Z Wikipedie, otevřené encyklopedie

Dolování z textu (anglicky *text mining*) je vědecká disciplína na pomezí dolování z dat, strojového učení a počítačové lingvistiky. Vyvíjí se především s potřebou automatického zpracování ohromného množství informací dostupných v podobě volného textu. Klasické metody dolování z dat totiž pracují pouze se strukturovanými daty (obsahujícími metadata důležitá pro zpracování) a většina informací jim tak zůstává nepřístupná.[1]

Typické úlohy

Určování druhu textů

Cílem určování druhu textů, anglicky *text categorization,* je přiřadit k danému textu jednu, či více kategorií z předem daného výčtu (např. sport, politika, krimi…). Typicky je prováděno na základě četností slov vyskytujících se v textech jednotlivých kategorií některou z technik strojového učení.

Shlukování textů/dokumentů

Shlukování textů/dokumentů, anglicky *document clustering,*[2] je úloha principiálně podobná určování druhu textů. Místo zařazování do předem daných kategorií jsou však jednotlivé texty dávány do skupin na základě jejich vzájemných podobností. Každý dokument je tak obvykle zařazen právě do jedné skupiny. Vytvořené skupiny mohou, ale nemusí odpovídat očekávaným kategoriím (burzovní zprávy a sportovní výsledky mohou například spadnout do jedné kategorie na základě faktu, že obsahují větší množství čísel).

Shrnutí textu

Shrnutí textu, anglicky *text summary.* K vytvoření krátkého shrnutí textu se dá přistupovat dvěma způsoby, buď z textu vybrat nejdůležitější pasáže (např. věty) a ty vhodně seřadit (*summary extraction*) anebo je možné text hlouběji analyzovat a na základě jeho sémantické reprezentace parafrázovat jeho obsah (*summary abstraction*). Druhý z obou přístupů by měl poskytovat lepší výsledky, ve skutečnosti však sémantická analýza ani následné

generování souvislého textu dosud není na takové úrovni, aby překonala výsledky prvně zmíněného přístupu.

Analýza sentimentu

Podrobnější informace naleznete v článku Analýza sentimentu.

Analýza sentimentu, anglicky *sentiment analysis.* Na základě výskytu citově zabarvených slov lze usuzovat na autorův pozitivní či negativní postoj k předmětu zprávy. To může být užitečná informace obzvláště ve spojení s tematicky zaměřenými diskusními fóry.

Extrakce konceptů; rozpoznání pojmenovaných entit

V angličtině *concept extraction* či konkrétněji *named-entity recognition (NRE).* Jde o určení entit, které jsou v textu zmíněny. Nástroj pro NRE by je měl identifikovat a zároveň klasifikovat. do předem definovaných kategorií.[3] Například v článku o V. Klausovi by tedy výrazy „Václav Klaus“ a „prezident“ měly být přiřazeny stejné entitě. Problém úzce souvisí s desambiguací slovních významů a tudíž patří k těm základním problémům zpracování přirozeného jazyka.

Určení vztahu mezi entitami

Dokážeme-li v textu určit pojmenované entity, můžeme na základě analýzy vět (např. pomocí rámců – FrameNet) určit jejich vztahy (např. z výrazu "Sarkozy se oženil s Bruinovou" je možné získat vztah, že Bruinová je manželkou Sarkozyho).

Odkazy

Reference

1. ↑ Unstructured Data and the 80 Percent Rule. *www.clarabridge.com* [online]. [cit. 2010-06-10]. Dostupné v archivu pořízeném dne 2010-07-02.
2. ↑ Wayback Machine. *web.archive.org* [online]. [cit. 2023-07-08]. Dostupné v archivu pořízeném z originálu dne 2023-07-08.

3. ↑ ROZPOZNÁVÁNÍ POJMENOVANÝCH ENTIT | Nový encyklopedický slovník češtiny. *www.czechency.org* [online]. [cit. 2022-04-10]. Dostupné online.

Externí odkazy

- Obrázky, zvuky či videa k tématu dolování z textu na Wikimedia Commons