Tabellen

query-product.csv

id	product_id	product_title	search term	relevance
1	100001	"Hampton Fan"	"ceiling fan"	3.0
2	100002	"3/8 in. drill Bit"	"drill bit"	2.33
3	100003	"AA Batteries"	"aa battery"	3.0

product_description

product_uid	product description
100001	"ceiling fan is great"
100002	"durable drill bit"
100003	"24-puch of AA alkaline"

Preprocessing:

- Merge tabellen

⇓

merged_df

id	uid	title	term	relevance	description
1	100001	"Hampton Fan"	"ceiling fan"	3.0	"ceiling fan is great"
2	100002	"3/8 in drill bit"	"drill bit"	2.33	"durable drill"

Tekstuele preprocessing:

- Hoofdletters weg
- Leestekens weg
- Array maken van woorden
- stopwoorden weg
- getallen, operatoren
- versimpelde woorden

⇒ nieuwe kolommen tokens:

search	title	desc
[ceil, fan]	[Hampton Fan]	[ceil, fan, great]
[drill, bit]	[drill bit]	[durable, drill, bit]

Tekstuele preprocessing:

- Hoofdletters weg
- leestekens weg
- Array maken van woorden
- stopwoorden weg
- getallen, operatoren
- versimpelde woorden

⇒

nieuwe kolommen tokens:

search	title	desc
[ceil, fan]	[Hampton Fan]	[ceil, fan, great]
[drill, bit]	[drill bit]	[durable, drill, bit]

word embeddings

woorden representeren als vectoren:

bijv "fan" → [0.13, -0.24, 0.76, ..., 0.02]

voor elk tekstveld, krijg je 1 vector van vaste lengte

search_vector = [0.25, 0.35, 0.45, ..., 0.02]
title_vector = [0.20, 0.40, 0.50, ..., -0.07]
desc. vector = [0.22, 0.37, 0.44, ..., 0.00]

- Features bedenken:
- Opsplitsen in train en test sets
- Regressie & classificatie