Zašto se traže ključne riječi s podacima?

U prošlosti se pretraživanje dokumenata provodilo tako da se tražilo točno podudaranje riječi unutar dokumenata (tzv. leksička pretraga). No, takav način pretraživanja nije uvijek bio učinkovit zbog bogatstva i složenosti različitih jezika

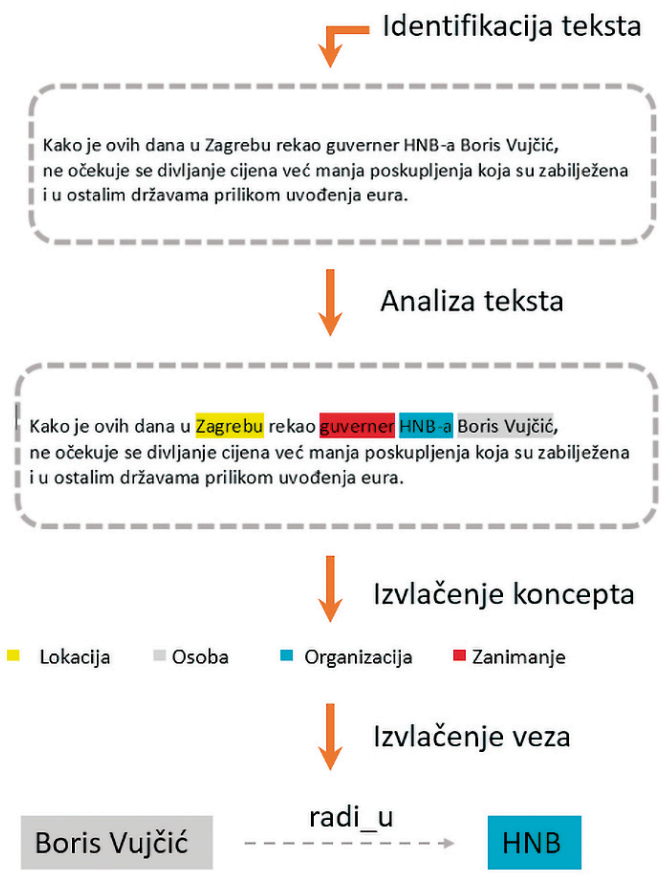
Tina Knežević, razvojni inženjer, Megatrend Poslovna Rješenja

Prelaskom na semantičku pretragu dokumenata, pretraživanje se obavlja, ne samo tražeći postojanost riječi u dokumentu, već uzimajući u obzir i značenje samih riječi. Cilj semantičkog pretraživanja je poboljšati točnost re-zultata pretraživanja razumijevanjem korisničkih namjera, kontekstualnog značenja unesenog termina te povezanosti samih riječi.

KOMPONENTE SEMANTIČKOG PRETRAŽIVANJA

Važna komponenta semantičkog pretraživanja, semantičko je označavanje. Ono obogaćuje sadržaj informacijama koje se strojno mogu obraditi povezivanjem postojećih informacija s izdvojenim pojmovima. Ti pojmovi izvučeni iz dokumenata nedvosmisleno su definirani i povezani jedni s drugima unutar i izvan sadržaja. Semantičko označavanje obavlja se korištenjem tehnika računalne obrade prirodnog jezika (engl. Natural Language Processing, NLP) koje pomažu pri prevođenju i pretvorbi teksta u strukturirane podatke.

Osim semantičkog označavanja, važna komponenta pri semantičkom pre-traživanju je i latentno semantičko indeksiranje (engl. Latent Semantic Indexing, LSI). LSI je tehnika u obradi prirodnog jezika koja analizira odnose između skupa dokumenata i pojmova koje oni sadrže, identificirajući skrivene kontekstualne odnose između riječi. Drugim riječima, LSI se zasniva na principu da riječi koje se koriste unutar istog konteksta, pretežito imaju isto ili povezano značenje, iako ne dijele iste znakove ili sinonime.



Proces semantičkog označavanja dokumenta

IBM WATSON DISCOVERY

IBM Watson Discovery IBM-ov je alat za inteligentno semantičko pretraživanje podataka pomoću umjetne inteligencije. Također, predstavlja i platformu za analizu teksta koja koristi NLP, kako bi otkrila korisne podatke iz složenih poslovnih dokumenata, web-stranica ili velikih skupova podataka te tako skratila vrijeme samog pretraživanja. IBM Watson Discovery omogućuje korisnicima da dodaju vlastite skupove dokumenata te nad njima primjenjuje algoritme koji obogaćuju umetnute podatke, izvlačeći ključne pojmove i entitete (poput lokacija, organizacija, osoba, itd.) te provodi semantičku analizu nad dokumentima.

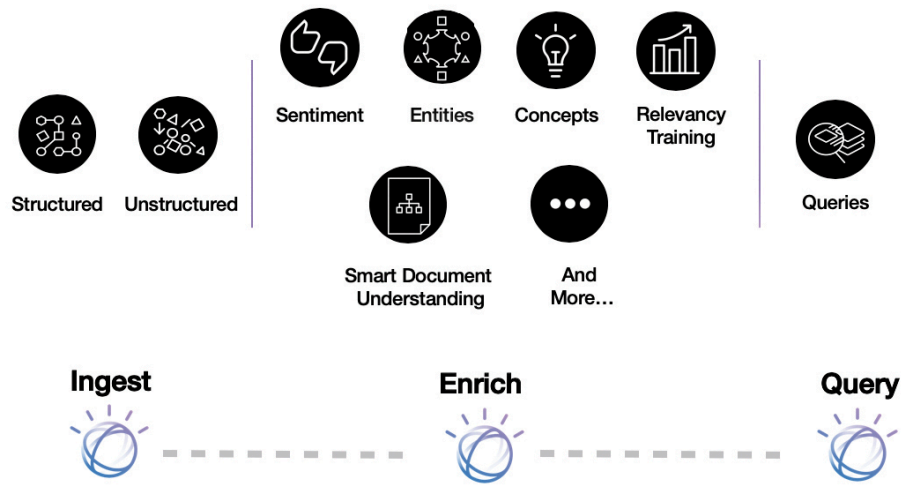
IBM Watson Discovery obogaćuje podatke do-davanjem metapodataka iz prikupljenih semantičkih informacija. Podaci se prikupljaju pomoću četiri glavne IBM Watsonove funkcije, a to su izdvajanje entiteta, analiza osjećaja, klasifikacija kategorija i označavanje koncepata. Također, integriran je i pametni alat za razumijevanje dokumenata (engl. Smart Document Understanding, SDU). SDU je alat baziran na algoritmima strojnog učenja, koji razbija dokumente u manje komade informacija te korisniku omogućuje da jednostavno kategorizira dijelove dokumenata, kako bi alat mogao izgraditi bolje razumijevanje kritičkih komponenata unutar danih dokumenata te poboljšati rezultate odgovora prilikom pretraživanja.

i što se namjerava učiniti

Nakon prijenosa podataka i njihova obogaćivanja, moguće je graditi upite te integrirati IBM Watson Discovery u vlastita rješenja ili s drugim IBM-ovim alatima, poput servisa IBM Watson Natural Language Understanding ili IBM Watson Assistanta. IBM Watson Discovery implementiran je u IBM Watson Assistant preko Search Skilla te omogućuje virtualnom asistentu da odgovara na složena pitanja, pregledavajući veliku bazu dokumenata. IBM Watson Discovery dostupan je na preko 20 svjetskih jezika, uključujući i hrvatski jezik, te ga je moguće koristiti *on-premise* ili na *cloudu*.

ZAKLJUČNO

Semantičko pretraživanje omogućilo je da se pretraživanje više ne temelji isključivo na postojanju riječi u dokumentima, već da se pretražuje značenje tih riječi, tj. upita. Drugim riječima, cilj semantičkog pretraživanja je znati zašto korisnik pretražuje upravo te ključne riječi, i što s dobivenim podacima namjerava učiniti. Takav način pretraživanja približava računalno pretraživanje dokumenata ljudskom načinu razmišljanja, uzimajući u obzir različite načine i tonove upita. Semantičko pretraživanje uvelo je revoluciju u tražilicama, što se pokazuje u njegovoj širokoj uporabi, ne samo kod web-pretraživača, već i kod pretraživača baza znanja specijaliziranih za različite znanosti i poslovanja.



Pregled usluge IBM Watson Discovery

Ilustracija IBM-ova alata SDU, implementiranog unutar aplikacije IBM Watson Discovery

PRIMJENE SEMANTIČKE PRETRAGE

- Web-tražilice
- Question Answering sustavi i virtualni asistenti
- Pretraživači baza znanja
  - Pretraživanje baza kandidata u procesu selekcije
  - Analiza kompetitivnosti
- Pretraživanje veće količine (ne)strukturiranih dokumenata u raznim formatima i iz različitih izvora
  - PDF format
  - Microsoftovi Office formati
  - Slikovne datoteke (OCR)
  - Web-stranice (web crawling/scraping)