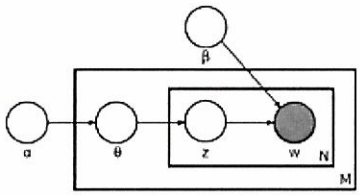
Topic Models for IR

Řešitel: Ing. Tomáš Tunys, FEL ČVUT
Vedoucí projektu: Ing. Jan Šedivý, CSc., FEL ČVUT

Modelování textových dokumentů a jejich klasifikace je důležitým a stále se rozvíjejícím oborem, který spadá do široké kategorie problémů z oblasti *information retrieval* (IR).

Jednou ze zajímavých alternativ ke klasické reprezentaci textových dokumentů jako neuspořádané kolekce slov (angl. *bag-of-words*) je tzv. *topic modeling*. Základem modelu témat je statistický model textu (angl. *topic model*) naučený na velké množině tematicky různorodých textů, který lze poté využít k identifikaci silně zastoupených témat v obsahu předložených dokumentů, které chceme klasifikovat.

Model lze tedy v nejjednodušší formě využít k reprezentaci dokumentů v podobě pravděpodobnostního vektoru, který lze interpretovat jako poměrné zastoupení určitého tématu v počtu slov dokumentů.



Model, který byl v základní fázi projektu vyzkoušen a který reprezentuje témata pomocí pravděpodobnostních distribucí nad slovy ze slovníku, se nazývá *Latent Dirichlet Allocation* (LDA). Důležité pro celý projekt bylo (automaticky) vytvořit velké množství dokumentů v dostatečně dobré kvalitě. K dosažení tohoto cíle byla využita česká Wikipedie. Pro testování na anglických dokumentech i byla použita i její anglická verze.

Jako základní řešení, vůči nimž byl měřen přínos zvoleného přístupu, byla naimplementována (či využita implementace již existujících) řada klasifikačních algoritmů mezi něž patří Multinomial Naive Bayes, Dirichlet Compound Multinomial, či Support Vector Machine.

Po dokončení základního řešení byla k původním vektorovým reprezentacím dokumentů (*bag-of-words*) přidána informace ze zjištěných témat a výsledky byly porovnány. Přestože byla kvalita textových dat vysoká, z výsledků, kterých bylo dosaženo, lze vyčíst, že přínos témat do klasifikace je zanedbatelný a ve většině případů škodlivý. Otázky, které by mohly vést k identifikaci toho, v čem ve zvolených přístupech obohacení reprezentací dokumentů spočíval problém, byly ponechány k budoucímu (a stále probíhajícímu) výzkumu.

Cíle projektu:
- vytvořit upravenou a čistou databázi českých dokumentů pro testování a výzkum v oblasti modelování témat (*topic modeling*);
- naimplementovat nejznámější algoritmy pro modelování témat a změřit jejich přínos v klasifikaci dokumentů;
- nalézt takovou posloupnost operací pro předzpracování dokumentů, která by zmenšila dimenzionalitu problému a vedla ke zlepšení (zrychlení) funkce klasifikačních modelů;
- připravit algoritmy pro základ budoucího výzkumu v dané oblasti;
- publikovat relevantní výsledky ve formě vědeckého článku.

Vytyčených cílů bylo dosaženo a část týkající se benchmarkování výše uvedených algoritmů byla prezentována na konferenci STAIRS 2014 v příspěvku s názvem „Empirical Study of Classification Models for Web Page Categorization".