DAR sessie 3 - TF, IDF, vector-space-model: exercises

30 april 2025

Inverse document frequency: idf(t) = log(N/df(t))

Term-frequency: w(t) = 1 + log(tf(t,d))

Weighting scheme: lnc for doc (l: log tf; n: doe niets met idf; c: cosine)

Weighting scheme: ltc for query (l: log tf; t: term weighting via idf; c: cosine)

Document: *car insurance auto insurance*
Query: *best car insurance*

Term	Query						Document				Prod
	tf-raw	tf-wt	df	idf	wt	n'lize	tf-raw	tf-wt	wt	n'lize	
auto	0	0	5000	2.3	0	0	1	1	1	0.52	0
best	1	1	50000	1.3	1.3	0.34	0	0	0	0	0
car	1	1	10000	2.0	2.0	0.52	1	1	1	0.52	0.27
insurance	1	1	1000	3.0	3.0	0.78	2	1.3	1.3	0.68	0.53

Exercise: what is *N*, the number of docs?

Doc length = $\sqrt{1^2+0^2+1^2+1.3^2} \approx 1.92$

Score = 0+0+0.27+0.53 = 0.8

1. Wat is N?
2. Verklaar idf(auto)
3. Verklaar kolom Query: wt
4. Verklaar Query: n'lize
5. Verklaar Doc: tf-raw
6. Verklaar Doc: tf-wt(insurance)
7. Verklaar Doc: wt
8. Verklaar Doc: n'lize(insurance)
9. Verklaar Product
10. Verklaar Score