Question
Timothy Leavitt · Jul 28, 2021

%SIMILARITY variations (Okapi BM25+) and iKnow/iFind

I'm working in an application that uses %SIMILARITY to find matches among a set of documents that vary greatly in length. It's generally good but I've noticed issues with ranking short partially-matching documents over longer documents that match the search string entirely.

Reading up on the Okapi BM25 ranking function (which is what %SIMILARITY / the %Text package use) at https://en.wikipedia.org/wiki/OkapiBM25 I see mention of the BM25+ modification, which "was developed to address one deficiency of the standard BM25 in which the component of term frequency normalization by document length is not properly lower-bounded; as a result of this deficiency, long documents which do match the query term can often be scored unfairly by BM25 as having a similar relevancy to shorter documents that do not contain the query term at all." This seems like exactly what I need.

I'm likely to go down the rabbit hole of implementing BM25+ in a %Text.English subclass and will share my results in an article when I do... but before I do that, I'm curious if iFind has some new-and-improved equivalent to %SIMILARITY, ideally that would just be a drop-in replacement for it. Has anyone worked with this sort of thing before?

#iFind #SQL #InterSystems Natural Language Processing (NLP, iKnow) #InterSystems IRIS
Product version: IRIS 2021.1

Source URL:https://community.intersystems.com/post/similarity-variations-okapi-bm25-and-iknowifind