**Пилотный проект по МАТЧИНГУ**

6 марта 2022

**Задача:** провести сопоставление номенклатуры товаров клиента («dogeat») и конкурента («pethop»), а именно: установить соответствие 13 739 SKU позиций клиента и 11 263 SKU позиций конкурента,

*где 13 739 SKU - это количество номенклатур клиента, которые не сопоставлены на текущий момент времени с номенклатурой конкурента (14710 SKU минус 971 SKU); пояснения цифр см. ниже;*

*11 263 SKU - это количество номенклатур конкурента, которые не сопоставлены на текущий момент времени с номенклатурой клиента (12234 SKU минус 971 SKU); пояснения цифр см. ниже;*

**Структура выходной информации (**приблизительная, будет уточняться в процессе работы над проектом):

1. Наименование номенклатуры клиента;
2. Наименование номеклатуры конкурента (ТОП1);
3. Показатель (метрика, определенная моделью) релевантности номенклатуры клиента (ТОП1) и номенклатуры конкурента;
4. Повторение пп. 2-3 в зависимости от заданного параметра – количество выдачи сопоставленных номенклатур (ТОП\_к).

**Критерий качества:** mean average precision [map@1 >= 0.6](mailto:map@k%3e=0.6) (необходимо уточниться будет, так как этот показатель не учитывает количество предложенных сопоставленных позиций; например, можно сопоставить 100 номенклатур клиента из 13739 и получить map=0.9; или можно сопоставить 10000 позиций и получить map=0.5; мое понимание, последний вариант лучше).

**Алгоритм проверки качества:** случайно сэмплируем из предложенных моделью сопоставлений N примеров (параметр N будет установлен в дальнейшем) и проверяем («ручная проверка») соответствие (соответствует/не соответствует) и считаем precision на этой сэмплированной подвыборке сопоставлений.

**Исходные данные:**

1. База данных номенклатуры товаров клиента (файл 'client\_data\_dogeat.xlsx')
   1. Количество строк в базе данных: 16294;
   2. Уникальных номенклатур: 14710 SKU;
   3. Список имен столбцов базы данных: ['dt', 'offer\_id', 'vendorCode', 'barcode', 'url', 'price', 'count' , 'picture', 'vendor', 'name', 'weight', 'description']
2. База данных номенклатуры товаров конкурента (файл ' shop\_data\_petshop.xlsx)
   1. Количество строк в базе данных: 53183;
   2. Уникальных номенклатур: 12234 SKU;
   3. Список имен столбцов базы данных: ['offer\_article', 'offer\_id', 'offer\_name', 'offer\_price', 'offer\_vendor', 'product\_breadcrumbs', 'product\_name', 'product\_url', 'url\_h1', 'url\_type']
3. База данных («размеченные данные») сопоставленной (частично!!!) номенклатуры клиента с конкурентами («4lapy», «magizzo», «petshop»)
   1. Количество строк в базе данных: 52028;
   2. Уникальных номенклатур клиента: 7535 SKU (51% от общего количества номенклатур клиента);
   3. Список имен столбцов базы данных: ['matching\_type', 'competitor', 'competitor\_price', 'competitor\_product\_url', 'competitor\_offer\_id', 'competitor\_offer\_article', 'competitor\_vendor', 'competitor\_offer\_full\_name', 'competitor\_offer\_weight', 'client\_offer\_name', 'client\_offer\_vendor', 'client\_offer\_weight', ‘client \_offer\_price', 'client\_offer\_id', 'client\_offer\_article', 'client\_offer\_barcode']
   4. Информация о сопоставленных номенклатурах:

|  |  |  |
| --- | --- | --- |
| Competitor | **competitor\_offer\_full\_name**  (SKU), штук | **client\_offer\_name**  (SKU), штук |
| 4lapy | 3 871 | 3 871 |
| magizoo | 11 548 | 11 548 |
| petshop | 971 | 971 |
| **Итого** | **16 390** | **16 390** |

* 1. Информация о количестве вариантов сопоставления для каждой позиции номенклатуры

|  |  |
| --- | --- |
| Сколько вариантов сопоставлено номенклатуры клиенты (т.е. с каким количеством номенклатур конкурентов сопоставлена каждая номенклатура клиента) | Количество SKU, штук |
| 1 | 2 373 |
| 2 | 2 720 |
| 3 | 1 641 |
| 4 | 474 |
| 5 | 223 |
| 6 | 89 |
| 7 | 13 |
| 8 | 1 |
| 10 | 1 |
| **ИТОГО** | **7 535** |

*Примечание: 57 позиций клиента в таблице сопоставления имеют значение NaN (они удалены из анализа)*