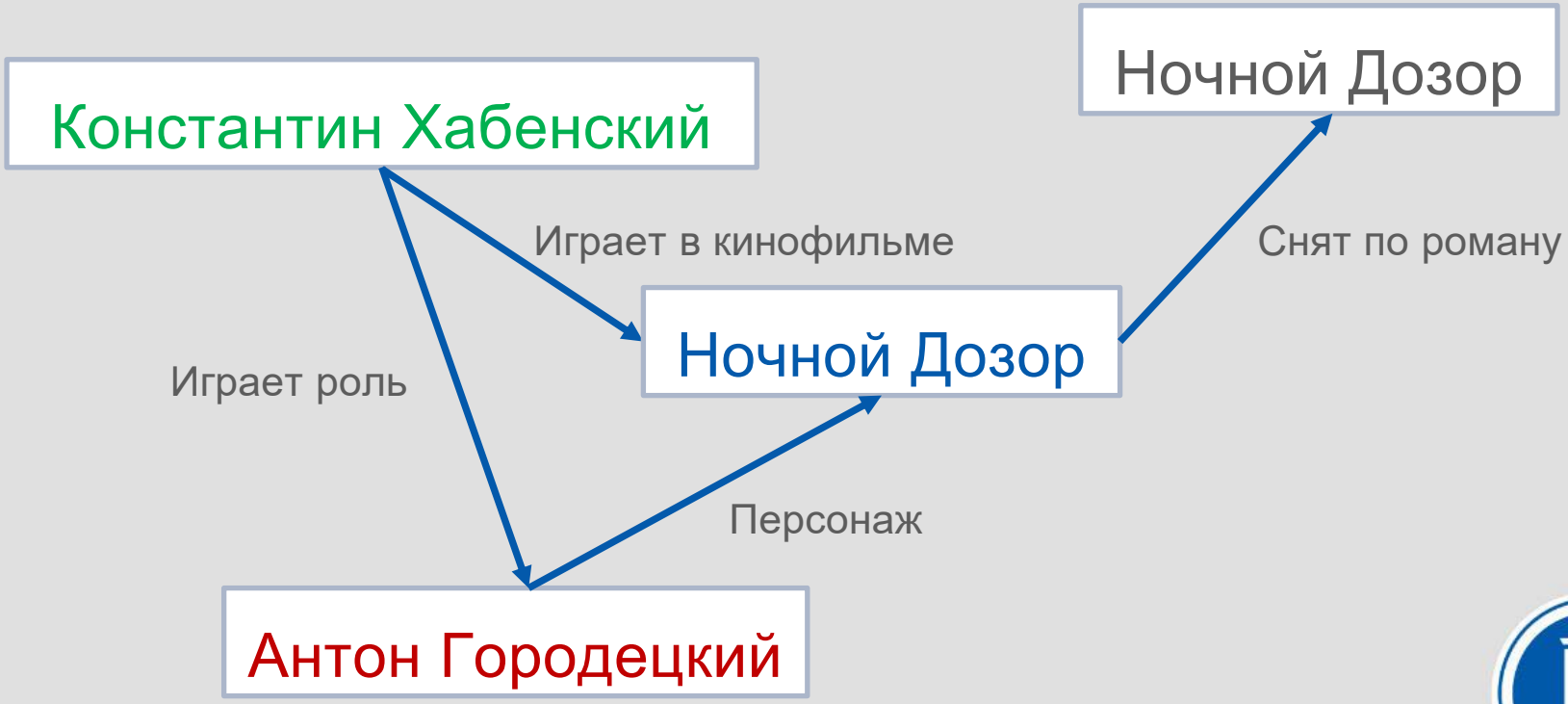
Где используется обработка естественного языка

На основе лекций Артемовой Екатерины https://openedu.ru/course/hse/TEXT/#

Обработка текстовой информации

- Значительная часть данных в мире представлена в текстовом виде
- Такие данные могут быть
  - структурированными (графы знаний, БД)



Обработка текстовой информации

- Значительная часть данных в мире представлена в текстовом виде
- Такие данные могут быть
  - структурированными (графы знаний, БД)
  - неструктурированными (сырые тексты)

Учебник по NLP.

Под авторством Сидорова Ивана Петровича

Настоящее учебное пособие …

Обработка текстовой информации

- Значительная часть данных в мире представлена в текстовом виде
- Такие данные могут быть
  - структурированными (графы знаний, БД)
  - неструктурированными (сырые тексты)
  - частично структурированными

```
{
Title: "Учебник по NLP«
Author: «Сидоров Иван Петрович"
Abstract: "Настоящее пособие …"
...
}
```

Обработка текстовой информации

- Методы обработки естественного языка (*Natural Language Processing*) успешно применяются во всех описанных случаях
- В основе NLP лежат
  - классическое машинное обучение
  - глубокое обучение
  - компьютерная лингвистика

Обработка текстовой информации

- Методы обработки естественного языка (*Natural Language Processing*) успешно применяются во всех описанных случаях
- В основе NLP лежат
  - классическое машинное обучение
  - глубокое обучение
  - компьютерная лингвистика
- NLP включает в себя две важных области (но не ограничивается ими!):
  - Понимание естественного языка (NLU)
  - Генерация естественного языка (NLG)

Особенности обработки языка

- Слово — базовая структурная единица языка
- Безотносительно контекста оно уже несёт много полезной информации
- В обработке языка обычно очень большие и разреженные признаковые пространства

Особенности обработки языка

- Текст без дополнительной разметки имеет внутреннюю структуру, определяемую языком:
  - Текст (дискурс, порядок реплик)
  - Предложение (синтаксис)
  - Словосочетания и слова (синтаксис, морфология)

Особенности обработки языка

- Текст без дополнительной разметки имеет внутреннюю структуру, определяемую языком:
  - Текст (дискурс, порядок реплик)
  - Предложение (синтаксис)
  - Словосочетания и слова (синтаксис, морфология)

Большое число сырых текстов

Наличие в них языковой структуры

Обучение больших общеязыковых моделей на сырых данных

Особенности обработки языка

- Текст без дополнительной разметки имеет внутреннюю структуру, определяемую языком:
  - Текст (дискурс, порядок реплик)
  - Предложение (синтаксис)
  - Словосочетания и слова (синтаксис, морфология)

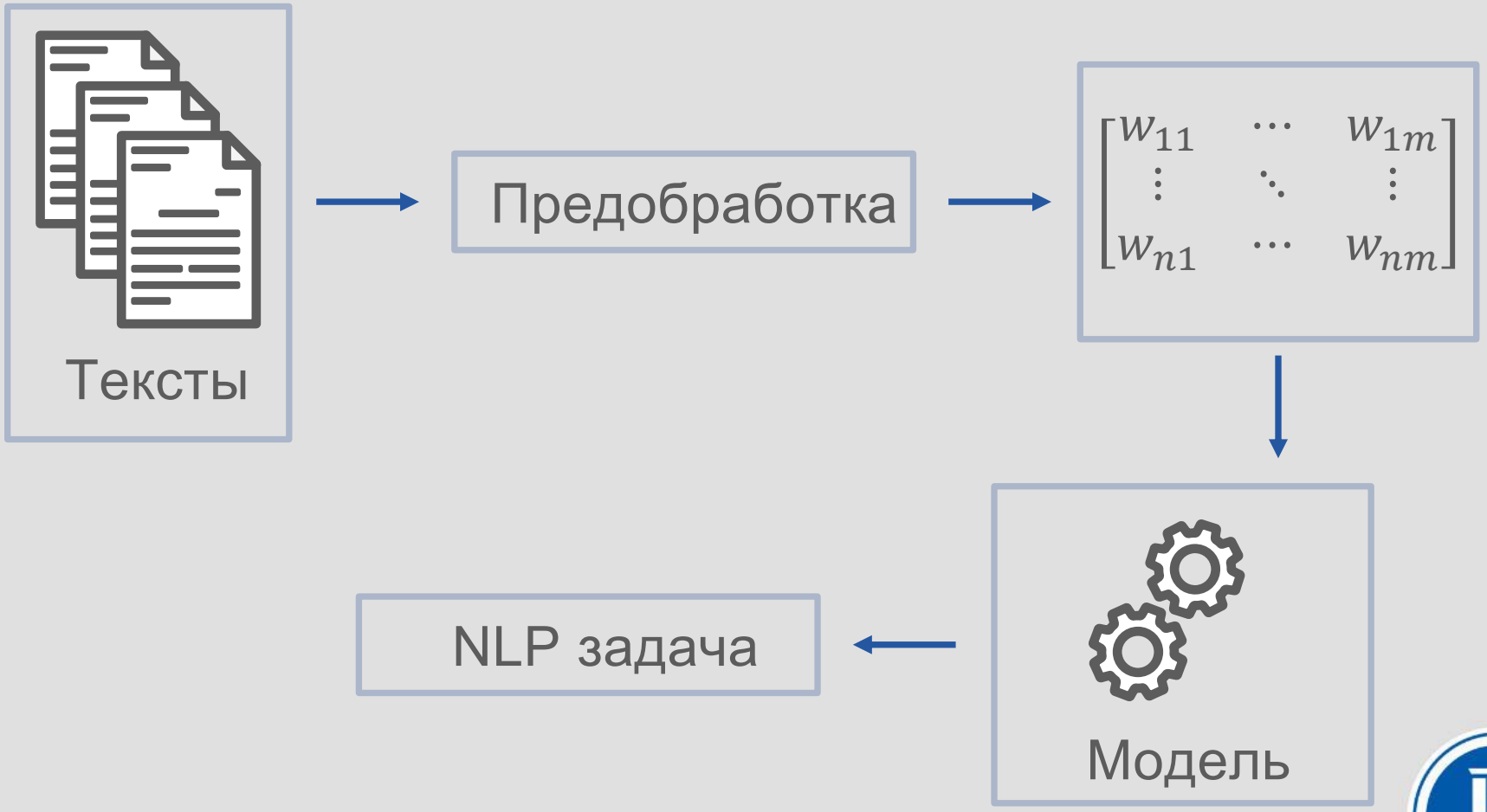
Большое число сырых текстов | Наличие в них языковой структуры

Обучение больших общеязыковых моделей на сырых данных

- Есть много лингвистических ресурсов, которые могут помочь при работе с текстами

Машинное обучение в обработке текстов

С середины 2010-х годов доминируют нейросетевые модели!



Пример задачи: классификация

- В работе многих интернет-сервисов используются классификаторы текстов для
  - фильтрации спама в почте
  - анализа тональности отзывов (Яндекс.Маркет)
  - категоризации новостей (Яндекс.Новости)
  - выявления некачественных текстов (Авито)

Никогда больше не стану покупать у вас пончики. В пончике должна быть ОДНА дырка посредине, а в вашем пончике я нашла 15! Ну это никуда не годится!

Обожаю пончики на завтрак! Это лучшее начало трудового дня, они заряжают энергией и хорошим настроением меня и всю мою семью. Гомер С.

Пример задачи: ранжирование

- Из множества текстов нужно выбрать те, которые наиболее хорошо подойдут в данной ситуации и отсортировать их по полезности
  - Информационный поиск — релевантность текста с сайта пользовательскому запросу
  - Рекомендации — близость текста статьи к интересам пользователя

Пример задачи: ранжирование

Google

Где находится Марианская Впадина?

Исследования Марианской Впадины…

Jan 23, 2020 - Марианская впадина находится на стыке двух литосферных плит. Вдоль ее оси происходит поддвиг Тихоокеанской литосферной ...

На дне Марианской впадины нашли то, что меньше всего ...

May 14, 2019 - Марианская впадина - самая глубокая точка на земле. Экспедиция Five Deeps под руководством Виктора Весково, 53-летнего ...

Марианский жёлоб — Википедия

Мариáнский жёлоб (или Мариáнская впáдина) — океанический глубоководный жёлоб на западе Тихого океана, самый глубокий из известных на Земле. Назван

Пример задачи: ранжирование

Google

Где находится Марианская Впадина?

Исследования Марианской Впадины…

Jan 23, 2020 - Марианская впадина находится на стыке двух литосферных плит. Вдоль ее оси происходит поддвиг Тихоокеанской литосферной ...

2

На дне Марианской впадины нашли то, что меньше всего ...

May 14, 2019 - Марианская впадина - самая глубокая точка на земле. Экспедиция Five Deeps под руководством Виктора Весково, 53-летнего ...

3

Марианский жёлоб — Википедия

Мариа́нский жёлоб (или Мариа́нская впа́дина) — океанический глубоководный жёлоб на западе Тихого океана, самый глубокий из известных на Земле. Назван

1

Пример задачи: ранжирование

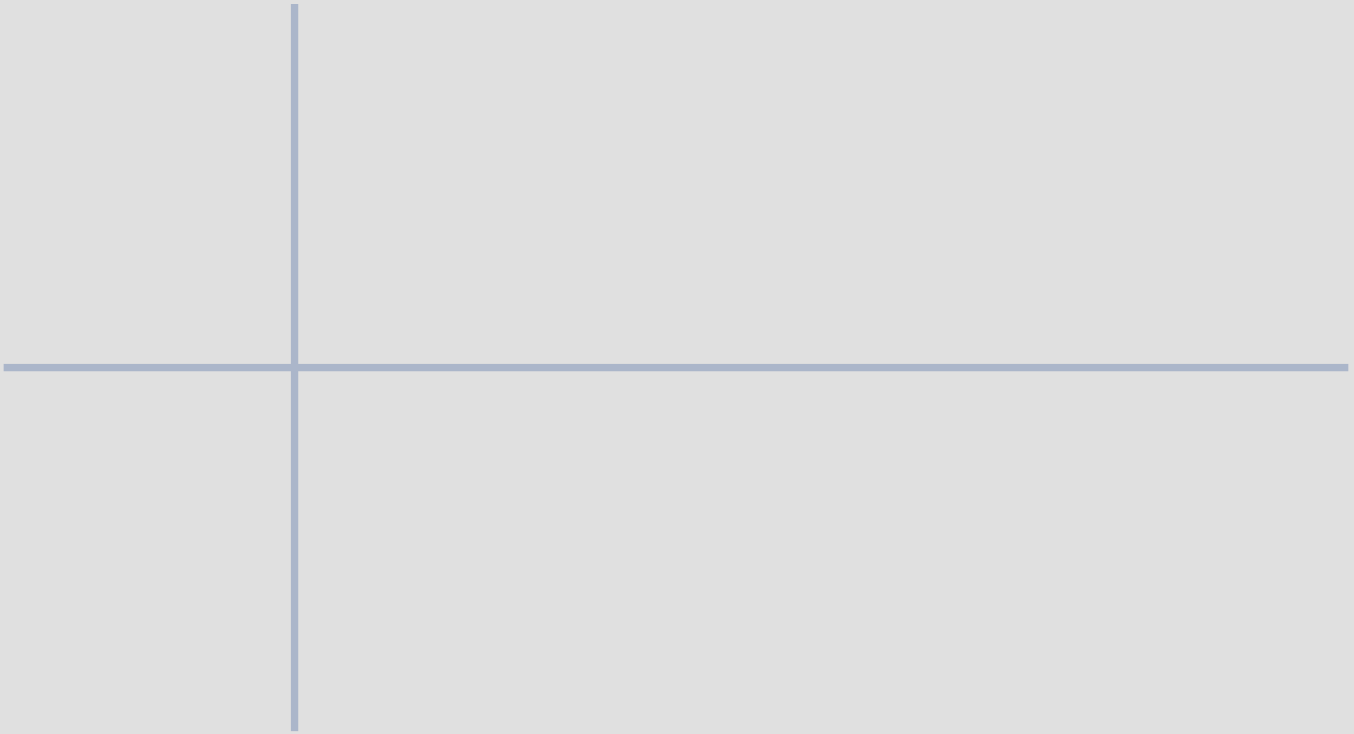
- Из множества текстов нужно выбрать те, которые наиболее хорошо подойдут в данной ситуации и отсортировать их по полезности
  - Информационный поиск — релевантность текста с сайта пользовательскому запросу
  - Рекомендации — близость текста статьи к интересам пользователя
- Примеры:
  - Google
  - Яндекс.Поиск
  - Яндекс.Дзен
  - Microsoft Bing

Пример задачи: машинный перевод

- Для текста на одном языке получить как можно более точный его перевод на другой язык
- Недостаточно просто перевести слова
- Важно учесть словосочетания, обороты речи, устойчивые выражения, семантику фразы и т.д.
- Примеры:
  - Google Переводчик
  - Яндекс.Переводчик

Пример задачи: машинный перевод

The shop owner caught the boy red-handed when he was stealing cigarettes.



Пример задачи: машинный перевод

The shop owner caught the boy red-handed when he was stealing cigarettes.

Хозяин магазина поймал парня краснoруким, когда он воровал сигареты.

Хозяин магазина поймал парня с поличным, когда он воровал сигареты.

Пример задачи: машинный перевод

The shop owner caught the boy red-handed when he was stealing cigarettes.

Хозяин магазина поймал парня краснoруким, когда он воровал сигареты.

Хозяин магазина поймал парня с поличным, когда он воровал сигареты.

Пример задачи: анализ и коррекция текста

- Проанализировать текст на предмет наличия
  - опечаток в словах
  - ошибок согласования слов
  - синтаксических ошибок в тексте

Пример задачи: анализ и коррекция текста

Когад-то в России и правда жило беспечальное юное поколение, которое улыбнулось к лету, морю и солнцу – и выбрали «Пепси».

Щас уже турдно установить, почему это произошло…

Пример задачи: анализ и коррекция текста

Когад-то в России и правда жило беспечальное юное поколение, которое улыбнулось к лету, морю и солнцу – и выбрали «Пепси».

Щас уже турдно установить, почему это произошло…

Пример задачи: анализ и коррекция текста

Когда-то в России и правда жило беспечальное юное поколение, которое улыбнулось лету, морю и солнцу – и выбрало «Пепси».

Сейчас уже трудно установить, почему это произошло…

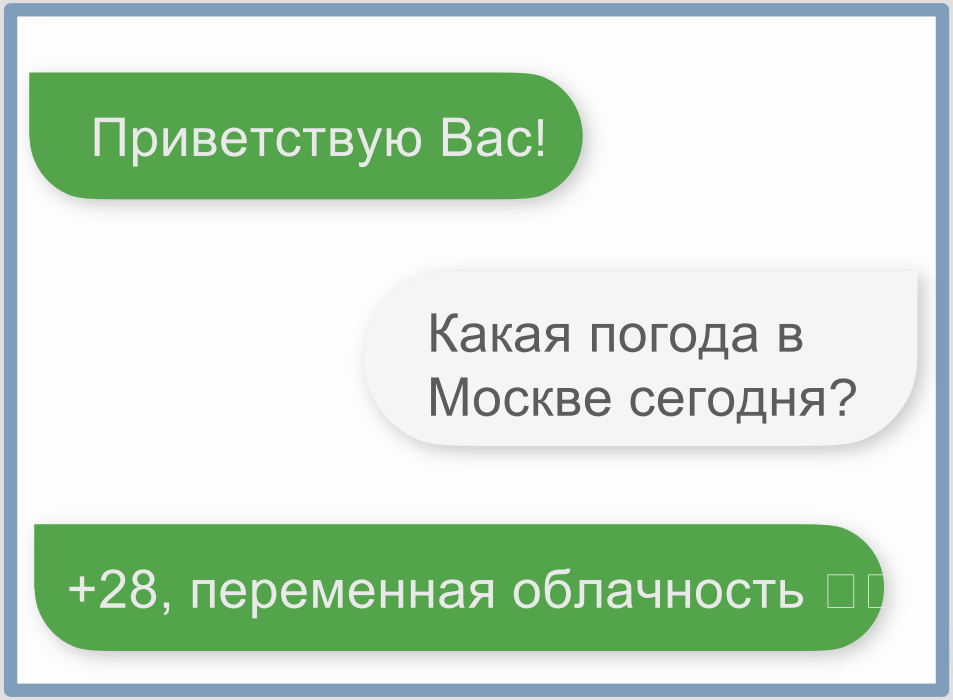
© В.О. Пелевин, “Generation П”

Пример задачи: анализ и коррекция текста

- Проанализировать текст на предмет наличия
  - опечаток в словах
  - ошибок согласования слов
  - синтаксических ошибок в тексте
- Оценить стиль текста, качество изложения, дать рекомендации по его улучшению
- Примеры:
  - Проверка грамматики в редакторе Microsoft Word
  - Проверка грамматики, стиля и качества текста в сервисе Grammarly

Пример задачи: ведение диалога

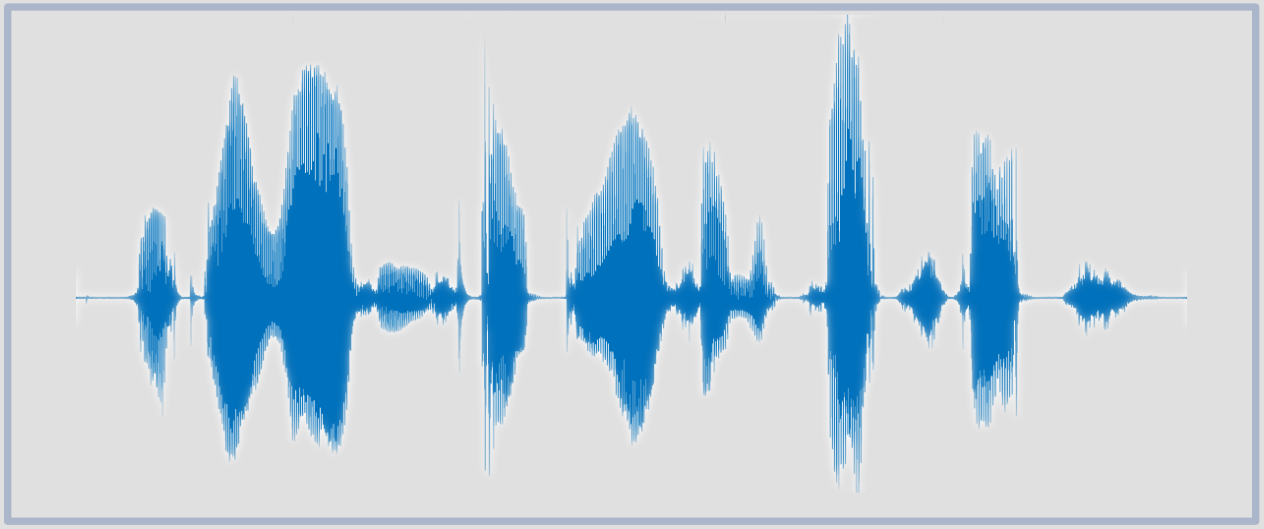
- Строится система, способная
  - обмениваться с человеком текстовыми сообщениями
  - анализировать входные реплики
  - на основе анализа генерировать адекватные ответы или производить нужные действия
- Примеры:
  - Google Ассистент
  - Amazon Alexa
  - Яндекс Алиса



Пример задачи: распознавание речи

- Диалоговые системы в виртуальных ассистентах предусматривают голосовой ввод
- Процесс перевода звукового сигнала в текст тесно связан с языковым моделированием
- Для заданной последовательности звуковых фрагментов нужно подобрать наиболее вероятный набор текстовых
- Примеры:
  - Google Cloud Speech API (Google Ассистент)
  - Amazon Transcribe (Alexa)
  - Яндекс SpeechKit API (Алиса)

Пример задачи: распознавание речи



…привет подскажи какая будет завтра погода…

➢ https://wiki.aalto.fi/display/ITSP/Windowing

Пример задачи: поиск ответов на вопросы

- Вопросно-ответные системы предназначены для поиска ответов на текстовые вопросы пользователя
- Вопросы могут быть как общего характера, так и, например, связанными с каким-то текстом
- QA-системы часто встроены в поисковые системы
- Примеры: *колдунщики* в поисках Яндекса и Google

Кто написал “Старик и Море”?

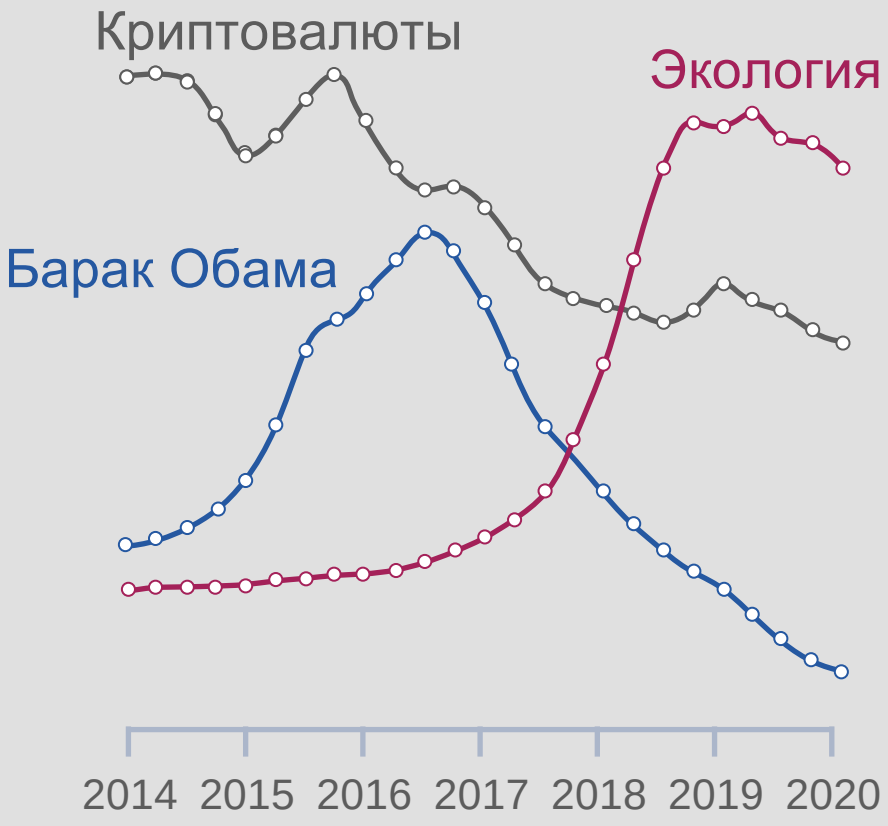
Эрнест Хемингуэй

Столица Соединённого Королевства?

Лондон

Пример задачи: тренд-аналитика

- В социальных и новостных интернет-ресурсах анализируется состав контента
- Отслеживается изменение этого состава в динамике
- Примеры:
  - Анализ популярности тем и хэштегов в Twitter и Instagram
  - Бренд-аналитика в сервисе YouScan



Пример задачи: суммаризация

- Для исходного текста необходимо сгенерировать краткое изложение
- Важно сохранить не только смысл, но и важные факты, содержащиеся в тексте
- Пример: сниппеты новостей в сервисе Яндекс.Новости

Пример задачи: суммаризация

Вот растение, которому в наш суматошный век истрёпанных нервов, изнурительных бессонниц и сдвинутой с места психики надо бы поставить
красивый памятник: валериана,
подобно матери, успокоит и усыпит,
вернет так необходимое всем нам душевное равновесие.

© В. Солоухин

Пример задачи: суммаризация

Валериане следовало бы поставить памятник: она успокаивает, помогает заснуть, возвращает душевное равновесие.

Основные выводы

- Огромное количество текстов в мире делает NLP одним из важнейших направлений в анализе данных
- Тексты — специфический вид данных, для работы с ними часто нужны особые подходы
- Тем не менее, многие модели классического ML и CV широко используются и при обработке текстов
- Существует масса прикладных задач, связанных с обработкой текстов
- Со многими из них мы познакомимся по ходу курса!

Предобработка текстовых данных

Что делать с данными

- Предположим, что нам дан набор данных D (коллекция), каждый элемент d которого является текстовым документом
- По сути, текст — это одна длинная строка из различных алфавитных и неалфавитных символов
- Работать с ним в таком виде иногда можно, но в большинстве задач это очень неудобно
- До выделения признаков и построения моделей данные надо привести к подходящему виду
- Этот процесс называется предобработкой

Какие бывают этапы предобработки

- К стандартным этапам предобработки обычно относят:
  - Токенизация
  - Приведение к нижнему регистру
  - Удаление пунктуации
  - Удаление стоп-слов
  - Фильтрация по длине/частоте/соответствию регулярному выражению
  - Лемматизация или стемминг
- Их можно применять в различных комбинациях или вообще не применять
- Разберём каждый подробно!

Токенизация

- Для анализа текст d удобно разбить на структурные единицы
- Обычно это предложения и/или слова
- Предложения можно выделять с помощью правил на основе регистра и пунктуации
- Можно обучать модели, предсказывающие границы предложений в последовательности слов

Токенизация

Но это стоит нервов, нервов и нервов. И сердца. И психики. Поэтому наряду с тишиной становится дефицитной на земном шаре и валерьянка.

© В. Солоухин

Токенизация

Но это стоит нервов, нервов и нервов. И сердца. И психики. Поэтому наряду с тишиной становится дефицитной на земном шаре и валерьянка.

© В. Солоухин

Токенизация

Но это стоит нервов, нервов и нервов.

И сердца.

И психики.

Поэтому наряду с тишиной становится дефицитной на земном шаре и валерьянка.

© В. Солоухин

Токенизация

- Для анализа текст d удобно разбить на структурные единицы
- Обычно это предложения и/или слова
- Разбиение на слова обычно делается по заданным символам или регулярному выражению
- Чаще всего на практике используется пробел

Токенизация

Но это стоит нервов, нервов и нервов. И сердца. И психики. Поэтому наряду с тишиной становится дефицитной на земном шаре и валерьянка.

© В. Солоухин

Токенизация

Но это стоит
нервов, нервов и
нервов. И сердца.
И психики. Поэтому
наряду с тишиной
становится дефицитной
на земном шаре и
валерьянка.

© В. Солоухин

Регистр и пунктуация

- Обычно для унификации текста все слова в нём приводятся к нижнему регистру, а большая часть пунктуации удаляется
- Удалять можно правилами или регулярными выражениями

Токенизация

Но это стоит нервов, нервов и нервов. И сердца. И психики. Поэтому наряду с тишиной становится дефицитной на земном шаре и валерьянка.

© В. Солоухин

Токенизация

но это стоит нервов нервов и
нервов и сердца и психики поэтому
наряду с тишиной становится
дефицитной на земном шаре и
валерьянка

© В. Солоухин

Регистр и пунктуация

- Есть задачи, для которых такая информация оказывается важной
- Например, выделение именованных сущностей (NER) с помощью правил работает намного лучше с информацией об исходном регистре
- В задаче анализа тональности существенное значение имеют текстовые смайлики

Регистр и пунктуация

- Есть задачи, для которых такая информация оказывается важной
- Например, выделение именованных сущностей (NER) с помощью правил работает намного лучше с информацией об исходном регистре
- В задаче анализа тональности существенное значение имеют текстовые смайлики *(emoticons, emoji)*, они могут конвертироваться в символы Unicode

Магазин у вас достаточно своеобразный:))))))

Магазин у вас достаточно своеобразный:((((((

Удаление лишних слов

- Часть слов в текстах часто является малоинформативной
- Такие слова могут мешать модели, и на работу с ними может тратиться значительная часть ресурсов

- К таким словам почти всегда относятся общеупотребительные стоп-слова:
  - Союзы
  - Предлоги
  - Вводные слова
  - Модальные глаголы
  - Местоимения

Удаление лишних слов

As we get older and when we think about our past we sometimes ponder the things that we should have done. And we also may regret those things we did badly and the mistakes we made. In reality, we can always learn from our mistakes and hope to never make them again. For example, if I failed a test because of a lack of study, the next test I will hope to pass because of hard work. Remember too that some regrets are not based in reality and we may waste time thinking that they are. Would I have really not have been involved in a car crash if I had been driving more slowly?

➢ https://www.bellenglish.com/blog/story-modal-verbs/

Удаление лишних слов

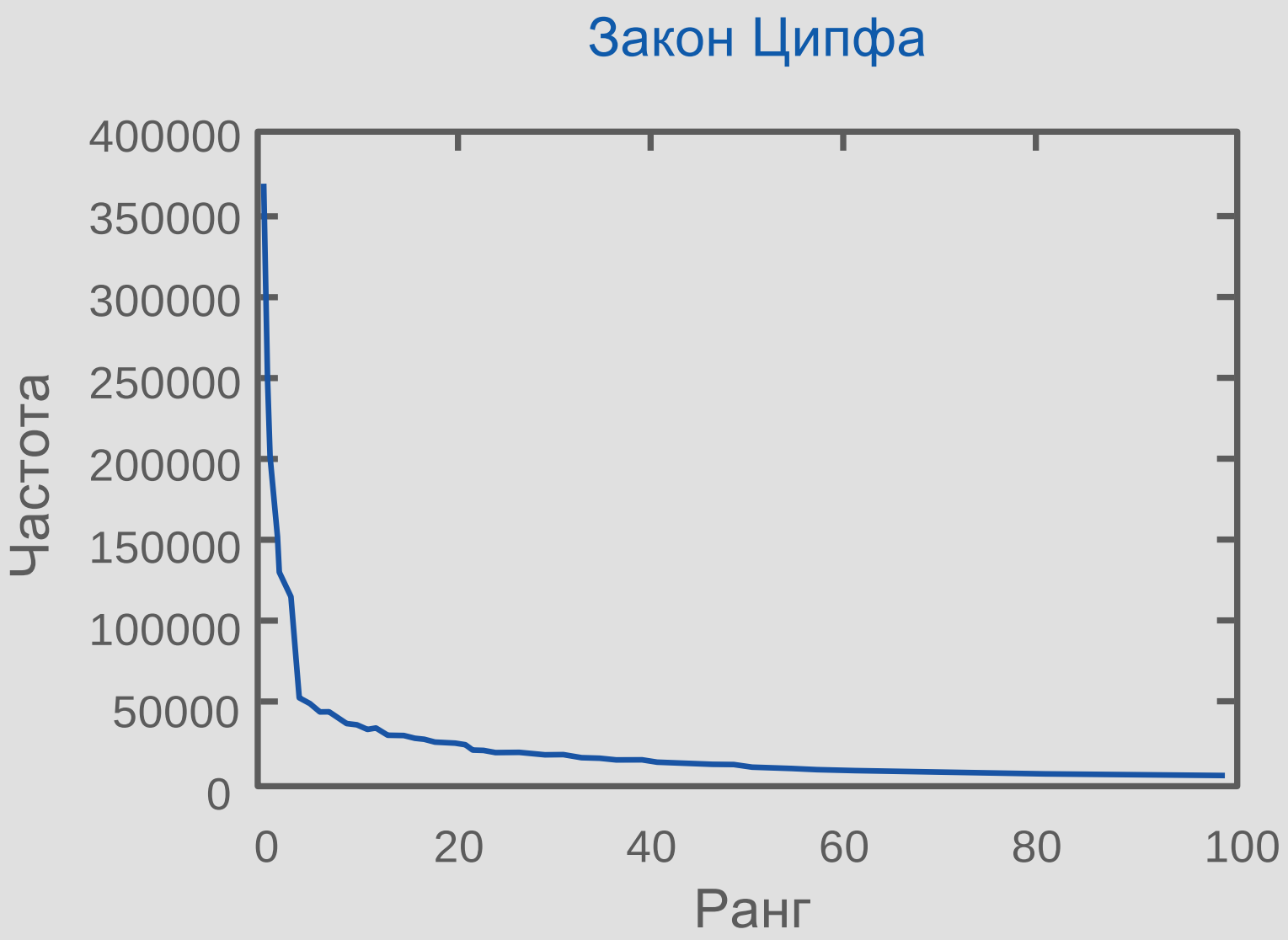
As we get older and when we think about our past we sometimes ponder the things that we should have done. And we also may regret those things we did badly and the mistakes we made. In reality, we can always learn from our mistakes and hope to never make them again. For example, if I failed a test because of a lack of study, the next test I will hope to pass because of hard work. Remember too that some regrets are not based in reality and we may waste time thinking that they are. Would I have really not have been involved in a car crash if I had been driving more slowly?

Удаление лишних слов

- Часть слов в текстах часто является малоинформативной
- Такие слова могут мешать модели, и на работу с ними может тратиться значительная часть ресурсов

- Часто такими могут являться слишком редкие/частые слова в анализируемом корпусе
- Полезно строить ранжированный график частот слов, обычно он удовлетворяет закону Ципфа

Удаление лишних слов



Регулярные выражения

- Регулярные выражения появились из теории автоматов и классификации формальных грамматик по Хомскому
- По факту регулярное выражение — это текстовый шаблон
- Из текста можно выделить фрагменты, соответствующие этому шаблону
- Регулярные выражения описываются специальным языком
- Иногда они могут быть сложными
- Это очень полезный и часто используемый инструмент

```
.*(от | за | на) [0-9]{1,2} (март | апрел | ма(я|й)).*
```

Нормализация слов

- Слова в тексте могут иметь различные формы, часто такая информация скорее мешает анализу
- Тогда применяется один из двух стандартных подходов:
  - Лемматизация
  - Стемминг
- Стоп-слова часто удаляются после нормализации!

Нормализация слов

- Слова в тексте могут иметь различные формы, часто такая информация скорее мешает анализу
- Тогда применяется один из двух стандартных подходов:
  - Лемматизация
  - Стемминг
- Лемматизация приводит слова к нормальной форме:

Туристам очень понравилась прогулка по мосту

Турист очень понравиться прогулка по мост

Нормализация слов

- Слова в тексте могут иметь различные формы, часто такая информация скорее мешает анализу
- Тогда применяется один из двух стандартных подходов:
  - Лемматизация
  - Стемминг
- Лемматизация приводит слова к нормальной форме
- Такой подход даёт более качественный результат, хотя часть слов может оказаться испорченной
- Лемматизаторы могут быть словарными или нейросетевыми
- Вторые работают существенно медленнее

Нормализация слов

- Слова в тексте могут иметь различные формы, часто такая информация скорее мешает анализу
- Тогда применяется один из двух стандартных подходов:
  - Лемматизация
  - Стемминг
- Стемминг приводит слова к псевдооснове (удаляет окончание и формообразующий суффикс, не удаляет приставки):

Туристам очень понравилась прогулка по мосту

Турист очень понрав прогулк по мост

Нормализация слов

- Слова в тексте могут иметь различные формы, часто такая информация скорее мешает анализу
- Тогда применяется один из двух стандартных подходов:
  - Лемматизация
  - Стемминг
- Стемминг приводит слова к псевдооснове (удаляет окончание и формообразующий суффикс, не удаляет приставки)
- Результат ощутимо хуже, чем у лемматизации
- Но работает достаточно быстро, что может быть важным в высоконагруженном сервисе
- Обычно для русского языка используется лемматизация, для английского — стемминг

Полезные библиотеки в Python

- Почти для всех описанных шагов есть готовые решения, которые можно и нужно использовать:
  - nltk — важнейший инструмент для обработки текстов в Python, содержит токенизаторы, стеммеры, списки стоп-слов и многое другое
  - pymorphy2 и pymystem3 — лемматизатор для русского языка
  - re и regex — библиотеки для построения регулярных выражений
- Полезные вещи для предобработки на разных этапах есть в стандартной библиотеке Python и модуле sklearn

Основные выводы

- Предобработка данных необходима практически в любой задаче, связанной с текстами
- Существует стандартный набор шагов, который можно урезать, выполнять многократно и в разном порядке
- Можно придумывать и использовать свою специфическую предобработку, если она полезна
- Для стандартных шагов есть готовые инструменты в Python, которыми нужно пользоваться, если нет острой необходимости делать специализированное решение

Выделение признаков из текстов

Признаковое описание объектов

- В стандартных задачах машинного обучения привычным представлением данных является матрица «объекты-признаки»:

Номер пациента	Пол	Возраст	…	Рост	Вес
1	М	43	…	178	78
…	…	…	…	…	…
N	Ж	21	…	176	54

- Признаки могут иметь различные типы значений
- Значения могут упорядоченными или нет

Признаки для объекта-текста

- В NLP объектами чаще всего являются тексты или их фрагменты
- Для них тоже требуется как-то строить признаковые описания и составлять матрицу «объекты-признаки»
- Пример: наличие в тексте слова «ужасно»:

Номер текста	Содержит «ужасный»
1	1
…	…
N	0

- Такой простой признак может быть неплох в задаче анализа тональности
- Но кажется, что не используется много полезной информации

Признаки для объекта-текста

- Базовыми элементами текста являются слова
- В описанном примере используется только информация о наличии/отсутствии одного слова
- Можно проверять наличие всех возможных слов из некоторого словаря:

Номер текста	содержит «абрикос»	…	содержит «ямал»
1	0	…	1
…	…	…	…
N	1	…	0

- Этот набор признаков уже несёт немало информации и полезен в разнообразных задачах

«Мешок слов»

- Допустим, что мы работаем с длинными текстами
- В них встречается много разнообразных слов
- Но первый текст явно связан с полуостровом Ямал, это название встречается в нём очень много раз
- Текущее признаковое описание это никак не отражает
- Попробуем учесть эту информацию:

Номер текста	вхождений слова «абрикос»	…	вхождений слова «ямал»
1	0	…	23
…	…	…	…
N	4	…	0

- Такое представление называется «мешком слов»
- Это стандартный тип признаков для текстов

Матрица TF-IDF

- «Мешок слов» часто используется на практике
- Но частота встречаемости — не самый информативный признак для слова
- Полезнее было бы знать для слова, насколько оно
  - часто встречается в документе d
  - отличает документ d от прочих

Матрица TF-IDF

- Заменим в описании d счётчик слова на значение TF-IDF:

$$tfidf_{wd} = tf_{wd} \times \log \frac{|D|}{df_w}$$

  - tf_{wd} — отношение числа вхождений слова w в d к общему число слов в d
  - df_w — число документов, содержащих w

Учёт связей между словами

- Признаки TF-IDF тоже часто используются на практике и показывают хорошие результаты
- Однако мы полностью упускаем информацию о порядке слов в текстах!
- Особенно это важно при обработке устойчивых сочетаний слов (коллокаций, N-грамм)
- Пример: «глубокая нейронная сеть»

Учёт связей между словами

- Признаки TF-IDF тоже часто используются на практике и показывают хорошие результаты
- Однако мы полностью упускаем информацию о порядке слов в текстах!
- Особенно это важно при обработке устойчивых сочетаний слов (коллокаций, словарных N-грамм)
- Пример: «глубокая нейронная сеть»
- Одно из решений:
  - Выделяем коллокации из текста
  - Объявляем каждую из них отдельным токеном
  - Формируем для них матрицу «мешка слов»/TF-IDF

Выделение коллокаций

- Грубо, но относительно просто, коллокации можно выделять с помощью статистических методов
- Для пары токенов вычисляется числовая оценка взаимной зависимости на основе совместной встречаемости
- Токеном может быть как слово (униграмма), так и последовательность слов (N-грамма)
- Если оценка высокая — токены объединяются в один

Оценка зависимости слов в коллокации

- Один из вариантов оценки — PMI *(Pointwise Mutual Information)*:

$$\mathrm{PMI}(u, v) = \log \frac{N_{uv}}{N_u N_v}$$

  - N_{uv} — число документов, содержащих оба слова u и v в окне заданного размера
  - N_u — число документов, содержащих слово u

Выделение коллокаций

PMI	w^1	w^2
4.4862	Стивен	Хокинг
4.4862	Чарли	Чаплин
4.4860	Альбрехт	Дюрер
4.4860	светлое	пиво
4.4860	холодное	оружие
2.3764	корзина	интерес
2.3541	гастроном	произведение
1.4875	базар	чай
1.2456	кресло	папироса
0.8752	задание	колодец

- Реализации различных методов выделения коллокаций можно найти в библиотеке nltk

Основные выводы

- Как и для любых объектов, для текстов можно строить признаковые описания
- Желательно учитывать в признаках как можно больше информации, содержащейся в тексте
- Стандартными признаковыми описаниями являются «мешок слов» и векторы TF-IDF
- N-граммные признаки обычно позволяют обучать более качественные модели, чем униграммные
- Есть другие стандартные виды признаков, более сложные и качественные, о которых будем говорить в дальнейшем