

Kopija __Data Quality Analysis

Šiame dokumente apžvelgiami gauti paslaugų ir paieškų failų duomenys.

Paslaugos

Failas "paslaugos_01_22-Formatuota.xlsx", parsisiųsta 2025-02-04.

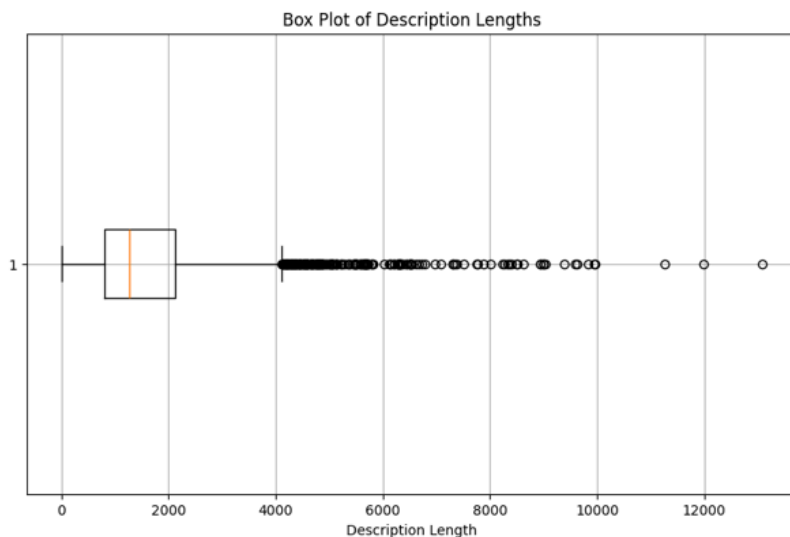
Faile iš viso pateikta 3642 paslaugų. 1413 iš jų turi unikalius pavadinimus, o tai reiškia, kad didelė dalis paslaugų turi tą patį pavadinimą, tik skiriasi teikėjas, aprašymas ar kt.

Toliau trumpai apžvelgiama paslaugų aprašymų ir pavadinimų statistika, kadangi šie duomenys svarbūs korektiškui semantinės paieškos veikimui.

Bendra paslaugų aprašymų statistika:

- 24 paslaugos neturi aprašymo
- Ilgiausio aprašymo ilgis: 13092 simboliai
- Ilgių vidurkis: 1676 simboliai
- Ilgių mediana: 1258 simboliai

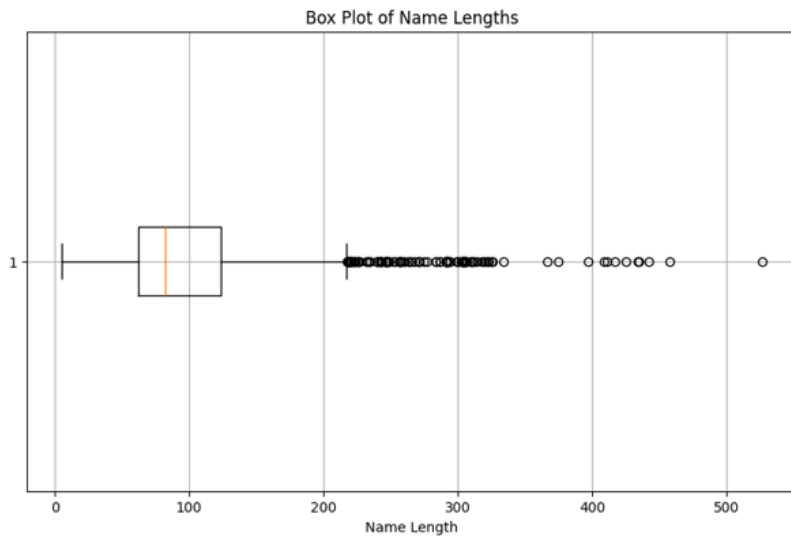
Aprašymų ilgių Box Plot:



Bendra paslaugų pavadinimų statistika:

- Trumpiausio pavadinimo ilgis: 5 simboliai
- Ilgiausio pavadinimo ilgis: 527 simboliai
- Ilgių vidurkis: 100 simbolių
- Ilgių mediana: 82 simboliai

Pavadinimų ilgių Box Plot:



Kadangi tiek skirtingi paslaugų aprašymai, tiek pavadinimai tarp skirtingų paslaugų skiriasi, jie nėra vienodai tinkami naudoti semantinei paieškai. Pvz., jei aprašymas/pavadinimas yra per trumpas arba per ilgas, DI modelis negali užfiksuoti reikiamos informacijos apie paslaugą, todėl tos paslaugos paieška neveiks efektyviai.

Vienas iš šios problemos sprendimų yra LLM paduodant paslaugos pavadinimą ir aprašymą sugeneruoti reikiamo ilgio sintetinius paslaugų aprašymus, tinkamus semantinei paieškai.

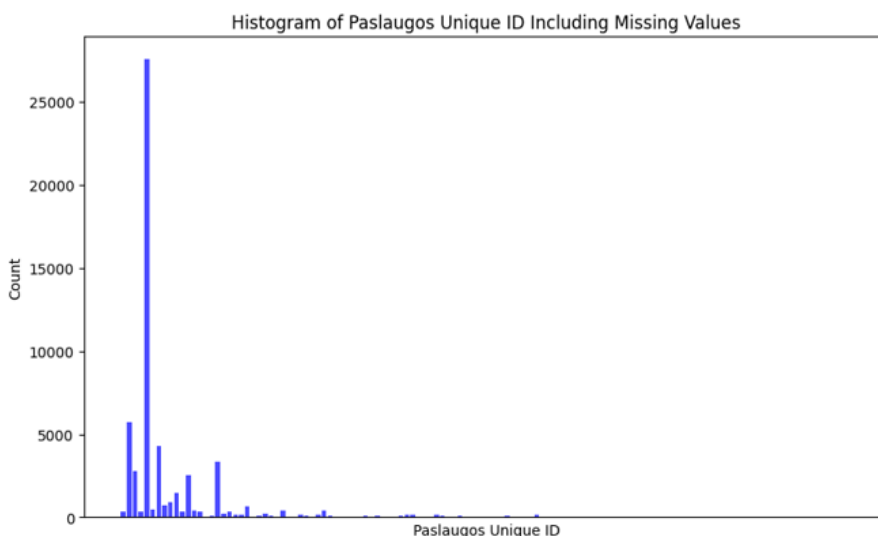
Paieškos

Failas "2024-paieskos-uzsakymai.csv", parsisiųsta 2025-02-04.

Failė iš viso pateikta 57 287 paieškų. Jos turi labai mažą pasiskirstymą: grupuojant pagal "paslaugos id" arba "paslaugos unique id" stulpelius paaiškėjo, kad tik 124 paslaugos turi bent vieną paieškos pavyzdį, t. y. tik 3,4 % visų paslaugų turi bent vieną paieškos pavyzdį. Be to, net paslaugos, turinčios bent po vieną paieškos paieškos pavyzdį, yra netolygiai pasiskirsčiusios.

Taip pat pastebėta, kad didelis kiekis paieškų yra ilgos ir tikėtina kopijuotos, pvz. "Gyvenamosios vietos deklaravimas, deklaravimo duomenų taisymas, keitimas ar naikinimas".

Paslaugų (kurios turi bent vieną priskirtą paiešką) ir joms priskirtų paieškų histograma:



Tai reiškia, kad nėra pakankamo kiekio kokybiškų duomenų apmokyti ir testuoti modelio gebėjimą rasti visas paslaugas pagal pateiktą paieškos frazę.

Vienas iš šios problemos sprendimų yra LLM paduodant paslaugos pavadinimą ir pateiktus raktažodžius sugeneruoti tam tikrą kiekį galimų paieškos frazių kiekvienai paslaugai ir jomis atlikti modelio mokymą bei testavimą.