Dataset CWRCzech vylepšuje výsledky hledání v češtině

21.4.2024 - Josef Vonášek | Seznam.cz

Náš nový článek přijatý na prestižní konferenci SIGIR 2024 přináší klíčový příspěvek v oblasti českého webového vyhledávání. Představuje CWRCzech, což je nový dataset pro hodnocení relevance vyhledávání obsahující 100 milionů párů dotaz-dokument v českém jazyce. Pojďme si ho představit.

Tento nekomerční dataset určený pro akademické použití je řádově větší než zahraniční anglické datasety a se skládá z klikaných dokumentů pro sémantické dotazy typu: apod. Zároveň náš článek ukazuje, jak daná data využít při trénování jazykových modelů pro přesnější vyhledávání.

Klikanost dokumentu je totiž řádově méně spolehlivá než manuálně vytvořená, ale cenově nákladná anotace. Tento vztah je dobře znázorněný na grafu níže, který porovnává kvalitu modelu naučeného čistě na:

- manuálních anotacích (DaReCzech, oranžová),
- klikanosti dokumentu (CWRCzech, modrá).

Mimo jiné z něj vyplývá, že 1 milion anotací je možné v našem případě nahradit přibližně 20 miliony klikanými dotazo-dokumenty. Naše metoda učení z klikaných dat tedy dokáže překonat stávající modely učené tradičním způsobem bez časově i finančně náročné tvorby anotovaného datasetu.

Klikaná data obsahují nejen informace o počtu kliků, ale také o času stráveném na stránce nebo pozici, na které se dokument zobrazil. V článku ukazujeme, jak se chovají modely naučené na každé z těchto informací odděleně a jak je možné informace kombinovat a vyplnit tak slepá místa u každé z nich. Platí totiž, že kvalitnější atributy s přesnější informací jsou zákonitě méně dostupné.

To se odráží i v následující statistice: Ačkoliv totiž 20 % dokumentů má alespoň jeden klik, čas strávený na stránce je známý jen pro 50 % z nich. Kombinace informací je tedy stěžejní pro maximální účinnost naší metody.

Další klíčový prvek naší metody spočívá v tom, jakým způsobem odstraňujeme selektivní zkreslení zanesené původem datasetu. Ten totiž obsahuje už poměrně relevantní dotazo-dokumenty pocházející z našeho stávajícího vyhledávání.

V článku proto ukazujeme, jak znovu vybalancovat poměr relevantních a nerelevantních dokumentů za pomoci uměle vytvořených nerelevantních dotazo-dokumentů a kontrastivního učení. Zároveň popisujeme, jak je možné kontrastivní učení upravit, aby lépe odpovídalo naší rankovací úloze. V základní formulaci totiž rozlišuje jen relevantní dokumenty od nerelevantních (tzn. klik vs. neklik), ale nepracuje už s mírou relevance jako takovou (tzn. počet kliků).

Modely použité v článku se řadí mezi transformery, v současnosti nejpopulárnější architekturu pro zpracování přirozeného jazyka. Naše metoda přináší signifikantní zpřesnění relevance jak pro naše menší (cca 20M parametrů) volně dostupné modely Small-E-Czech a sémantické RetroMAE, tak i pro větší (cca 160M parametrů) externí modely typu FERNET.

Modely jsou standardně předučené na českém jazyce a poté naučené naší metodou na klikaných dokumentech Seznamu. Již v tomto kroku dosahujeme zásadního zlepšení. Na závěr pak použijeme menší množství anotací k dotrénovaní modelu, abychom z našich dat vytěžili maximum.

Celkově je náš přístup využití klikových dat zásadním krokem pro kvalitnější vyhledávání v českém jazyce a ilustruje, jak je možné efektivně pracovat s tímto typem dat, která jsou pro firmy často jednoduše dostupná.

https://blog.seznam.cz/2024/04/dataset-cwrczech-vylepsuje-vysledky-hledani-v-cestine