МЕТОДЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТА (искусственный интеллект)

МЕТОДЫ АОТ

- Статистические (квантитативные)
- Метаразметка (аннотирование)
- Глубокое обучение (нейронных сетей)
- Моделирование

Все группы методов взаимосвязаны и включают элементы других групп.

СТАТИСТИКА НА ВСЕ РЕМЕНА

Статистический учёт вёлся и в Древнем мире.

В науку термин «статистика» ввел немецкий ученый Готфрид Ахенвалль в 1746 году

СТАТИСТИКА ДЛЯ ВСЕХ ОДНА

Один и тот же тот же математический аппарат используют:

- экспериментальная фонетика
- психолингвистика
- лингвистическая география
- физика, социология, генетика
- и т.д. и т.п

СТАТИСТИКА

- Статистика - совокупность методов сбора, анализа представления и интерпретации статистических данных.
- Статистические данные — значения свойств и отношений объектов.
- Значения могут быть количественными и качественными (указание на категорию).
- Мониторинг объекта дает вектор…

ПРИКЛАДНАЯ СТАТИСТИКА

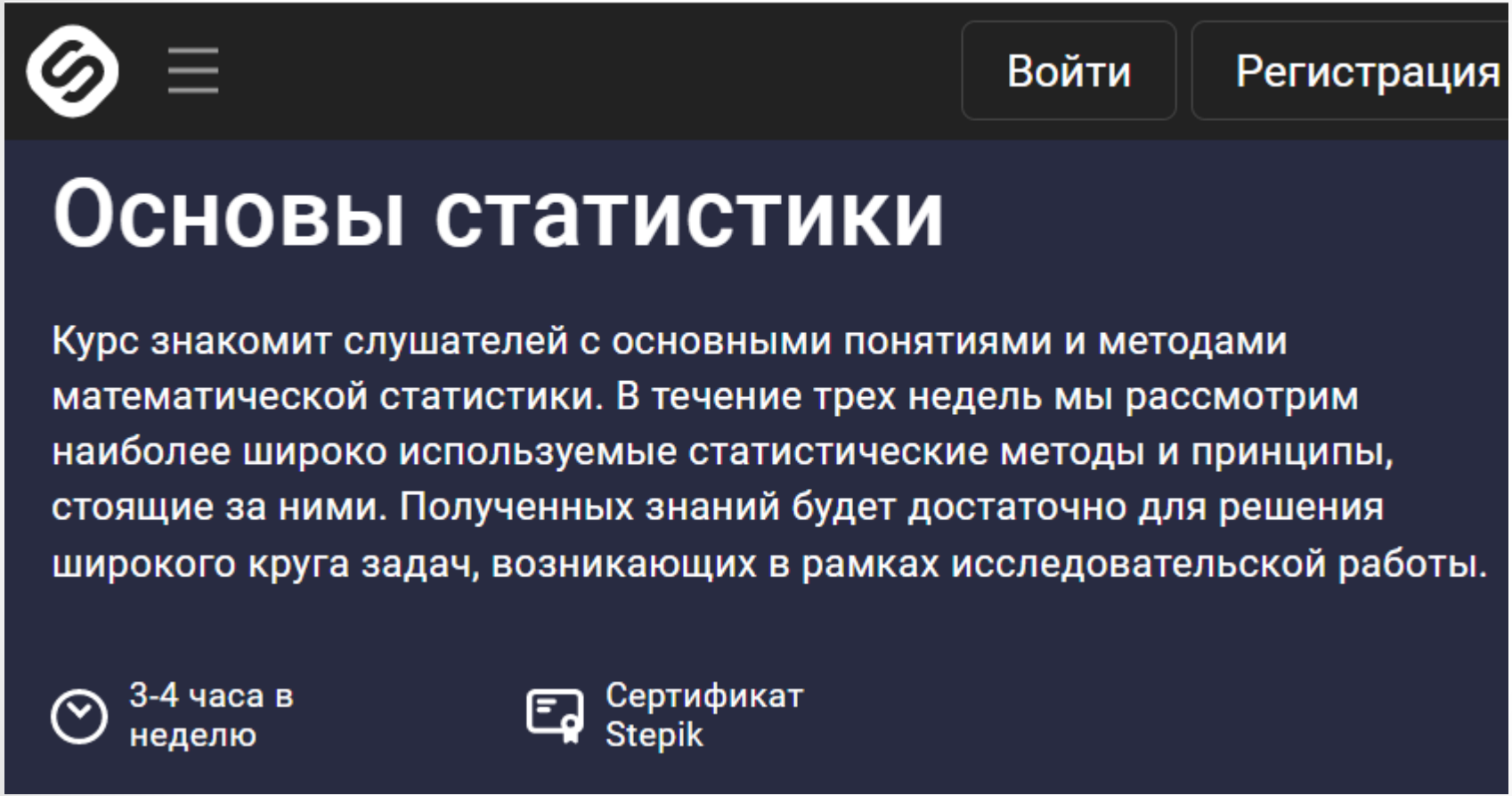
- Числовая и нечисловая.
- Числовые статистические данные — это числа, вектора, функции. Их можно складывать, умножать на коэффициенты. Математический аппарат анализа сумм случайных элементов выборки — это (классические) законы больших чисел и центральные предельные теоремы.

НЕЧИСЛОВЫЕ ДАННЫЕ

- Нечисловые статистические данные — это вектора разнотипных признаков, бинарные отношения, множества, группы, кольца, и др. элементы нечисловых математических пр-в.
- Математический аппарат анализа нечисловых статистических данных основан на использовании мер близости и показателей различия в таких пространствах.

СТАТИСТИЕСКИЕ МОДЕЛИ

- Статистические модели изначально были линейны.
- Разработка численных алгоритмов и рост вычислительной мощности компьютеров позволили перейти к нелинейным моделям.
- Существует разнообразное статистическое программное обеспечение общего и специализированного назначения.


Войти
Регистрация
Основы статистики
Курс знакомит слушателей с основными понятиями и методами математической статистики. В течение трех недель мы рассмотрим наиболее широко используемые статистические методы и принципы, стоящие за ними. Полученных знаний будет достаточно для решения широкого круга задач, возникающих в рамках исследовательской работы.
3-4 часа в неделю
Сертификат Stepik

Anatoliy Karpov

Работал ведущим аналитиком в VK и Stepik

Руководил командой аналитики в отделе бизнеса и рекламы VK. Специализируется на статистике, A/B-тестировании, машинном обучении и построении аналитических хранилищ данных. Автор онлайн-курсов по анализу данных на платформе Stepik

Бесплатно

Поступить на курс

Хочу пройти

Учиться можно сразу

Введение в Data Science и машинное обучение

Bioinformatics Institute

Бесплатно ★★★★

Нейронные сети

Bioinformatics Institute

Бесплатно ★★★★

В курс входят

29 уроков

4 часа видео

105 тестов

Программа курса

Последнее обновление 29.08.2022

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА

- исследует язык статистическими методами;
- задача — сформулировать законы функционирования языка
- цель — построить общую теорию языка в виде совокупности взаимосвязанных законов функционирования языков.

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА

Эмпирически основывается на результатах языковой статистики,

Может интерпретироваться как:

- статистика лингвистического объекта…
- статистика языка

Селегей Владимир Павлович

директор по лингвистическим исследованиям ABBYY; зав. кафедрами "Компьютерной лингвистики" в РГГУ и МФТИ; председатель Оргкомитета «Диалог».

Селегей Владимир Павлович

- В лингвистике нельзя обойтись одной лингвистикой…
- На современном этапе компьютерная лингвистика преимущественно основывается не на лингвистических, а на статистических подходах

Frequency Dictionary

Частотный словарь:

- список элементов (слов), вместе с информацией об их частотности (в тексте, языке, корпусе).
- элементы множества (слов) упорядоченные в соответствии с некоторой статистической величиной (обычно,

Frequency Dictionary

Частотный словарь - основа любых статистических методов и моделей NLP

Bag of words

Простейшая статистическая модель NLP – «мешок слов»:

- учитывается только количество вхождений конкретных слов в тексте
- игнорируются все прочие характеристики:
    - порядок слов в документе,
    - морфологические формы слов,
    - …..

Bag of words

Простейший (?) элемент модели
Bag of words — слово с единственным атрибутом, частотой встречаемости этого слова в тексте (корпусе текстов).

Bag of words является мощным ээффективным инструментом (в умелых руках, конечно)

Солнце Любовь Мама

Я возьму с собой в дорогу
Десять тысяч лучших слов....
Ольга Коваль

- Собираем в дорогу 100 слов
- Пишем ровно 100 слов
- Без знаков пунктуации
- Без ошибок
- На русском языке

Инфографика

Автор? Название?

Таямнічае «што» да «як»...

Японские пословицы

Русские пословицы

… с частотной лексикой

14.10.2015 г. в РУДН представлены словари с **наиболее частотной лексикой** для мигрантов по истории и законодательству России: таджикский, узбекский, киргизский, молдавский, китайский, вьетнамский, корейский и турецкий.

Костомаров Виталий Григорьевич

рос. лингвист, д.ф.н., проффессор
президент АПН СССР, директор НМЦ русского языка МГУ, инициатор создания и ректор гос. института русского языка им. А. С. Пушкина.
Президент Международной ассоциации преподавателей русского языка и литературы.
Главный ред. журнала «Русская речь»

Костомаров Виталий Григорьевич

- Идея учебного словаря, по-настоящему она была сформулирована именно у нас в Институте.
- Мы впервые посмотрели на словарь как на словарь обучающий, и поэтому получилось, что, например, англо-русских словарей может быть несколько.

Костомаров Виталий Григорьевич

- Раньше частотные словари создавались только в научных целях, а мы впервые подвели его под рамки учебного.
- Поэтому у нас такой словарь стал сопрягаться с идеей минимизации состава словаря.
- Например, был выпущен словарь под названием «Минимумы русской лексики».

Вспоминая детство….

Текст в информатике

ТЕКСТ – цепочка символов некоторого конечного алфавита

ТЕКСТ – МНОЖЕСТВО ЭЛЕМЕНТОВ

(произвольной природы = любых)

Bag of term

- модель текста как суммативное единство любых (произвольных) составляющих текст элементов (слов, лексем, n-грамм, строк, термов, знаков пунктуации, коллокаций, словоформ, грамматических связей и др...)

Для каждого слова из набора модели «мешок слов» может указывается некоторый «вес» (формальный/содержательный)

Атрибуция текста

- определение подлинности или подложности текста и установление его автора.
- основана на статистических методах и средствах автоматизации...

Ніна Федорівна Клименко

член-кореспондент НАН :

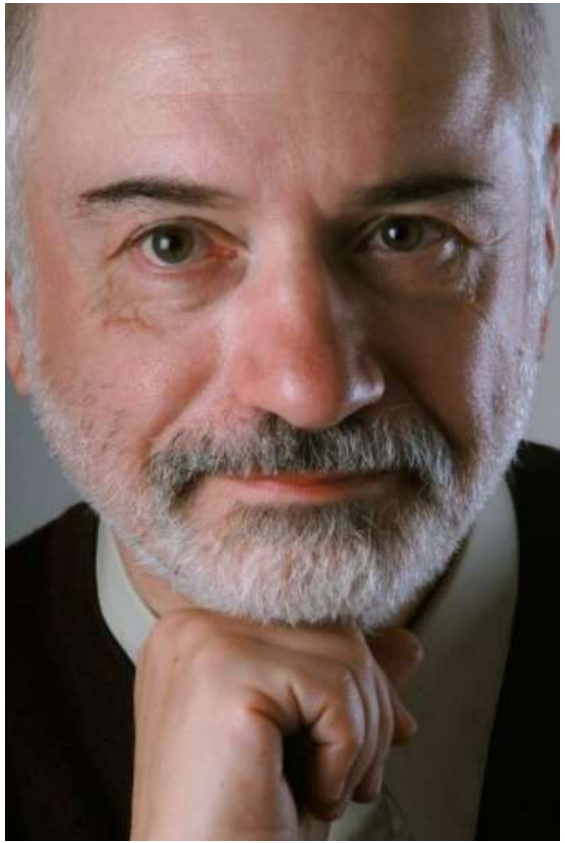
- чому такі службові слова (прийменники, сполучники) такі частотні в текстах?

Ніна Федорівна Клименко

- ми їх називаємо кріпильним матеріалом зв’язного тексту. Без цього жодного зв’язного тексту немає. І тому в усіх мовах світу за даними частотних словників найбільше, вони потрапляють у тисячу першу найчастотніших слів будь-якого тексту

Эпштейн Михаил Наумович

философ, культуролог, литературовед, лингвист, эссеист, профессор университета Эмори (Атланта, США) и Центра гуманитарных инноваций Даремского университета (Великобритания), Академии российской современной словесности.



Михаил Эпштейн

- …в русском языке предлог «в» встречается 1 раз на каждые 23 слова
- в английском the - каждое из 16 слов.
- достаточно взять частоту употребления этих слов, умножить на 23 или на 16 – и можно получить приблизительный объем Рунета и англоязычного Интернета.
- Получается, что англоязычный Интернет по объему слов примерно в …..

Долгин Александр Борисович

- Интернет располагают отличной базой для аналитики современного языка и позволяет получать любые срезы: социальные, вкусовые...
- В системе зарыто колоссальное количество ответов на эти и другие незаданные вопросы.

Big Data

Большие данные это данные, которые:

- огромны
- многообразны
- взаимосвязаны
- стремительно растут
- быстро меняются

Большие данные

- Big Data не о данных
- Big Data о выживании в **пост**информационном обществе
- Big Data о новой «компьютерной» парадигме мышления

Лексикон человека

- Два года - 200 слов
- Три года - 1000 слов
- Четыре года – 1500 слов
- Пять лет – 2200 слов
- Выпускника средней школы - 4000
- Выпускник вуза - 8000-12 000 слов.

Лексикон гениев

- В "Словаре языка **А.С. Пушкина**" в 4-х томах (М., 1956-1961)
  21 191 слово.
- В произведениях **Вильяма Шекспира**
  29 066 лексем.
- В текстах **Иоганна Вольфганга Гёте**
  17 000 слов.

21.09.2022

Английская Википедия 6 552 985 статей

Русская Википедия 1 854 483 статьи

Большая энциклопедия Терра 160 000

ЭС Брокгауза и Ефрона 121 240 статей

Большая советская энциклопедия 95 279

Urban Dictionary

UD - 7 500 000 вокабул
Oxford ED - 600 000 вокабул.
Global Language Monitor - 1,5 млн
Английский язык - филологический диалект Urban Dictionary?

Aaron Peckham

Grammar nazi

Абсолютизация нормативного подхода к языку

Где поставить ударение?

Робот сначала обращается

- к составленным ручную словарям
- к правилам из академических справочников
- статистическим правилам.

ЗВОНИТ ИЛИ ЗВОНИТ

Ударение «звОнит»:

- лингвистами воспринимается спокойно,
- соответствует законам языка,
- не одобряется «большинством образованных носителей», поэтому нормативным пока не признаётся.

Как будет дальше? Посмотрим.

Было время...

«платИт», «курИт», «объявИт», «красИт»

у глагола «звонить» происходит переход ударения в личной форме с окончания на корень

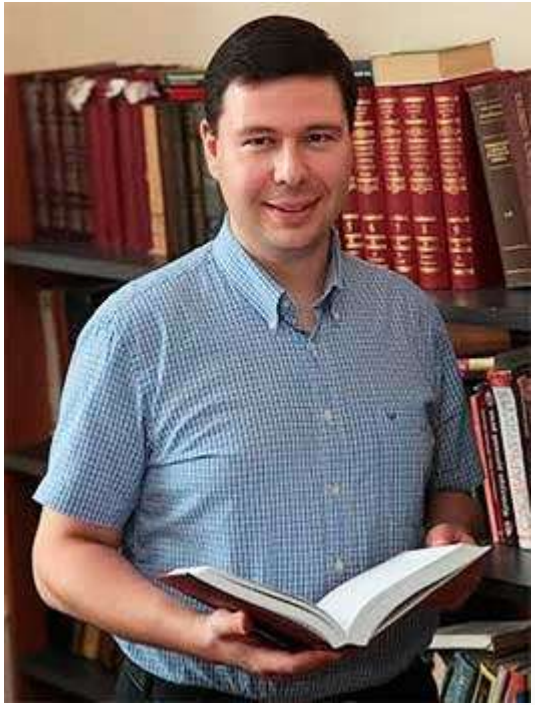
корень басит бесит бузит бурит винит водит вопит вялит гноит губит дымит душит

Владимир Маркович Пахомов

- главный редактор Грамота.ру, кандидат филолог-х наук

Основатель проекта
Алексей Кормилицын
(1961-2013).

Грамота.ру на 997 месте по популярности в России



Логика и фанатизм

Владимир Пахомов (главный редактор справочно-информационного портала «Грамота.ру»):

- логика и законы языка подсказывают: кофе должно быть среднего рода.
- Несклоняемые неодушевлённые иноязычные слова, оканчивающиеся на гласный, в подавляющем большинстве случаев относятся к среднему роду

Логика и фанатизм

Владимир Пахомов

В 1920-х гг. можно прочитать

- «Открылся новый кино» (кинотеатр).
- «В пролёты улиц вас умчал авто».
- «Метро сверкнул перилами дубовыми».

Почему кто-то считает, что, если слово кофе станет среднего рода, это нанесет какой-то вред языку?

Интернет

В Русском орфографическом словаре ИРЯ им. В.В. Виноградова РАН - М., 1999. С. 344):

- ИНТЕРНЕТ - Произносится [интэрнЭт].
- Имя собственное, мужского рода.
- Пишется с заглавной “и”: “Интернет” (как система, всемирная сеть) и со строчной - как сокращение от «доступ в интернет»).

Векторная модель текста

Модель текста «мешок слов» является векторной моделью.

Векторная модель (vector space model) — представление текстов векторами из одного общего для всей коллекции текстов векторного пространства.

Например: частотный словарь текста…

Кортеж и вектор

- Кортеж — упорядоченный конечный набор длины n (где n∈N∪{0}), каждый из элементов которого fi принадлежит некоторому множеству F.
- Вектор — кортеж однородных элементов (скаляров).
- Элементы кортежа называются его компонентами, или координатами.

Кортеж и вектор

- Совокупность векторов образует векторное пространство V над линейным пространством (полем) F.
- Перечень свойств вектора моделирует принятое в теории систем определение класса и состояния объекта.
- Изучаются векторным исчислением …

Векторная модель текста

Векторная модель - основа решения задач:

- информационного поиска,
- атрибуции текста,
- статистического перевода,
- классификация документов,
- распознавания образов,
- кластеризации документов и др...

Латентно-семантический анализ

Модель "мешок слов" используется в LSA

LSA – обработка вербальной информации, устанавливающая взаимосвязь между текстами (документами) на естественном языке и терминами в них встречающимися.

Предваряя LSA

- исключение стоп-символов (стоп-слов, стоп-термов - не несущих специальной смысловой нагрузки предлогов, причастий, междометий, частиц ....).
- нормализация (лемматизация, стемминг) текста - удаление грамматической информации (падежи, числа, глагольные виды и времена, залоги причастий, род и так далее).
- исключение малочастотных термов (сильно упрощает математические вычисления)....

Пунктуационная модель текста

Пунктуация (от лат. punctum — точка) — система знаков препинания в письменности (конкретного языка).

Пунктуация не безразлична к мировоззрению автора и эпохи, способна рассказать о них не меньше, чем содержательные аспекты текстов...

Пунктуационная модель текста

Эпоха энциклопедий (18 век) нашла свое синтаксическое выражение в двоеточии...

Феноменология (19 век) предпочитает "заключение в скобки"

Синтаксическая революция осуществлена введением восклицательного знака.

Пунктуационная модель текста

Фрейдизм построен на многоточиях...

Мартин Хайдеггер главную роль отдал вопросительному знаку и дефисам

Кавычки стали основным пунктуационным символом философии постмодернизма...

Пунктуационная модель текста

Описаны семантические значения:

- тире, у Максима Горького
- переноса, у Иосифа Бродского,
- точки, у Исаака Бабеля,
- запятой, у Вирджинии Вульф,
- небрежения пунктуацией у Джеймса Джойса...

Подсчет числа точек, как проблема

- когда мы ищем точку, то надо нажать кнопку "пробела", потом "точка" и снова "пробел"

- число точек х 3 - число многоточий.

Подсчет числа точек, как проблема

Возможное решения алгоритма неоднозначности точек и многоточия:
1) Находим все точки в тексте: Ctrl+F '.'
2) Находим все многоточия в тексте: Ctrl+F '...'
3) Умножаем количество многоточий на 3 и вычитаем из общего количества точек
4) Находим все двоеточия в тексте: Ctrl+F '..' (туда входят все элементы типа: '!..', '?..', '..'; элементы многоточия не определяются по данному запросу)
5) Умножаем количество двоеточий на 2 и вычитаем из разности которую мы получили на 3 шагу
6) Получаем необходимый результат
P. S. Если ваш блокнот даёт другой результат по данному результату,то возможно дело в КОДИРОВКЕ, в которой сохранён текст
P. P. S. Или у возможно это пиратский блокнот

1 2 3 @# + 4:

- согласно статистике Microsoft Word – это 6 слов,
- по анализу программы Trados – 0 слов.

Позиция 1: раз их не нужно переводить, то и в подсчете их учитывать не следует,

Позиция 2: переводчику все равно приходится просматривать и проверять каждое числительное и символ, а значит, они должны быть включены в подсчет.

Для переводчика

	Числительные и символы	Ссылки и примечания	Колонтитулы	Текстовые блоки	Внедренные OLE-объекты	Теги
MS Word	да	нет	нет	нет	нет	да
Trados	нет	да	да	да	нет	нет
WordFast	нет	да	да	да	нет	да
PractiCount and Invoice	да	да	да	да	да	да

Суета вокруг богатства

Богатство речи =
число разных слов / число всех слов

- Фёдор М. Достоевский
- Лёв Н. Толстой
- Валерия Гаврилова (манекенщица, фотомодель, «писатель»)

Доверие машине и беззаботность

параметр	В. Гаврил	Л.Н.Толстой	Ф.М.Достоев
всего форм	43210	589803	994030
всго разных форм	11462	15	15
богатство речи	0,265	0,00002543	0,00001509
Вывод: богатство речи Достоевского на порядок выше!!!			

Просто счёт

- Страница простого неформатированного текста (2000 знаков) имеет информационный объем равный…. 2Кбита?

St2: лексемы к концепту

	смех
	бессмертной
	смейся
	смело
	смелее
	бессмертия
	бессмертная
Смерть	бессмертных

АКТУАЛЬНЫЕ МОДЕЛИ

- Скрытые марковкие модели (СММ).
- Байесовские алгоритмы
- Сэмплирование по Гиббсону
- Бутстреппинг (bootstrapping)

CMM

- статистические модели,
- имитирующие работу процесса, похожего на марковский процесс с неизвестными параметрами,
- и разгадывание неизвестных параметров на основе наблюдаемых.

Марков Андрей Андреевич

Русский математик (1856-1922), Огромный вклад в теорию вероятностей, математический анализ и теорию чисел.

Первооткрыватель стохастических процессов:

- следующее состояние процесса зависит, вероятностно, только от текущего состояния.

МАРКОВСКИЙ ПРОЦЕСС

- AR(1): $x_t=\psi_1*x_{t-1}+\varepsilon_t$
- случайный процесс, эволюция которого после любого заданного значения временно́го параметра t не зависит от эволюции, пред шествовавшей t, при условии, что значение процесса в этот момент фиксировано («будущее» процесса не зависит от «прошлого» при известном «настоящем».

ЦЕПИ МАРКОВА

- последовательности случайных событий с конечным или счётным числом исходов, характеризующиеся тем, что при фиксированном настоящем будущее независимо от прошлого.
- Основное применение: в распознавания речи, текста, образов, при криптоанализе, в машинном переводе, в биоинформатике, в робототехнике, автоматизированном управлении, экономике и производстве….

Скрытая марковская модель

Скрытая модель Маркова — это вероятностная модель множества случайных переменных $\{O_1, \ldots, O_t, Q_1, \ldots, Q_t\}$. Переменные O_t — известные дискретные наблюдения, а Q_t — «скрытые» дискретные величины. В рамках скрытой модели Маркова есть два независимых утверждения, обеспечивающих сходимость данного алгоритма:

1. t-я скрытая переменная при известной $(t-1)$-ой переменной независима от всех предыдущих $(t-1)$ переменных, то есть
   $P(Q_t \mid Q_{t-1}, O_{t-1}, \ldots, Q_1, O_1) = P(Q_t \mid Q_{t-1})$
2. t-е известное наблюдение зависит только от t-го состояния, то есть не зависит от времени, $P(O_t \mid Q_t, Q_{t-1}, O_{t-1}, \ldots, Q_1, O_1) = P(O_t \mid Q_t)$

Скрытая марковская модель

- **СММ** — статистическая модель, имитирующая работу процесса, похожего на Марковский процесс с неизвестными параметрами, и задачей ставится разгадывание неизвестных параметров на основе наблюдаемых.
- СММ применяются в области распознавания речи, письма, движений, машинном переводе, биоинформатике, криптоанализе.

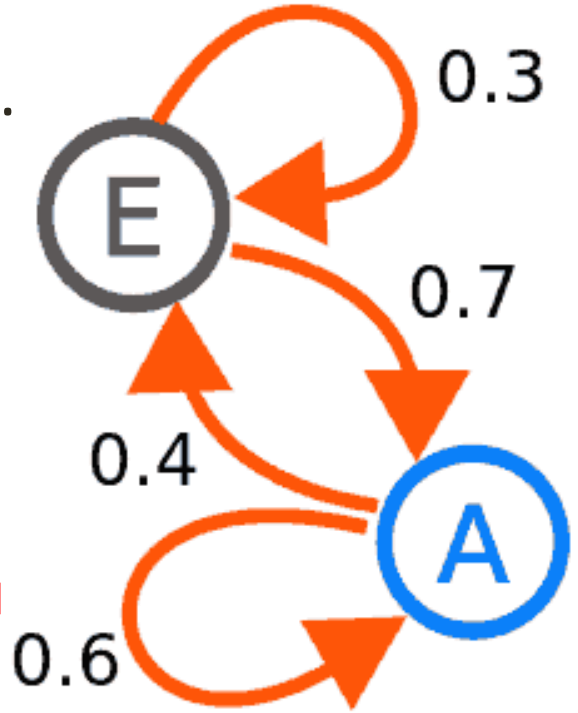
Примеры алгоритмов СММ

- Алгоритм Витерби: даны параметры модели, требуется определить наиболее подходящую последовательность скрытых узлов, наиболее точно описывающую данную модель (помогает при решении данной задачи).
- Алгоритм Баума-Велша: дана выходная последовательность (или несколько) с дискретными значениями, требуется «потренировать» СММ на данном выходе.

Пример ССМ

- по телефону о дневных делах.
- прогулка, за покупками, ВКонтакте
- выбор основывается лишь на погоде основываясь на его решениях,

Какая была погода (солнечно или дождливо) ?



N-грамма

N-грамма — последовательность из n элементов (звуков, слогов, слов, букв, коллокаций и др...).

N-граммная модель рассчитывает вероятность последнего слова N-граммы, если известны все предыдущие (предполагается, что появление каждого слова зависит только от предыдущих слов).

N-грамма

N-граммы используется для:

- предугадывания на основе вероятностных моделей,
- поиска плагиата
- категоризации текста и языка.
- получения знания из текстовых данных
- поиска кандидатов, чтобы заменить слова с ошибками правописания

Вероятностные модели

- Вероятность — степень возможности наступления некоторого события.
- Вероятность случайного события A - отношение числа m несовместимых равновероятных элементарных событий, составляющих событие A, к числу всех возможных элементарных событий n:

$$P(A) = \frac{m}{n}$$

Томас Байес

Thomas Bayes (1702 —1761)

- английский математик
- пресвитерианский священник,
- член Лондонского королевского общества (1742).

Формула Томаса Байеса

- P(A) - вероятность наступления события A.
- P(AB) - вероятность наступления обоих событий вместе. P(AB)=P(BA).
- P(A|B) вероятность наступления события A, если событие B произошло.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

БАЙЕСОВСКИЕ АЛГОРИТМЫ

- графические вероятностные модели множеств переменных и их вероятностных зависимостей.
- Математический аппарат байесовых сетей создан Джудой Перлом

Джуда Перл

лауреат Премии Тьюринга (2011 года)

за «фундаментальный вклад в искусственный интеллект посредством разработки исчисления для проведения вероятностных и причинно-следственных рассуждений».

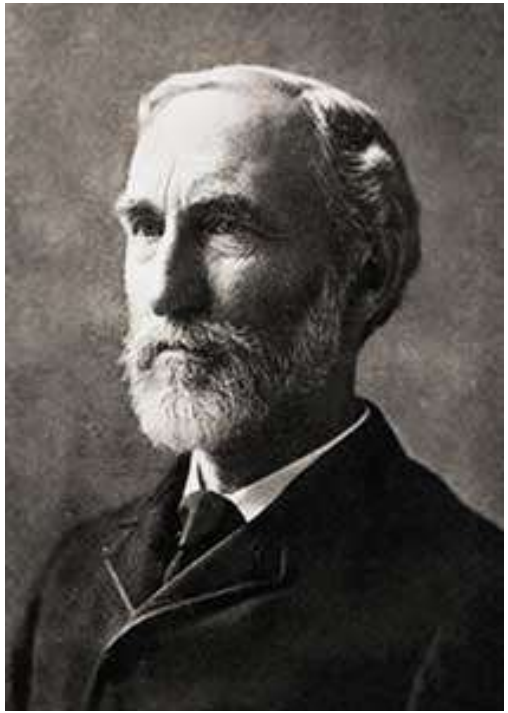
СЕМПЛИРОВАНИЕ ПО ГИББСУ

- алгоритм для генерации выборки совместного распределения множества случайных величин.
- использует условные вероятности для каждой переменной, входящей в распределение.
- последовательность получаемых значений образуют возвратную цепь Маркова, устойчивое распределение которой является как раз искомым совместным распределением.

Джозайя Уиллард Гиббс

- 1839—1903) - физик, химик, математик, основатель векторного анализа.

Вектор (несущий) — элемент линейного (векторного) пространства.



БУТСТРЕПИНГ

Компьютерный метод определения статистик вероятностных распределений, основанный на многократной генерации выборок методом Монте-Карло на базе имеющейся выборки.

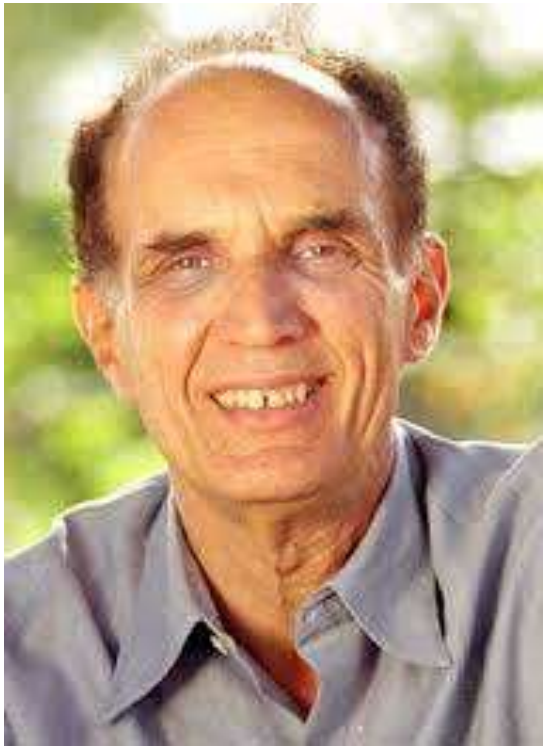
Позволяет оценивать доверительные интервалы, разброс, устойчивость, дисперсию, корреляцию и др. для сложных моделей и для каждой полученной псевдовыборки.

БУТСТРЕПИНГ

Предложен в 1977 г.

Брэдли Эфроном,

ам. статистиком, проф. Стэнфордского университета, создателем кубиков Эфрона — нетранзитивных игральных костей…



ТОГДА И ТЕПЕРЬ

- 1899 год
  Статистический метод в языкознании применялся и применяется довольно редко, что объясняется сравнительно незначительными результатами, полученными до сих пор при его помощи.
- 2014 год
  квантитативный анализ широко применяется к единицам любого уровня. Основным объектом является текст

ПРИЛОЖЕНИЯ КЛ:

- атрибуция текста
- статистическая стилистика
- классификация текстов
- лингвогеография….
- ……
- открывают путь к изучению самого языка, поскольку сегменты текстов, являющиеся объектами подсчетов, соотнесены с единицами языка.

Reinhard Köhler

Reinhard Köhler ist ein deutscher Sprachwissenschaftler. Er ist seit 1990 Professor für Linguistische Datenverarbeitung (Computerlinguistik) an der Universität Trier.

Кватитативные законы

Reinhard Köhler:

- Эти законы стохастической природы; они не соблюдаются в каждом отдельном случае; они скорее определяют вероятности событий или количественные отношения изучаемых явлений. Легко найти противоположные примеры каждому из упомянутых выше примеров…

ЗАКОНЫ КЛ:

- Köhler : свойства лингвистических элементов и отношений между ними подчиняются универсальным законам, которые могут быть сформулированы строго математически также как и законы естественных наук. Нужно иметь в виду в данном контексте, что эти законы стохастической природы; они не соблюдаются в каждом отдельном случае; они скорее определяют вероятности событий или количественные отношения изучаемых явлений. Легко найти противоположные примеры каждому из упомянутых выше примеров…

ПРИМЕРЫ ЗАКОНОВ КЛ:

- Закон диверсификации: Если лингвистические категории (такие, например, как части речи или грамматические окончания) появляются в различных формах, то можно сказать, что частоты их появления в текстах подчиняются определенным распределениям.
- Закон распределения длин морфов;
- Закон распределения длин слогов;
- Закон распределения длин слов…….

ПРИМЕРЫ ЗАКОНОВ КЛ:

- Закон текстового блока: Лингвистические единицы (слова, буквы, синтаксические конструкции) демонстрируют определенное распределение частоты в одинаково больших блоках текстов.
- Закон Ципфа: Частота слова обратно пропорциональна его порядковому номеру в списках частотности.
- Закон Менцерата: размеры составляющих конструкции уменьшаются с увеличением самой изучаемой конструкции.

ПРИМЕРЫ ЗАКОНОВ КЛ:

- Закон текстового блока: Лингвистические единицы (слова, буквы, синтаксические конструкции) демонстрируют определенное распределение частоты в одинаково больших блоках текстов.
- Закон Ципфа: Частота слова обратно пропорциональна его порядковому номеру в списках частотности.
- Закон Менцерата: размеры составляющих конструкции уменьшаются с увеличением самой изучаемой конструкции.

If it bleeds, it leads?

Учёные учли миллиарды слов «Нью-Йорк таймс» за 20 лет + книги Google Books (с 1520 года) + твитты + тексты популярных песен за 50 лет.

- Все источники: слова с положительными коннотациями встречаются чаще.
- Причём их больше и среди пяти тысяч самых употребительных, и среди менее обиходных.
- оптимизм англоязычного мира за последние годы сильно упал.

Twitter в лингвогеографии

Используя автоматическую программу определения языков, исследователи проанализировали 10% всех твитов.

- выявление локальных центров испанского, корейского, португальского, японского, русского, датского и инденезийского языков в Нью-Йорке.
- 5 самых частотных языков, на которых пишут в Twitter из Таиланда: тайский, английский, японский, малазийский и русский.

Латентные изменения

В английском языке зафиксированы неуловимые изменения

… одни конструкции и варианты становятся более частотными, а другие постепенно исчезают...

- they started walking вместо they started to walk
- It is being held вместо It is held
- to get вместо to be

Поиск призраков

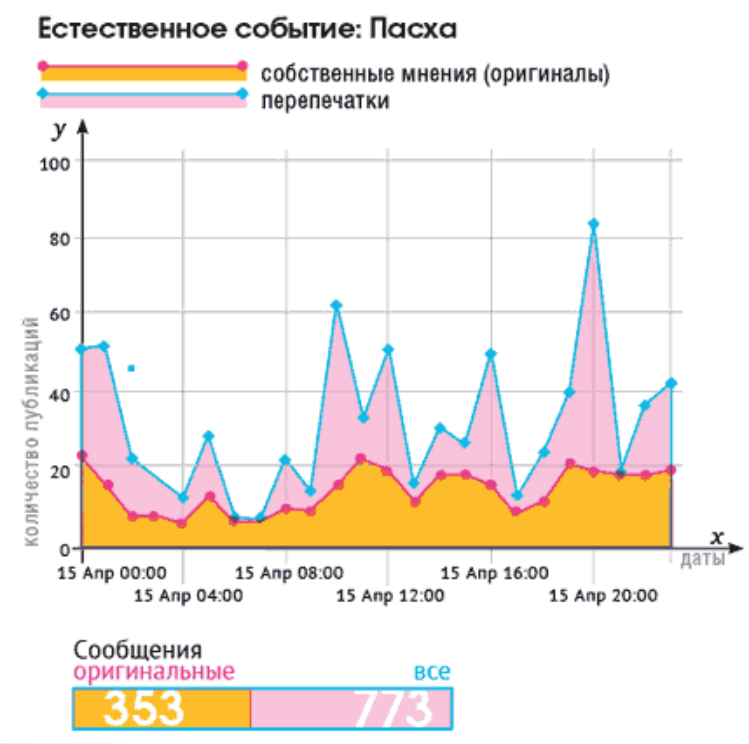
Литературный негр (фр. nègre littéraire) — автор, за вознаграждение пишущий за другое, как правило, известное, лицо, в том числе за того, кто известен как писатель).

- Писатель-призрак (ghost writer)
- Книггер (рус. «книжный» + «ниггер»).

Игорь Станиславович Ашманов

Орфо
Контекст
Спамтест
Поисковые технологии
Поисковая оптимизация
Автоконтекст
Наносемантика
Диалоговые боты
Интеллектуальные агенты

Выявление вбросов



Выявление вбросов