Занятие №1

> **Мастер-класс: обработка естественного языка.**
> Установка и настройка окружения, обсуждение задачи.

Вступление

Еще всего несколько десятилетий назад главным источником новостей были газеты, но уже сегодня все новости мира попадают к человеку буквально за считанные минуты через экран телефона, планшета или компьютера. Посмотрим на список таких сайтов-ресурсов с новостями.

Название	Вид ресурса
Яндекс.Новости	Агрегатор СМИ
РИА Новости	Обычное СМИ
Лента	Обычное СМИ
Дзен	Агрегатор СМИ
«Рамблер/новости»	Агрегатор СМИ
Russia Today	Обычное СМИ

Заметим, что некоторые ресурсы являются не обычными СМИ, а `агрегаторами`. `Агрегаторы` собирают информацию с нескольких новостных ресурсов и выдают её пользователю по его интересам или важности самой новости. Это востребовано, потому что в современном мире, где информация находится буквально на каждом шагу, её становится все сложнее отсеивать. Новостной `агрегатор` выполняет роль “секретаря” человека, самостоятельно изучая все актуальные новости и передавая только самое важное или интересное.

Например, у Людмилы каждый сезон должна быть актуальная коллекция, поэтому для неё важно узнать о новых трендах в индустрии моды. А Олег увлекается гонками формулы-1, поэтому он с удовольствием следит за всеми событиями из этой области. Новостной `агрегатор` это единое место, в

котором найдут себе интересные новости и Людмила, и Олег. Но при этом информация о том, что в стране проходит важное мероприятие узнают оба.

Статьи про моду 80-х.

Информация по гонкам Формула-1.

Для этого новостные ресурсы задействую алгоритмы с использованием машинного обучения и нейронных сетей. Таким образом, любой желающий может создать свой собственный новостной `агрегатор` для подборки идеальных новостей.

Немного о задаче

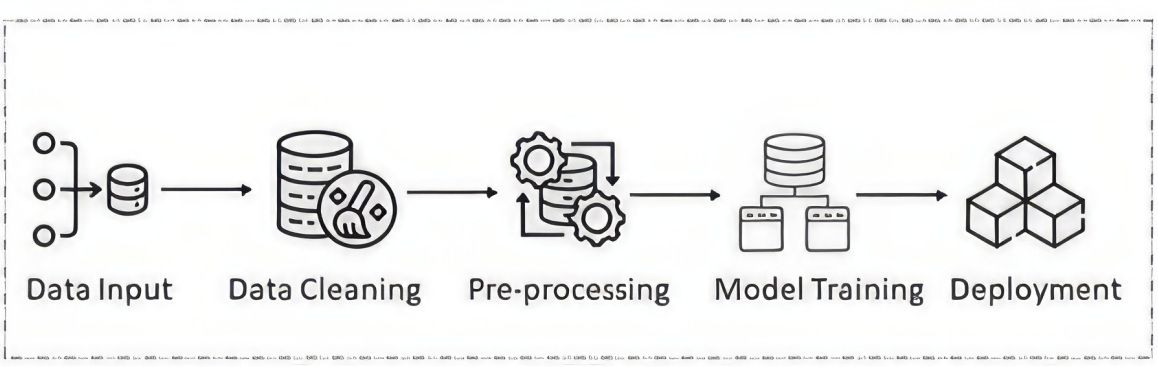
Вы решили написать свой собственный новостной агрегатор, который будет брать новости с популярных ресурсов и выдавать их конечному пользователю. Ключевая идея заключается в том, что на нашем новостном портале не будет негативных новостей. С помощью искусственного интеллекта мы будем фильтровать тексты и выдавать только позитивные или нейтральные.

План действия будет следующий:

0. (1 занятие) Установим все необходимые программы, библиотеки и настроим их;
1. (1 занятие) Выберем новостные порталы, которые будем обрабатывать (лента.ру и РБК);
2. (1-2 занятие) Напишем программу-парсер для скачивания данных с выбранных порталов;

3. (2 занятие) Обучим алгоритм предсказывать вид новости по её заголовку и тексту;
4. (2-3 занятие) Напишем простой сайт и загрузим туда новости с фильтрацией в реальном времени;
5. (3 занятие) Попробуем улучшить точность предсказания, применив другие подходы.

> 💡 Важно заметить, что все этапы это важные звенья в цепи разработки конечного ML-продукта, обобщенный процесс можно посмотреть на следующем изображении.



Процесс создания ML-продукта.

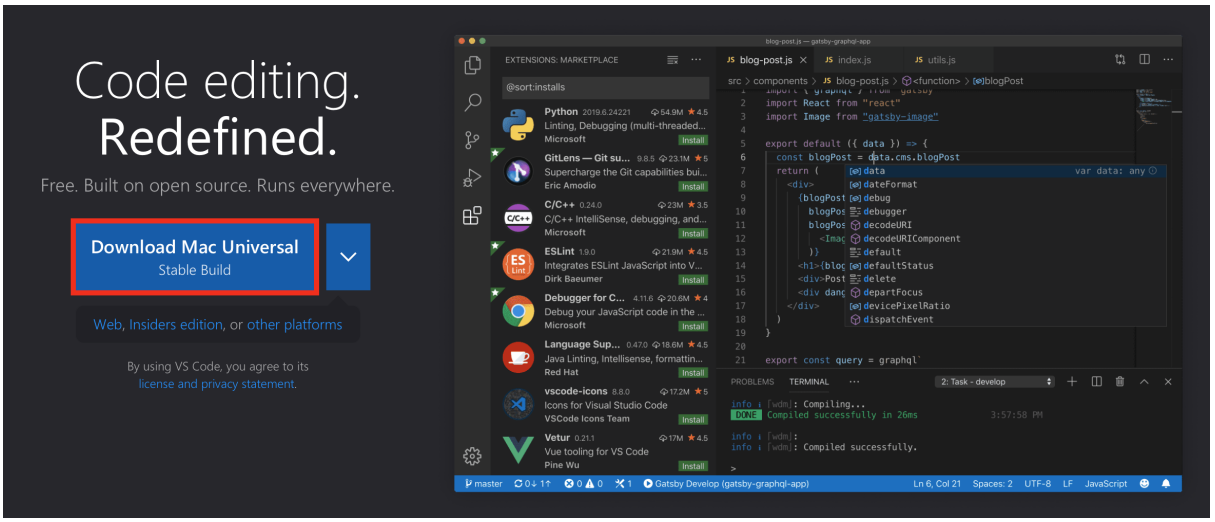
Итого, в рамках этих занятий мы рассмотрим основные этапы создания новостного `агрегатора` с использованием искусственного интеллекта. Изучим различные методы анализа текста и классификации новостей, а также научимся строить модели, способные автоматически определять релевантность новостных материалов.

Давайте начнем этот увлекательный процесс создания собственного новостного `агрегатора` и узнаем, как использовать искусственный интеллект для повышения качества информации, которую мы получаем!

Установка `Visual Studio Code`

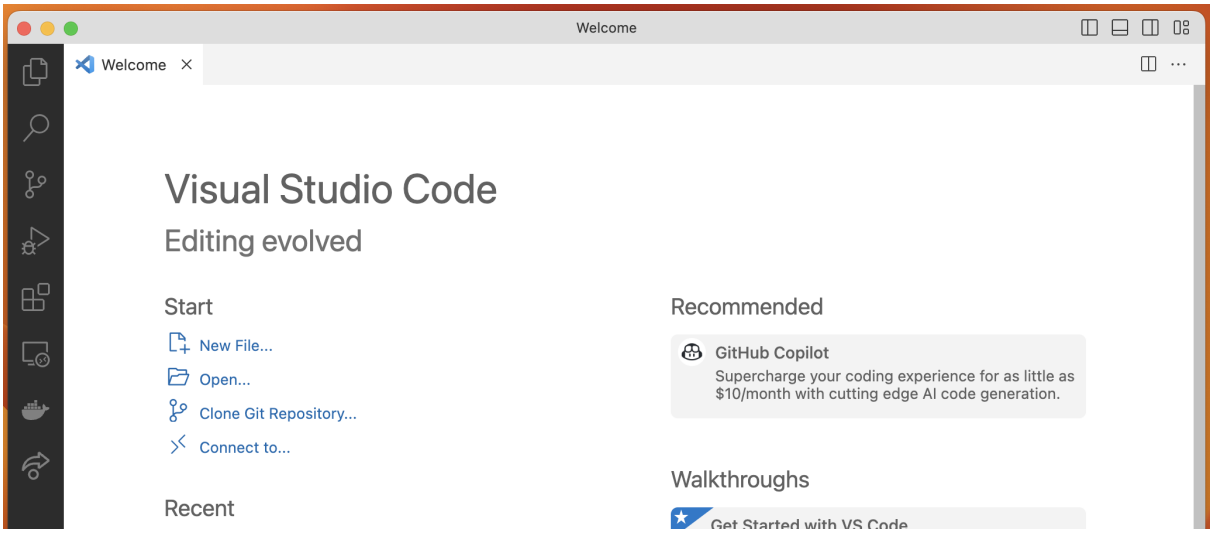
Нам необходим текстовый редактор для написания кода, но обычный блокнот не подойдет. Мы воспользуемся средой разработки, особым редактором, предназначенным специально для программистов.

Наш выбор пал на `Visual Studio Code` из-за его универсальности, легкости и удобства. Для того, чтобы его скачать необходимо перейти на официальный сайт и загрузить установщик.



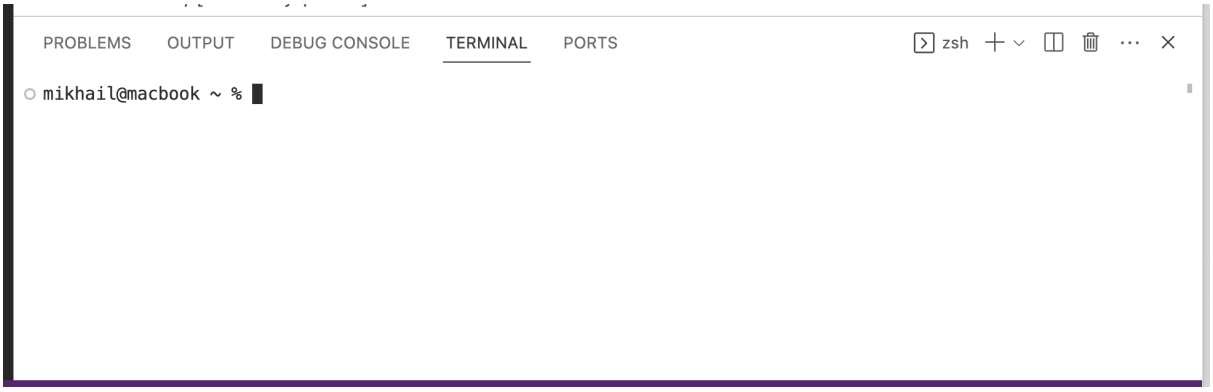
Далее нужно проследовать по всем пунктам установки, пока приложение не появится на рабочем столе.

Откроем программу, вы увидите следующее.



Теперь создайте папку проекта в любом месте на компьютере, где вам удобно будет с ней работать, и откройте её с помощью `Visual Studio Code`. Для этого нужно нажать `Файл → Открыть папку…`, после этого выберете нужную папку.

Далее откроем терминал для работы, для этого нажмем `Терминал → Новый терминал`. Все команды нужно будет вводить в нём.



Если возникнут какие-то проблемы с установкой `Visual Studio Code`, то возможно ответ будет в официальной инструкции с установкой программы.

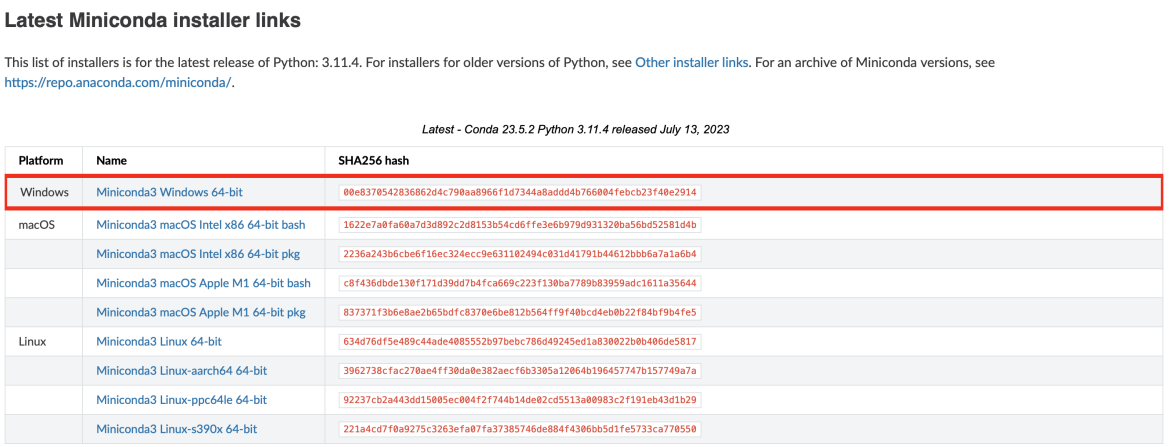
Установка и настройка окружения

Обычно чтобы начать программировать на Python необходимо установить сначала дистрибутив языка, а потом все необходимые библиотеки. Однако исследователи данных придумали более простое решение, собрали всё необходимое для работы в одном месте, дистрибутиве `Anaconda`.

Установка `Anaconda`

Инструкцию с установкой `Anaconda'ы` можно посмотреть на сайте с документацией. Существует два варианта установки: графический и через терминал.

Графический метод. Если вы новичок, то лучше воспользоваться этим способом. Про графический метод всё понятно, необходимо скачать исполняемый файл в зависимости от вашей операционной системы, запустить его и проследовать по всем шагам.



Latest Miniconda installer links

This list of installers is for the latest release of Python: 3.11.4. For installers for older versions of Python, see Other installer links. For an archive of Miniconda versions, see https://repo.anaconda.com/miniconda/.

Latest - Conda 23.5.2 Python 3.11.4 released July 13, 2023

Platform	Name	SHA256 hash
Windows	Miniconda3 Windows 64-bit	00e837054283686 2d4c790aa8966f1d7344a8addd4b766004febcb23f40e2914
macOS	Miniconda3 macOS Intel x86 64-bit bash	1622e7a0fa60a7d3d892c2d8153b54cd6ffe3e6b979d931320ba56bd52581d4b
	Miniconda3 macOS Intel x86 64-bit pkg	2236a243b6cbe6f16ec324ecc9e631102494c031d41791b44612bbb6a7a1a6b4
	Miniconda3 macOS Apple M1 64-bit bash	c8f436dbde130f171d39dd7b4fca669c223f130ba7789b83959adc1611a35644
	Miniconda3 macOS Apple M1 64-bit pkg	837371f3b6e8ae2b65bdfc8370e6be812b564ff9f40bcd4eb0b22f84bf9b4fe5
Linux	Miniconda3 Linux 64-bit	634d76df5e489c44ade4085552b97bebc786d49245ed1a830022b0b406de5817
	Miniconda3 Linux-aarch64 64-bit	3962738cfac270ae4ff30da0e382aecf6b3305a12064b196457747b157749a7a
	Miniconda3 Linux-ppc64le 64-bit	92237cb2a443dd15005ec004f2f744b14de02cd5513a00983c2f191eb43d1b29
	Miniconda3 Linux-s390x 64-bit	221a4cd7f0a9275c3263efa07fa37385746de884f4306bb5d1fe5733ca770550

Пример для операционной системы `Windows`.

Установка через терминал. Мы воспользуемся установкой через терминал/командную строку. Шаги будут различаться в зависимости от операционной системы.

Windows. Откроем командную строку и напишем команды:

```
curl https://repo.anaconda.com/miniconda/Miniconda3-latest-Wi
start /wait "" miniconda.exe /S
del miniconda.exe
```

macOS. Откроем терминал и пропишем команды:

```
mkdir -p ~/miniconda3
curl https://repo.anaconda.com/miniconda/Miniconda3-latest-Ma
bash ~/miniconda3/miniconda.sh -b -u -p ~/miniconda3
rm -rf ~/miniconda3/miniconda.sh

~/miniconda3/bin/conda init bash
~/miniconda3/bin/conda init zsh
```

Linux. Откроем терминал и пропишем команды:

```
mkdir -p ~/miniconda3
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Li
bash ~/miniconda3/miniconda.sh -b -u -p ~/miniconda3
rm -rf ~/miniconda3/miniconda.sh
```

```
~/miniconda3/bin/conda init bash
~/miniconda3/bin/conda init zsh
```

В итоге у нас будет установлена Anaconda, чтобы ею воспользоваться необходимо прописать в терминал `conda`.

> ⚠ Если у вас система отличная от `x86_64`, то вам необходимо сделать соответствующую замену в команде установки.

Активация/деактивация виртуального окружения

В процессе работы мы неизбежно будем устанавливать новые пакеты (подпрограммы), они нам пригодятся при работе с моделями машинного обучения. Но если, например, мы захотим после этого заняться разработкой веб-сайта на Python, придется устанавливать уже другие пакеты. Они могут и будут конфликтовать друг с другом, поэтому существуют виртуальные окружения, которые изолируют пакеты для разных задач. В `Anaconda` есть виртуальное окружение по умолчанию.

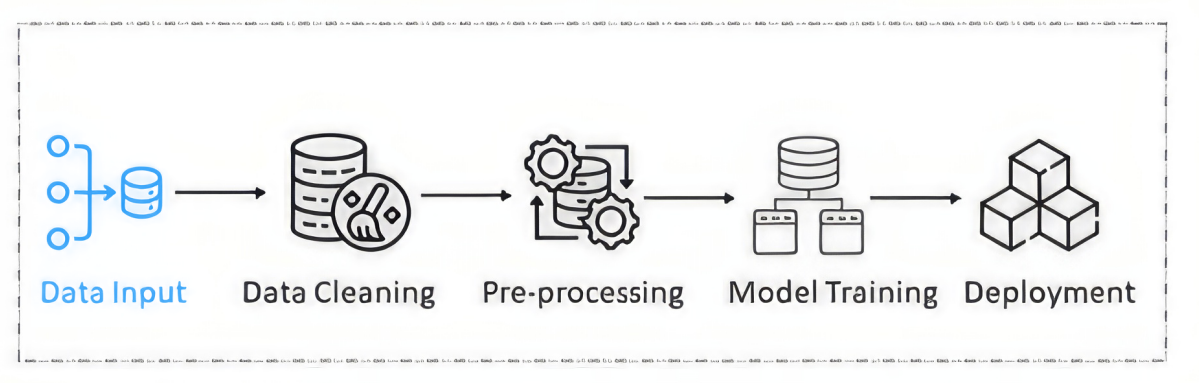
Чтобы активировать и деактивировать окружение необходимо прописать команду в терминале.

```
conda activate # Активировать окружение
conda deactivate # Деактивировать окружение
```

Активируем окружение, после активации в начале командной строки появится соответствующий индикатор с названием окружения.

В начале командной строки в круглых скобках указано название окружения.

Шаг №0. Поиск источника данных



Этапы создания ML-продукта.

Основа всего машинного обучения это данные, поэтому необходимо начать именно с них. Данные окружают нас повсюду, особенно в интернете, при этом большая часть находится в свободном доступе.

> ⚠ Чаще всего, данные предоставляются компанией, в которой работает разработчик.

Если же компания небольшая, то данные необходимо либо покупать, либо находить самостоятельно. Основными источниками данных тогда выступают платформы для проведения соревнований по искусственному интеллекту, на них размещены уже готовые наборы данных на различную тематику (в том числе и по нашей задаче). Но это все уже готовые данные, однако данные можно брать и напрямую из косвенных источников, таких как новостные порталы, социальные сети, поисковики и прочие источники информации. Такие данные называются сырыми и требуют дополнительной обработки.

В образовательных целях мы не будем брать готовый набор данных, а сделаем свой самостоятельно. Для этого первым делом нужно найти источник данных, для этого вернемся к табличке, которая была в самом начале занятия.

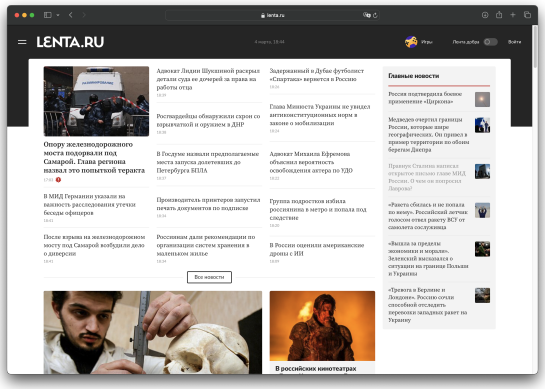
Название	Вид ресурса
Яндекс.Новости	Агрегатор СМИ
РИА Новости	Обычное СМИ
Лента	Обычное СМИ
Дзен	Агрегатор СМИ

Название	Вид ресурса
«Рамблер/новости»	Агрегатор СМИ
Russia Today	Обычное СМИ

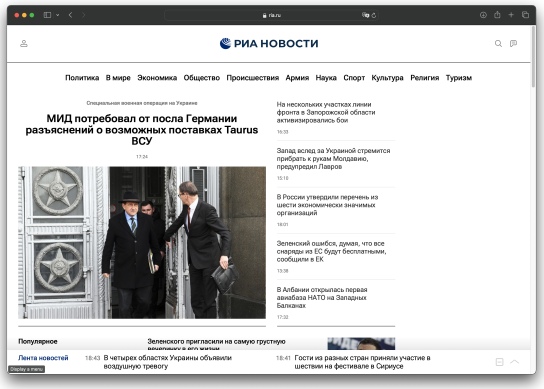
При выборе источника нам важно понимать следующее: это первоисточник (то есть обычное СМИ), мы можем брать данные ресурса с правовой точки зрения, удобно ли получать данные с помощью программы (и есть ли вообще такая возможность), это релевантные данные (то есть они похожи на те, что мы будем пытаться классифицировать в дальнейшем). Ответив на каждый вопрос, мы можем решить какой источник нам подходит, а какой — нет.

> !? Важно понимать, что написанная статья это чья-то собственность и на её использование (особенно в коммерческих целях) необходимо разрешение. Более того, если вы хотите открыть свой агрегатор новостей, то нужно зарегистрировать его в соответствующем реестре новостных агрегаторов, а также делать какие-то отчисления ресурсам, которые вы цитируете.

В рамках данного мастер-класса мы рассмотрим, как получить данные с ресурсов ”Лента” и “РИА Новости”.



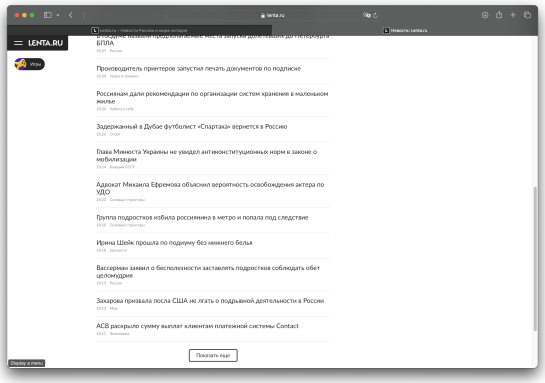
Сайт новостного портала “Лента”.



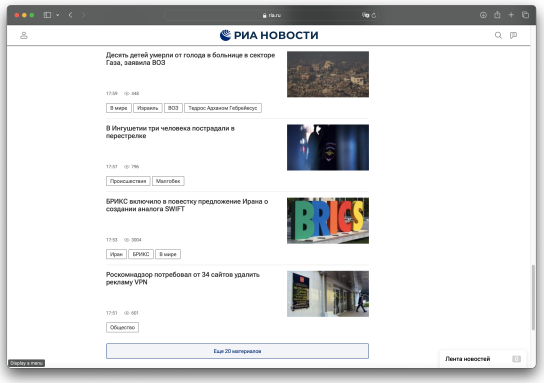
Сайт новостного портала “РИА Новости”.

Теперь нам нужно найти на каждом из сайтов, где находится страница со списком новостей, чтобы их можно было удобно загружать. Если на первый взгляд такой страницы нет, то можно воспользоваться поисковиком. С сайтом “Ленты” все сразу понятно, там есть кнопка “Все

новости”, она и ведёт на искомую страницу. А вот с сайтом “РИА Новостей” не все так просто, там на главной странице нет ссылки, но вот снизу есть карта сайта, где и можно найти необходимую нам страницу.



Страница-лента новостей на “Ленте”.



Страница-лента новостей на “РИА Новости”.