Рекомендательная система онлайн-магазинов

Жестяников А. (Веб-приложение),
Копчев В. (Скрейпинг, БД)

20.12.2021

Описание проекта

Цель, сценарии использования

- Требуется разработать рекомендательную систему
- Сейчас: систему сбора данных о категориях с сайта vons.com
- Три уровня категорий, родительские категории в БД
- Построить БД и веб-сервис
- Внешние данные: vons.com и категории с субкатегориями на их сайте
- Основные сценарии использования: добавление и удаление категорий
- Бейзлайн для более крупного проекта, над которым работаем с ноября

Словарь сущностей

Описание сущностей и связей (исходя из проекта рекомендательной системы)

• **Site - сайты, с которых надо собирать данные**
  ○ id
  ○ Title - название сайта для сбора данных
  ○ url
  ○ Comment - комментарии
  ○ Options - особенности для сбора данных
	• **Experiments_run** - запуски краулера
	  ○ id
	  ○ Date_start - время начала сбора
	  ○ Date_stop - время завершения сбора
	  ○ Options - параметры запуска клаулера
		• **Product Type (бананы)**
		  ○ Id

• **Product (бананы в Магните)**
  ○ Description - Описание (текст в произвольной форме)
  ○ Specification - Спецификация данные с сайта
  ○ nutritional - Ингридиенты
  ○ category_id -> FK supplier_category_id
  ○ app_category_id -> FK app_category
  ○ catalog_product_id - FK к каталогу продуктов
  ○ price
	• **Category** - Категории (таксономия)
	  ○ Id - int
	  ○ Title - название категогии
		• **Supplier_category** - Таксономия продавца
		  ○ supplier_category_id - категории на сайте
		  ○ supplier_parent_category_id
		  ○ app_category_id -> FK category

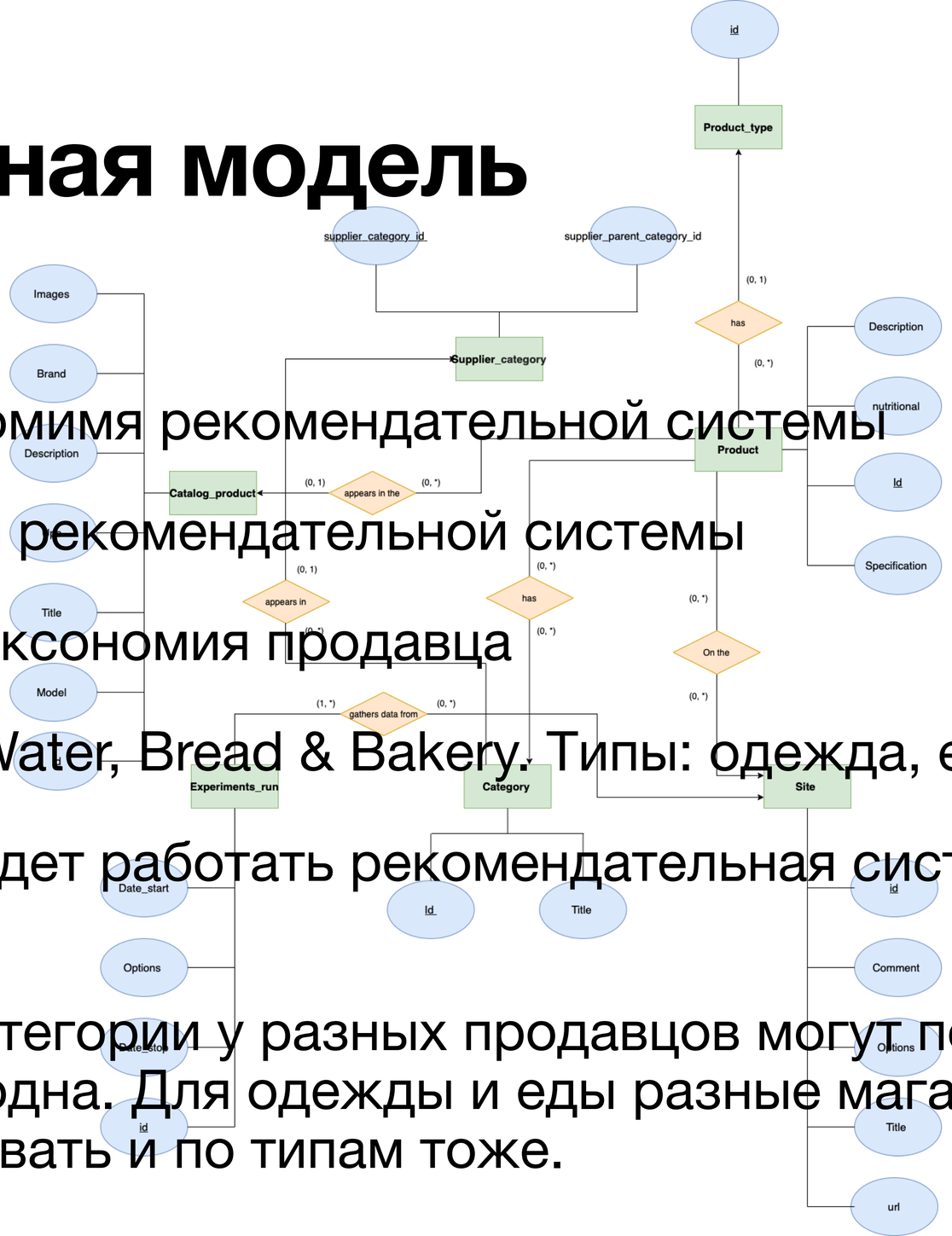
• **Catalog_product** - Каталог продуктов на сайте
  ○ Title - название типа
  ○ Upc - код
  ○ Description - описание
  ○ Brand - Производитель
  ○ Model - модель
  ○ Images - Ссылка на картинки
	• Product appears in the catalog_product: продукт находится в некотором каталоге
	• Product has product_type: продукт имеет некоторый тип в таксономии рекомендательной системы
	• Product on the site: продукт находится на некотором сайте
		• Product has category: продукт относится к категории в таксономии рекомендательной системы
		• сайта
		• Experiment_run gathers data from site: запуск программы собирает данные с некоторого сайта
		• Category appears in suppliers_category: продукт относится к категории в таксономии сайта

Концептуальная модель

Концептуалы

- 7 сущностей
- 6 связей
- Есть связи
- многие-ко-многим

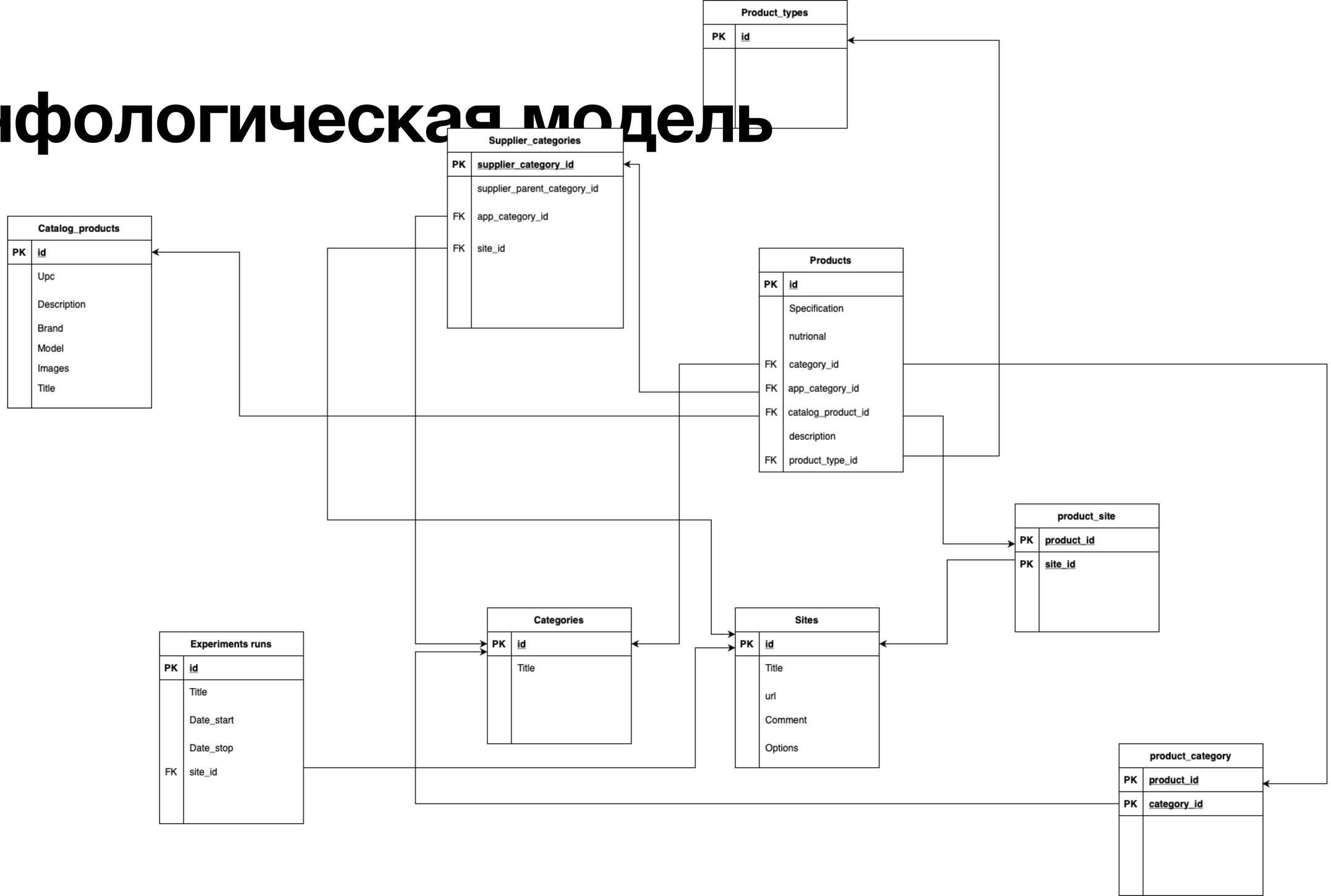
Концептуальная модель



- Product_type: таксономимя рекомендательной системы
- Category: таксономия рекомендательной системы
- Suppliers_category: таксономия продавца
- Категории: Coconut Water, Bread & Bakery. Типы: одежда, еда
- БД объясняет, как будет работать рекомендательная система — сейчас это скорее прообраз
- Зачем разделять? Категории у разных продавцов могут по-разному называться, но суть одна. Для одежды и еды разные магазины, поэтому товары надо сортировать и по типам тоже.

Инфологическая модель

Инфологическая модель



DDL-код

```
CREATE TABLE Product_types (
    id int PRIMARY KEY
);
```

```
CREATE TABLE Catalog_products (
    id int PRIMARY KEY,
    upc text NOT NULL,
    description text NOT NULL,
    brand text NOT NULL,
    model text NOT NULL,
    images text NOT NULL,
    title text NOT NULL
);
```

```
CREATE TABLE Categories (
    id int PRIMARY KEY NOT NULL,
    title text NOT NULL
);
```

```
CREATE TABLE Sites (
    id int PRIMARY KEY,
    title text NOT NULL,
    url text NOT NULL,
    comment text NOT NULL,
    options text NOT NULL
);
```

```
CREATE TABLE Supplier_categories (
    supplier_category_id int PRIMARY KEY NOT NULL,
    supplier_parent_category_id int,
    app_category_id int REFERENCES Categories(Id) ON DELETE CASCADE ON UPDATE RESTRICT,
    site_id int REFERENCES Sites(Id)  ON DELETE CASCADE ON UPDATE RESTRICT
);
```

```
CREATE TABLE Products (
    id int PRIMARY KEY,
    specification text NOT NULL,
    nutrional text NOT NULL,
    description text NOT NULL,
    price int CHECK (PRICE > 0),
    category_id int REFERENCES Categories(Id),
    app_category_id int REFERENCES Supplier_categories(supplier_category_id) ON DELETE CASCADE ON UPDATE RESTRICT,
    catalog_product_id int REFERENCES Catalog_products(Id) ON DELETE CASCADE ON UPDATE RESTRICT,
    product_type_id int REFERENCES Product_types(Id) ON DELETE CASCADE ON UPDATE RESTRICT
);
```

```
CREATE TABLE Experiments_runs (
    id int PRIMARY KEY,
    title text NOT NULL,
    date_start text,
    date_stop text,
    site_id int NOT NULL
);
```

```
CREATE TABLE Product_site (
    product_id int REFERENCES Sites(Id),
    site_id int,
    PRIMARY KEY (product_id, site_id)
);
```

```
CREATE TABLE Product_category (
    product_id int,
    category_id int REFERENCES Categories(Id) ON DELETE CASCADE ON UPDATE RESTRICT,
    primary key (product_id, category_id)
);
```

DDL-код

```
INSERT INTO Categories VALUES (0, 'Baby Care');
INSERT INTO Categories VALUES (661, 'Beverages');
INSERT INTO Categories VALUES (2, 'Bread & Bakery');
INSERT INTO Categories VALUES (3, 'Breakfast & Cereal');
INSERT INTO Categories VALUES (4, 'Canned Goods & Soups');
INSERT INTO Categories VALUES (5, 'Condiments, Spice & Bake');
INSERT INTO Categories VALUES (6, 'Cookies, Snacks & Candy');
INSERT INTO Categories VALUES (7, 'Dairy, Eggs & Cheese');
INSERT INTO Categories VALUES (8, 'Deli');
INSERT INTO Categories VALUES (102, 'Flowers');
INSERT INTO Categories VALUES (10, 'Frozen Foods');
INSERT INTO Categories VALUES (11, 'Fruits & Vegetables');
INSERT INTO Categories VALUES (12, 'Grains, Pasta & Sides');
INSERT INTO Categories VALUES (13, 'International Cuisine');
INSERT INTO Categories VALUES (14, 'Meat & Seafood');
INSERT INTO Categories VALUES (15, 'Paper, Cleaning & Home');
INSERT INTO Categories VALUES (16, 'Personal Care & Health');
```

DDL-код

- DDL базы написан вручную с использованием СУБД PostgreSQL, диалект PL/pgSQL
- DDL значений был написан с помощью Python-скрипта
- Был написан Python-скрипт для скрейпинга, а также Python-скрипт, который по собранным данным создает DDL-код, вставляющий эти данные в БД.
- Связи между id, поэтому везде ограничения целостности одинаковые: ON DELETE CASCADE ON UPDATE RESTRICT

Словарь данных

- Составлю таблицу для
- Случаев, где информация
- О данных не очевидна
- Из словаря сущностей

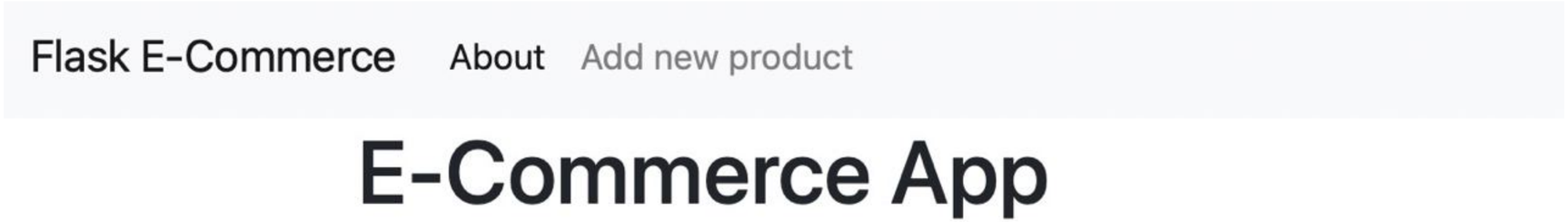
Источник данных	Идентификатор данных	Назначение данных	Диапазон значений	Тип данных
vons.com	Product.Description	Описание продукта на сайте (К нему будут применены алгоритмы NLP)	-	Text
Администратор БД	Category.Title	Различать категории не только по id	-	Text
Администратор БД	Experiments_run.Options	Параметры запуска программы для сбора данных	-	Text
Администратор БД	Site.options	С какими параметрами запускать сбор данных с этого сайта	-	Text
vons.com	Product.price	Цена товара на сайте	> 0	Int

Гайд по XPath

- В планах использовать XPath вместо BeautifulSoup. Для этого был написан небольшой гайд по данному языку запросов к XML-файлу в отчете

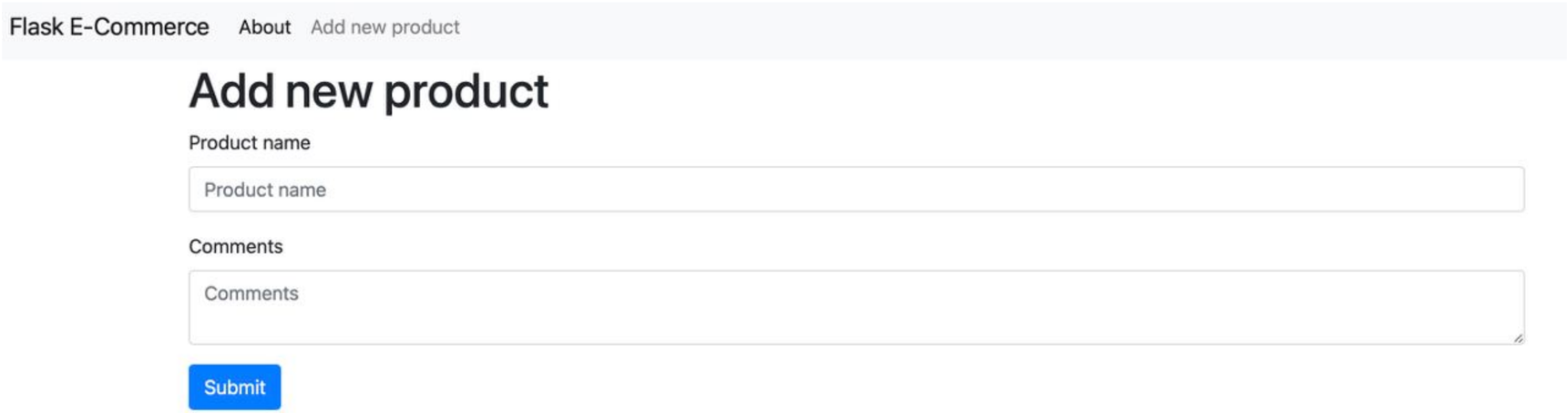
Клиентское приложение

Интерфейс



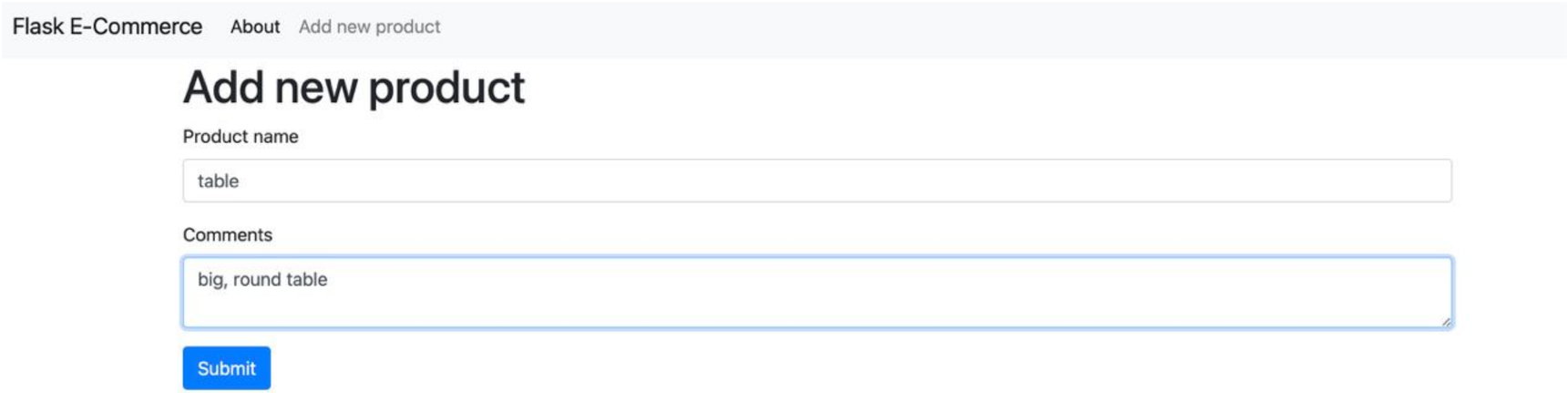
Клиентское приложение

Изменение данных



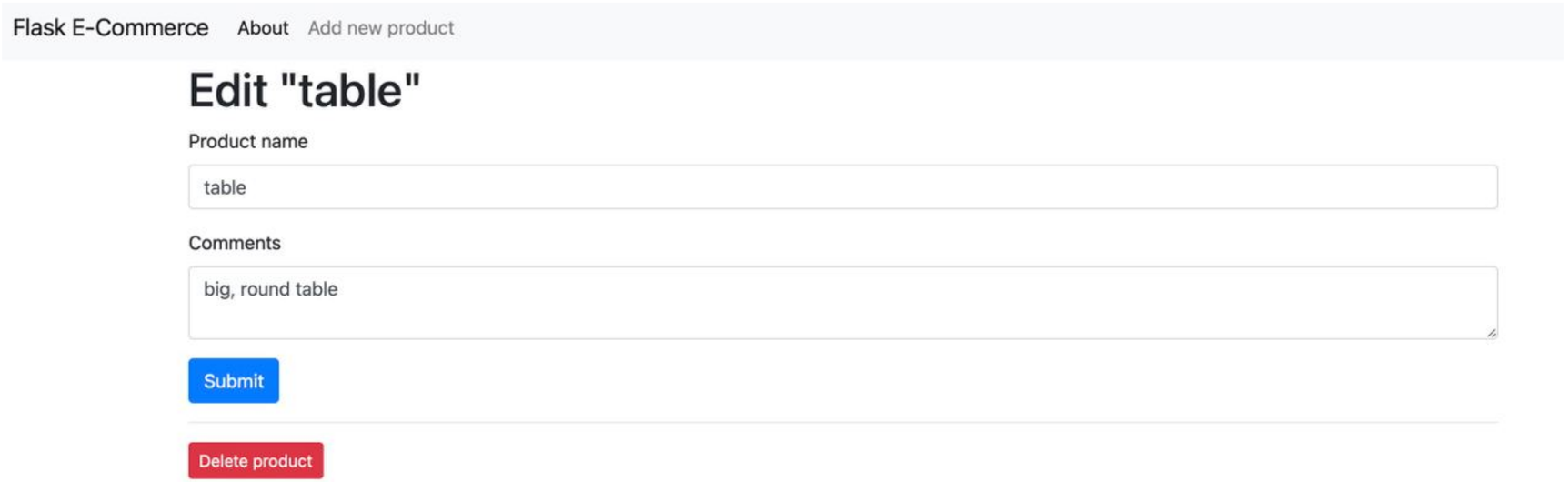
Клиентское приложение

Изменение данных



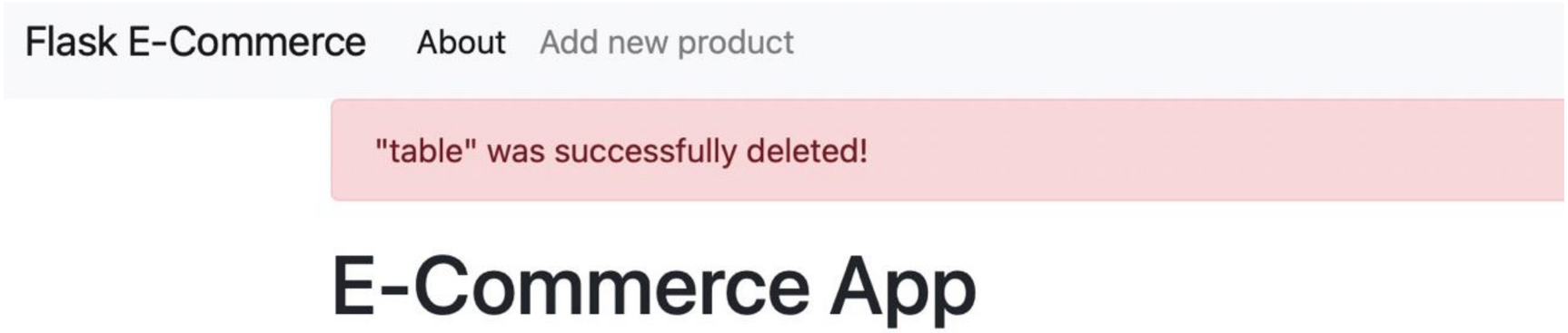
Клиентское приложение

Изменение данных



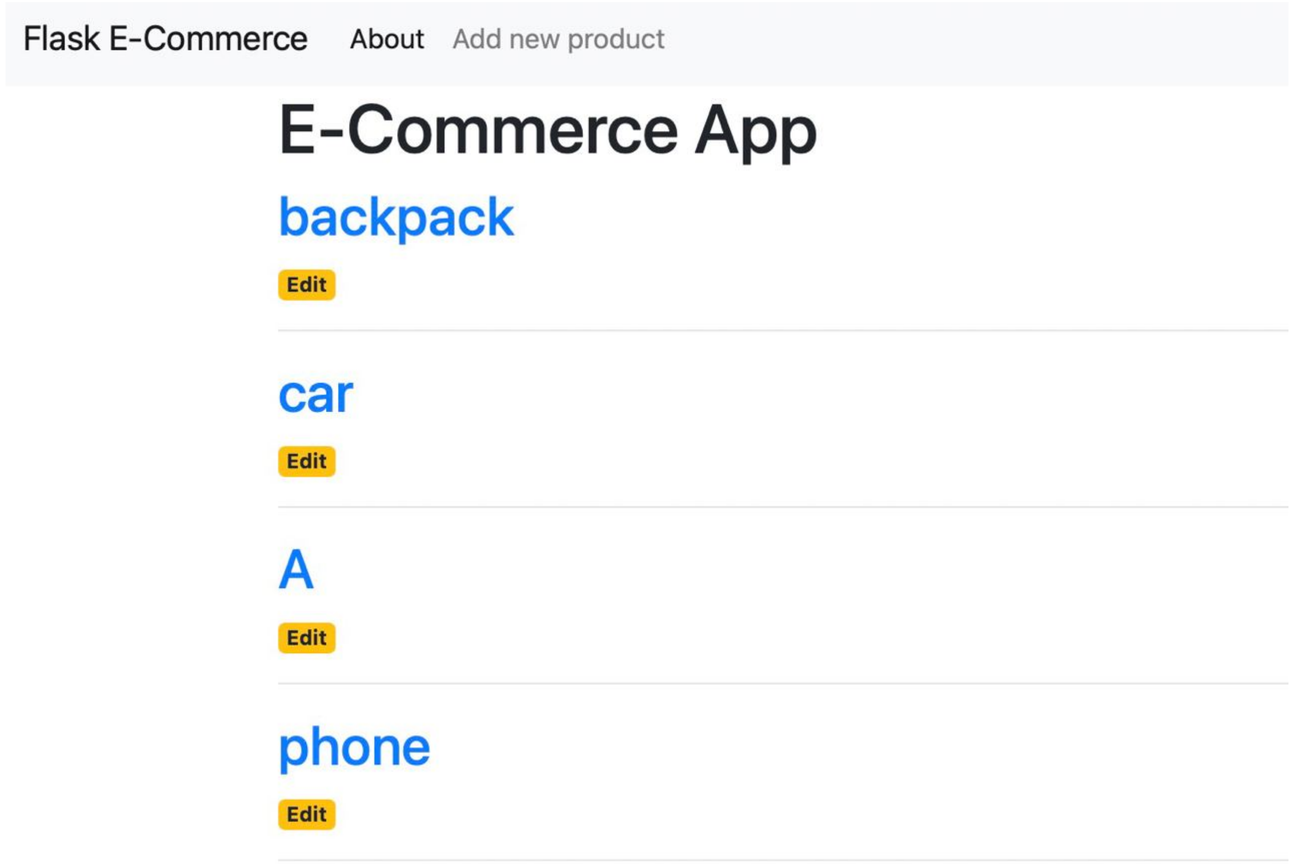
Клиентское приложение

Изменение данных



Клиентское приложение

Просмотр данных



Заключение

- На этом работа не оканчивается — это только бейзлайн
- Командный проект. Мои задачи: собрать категории с других сайтов, собрать данные о товарах, произвести первичную обработку собранных данных, перевести код с BeautifulSoup на XPath, привести функции к виду, соответствующему шаблону класса рекомендательной системы
- Задачи других: произвести обработку данных с помощью алгоритмов NLP, разработать рекомендательную систему. Андрей продолжит разработку БД.