

# Инструменты и методы работы с большими массивами текстовой информации

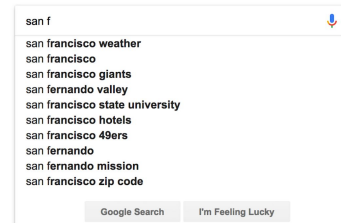
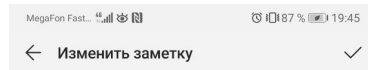
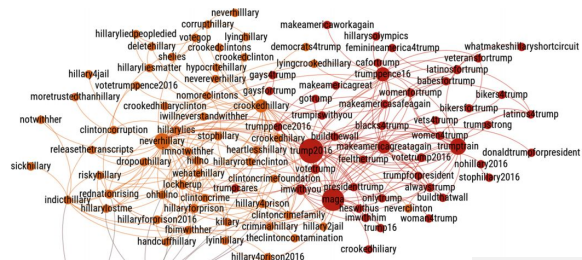
Саркисян Вероника

Лаборатория моделей и методов вычислительной прагматики

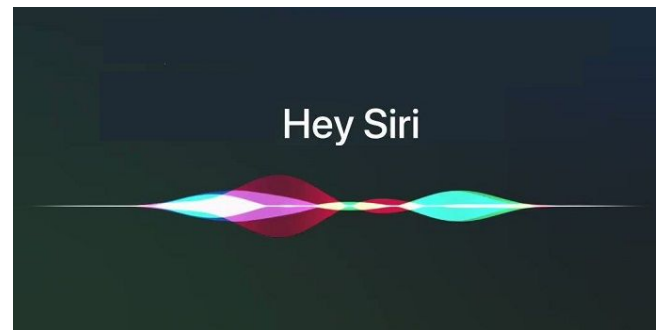
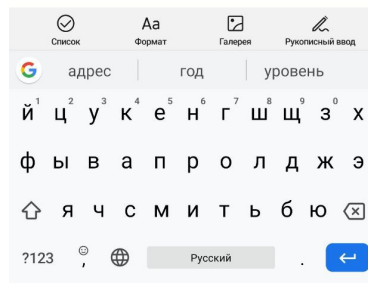
# План на сегодня

1. О чем этот курс?
2. Как курс будет устроен?
3. Natural Language Processing: основные задачи
4. Частотный анализ текста
5. Дистрибутивная семантика

1. О чем этот курс?
2. Как курс будет устроен?
3. Natural Language Processing: основные задачи
4. Частотный анализ текста
5. Дистрибутивная семантика



Яндекс.Алиса



# План курса

- Предобработка текста, частотный анализ, извлечение ключевых слов
- Дистрибутивная семантика
- Тематическое моделирование
- Классификация текстов
- Синтаксический парсинг
- Языковые модели
- Машинный перевод
- Активное обучение
- Чат-боты
- ... темы по заявкам

1. О чем этот курс?
2. Как курс будет устроен?
3. Natural Language Processing: основные задачи
4. Частотный анализ текста
5. Дистрибутивная семантика

Экзамена по курсу **не будет**, оценка складывается из:

- Тестов после лекций (30%)
- Домашних задании на программирование (70%)

1. О чем этот курс?
2. Как курс будет устроен?
3. **Natural Language Processing: основные задачи**
4. Частотный анализ текста
5. Дистрибутивная семантика



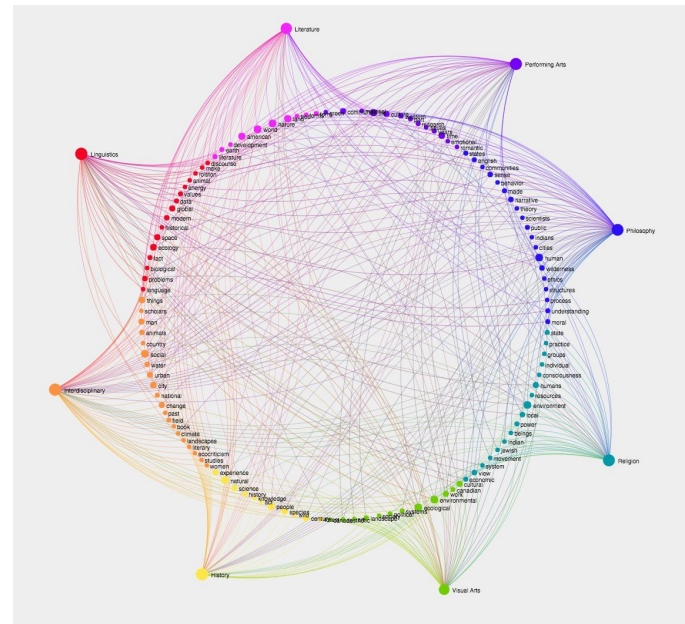
# Классификация текстов

- Классификация по тональности (sentiment analysis)
- Классификация по темам (например, проставление тегов к статьям)
- Классификация по жанрам и стилю текста
- Определение интента в диалоговой системе
- Фильтрация спама



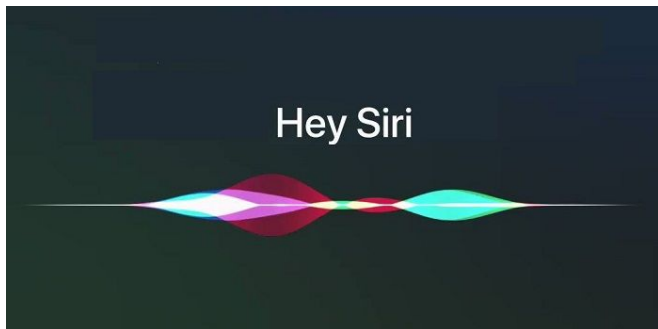
# Кластеризация текстов

- Тематическое моделирование: в данной коллекции текстов выделить темы и представить каждый текст как совокупность подмножества тем
- Обычная задача кластеризации (выделения групп текстов, похожих внутри одной группы и различающихся между группами)



# Диалоговые системы

- Chit-chat (болталки)
- Голосовые помощники
- Роботы-юристы, роботы-психологи, техническая поддержка и т.д.
- Искусственный интеллект для игр

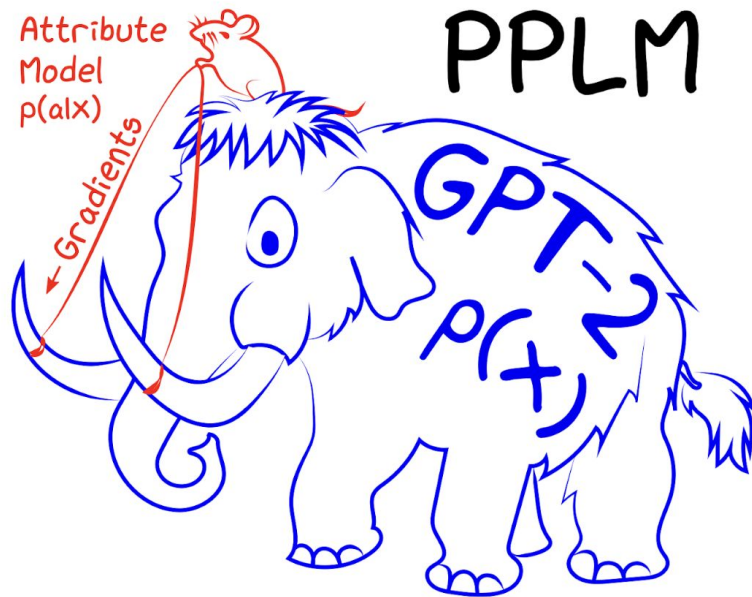


# Генерация текста

**[-]** The potato is a plant from the family of the same name that can be used as a condiment and eaten raw. It can also be eaten raw in its natural state, though...

**[Negative]** The potato is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you...

**[Positive]** The potato chip recipe you asked for! We love making these, and I've been doing so for years. I've always had a hard time keeping a recipe secret. I think it's the way our kids love to eat them...



# Основные трудности

- Неоднозначность
  - Лексическая неоднозначность: орган, парить, рожки, атлас
  - Морфологическая неоднозначность: Хранение денег в банке. Что делают белки в клетке?
  - Синтаксическая неоднозначность: Мужу изменять нельзя. Его удивил простой солдат.
- Неологизмы: печеньки, заинстаграммить, репостнуть, расшарить, затащить, килорубли
- Разные варианты написания: Россия, Российская Федерация, РФ
- Нестандартное написание и опечатки: *каг дила?*



1. О чем этот курс?
2. Как курс будет устроен?
3. Natural Language Processing: основные задачи
4. **Частотный анализ текста**
5. Дистрибутивная семантика

# Токенизация

**Сколько слов в этом предложении?**

*На дворе трава, на траве дрова, не руби дрова на траве двора.*

**12 токенов** : На, дворе, трава, на, траве, дрова, не, руби, дрова, на, траве, двора

**8 - 9 типов** : Н/на, дворе, трава, траве, дрова, не, руби, двора.

**6 лексем** : на, не, двор, трава, дрова, рубить

**Токен и тип**

**Тип** – уникальное слово из текста

**Токен** – тип и его позиция в тексте

## Обозначения

$N$  = число токенов

$V$  = словарь (все типы)

$|V|$  = количество типов в словаре

Как связаны  $N$  и  $|V|$  ?

## Закон Цифпа

В любом достаточно большом тексте ранг типа обратно пропорционален его частоте:  $f = a / r$

$f$  – частота типа,

$r$  – ранг типа,

$A$  – параметр, для славянских языков – около 0.07

## Закон Хипса

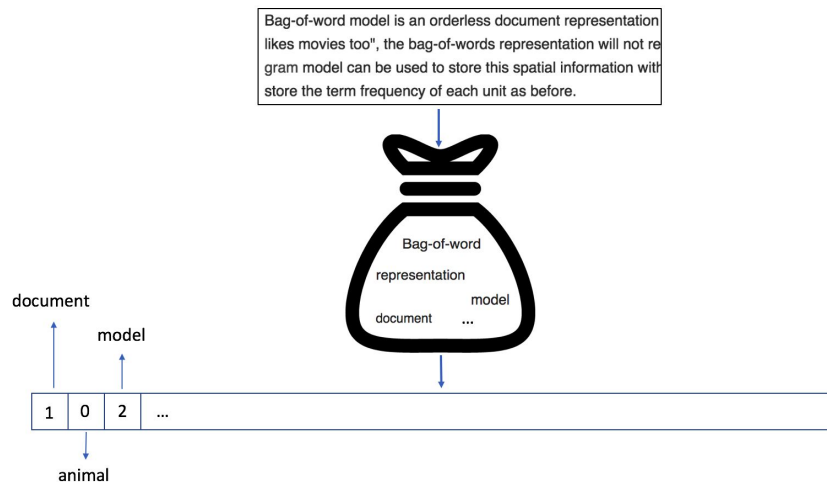
С увеличением длины текста (количества токенов), количество типов увеличивается в соответствии с законом:  $|V| = K * (N^b)$

$K, b$  – параметры, обычно  $K \in [10, 100]$ ,  $b \in [0.4, 0.6]$



# Модель мешка слов (Bag of words)

Документ - это набор слов, которые в нем содержатся.



Проблемы:

- Векторы получаются разреженные
- Не все слова одинаково важны

# Как уменьшить размер словаря?

**Лемматизация** (приведение слова к начальной форме):

ломала -> ломать

**Стемминг** (выделение основы слова):

ломала -> лом

# TF-IDF

Для каждой пары токен-документ вычислим, насколько важен этот токен в этом документе:

$$w_{x,y} = \text{tf}_{x,y} \times \log \left( \frac{N}{\text{df}_x} \right)$$

**TF-IDF**

Term  $x$  within document  $y$

$\text{tf}_{x,y}$  = frequency of  $x$  in  $y$

$\text{df}_x$  = number of documents containing  $x$

$N$  = total number of documents

1. О чем этот курс?
2. Как курс будет устроен?
3. Natural Language Processing: основные задачи
4. Частотный анализ текста
5. **Дистрибутивная семантика**