Výzva 1 – párování produktu

ZADÁNÍ

Background:

Nový manažer v nejmenované firmě se sportovním zbožím by rád do zkostnatělé struktury zavedl prvky moderní práce s daty. Rád by věděl, jak si se svými výrobky stojí oproti konkurenci, a proto si objednal službu, která mu dodá stáhnutá data o výrobcích z cenového srovnávače. Tato data (*složka scraped\_data*) předal svým kolegům z analytického oddělení a těšil se, jak bude moci brzy vidět, jak si jeho výrobky stojí oproti konkurenci. Zejména aktuální kolekce z několika posledních let, ta ho zajímá nejvíce.

Bohužel narazil na problém – analytici mu dodaná data vrátili s tím, že je nedokáží napárovat na produkty. K obdrženým záznamům potřebují vědět, k jakému produktovému kódu záznam patří. Poslali manažerovi nově vytvořený produktový katalog (*Excelová tabulka Produkty*), ať jim k jednotlivým záznamům dodá číslo produktu. Prý ještě není zcela kompletní, tak pokud by nemohl použít tu, ať použije zastaralý katalog artiklů (*Excelová tabulka Artikly)*, oni je umí propojit.  
  
Manažer rychle zjistil, že dělat toto mapování ručně je nad jeho síly. Spousta záznamů vůbec neodpovídala žádnému zboží z katalogu, což tak nějak očekával. Po chvíli mravenčí práce se mu některé napárovat podařilo, např. záznam z tabulky dataset\_sport-product označený product\_id n449935180 odpovídá nejspíše produktu s produktovým kódem 7133024274. Ale kolik takových dvojic najde?

FAQ:

1. Co máme dělat?

* Namapovat co nejvíce výrobků ve složce scraped\_data na interní kódy.

1. Jak?

* To je na vás. Jsme na datovém hackatonu, takže upřednostňujeme řešení, kde předvedete vaši schopnost pracovat s daty.

1. Jak má vypadat výstup?

* Výstup má tři části:
  + Tabulka, ze které bude poznat, které výrobky (podle product\_id, anebo třeba podle url) jste přiřadili ke kterému internímu produktu (kódu výrobku)
  + Technické řešení. Klient by si rád ušetřil práci jednou provždy, takže pokud dostane nástroj, který podobné úlohy vyřeší v budoucnu za něj, bude vděčný.
  + Vysvětlení klientovi, jak technické řešení funguje.

1. Dostaneme k datům nějaké další informace?

* Můžete se zeptat, ale data jsou buď samovysvětlující (tabulky Artikly a Produkty) nebo o nich klient sám nic neví (scraped\_data).

1. Je možné, že jsou v datech chyby?

* Je to zcela jisté.

1. Je možné, že jsou v datech položky, které s těmi v katalogu vůbec nesouvisí, případně obráceně?

* Je to zcela jisté.

1. Je možné, že je v zadání nějaká chyba či opomenutí, které se během dne upřesní?

* Je to dost pravděpodobné.

1. Dostaneme nějaká testovací data, na kterých si můžeme náš postup ověřit?

* Ne.

1. Jak se práce hodnotí?
   1. **20 % celkového hodnocení tvoří počet namapovaných produktů**Nejde jen o čistý počet, ale i o strukturu, rozmanitost atd. podle preferencí klienta. Pokud v katalogu přesně stejný produkt není, „velmi podobný“ je lepší než žádný.
   2. **40 % celkového hodnocení tvoří technické řešení**

Kvalitní technické řešení je funkční, vhodné, spolehlivé, elegantní, znovu použitelné, robustní. Nebojte se pochlubit zajímavým uvažováním či postupem, i když třeba v tomto konkrétním případě nepovede ke skvělým výsledkům.

* 1. **30 % celkového hodnocení tvoří komunikace s klientem**

Jedná se zejména o závěrečné vysvětlení, jak vaše řešení vypadá. Forma je na vás, můžete zvolit ústní vysvětlení, písemnou dokumentaci, Powerpointovou prezentaci, Tableau vizuál, muzikální vystoupení, maňáskové divadlo – cokoli vám přijde vhodné. (Hodnotíme ovšem efektivitu předání informace, ne originalitu)

Zároveň je součástí hodnocení i průběžná komunikace – informování o postupu, kladení inteligentních dotazů, respektování faktu, že ne vždy bude mít klient čas zrovna na vás atd.

* 1. **10 % celkového hodnocení tvoří použití preferovaných nástrojů**

Klientův technologický stack tvoří Keboola, Tableau, PowerBI, Python a SQL, preferuje tedy jejich použití.

1. Could you describe all of this in English?

* I will when someone asks me to.