BigData, ML, AI простыми словами

Дмитрий Колодезев,
ООО Промсофт

План А

- Привет. Мы находимся здесь.
- Идеальный специалист.
- Ложь, Наглая ложь и Статистика.
- Tlön, Uqbar, Orbis Tertius.
- Машинное обучение.
- Искуственный интеллект.
- Простой классификатор текстов.

План Б

- Анализ тональности на счетах.
- Скрытые признаки и тематические модели.
- Запомни свою фамилию, потом скажешь!
- Word2Wec, или Слово Узнаешь По Соседям Его.
- LSTM, Трансформеры, BERT — без деталей.
- GPT3. Робот-агитатор.
- Амплификация текстовых корпусов.
- Что считать за текст?

Идеальный специалист

- Подобен двустороннему флюсу.
- T-shaped ⇒ П-shaped.
- Широкий кругозор.
- Computer Science:
  - вторая нога.
  - соединяющая перекладина.
- Comb-shaped

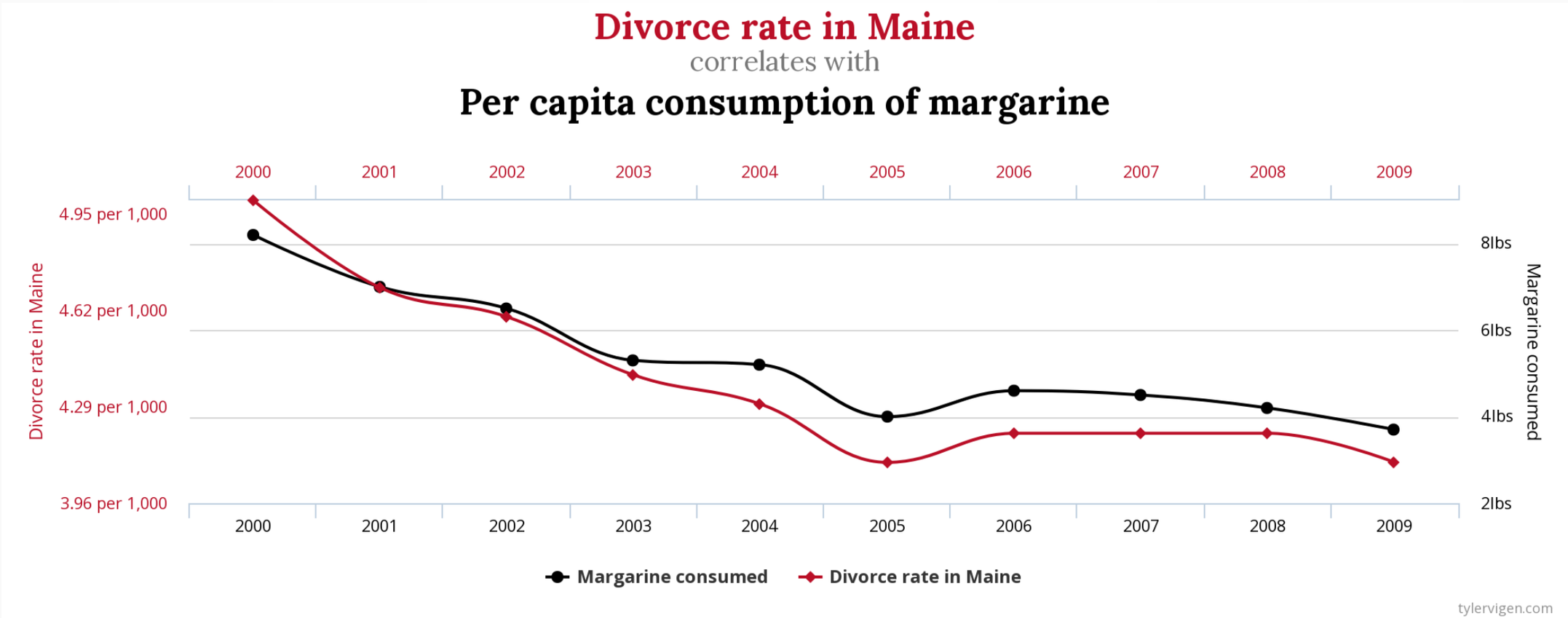
Ложь, Наглая Ложь и Статистика

- Оценить то, чего не можешь счесть.
- Если данных много — достаточно арифметики.
- Если данных мало — статистика не работает.
- Хороший статистик похож на сапера.
- Ошибка первого рода — нашли то, чего нет
  см https://xkcd.com/882/
- Ошибка второго рода — не нашли то, что есть
  см. https://xkcd.wtf/2303/

Привет от Борхеса

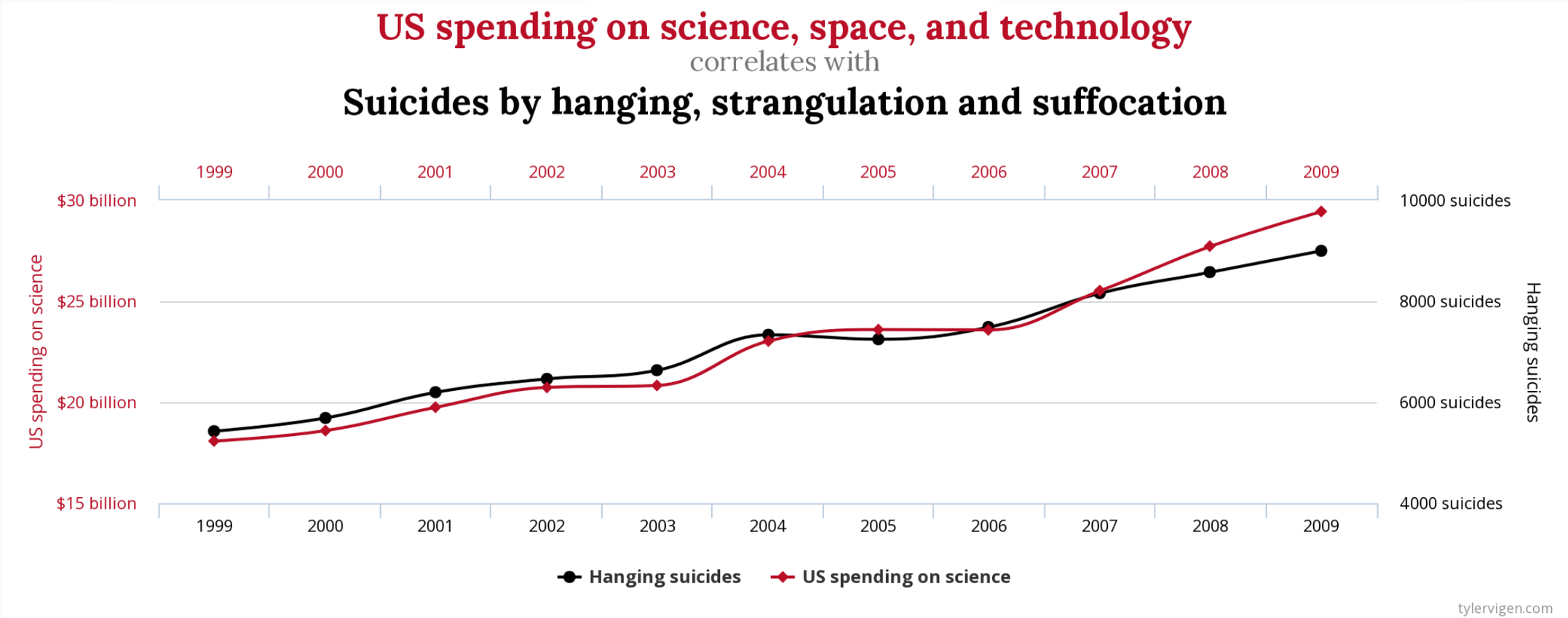
- Тлен, Укбар, Orbis tertius.
- «хрёнир» - копия реально существующего объекта, найденного по ошибке вместо него.
- «Ур» – объект, извлеченный из небытия надеждой.
- Есть некоторая вероятность, что данные случайно легли так, что на них можно найти нужную нам зависимость. P-value, обычно 1/20
- Если сделать 13 попыток, шансы будут 1:1

Поосторожнее с маргарином



https://www.tylervigen.com/spurious-correlations

И вообще с технологиями



https://www.tylervigen.com/spurious-correlations

Машинное обучение

- Универсальный подход к решению задач:
  - Найти или построить такое описание проблемы, из которого решение очевидно следует.
  - Пример — выбор работы.
- Машинное обучение:
  - Автоматически построить такое описание проблемы, в котором решение сводится к арифметике.

Искуственный интеллект

- Искуственный интеллект — программа, которая умеет все, что компьютеры сегодня не умеют делать.
- Сильный искуственный интеллект — программа, которая умеет делать все, что сегодня никто не умеет.
- Шахматы, распознавание речи, Го, DOTA2.
- Расходимся.

Многие Нейронные Сети Плачут, Пытаясь Побить Этот Классификатор

- Дано: Корпус публикаций, с разметкой «серьезная пресса» или «желтая пресса».
- Классификатор: «Желтая» vs «Серьезная»
- Составим словарь слов.
- Для каждого слова автоматически подберем коэффициент «желтизны».
- Сложим коэффициенты для всех используемых слов, использованных в той или иной публикации.
- Всех выше какого-то порога, считаем «желтой прессой».
- См например https://habr.com/ru/company/oleg-bunin/blog/352614/

Биграммы

- Сверхмощный космический
- Ряд экспертов
- Сверхцивилизация
- Выдвинуть версию
- Кто-то очень

Космический корабль

Ряд экспертов полагает, что раз Луна является искусственным спутником Земли и наблюдательным пунктом за нашей планетой, то ничто не мешает выдвинуть версию о том, что на самом деле Луна – действительно сверхмощный космический корабль, который предназначен для того, чтобы в минуту опасности для Земли попытаться сохранить нашу планету.

Какая-то сверхцивилизация провела «планетную инженерию» – так подкорректировала параметры Земли, чтобы создать здесь благоприятные условия для жизни. Представьте на секунду, что сегодня в сутках 11 часов, а наклон земной орбиты 10 градусов. Жизнь либо умрёт, либо изменится до неузнаваемости. А 21% кислорода в земной атмосфере – это случайность? Кто-то очень точно рассчитал и этот параметр

http://ibirzha.kz/pod-vlastyu-seleny/

Биграммы

- Космических аппаратов
- Ключевых характеристик
- Относительно простой
- Сводится к ряду
- Ключевых характеристик

Кроме того, запуски космических аппаратов с Луны энергоэффективнее и, следовательно, дешевле — гравитация спутника составляет лишь 17% от земной, кроме того, не придется преодолевать сопротивления воздуха из-за отсутствия атмосферы.

Новый дом

Сооружение даже относительно простой жилой конструкции на Луне, представляющей собой форпост человечества на спутнике, потребует максимума ресурсов и беспрецедентных мер. Проект строительства лунной базы сводится к ряду ключевых характеристик: место, безопасность, продовольствие.

Грустно или весело?

- Словарь «грустных» и «веселых» слов:
  - Я: 0
  - Буквально: 0
  - Подпрыгивал: +0,02
  - От: 0
  - Радости: +0,5
  - Когда: 0
  - Узнал: 0
- Итого тональность текста +0,52 - веселый

Ирония

- Например:
  - Ну: 0
  - Прям -0,1
  - Щас: -0,1
  - От: 0
  - Радости +0,5
  - Запрыгаю 0,01
- Итого тональность текста +0,31 - веселый

Серьезно про тональность

- https://habr.com/ru/company/vk/blog/516214/
- https://habr.com/ru/company/vk/blog/516726/
- https://habr.com/ru/company/vk/blog/516730/

Скрытые признаки и LDA

- Дано: Корпус документов.
- Вводная:
  - Есть несколько тем, затронутых в разной мере в некоторых из этих документов.
  - Темы различаются частотой словоупотребления
  - Мы не знаем темы
- Давайте найдем 10 (20, 30) тем и слова, которые их описывают.

Тематическое моделирование

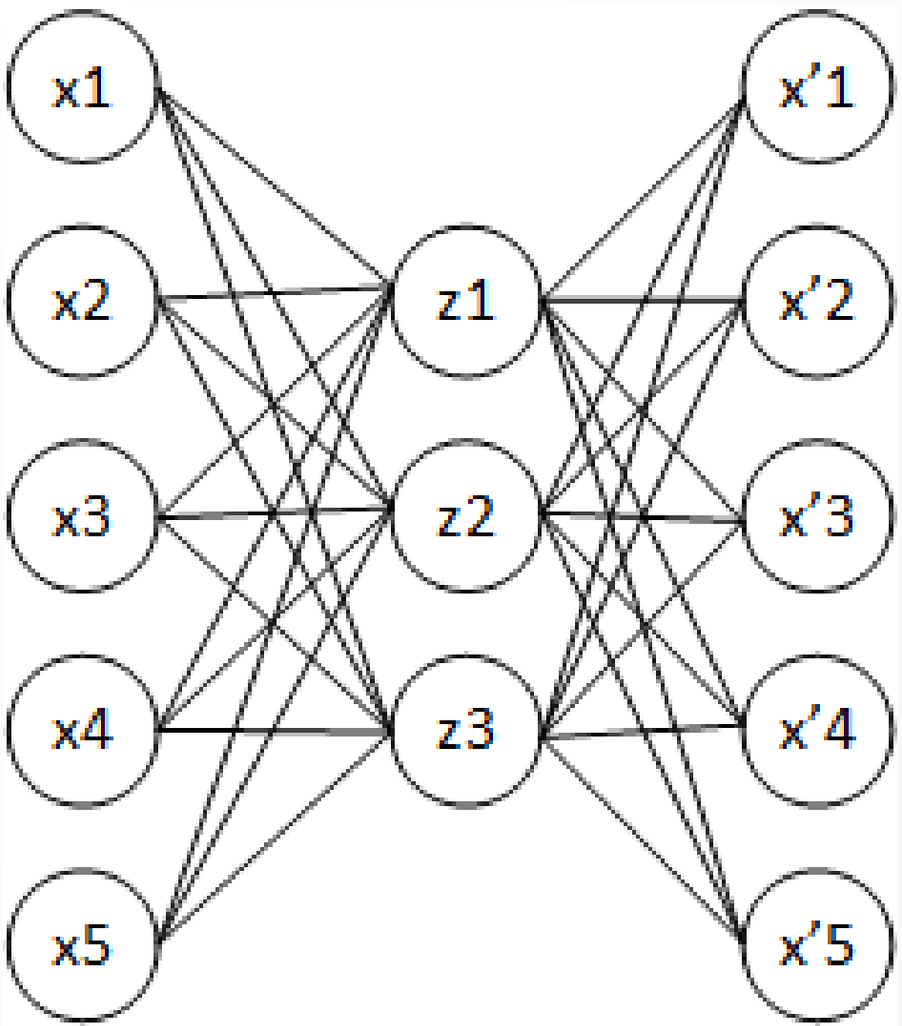
- Составим матрицу встречаемости слов по документам.
- Факторизуем (разложим в произведение)
- Темы*Документы = Матрица терминов
- BigARTM
- Тематическое моделирование
- Gensim

Например

- Мама мыла раму (Семья 0.3, Дом 0.7, Техника 0)
- Папа чинил машину (Семья 0.3, Дом 0, Техника 0.7)
- Темы:
  - Техника (машину 0.5, чинить 0.5)
  - Дом (раму 0.5, мыть 0.5)
  - Семья (папа 0.5, мама 0.5)

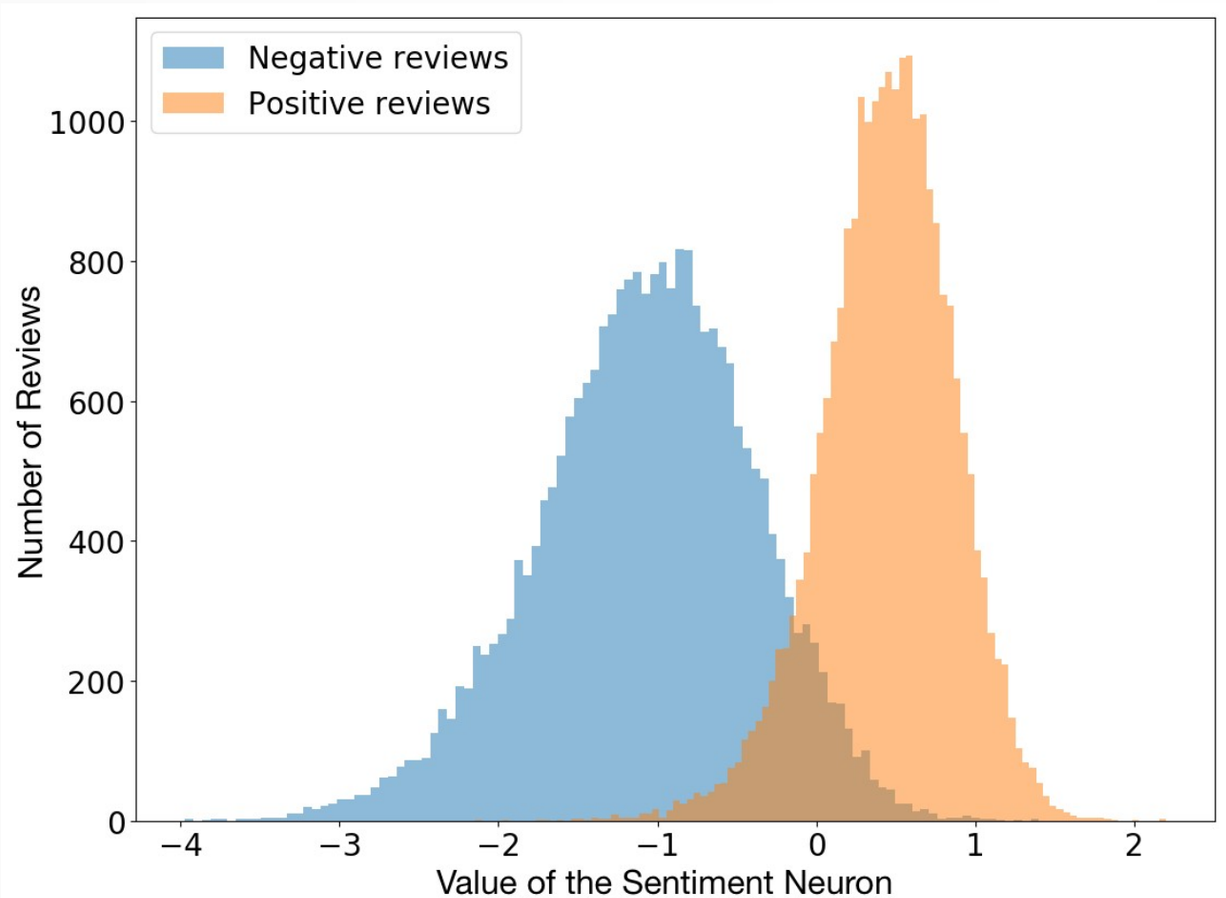
Автоэнкодер

- На вход текст
- На выходе он же
- Бутылочное горло
- Вынужден обобщать



https://ru.wikipedia.org/wiki/Автокодировщик

Сентиментный нейрон



https://openai.com/blog/unsupervised-sentiment-neuron/

Генерация текстов

SENTIMENT FIXED TO POSITIVE	SENTIMENT FIXED TO NEGATIVE
Just what I was looking for. Nice fitted pants, exactly matched seam to color contrast with other pants I own. Highly recommended and also very happy!	The package received was blank and has no barcode. A waste of time and money.
This product does what it is supposed to. I always keep three of these in my kitchen just in case ever I need a replacement cord.	Great little item. Hard to put on the crib without some kind of embellishment. My guess is just like the screw kind of attachment I had.
Best hammock ever! Stays in place and holds it's shape. Comfy (I love the deep neon pictures on it), and looks so cute.	They didn't fit either. Straight high sticks at the end. On par with other buds I have. Lesson learned to avoid.
Dixie is getting her Doolittle newsletter we'll see another new one coming out next year. Great stuff. And, here's the contents - information that we hardly know about or forget.	great product but no seller. couldn't ascertain a cause. Broken product. I am a prolific consumer of this company all the time.
I love this weapons look . Like I said beautiful !!! I recommend it to all. Would suggest this to many roleplayers, And I stronge to get them for every one I know. A must watch for any man who love Chess!	Like the cover, Fits good. . However, an annoying rear piece like garbage should be out of this one. I bought this hoping it would help with a huge pull down my back & the black just doesn't stay. Scrap off everytime I use it.... Very disappointed.

Слово Узнаешь По Соседям Его

- Word2wec
  - Давайте предскажем, с какими словами это слово может встречаться
  - Вариант: Двайте предскажем, какое слово уместнее всего в этом контексте?
- Мама мыла ...
- Курс ... к евро падает
- https://en.wikipedia.org/wiki/Word2vec

Дистрибутивная семантика

- Семантика, найденная автоматически из статистических закономерностей текста
- Чтобы свести семантику к арифметике
- Man — woman + king = queen
- https://rusvectores.org/ru/calculator/

BERT GPT и их друзья

- Соберем все тексты интернета
- Обучим продолжать начатую фразу
- Спросим что-нибудь
- Яндекс-Балабола
- https://yandex.ru/lab/yalm

Археологи нашли следы древней цивилизации: Что же стало причиной катастрофы легендарного корабля?

Как GPT спускали с цепи

SYSTEM PROMPT (HUMAN-WRITTEN)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)	The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

https://openai.com/blog/better-language-models/

Амплификация корпусов

- Покажите мне котика ;)
  - Он любит котиков!
- Выбираю из одних котиков ;(

- В этом году половина текста в интернете будет сгенерирована роботами.
- Которые участся на текстах из Интернета.
- Пока непонятно, как быть.

Что такое текст

- Действия как текст:
  - Что он имел в виду, сделав это?
- Финансовые транзакции
- Клики по рекламе / пути по сайту
- Геотеги в соцсетях
- Эволюция коммерческой недвижимости

Разное полезное

- https://yandex.ru/lab/yalm
- https://rusvectores.org/ru/visual/
- https://natasha.github.io/
- https://github.com/sberbank-ai/ru-gpts
- https://sysblok.ru/knowhow/obuchaem-word2vec-praktikum-po-sozdaniju-vektornyh-modelej-jazyka/
- https://lena-voita.github.io/nlp_course.html
- https://plainrussian.ru/
- https://habr.com/ru/company/infoculture/blog/238875/

Вопросы

Слайды тут

dkolodezev

promsoft

dkolodezev

d_key

dmitry_kolodezev

https://kolodezev.ru/download/slides-media-2021.pdf