

MEATEST spol. s r.o. - sdělení pověření pro ochranu osobních údajů číslo **73** ze dne **24.6.2024**

Vytěžování informací z dokumentů pomocí AI

V dnešní době je velká část informací uložena ve formě dokumentů, jako jsou faktury, smlouvy, formuláře, zprávy, články, knihy a další. Tyto dokumenty obsahují cenná data, ale je potřeba je umět zpracovat. Nejprve je ale musíme z dokumentů získat – vytěžit. Vytěžování dat z dokumentů je proces, který se snaží extrahovat a strukturovat data z dokumentů pomocí různých metod a technik. Nástup umělé inteligence tuto oblast výrazně zjednodušil a umožnil provozovat ve velkém měřítku. Podívejme se, jak to vlastně funguje.

Co je to vytěžování dat z dokumentů?

Podívejme se nejprve, co to vlastně je vytěžování dat z dokumentů. Je to schopnost získávat data z různých typů a formátů dokumentů, jako jsou PDF, obrázky, skeny, e-maily, tabulky, prezentace, webové stránky a další. Cílem je převést neuspořádaná a nehomogenní data do uspořádané a strukturované formy, která umožňuje jejich snadné zpracování, analýzu a využití v dalších systémech (ERP, CRM, DMS a Workflow...). Vytěžování dat z různých dokumentů je důležitou součástí digitalizace a automatizace procesů.

Nejen faktury

V současné době se s pojmem vytěžování dat setkáme nejčastěji v souvislosti s fakturami a účtenkami. Je to důsledkem toho, že faktury jsou nejčastěji rutinně a ručně zpracováváné dokumenty v rámci firem. Ruční přepis těchto dat často zaměstnává i několik účetních, jejichž kapacita je tak mrhána na rutinní činnosti, místo aby se věnovali smysluplnější a odbornější práci. Vytěžování dat z faktur umožňuje extrahovat důležité informace, jako je datum, číslo, dodavatel, položky, ceny, DPH a další, a automaticky je zaúčtovat do účetního systému, či například spustit schvalování. Tím se zrychlí a zjednoduší celý proces zpracování faktur, sníží se chybovost a ušetří se čas a náklady.

Vytěžování dat z různých dokumentů však není omezeno jen na faktury a účtenky. Existuje mnoho dalších možných scénářů využití této technologie nejen ve firemní sféře. Například:

- Vytěžování dat z dodacích listů a skladových dokladů umožňuje efektivněji sledovat pohyb zboží, stav zásob či vyřizování objednávek.
- Vytěžování dat ze smluv, objednávek a nabídek umožňuje extrahovat klíčové podmínky, jako jsou lhůty, ceny, slevy, garance, sankce a další. Tato data lze pak automatizovaně zpracovat v obchodních či CRM systémech a zefektivnit tak kontrolu, vyhodnocování a plánování obchodních aktivit.
- Vytěžování dat z životopisů, pracovních smluv a hodnocení umožňuje získat informace o zaměstnancích, jako jsou jejich kvalifikace, zkušenosti, kompetence, plnění cílů a další. Data z těchto dokumentů lze využít pro nábor, řízení lidských zdrojů, motivaci a odměňování.
- Vytěžování dat z lékařských zpráv lze využít k rychlejší a automatizované detekci nestandardních situací – například sekundárních nemocničních infekcí.
- Vytěžování dat z finančních dokumentů, jako jsou výkazy, transakce nebo auditní zprávy, umožňuje získat informace o finanční situaci, výkonnosti, rizicích nebo příležitostech. Data z těchto dokumentů lze využít pro finanční analýzu, obchodní aktivity nebo detekci podvodů.
- Vytěžování dat z právních dokumentů, jako jsou smlouvy, rozsudky, žaloby nebo patenty, umožňuje

MEATEST spol. s r.o. - sdělení pověření pro ochranu osobních údajů číslo **73** ze dne **24.6.2024**

získat informace o právních předpisech, precedencích, argumentech nebo inovacích. Data z těchto dokumentů lze využít pro právní analýzu, poradenství, vymáhání práva nebo ochranu duševního vlastnictví.

Tyto a mnoho dalších příkladů nejsou většinou žádné objevy ani průlomová nová řešení. Většinu těchto dat již máme. Přínosem je ale automatizace a rychlost získání těchto dat ve strukturované podobě pro jejich další zpracování. To, že nejsme závislí na ručním přepisování dat, které někdy vytváří chyby, někdy zpomaluje proces a někdy se vůbec nestane, přináší možnost spustit řadu automatických procesů a především se na ně více spolehnout. I když i kontrola těchto dat, z principu využití AI, musí být pochopitelně součástí procesu.

Jak to ta AI dělá?

Nástroje na bázi umělé inteligence využívají různé metody, jako je strojové učení, zpracování přirozeného jazyka nebo optické rozpoznávání znaků, aby extrahovaly data z textu, obrázků, tabulek nebo grafů. Již řadu let se pro získávání dat z dokumentů, nejčastěji třeba právě faktur a účtenek, používají nástroje na bázi strojového učení (Machine Learning – ML). Data se nejprve z obrazové podoby pomocí OCR převedou na text tak, že obsahují i informace o umístění konkrétních slov či čísel v rámci dokumentu (koordináty). Tyto informace následně zpracuje ML model, naučený na historických datech, který určí význam konkrétních dat a ta přiřadí do definované struktury. Tato data jsou pak buď uložena do databáze, nebo předána zpět v určené struktuře (nejčastěji jako XML/JSON soubor). Výhodou těchto modelů je jejich rychlost a možnost učení a úprav pro konkrétní typy dokumentů či požadovaných informací. Právě ale ta možnost učení je současně i určitou nevýhodou – tyto modely musíte na požadovaný typ dokumentů naučit. Současně se tento přístup hodí více na různé strukturované dokumenty (faktury, dodací listy, formuláře...) než na zpracování delších textů.

Co na to populární GenAI?

Některé nástroje na bázi generativní umělé inteligence umí také docela dobře vytáhnout strukturovaná data. Když vyzkoušíte vytěžit fakturu některou z těchto služeb, dostanete celkem dobré výsledky, často na první pohled srovnatelné se specializovanými modely. Nicméně pošlete tam stejnou fakturu znovu. Výsledek, především z pohledu struktury, bude jiný. Museli byste tak před generativní AI postavit propracovaný prompting nástroj, který by se snažil zajistit, abyste dostali stále stejnou strukturu. Současně by tento model nemusel být úplně levný na využití.

Nicméně generativní AI má velkou výhodu při vytěžování textových informací. Její možnosti „pochopení“ textu a získání souhrnu či klí-čo-vých bodů se uplatní při vytěžování dokumentů typu smlouvy, na-bíd-ky, články, případně poptávky. A toto je velká nová oblast využití vytěžování dat, kde můžeme čekat řadu zajímavých nových řešení.

Jak dál?

Vytěžování dat z dokumentů je aktuální disciplína, která zažívá svou renesanci právě i díky trendu umělé inteligence. Vytěžování faktur a účtenek se pomalu stává běžnou komoditou a součástí běžného SW. Nicméně i zde je stále prostor k posunu a narovnání některých oče-ká-vá-ní.

Zdroj: AVcentrum ČR