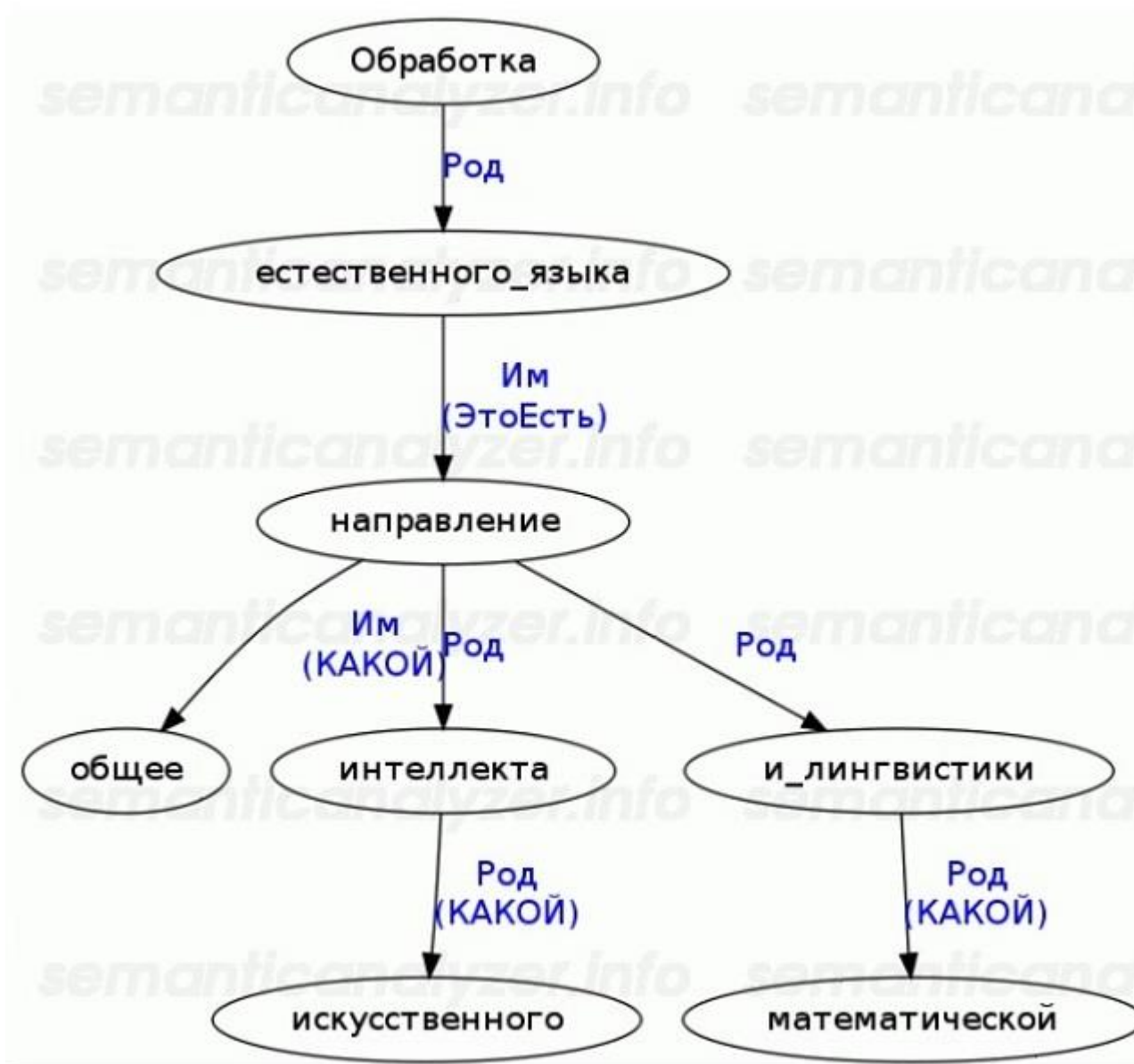


# Фрагменты дерева префиксов для слов «вектор», «лектор», «мама»,

А Б В Г Д Е Ё Ж З И К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Э Ю Я

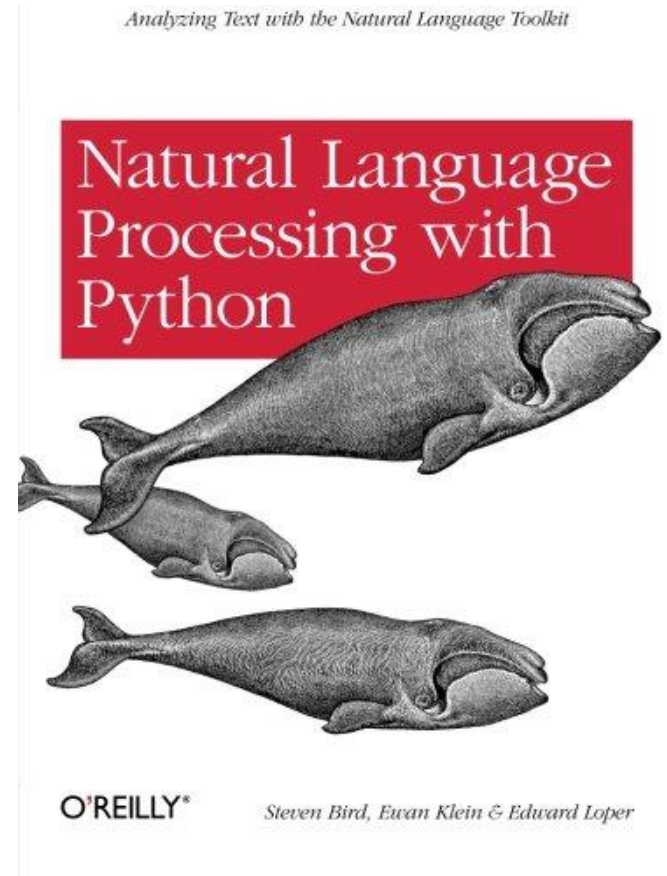
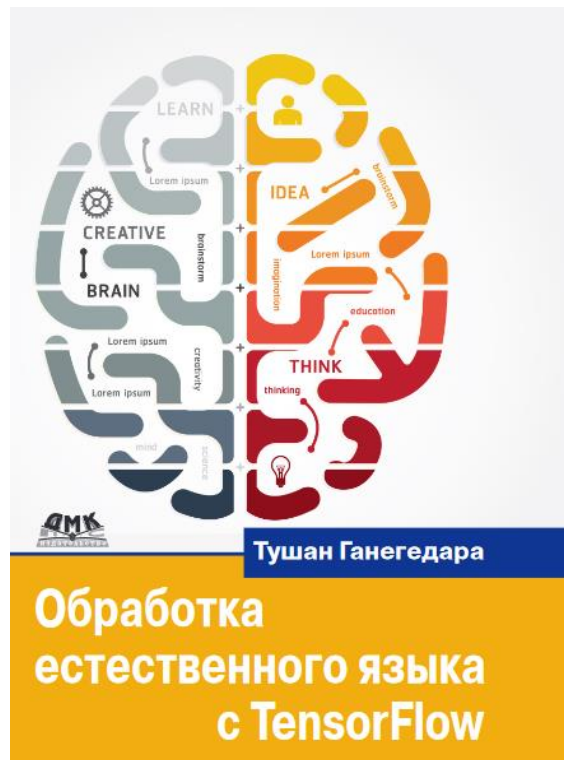


# Пример синтаксического анализа



# Библиотеки Пайтона:

- Библиотека на Python: nltk (Natural Language Toolkit), Стивен Берд, 2007 г.
- `keras.preprocessing.text`





## NLTK в действии:

- Токенизация по предложениям и словам
- Нормализация и чистка (единый регистр, удаление стоп-слов)
- Стемминг и лемматизация
- Регулярные выражения
- Мешок слов

# Токенизация

- в nltk Функция: word\_tokenize()
- Или from keras.preprocessing.text import Tokenizer
- Токенами могут быть не только слова, но и знаки пунктуации (!)

```
# import nltk
# nltk.download('punkt')

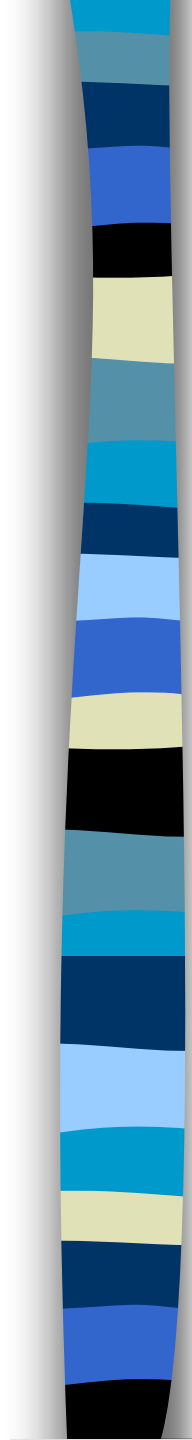
from nltk.tokenize import word_tokenize

tokens = nltk.word_tokenize(cleaned_review)

print(cleaned_review)
print(tokens)
```

```
a touching movie it is full of emotions and wonderful acting i could have sat through it a second time
['a', 'touching', 'movie', 'it', 'is', 'full', 'of', 'emotions', 'and', 'wonderful', 'acting', 'i', 'could', 'have', 'sat', 'th
rough', 'it', 'a', 'second', 'time']
```





Лемматизация и стемминг – это частные случаи нормализации и они отличаются.

Стемминг – это грубый эвристический процесс, который отрезает «лишнее» от корня слов, часто это приводит к потере словообразовательных суффиксов.

Лемматизация – это более тонкий процесс, который использует словарь и морфологический анализ, чтобы в итоге привести слово к его канонической форме – лемме.

Слово good – это лемма для слова better. Стеммер не увидит эту связь, так как здесь нужно сверяться со словарем.

Слово play – это базовая форма слова playing. Тут справятся и стемминг, и лемматизация.

## Стемминг

- процесс приведения слова к его корню/основе, обычно усечением
- Джули Бет Ловинс, МИТ 1968 г., американский лингвист (японский язык)
- Мартин Портер, 1980 и 2000 гг.
- Алгоритмы усечения – хороши для английского (!) (кровать->крова)



# Стемминг

- различные вариации слова (например, "help", "helping", "helped", "helpful") приводится к начальной форме (например, "help")

```
from nltk.stem import PorterStemmer
```

```
stemmer = PorterStemmer()
```

```
stemmed_review = [stemmer.stem(word) for word in filtered_review]
```

```
print(stemmed_review)
```

```
['touch', 'movi', 'full', 'emot', 'wonder', 'act', 'could', 'sat', 'second', 'time']
```



## Лемматизация – для флективных языков (английский – аналитический язык!)

- Аналитические языки более поздние, развились из флективных, в них предложение строится через порядок слов и предлоги, а не через изменение форм слов
- Лемма – нормальная словарная форма (кошками→кошка, бежал→бежать)
- MyStem, Илья Сегалович 1998, Яндекс
- Stemka, Андреем Коваленко 2002 г.

## Лемматизация: MyStem



- находим границу между основой и суффиксом (дерево суффиксов)
- Поиск в словаре (дерево основ)
- Если нет в словаре, генерируется гипотетическая модель изменения данного слова. Гипотеза запоминается, а если она уже была построена ранее — увеличивает свой вес. Если слово так и не было найдено в словаре — длина требуемого общего окончания основ уменьшается на единицу, идёт просмотр дерева на предмет новых гипотез. Когда длина общего «хвоста» достигает 2, поиск останавливается и идёт ранжирование гипотез по продуктивности: если вес гипотезы в пять и более раз меньше самого большого веса — такая гипотеза отсеивается.
- Результатом работы алгоритма является получившийся набор гипотез для несуществующего или одна гипотеза для словарного слова



## Ошибки: overstemming' и understemming

- флективные слова ошибочно относят к одной лемме («universal», «university» и «universe» - все «univers»)
- морфологические формы одного слова относят к разным леммам («alumnus» → «alumnu», «alumni» → «alumni», «alumna»/«alumnae» → «alumna»: выпускник, выпускников, выпускница оказались с разными основами)

## Лемматизация в английском

- Лемматизация похожа на стеммизацию в том, что она приводит слово к его начальной форме, но с одним отличием: в данном случае корень слова будет существующим в языке словом. Например, слово "caring" превратится в "care", а не "car", как в стеммизации.
- WordNet — это база существующих в английском языке слов. Лемматизатор из NLTK WordNetLemmatizer() использует слова из WordNet.

```
# nltk.download('wordnet')
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

lemm_review = [lemmatizer.lemmatize(word) for word in filtered_review]

print(lemm_review)
```

```
['touching', 'movie', 'full', 'emotion', 'wonderful', 'acting', 'could', 'sat', 'second', 'time']
```

# Стоп слова

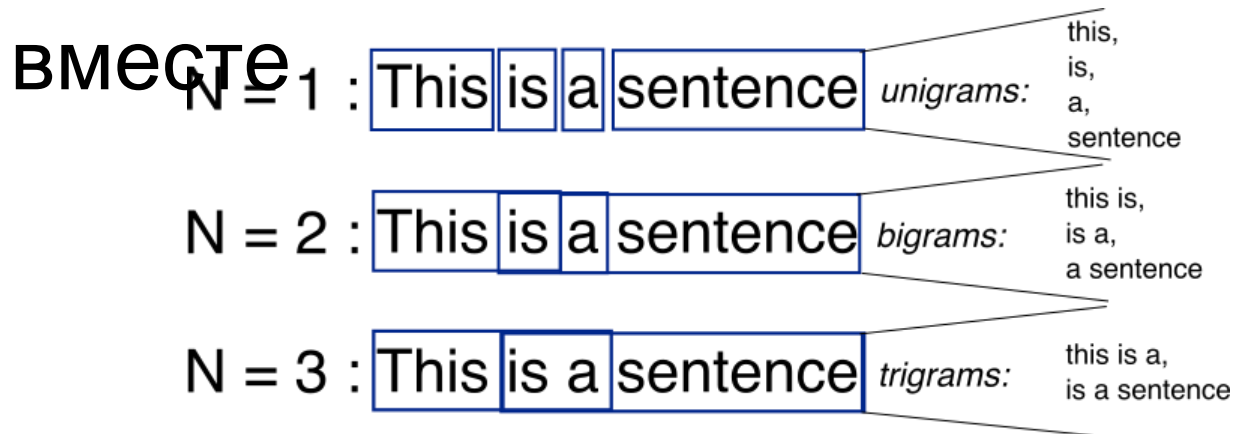


- Стоп-слова – это слова, которые выкидываются из текста до/после обработки текста. Когда мы применяем машинное обучение к текстам, такие слова могут добавить много шума, поэтому необходимо избавляться от нерелевантных слов.
- Стоп-слова это обычно понимают артикли, междометия, союзы и т.д., которые не несут смысловой нагрузки. При этом надо понимать, что не существует универсального списка стоп-слов, все зависит от конкретного случая.
- В NLTK есть предустановленный список стоп-слов. Перед первым использованием вам понадобится его скачать: `nltk.download("stopwords")`. После скачивания можно импортировать пакет `stopwords` и посмотреть на сами слова:



# N-граммы:

- N-граммы — это комбинации из нескольких слов, использующихся



## Мешок слов

- Bag of Word (в противовес словам Зеллига Харриса, 1954 г. : "язык – это не мешок слов")

"This is how you get ants."

tokenizer

`['this', 'is', 'how', 'you', 'get', 'ants']`

Build a vocabulary over all docum

`['aardvak', 'amsterdam', 'ants', ... 'you', 'your', 'zyxst']`

Sparse matrix encoding

aardvak	ants	get	you	zyxst
[0, ..., 0, 1, 0, ... , 0, 1 , 0, ..., 0, 1, 0, ..., 0 ]				

## Bag of Word : улучшения

- теряется порядок слов – использовать n-граммы
- использовать n-граммы не слов, а букв
- tf-idf (term frequency–inverse document frequency):

$\text{idf}(t,D) = \log(\text{число документов} / \text{число документов с } t)$

$\text{tf}(t,d) = \text{частота } t \text{ в } d / \text{общее число слов в } d$

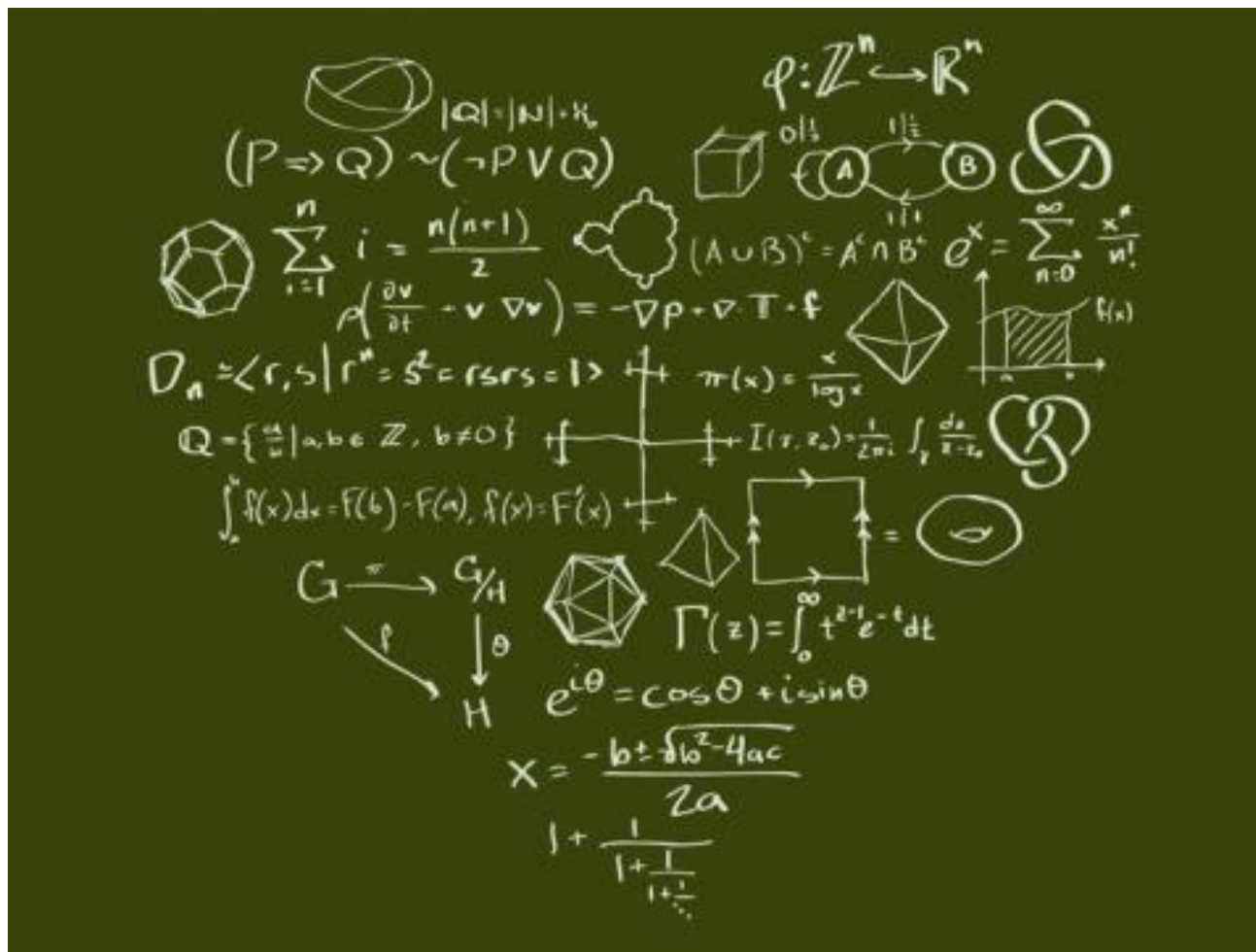
$\text{tfidf}(t,d,D) = \text{tf}(t,d) * \text{idf}(t,D)$

Большой вес получают слова, которые встречаются в документе чаще, чем во всем остальном корпусе

- Word2Vec (Google, 2013). Вектора слов учатся по предсказаниям контекстов, получают вектора, которые решают задачу аналогий:

king – man + woman = queen.  $(\cos(b-a+a',x) \rightarrow \max)$

# Математика поможет:



Спасибо за терпение!