T A
Č R

Program Éta

Průběžná evaluace Programu ÉTA

Příloha 6

Metodika analýzy afinity/(obsahové) podobnosti mezi textově reprezentovanými entitami

Vypracovala: Technologická agentura ČR
Září 2019

Základní ideje a cíle

Máme-li kolekci entit, které jsou reprezentovány sadami textů, můžeme pracovat s podobnostmi těchto jejich reprezentací. V rámci evaluace program ÉTA se budeme zabývat zejména:

- **projekty,** které budou reprezentovány svými názvy, klíčovými slovy a cíli (z důvodu snadnějšího zpracování v AJ),
- **komentáři hodnotitelů** (entitou je samotný text komentáře).

Cílem je získat a popsat *shluky (clustery, komunity, …)* podobných projektů/komentářů za účelem hlubšího vhledu do obsahu popisovaného prostředí – v případě projektů: která témata jsou pokryta programem a která nikoliv, jaká témata jsou charakteristická pro nepodpořené projekty. V případě komentářů hodnotitelů může jít o určení typických chyb, jichž se dopouštějí předkladatelé návrhů projektů atp.

Pozn.: *Korpusem* budeme pro účely tohoto dokumentu nazývat kolekci dokumentů, z nichž každý reprezentuje danou entitu (v případě projektů to může být sada dokumentů, které obsahují název, klíčová slova a cíle projektu). Vzhledem k účelu korpusu (určení obsahové podobnosti) je potřeba korpus nejprve předzpracovat (*preprocessing*).

Předzpracování

Preprocessing korpusu obsahuje tyto kroky:

- **lowercasing** (převod na malá písmena) – slova (tokeny) “projekt” a “Projekt” pro nás jsou rovnocenné
- odstranění **interpunkce**
- v případě českých textů i **lemmatizace** (převod na základní tvar, např.: “projektem” → “projekt”
- **odstranění stop-slov** (slov nenesoucích význam, např. spojky, zájmena, …, neplnovýznamová slova)
- **odstranění číslic**
- **odstranění bílých znaků** (nadbytečných mezer, tabulátorů, ...)
- **odstranění krátkých slov** (délky 1–2)

Cílem těchto kroků je odstranit slova či řetězce, které z hlediska analýzy obsahu nejsou relevantní či jsou přímo zavádějící.

Výpočet podobnosti

Po provedení předzpracování, je korpus reprezentován jako document-term matrix – řádky matice odpovídají dokumentům čili entitám (např. projektům), sloupce odpovídají slovům (termům). V případě, že se dané slovo v dokumentu nevyskytuje, je na průsečíku příslušného řádku a sloupce číslo nula, pokud se vyskytuje, použije se kladné reálné číslo, které vyjadřuje jak četnost výskytu daného slova v daném dokumentu (“četnější je důležitější”), tak četnost výskytu napříč korpusem (“slova, která jsou ve velkém množství dokumentů nejsou tak důležitá”) – přesněji řečeno, využíváme

tf-idf weighting. Na každém řádku matice tedy najdeme vektor, který reprezentuje daný dokument (složky vektoru odpovídají slovníku, který máme k dispozici: vznikl z celého korpusu). Tento způsob reprezentace dokumentů patří mezi vektorové reprezentace.

Pozn. Existují i další možné vektorové reprezentace dokumentů, např. ty, které vycházejí ze zpracování textových dat pomocí neuronových sítí (word embeddings, potažmo sentence či document embeddings), ty však naše účely v tuto chvíli nepoužíváme (třebaže jejich využití plánujeme), navíc následný způsob zpracování se *neodlišuje* v závislosti na tom, jak vektory reprezentující daný dokument byly získány.

Máme-li k dispozici pro každý dokument (odpovídající např. projektu v kolekci projektů daného programu), lze měřit vzájemnou vzdálenost těchto dokumentů, resp. jejich vektorových reprezentací. K tomuto účelu používáme kosinovou podobnost (cosine similarity) – ta vyjadřuje afinitu/podobnost mezi každou dvojicí (projektů, příp. komentářů). Tento proces vede implicitně k neorientovanému grafu s váženými hranami (vrcholy jsou projekty, šířka hrany/spojnice mezi nimi odpovídá míře jejich podobnosti).

Pro získání shluků byl použit algoritmus WalkTrap. Výsledek může být reprezentován ve formě grafu se zvýrazněnými shluky. (Konkrétní shluk projektů můžeme pak následně vizualizovat např. pomocí wordcloudu sestaveného z klíčových slov a názvů projektů ve shluku.)

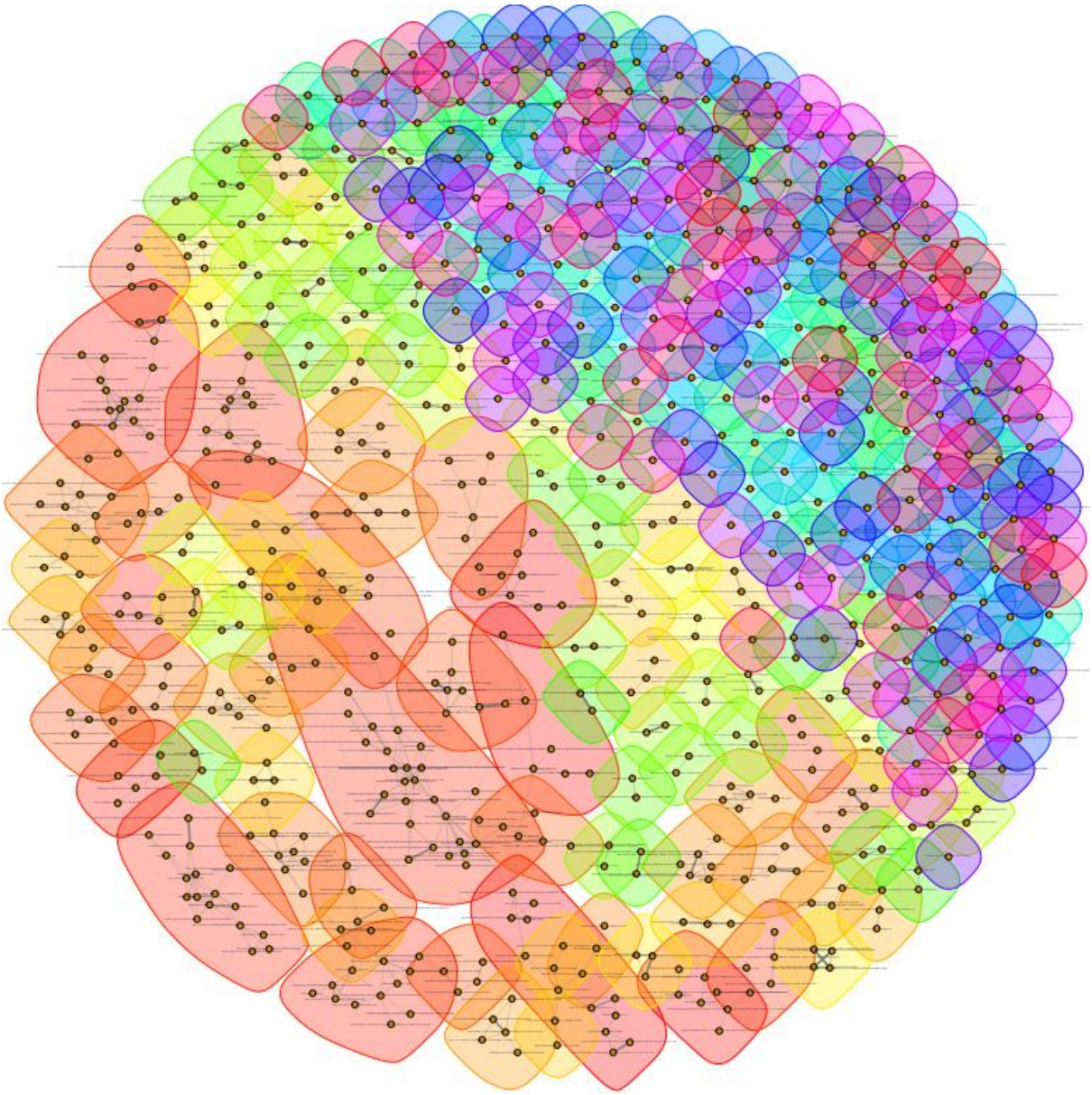
Výstup

Výstupem je **tabulka "klíč-hodnota",** v našem případě např. ID projektu a číslo shluku, do kterého daný projekt patří (případně text komentáře a opět číslo shluku, do kterého komentář patří).

Implementační poznámky

Celý proces včetně vizualizace (s výjimkou wordcloudů) byl implementován skripty v jazyce R/RStudio s využitím knihoven `tm`, `lsa`, `igraph`.

Ukázka vizualizace sady projektů s vyznačenými shluky



Reference

Pozn. tento přístup byl použit k prozkoumávání medicínského kurikula, podrobně popsáno v článku:

Komenda, M., Víta, M., Vaitsis, C., Schwarz, D., Pokorná, A., Zary, N., & Dušek, L. (2015). Curriculum mapping with academic analytics in medical and healthcare education. PloS one, 10(12), e0143748.