**ТЗ**

**Общая структура собираемых нами данных представлена ниже:**

|  |
| --- |
| **Данные** |
| 1. город отправления |
| 1. город прибытия |
| 1. дата поиска |
| 1. дата сбора данных |
| 1. дней до вылета |
| 1. класс обслуживания (Y – Эконом, С – Бизнес) |
| 1. авиакомпания |
| 1. код авиакомпании |
| 1. номер рейса |
| 1. индикатор рейса (прямой / с пересадкой, не более 1 пересадки) |
| 1. дата и время вылета (местное) |
| 1. дата и время вылета (Московское) |
| 1. дата и время прилета (местное) |
| 1. дата и время прилета (Московское) |
| 1. классы тарифов — тарифные семейства |
| 1. свободные места |
| 1. стоимость для каждого тарифного семейства |

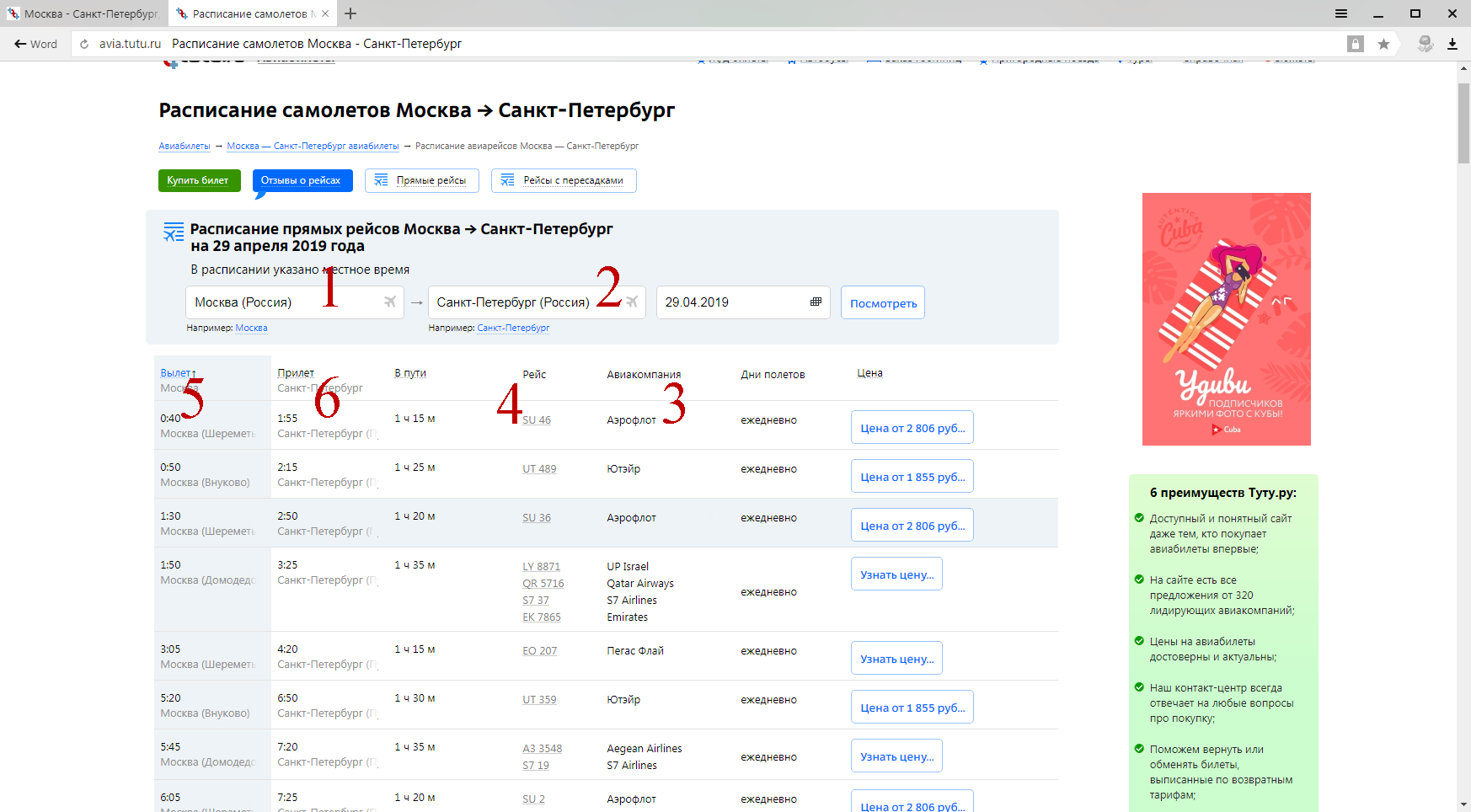
Сбор этих данных будет проводиться в основном с сайта tutu.ru в 3 этапа. На 1 и 2 этапе будут собираться условно-постоянные данные, на 3 – переменные.

**Этап 1. Сбор информации о рейсах**

<https://avia.tutu.ru/расписание_самолетов/Москва__Санкт-Петербург/>

Зеленым в таблице ниже выделены те данные, которые собираются на этом 1-м этапе

|  |  |
| --- | --- |
| **Данные** | **Примечание** |
| 1. город отправления | задается нами в парсинговом запросе (1 на рис.ниже) |
| 1. город прибытия | задается нами в парсинговом запросе (2 на рис.ниже) |
| 1. дата поиска |  |
| 1. дата сбора данных |  |
| 1. дней до вылета |  |
| 1. класс обслуживания (Y – Эконом, С – Бизнес) |  |
| 1. авиакомпания | Парсится (3 на рис.ниже) |
| 1. код авиакомпании | Берется из первых 2 символов номера рейса (4) |
| 1. номер рейса | Парсится (4) |
| 1. индикатор рейса (прямой / с пересадкой, не более 1 пересадки) | Парсится (подставляется для рейсов с пересадками) |
| 1. дата и время вылета (местное) | Время парсится (5), а даты вылета транслируются на следующие 180 дней на указанные в расписании дни недели |
| 1. дата и время вылета (Московское) | корректируется с учетом часового пояса аэропорта вылета |
| 1. дата и время прилета (местное) | Время парсится(6), а даты вылета транслируются на следующие 180 дней на указанные в расписании дни недели |
| 1. дата и время прилета (Московское) | корректируется с учетом часового пояса аэропорта прибытия |
| 1. классы тарифов — тарифные семейства |  |
| 1. свободные места |  |
| 1. стоимость для каждого тарифного семейства |  |



Если между городами из перечня нет авиасообщения (ни прямого, ни с пересадкой), то в базу должна вноситься метка об отсутствии авиасообщения.

*Создание парсера и сбор данных по 1-му этапу завершены.*

*Парсер – в «Парсер по 1-му этапу.zip»*

*Дамп собранной базы данных- в «Дамп по 1-му этапу.sql»*

**Этап 2. Сбор данных о классе обслуживания, классах тарифов, стоимости тарифов.**

Синим в таблице ниже выделены те данные, которые собираются на этом 2-м этапе

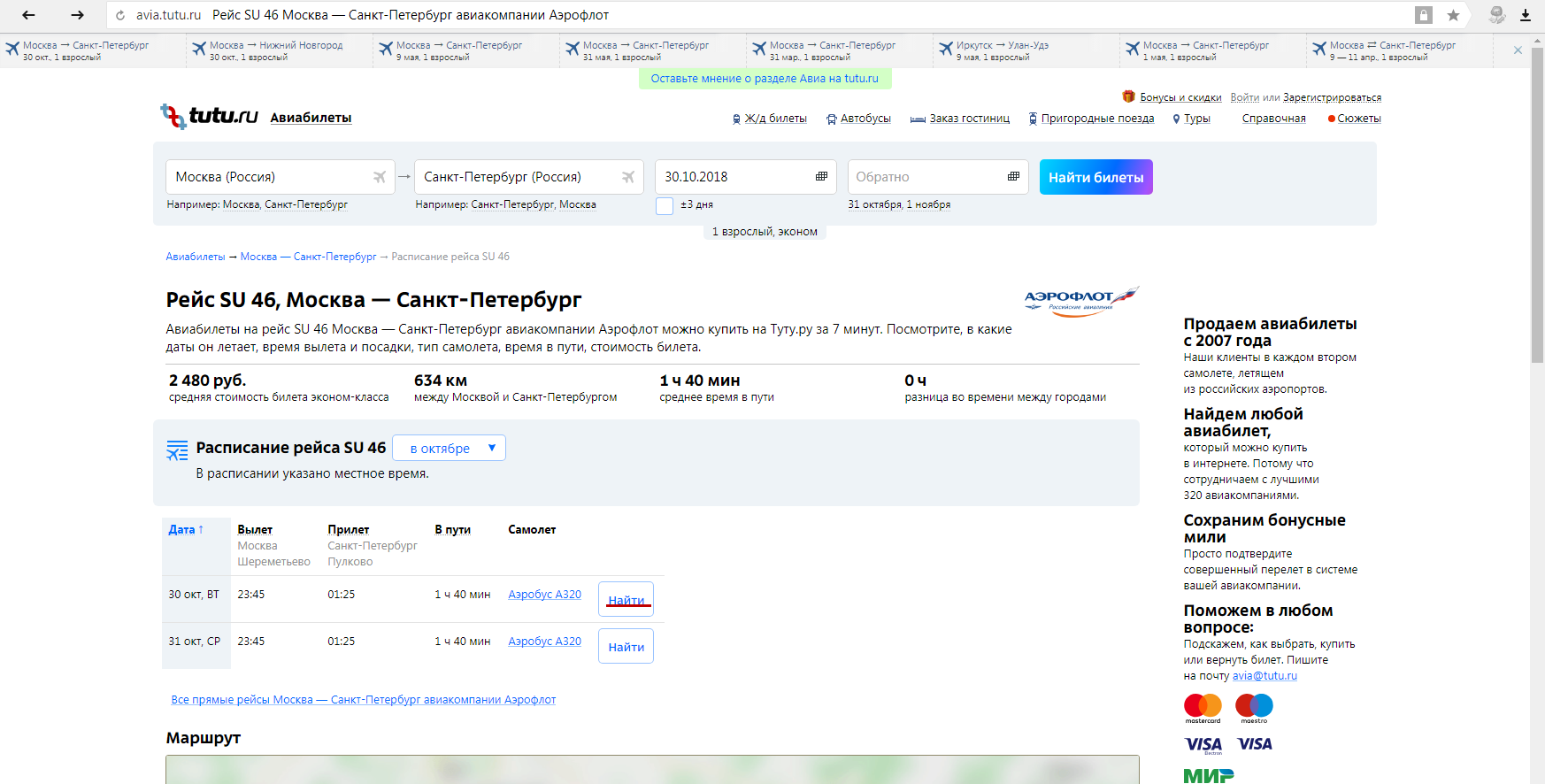
|  |  |
| --- | --- |
| **Данные** | **Примечание** |
| 1. город отправления |  |
| 1. город прибытия |  |
| 1. дата поиска |  |
| 1. дата сбора данных |  |
| 1. дней до вылета |  |
| 1. класс обслуживания (Y – Эконом, С – Бизнес) | парсится |
| 1. авиакомпания |  |
| 1. код авиакомпании |  |
| 1. номер рейса |  |
| 1. индикатор рейса (прямой / с пересадкой, не более 1 пересадки) |  |
| 1. дата и время вылета (местное) |  |
| 1. дата и время вылета (Московское) |  |
| 1. дата и время прилета (местное) |  |
| 1. дата и время прилета (Московское) |  |
| 1. классы тарифов — тарифные семейства | парсится |
| 1. свободные места |  |
| 1. стоимость для каждого тарифного семейства | парсится |

После завершения 1-го этапа сформирована база, содержащая в том числе и номера авиарейсов.

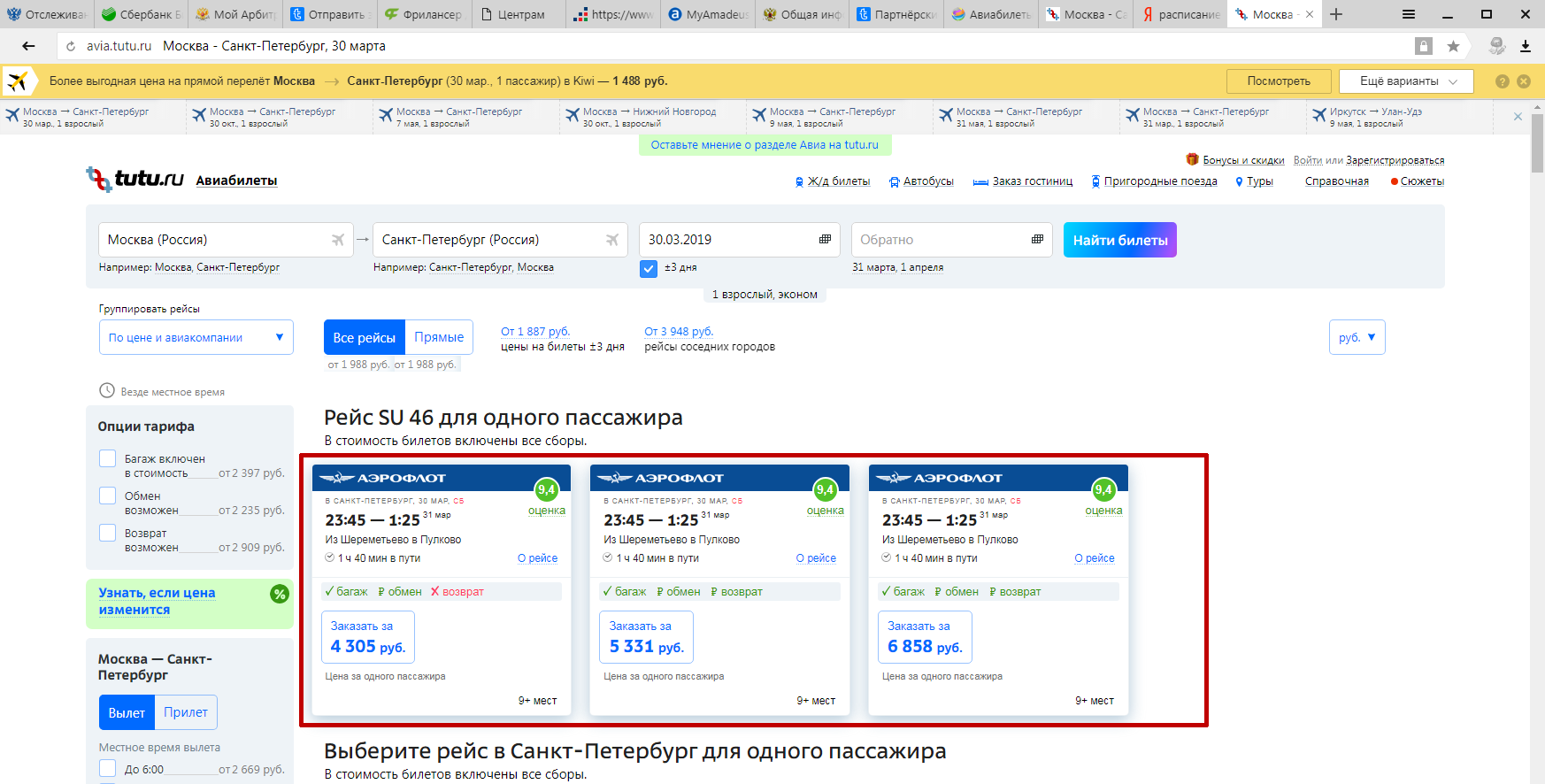
Используя эти номера, будем парсить данные для 2-го этапа.

Данные, например, по рейсу Аэрофлота su 46, находятся на странице <https://avia.tutu.ru/reys/su_46/> (номер рейса добавляется в урл страницы).

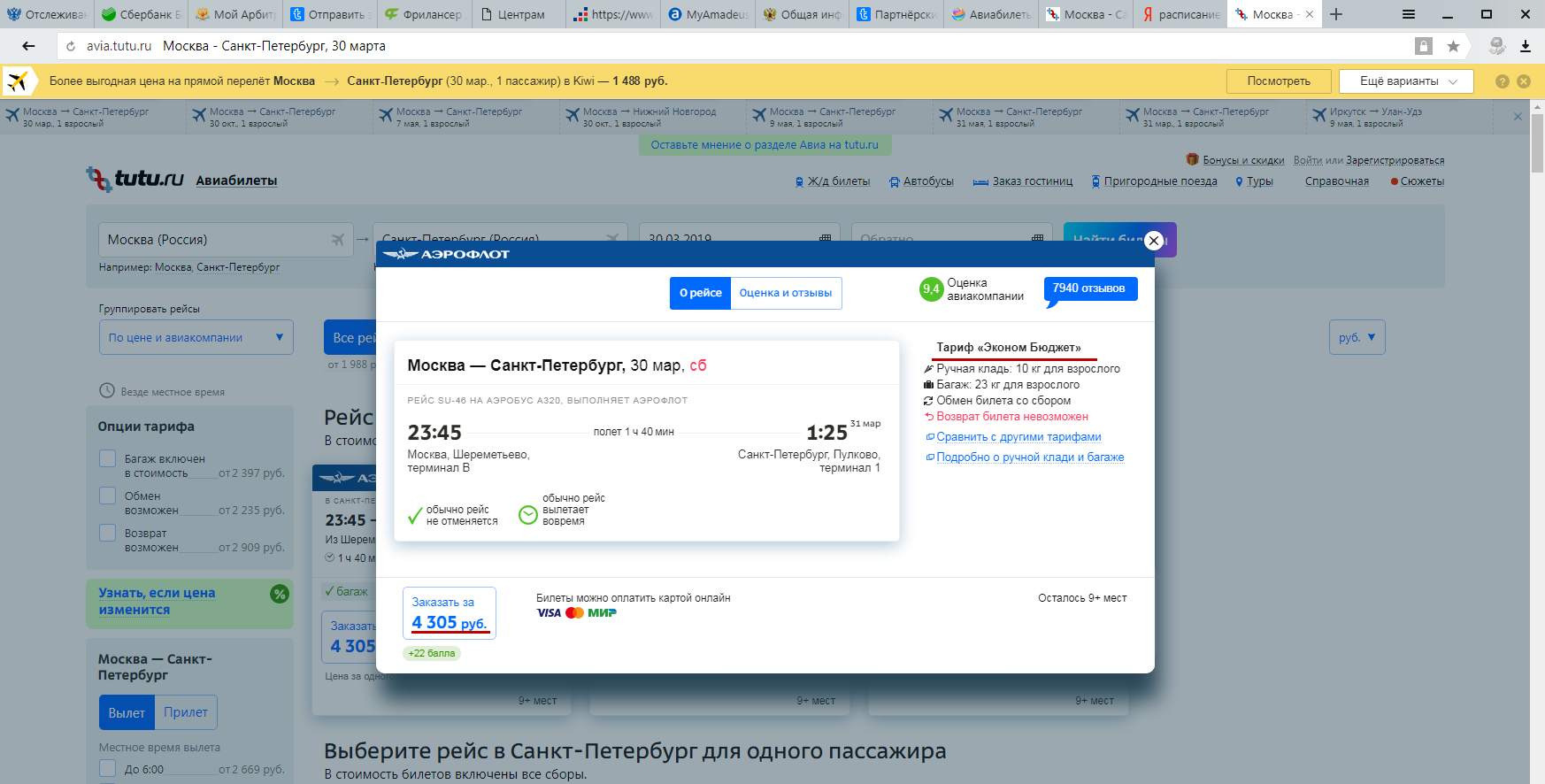
На этой странице нажимаем нижнюю кнопку «Найти»[[1]](#footnote-1):



Собираем данные обо всех билетах этого рейса:



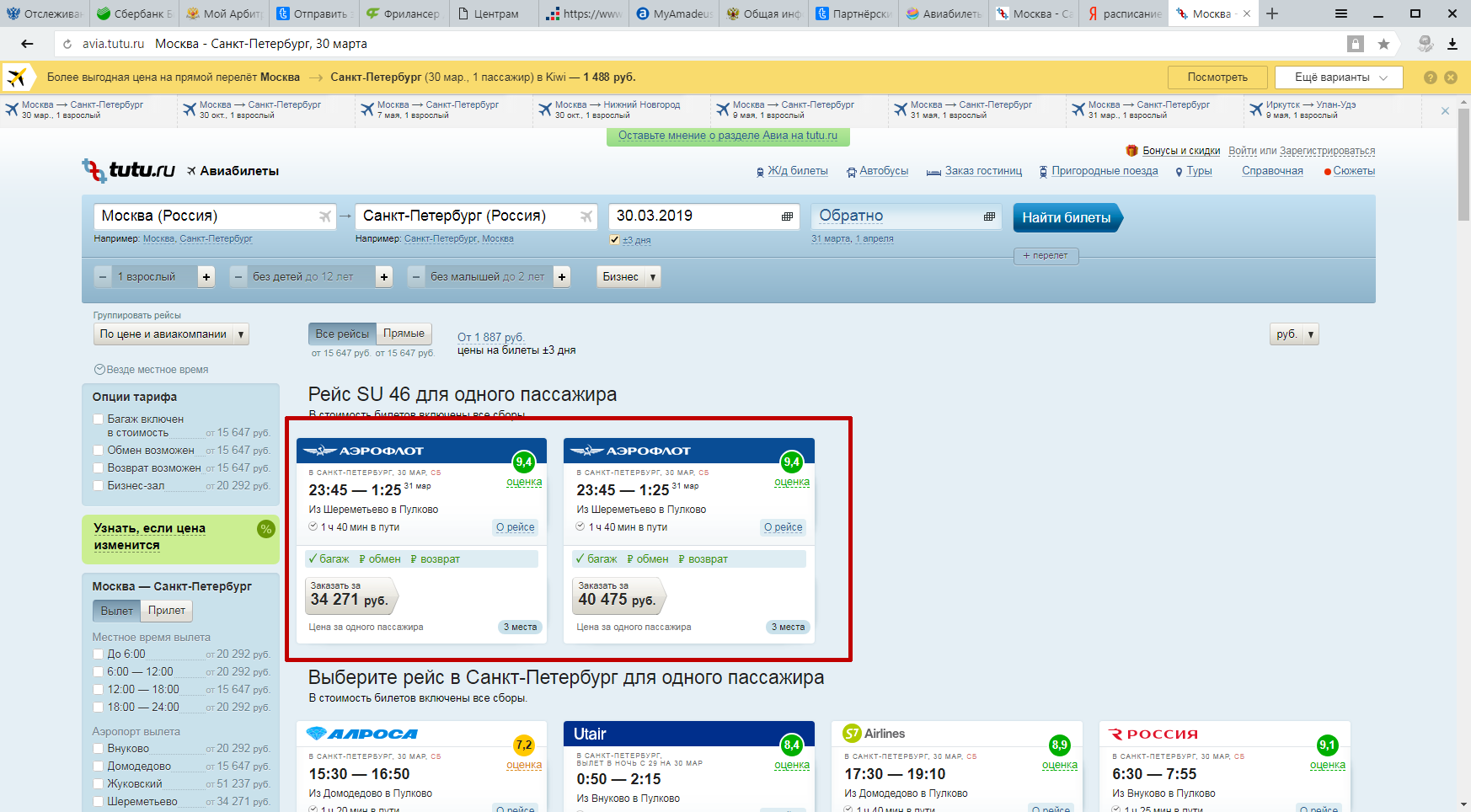
Затем нажимаем в блоке каждого билета ссылку «О рейсе» и берем во всплывшем окне данные о стоимости и названии тарифа:



Меняя в запросе букву Y на С по этому рейсу,

[https://avia.tutu.ru/offers/?passengers=100&flight[0]=SU-46&threeDays=false&class=C&changes=all&route[0]=491-30032019-75](https://avia.tutu.ru/offers/?passengers=100&flight%5b0%5d=SU-46&threeDays=false&class=C&changes=all&route%5b0%5d=491-30032019-75)

попадаем на страницу билетов бизнес-класса:



Собираем информацию о стоимости и названии тарифов бизнес- класса, заносим их в нашу базу.

Далее - переходим к следующему номеру рейса, который попал в нашу базу на 1-м этапе, и повторяем то же самое.

По завершению 2-го этапа у нас должна сформировать база, содержащая условно-постоянные данные. Данные по 1-му и по 2-му этапу заносятся в базу единоразово и в последующем не обновляются.

**3 этап. Сбор информации о наличии билетов**

Фиолетовым отмечены данные, которые будут собираться на этом 3 этапе:

|  |  |
| --- | --- |
| **Данные** | **Примечание** |
| 1. город отправления |  |
| 1. город прибытия |  |
| 1. дата поиска | генерируется автоматически |
| 1. дата сбора данных | дата актуализации информации; может совпадать с датой поиска |
| 1. дней до вылета | рассчитывается автоматически, может принимать значение от 0 до 180 |
| 1. класс обслуживания (Y – Эконом, С – Бизнес) |  |
| 1. авиакомпания |  |
| 1. код авиакомпании |  |
| 1. номер рейса |  |
| 1. индикатор рейса (прямой / с пересадкой, не более 1 пересадки) |  |
| 1. дата и время вылета (местное) |  |
| 1. дата и время вылета (Московское) |  |
| 1. дата и время прилета (местное) |  |
| 1. дата и время прилета (Московское) |  |
| 1. классы тарифов — тарифные семейства |  |
| 1. свободные места | парсится |
| 1. стоимость для каждого тарифного семейства |  |

На этом этапе планируется опрос сайта <https://avia.tutu.ru> по тем направлениям, которые были опознаны на 1-м этапе, как поддерживающее авиационное сообщение.

Подробнее порядок сбора данных по этому этапу будет описан позднее.

1. Нажимать все кнопки «Найти» на странице нет смысла, поскольку нас интересует справочная информация о рейсе, которая одинакова для всех дат, на которые есть свободные билеты. Нижняя же кнопка «Найти» будет вести на самую отдаленную от сегодняшней дату, вероятность наличия свободных билетов на которую является самой высокой и, соответственно, информация о тарифах- самой полной [↑](#footnote-ref-1)