Анализ рыночной корзины

Анализ рыночной корзины(market basket analysis) -это семейство методов в BI для извлечения ассоциаций между событиями из данных. Анализ рыночной корзины -это конкретное приложение анализа сродства, в котором заказы клиентов (например, рыночные корзины супермаркетов) анализируются с целью выявления общих моделей покупок, то есть продуктов, которые обычно покупаются вместе.

Эти шаблоны затем можно использовать для принятия бизнес решений.

Например, если 70% людей, которые покупают макароны, также покупают томатный соус, может быть хорошей идеей разместить их рядом друг с другом, чтобы повысить вероятность того, что покупательмакаронных изделий также купит томатный соус. Или, если 99% людей, которые покупают утренние хлопья, также покупают молоко, они могут быть размещены дальше друг от друга, чтобы покупатели, которые хотят купить и то и другое, прошли длинный путь через супермаркет и увидели множество других продуктов, которые они, возможно, решат купить. На практике интересны даже более «слабые» ассоциации (например, 2%) из-за высокой вариативности покупательских привычек.

В этом упражнении вы реализуете простую версию анализа рыночной корзины и получите информацию из заказов клиентов. В частности, мы будем искать покупательские закономерности между парами товаров.

Данные

Набор данных предоставляется в одном текстовом файле, где:

●

Каждая строка представляет собой отдельный заказ клиента (для

удобства чтения заказы разделены двумя строками).

●

В каждом заказе содержится список продуктов, приобретенных в

этом заказе, разделенных символом «@@@».

Например, вот строка из набора данных:

Bag of Organic Bananas@@@Celery Sticks@@@Sparkl

ing

Water@@@Organic Strawberries

В этой строке описывается заказ со следующими четырьмя товарами:

● Bag of Organic Bananas

● Celery Sticks

● Sparkling Water

● Organic Strawberries

Требование к решению

В этом упражнении мы хотим выявить покупательские ассоциации между парами продуктов, например

-

Pears => Clementines.

Мы будем использовать два критерия для определения интересных ассоциаций. Для двух продуктов p1 и p2 определим:

1.

Уверенность

: процент заказов, содержащих оба продукта, от заказов, содержащих p1. Например, если 200 заказов содержали

apples, а 30 из них содержали апельсины, то уверенность

ассоциации apples => oranges составляет 15% (30/200).

2.

Поддержка:

Абсолютное количество заказов, содержащих оба

продукта; в приведенном выше примере поддержка равна 30.

Почему мы используем два критерия?

●

Редкие продукты могут иметь высокую степень достоверности, но

на самом деле это не указывает на значительную связь.

Например, если было только 2 заказа, содержащих «Aunt Esme’s

Special Apple Sauce», и один из них содержал Tomatoes,

достоверность составила бы 50% (что считается очень высоким).

●

Очень распространенные продукты могут иметь высокую поддержку из-за своей популярности, но, опять же, это не обязательно указывает на значительную связь. Например, если есть 1 миллион заказов, содержащих «bottled water», и 2 миллиона заказов, содержащих «cucumbers», у нас может быть 10 000 заказов, содержащих и то, и другое, высокое абсолютное число, но в контексте это не показывает сильной взаимосвязи.

Ваше задание

-

написать функцию, которая при минимальной уверенности

и поддержке печатает все парные ассоциации, которые соответствуют критериям, т.е. имеют больше, чем минимальную уверенность и поддержку.

Инструкции и руководство

1.

Сначала напишите код для чтения и анализа данных заказа

-разделение заказов по строкам и продуктов по «@@@».

2.

Затем напишите код для расчета поддержки и уверенности для

каждой пары продуктов.

a.

Создайте два словаря -один для подсчетакаждого продукта и один для подсчета каждой пары продуктов.

b.

Переберите все заказы и обновите два словаря для каждого заказа по мере необходимости.

c.

Словарь подсчета пар содержит поддержку каждой пары элементов.

d.

Используйте два словаря, чтобы рассчитать уверенность для

каждой пары элементов.

e.

Обратите внимание, что поддержка симметрична, а уверенность

-нет. Например, если есть 100 заказов с “bananas”, 200 заказов с «peanut butter» и 30 заказов с обоими, тогда уверенность bananas => peanut butter составляет 3

0%, тогда как уверенность peanut butter => bananas -15 %.

f.

Помните, что нельзя рассчитывать уверенность продукта в

себе (она всегда будет 100%, но это не интересно).

g.

Наконец, учитывая минимальные требования к уверенности и

поддержке, переберите все пары продуктов и напечатайте

пары, соответствующие критериям, в следующем формате:

Chunky Guacamole => Organic Tortilla Chips (5.37%

confidence), 38 support

Удачи

Для следующих упражнений загрузите набор данных фильма по этой ссылке  
(откройте ссылку и используйте «Сохранить как», чтобы загрузить файл).  
2. Загрузите набор данных фильмов в pandas.  
3. Перечислите все столбцы набора данных и изучите их типы. Изучите статистику  
по различным областям. Опишите, какие данные у нас есть.  
4. Сколько всего фильмов в наборе данных?  
5. Сколько фильмов содержится в наборе данных за каждый год?  
6. Покажите подробную информацию о наименее и наиболее прибыльных  
фильмах в наборе данных.  
7. Значение «Жанр» временами кажется непоследовательным; попробуйте найти  
эти несоответствия и исправить их.  
8. Сохраните (в новый файл CSV) 10 лучших комедий по количеству зрителей;  
покажите только название фильма, год и студию.  
9. Используйте pip для установки двух библиотек: lxml, MySQL-connector-python.

