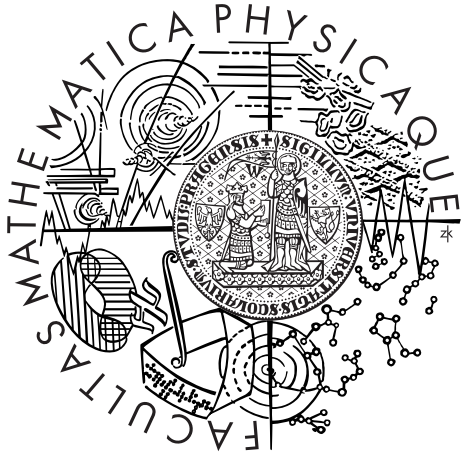
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

**BAKALÁŘSKÁ PRÁCE**




Karel Bílek

# Sledování témat v elektronickém zpravodajství

Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: RNDr. Ondřej Bojar, Ph.D.
Studijní program: Informatika
Studijní obor: Obecná informatika

Praha 2011

Na tomto místě bych chtěl poděkovat Ondřejovi Bojarovi za vedení této práce. Také bych chtěl poděkovat Josefu Šlerkovi ze Studia nových médií na FF UK za některé nápady, byť se nakonec ne všechny materializovaly.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Sledování témat v elektronickém zpravodajství

Autor: Karel Bílek

Katedra: Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: RNDr. Ondřej Bojar, Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt: V této práci se snažím nalézt definici zpravodajského tématu tak, aby byla detekce těchto témat v textu implementovatelná a kvalita této detekce měřitelná. Popisuji možné metody — „prosté“ počítání slov, případně se zavedením stopslov; TF-IDF; dále popisuji problém textové klasifikace, mírně se dotknu text clusteringu. Dále popisuji přístupy, nazvané latent semantic indexing a latent Dirichlet allocation. Také popisuji experimenty s „prostým“ počítáním slov, TF-IDF a textovou klasifikací na databázi článků z několika elektronických zdrojů; vznik této databáze v práci popisuji rovněž. Ke způsobu řešení pomocí textové klasifikace uvádím metriku pomocí měření přesnosti a úplnosti; podle těchto metrik měřím několik variant textové klasifikace.

Klíčová slova: Zpravodajství, články, témata, klíčová slova

Title: Topic monitoring in online news articles

Author: Karel Bílek

Department: Institute of Formal and Applied Linguistics

Supervisor: RNDr. Ondřej Bojar, Ph.D., Institute of Formal and Applied Linguistics

Abstract: In this thesis, I try to find a definition of a news topic to make topic detection implementable and its quality measurable. I describe various methods — a “simple” words counting, optionally with stopwords. I also describe TF-IDF and the text categorization problem. I touch the subject of text clustering. Then I briefly describe approaches called latent semantic indexing and latent Dirichlet allocation. The thesis includes my experiments with “simple” words counting, TF-IDF and text categorization on database of articles from several online news websites; I also describe the creation of this database. Precision and recall are used as a metric to text categorization approach.

Keywords: News, articles, topics, keywords

# Obsah

# Úvod

Úkolem bakalářské práce je sledovat elektronické zdroje zpravodajství, na základě opakujících se slov zařazovat články do témat, tato témata sledovat v čase a případně na základě tohoto sledování detekovat něco, jako je „okurková sezóna“.

V průběhu přípravy práce jsem zjistil, že úkol detekce tématu je poměrně častý v oboru *text mining*, respektive *information retrieval*. Zadání práce se mi částečně splnit podařilo, součástí práce je také shrnutí dalších možných přístupů a pohledů na danou problematiku.

Většina zpravodajských médií dnes používá pro vydávání svých článků síť *World Wide Web*, ať už primárně (např. server `http://aktualne.cz`), nebo sekundárně jako doplněk např. k tištěným novinám (např. server `http://idnes.cz`). Pro oba případy ale platí, že zpravodajský server lze vzít jako pravidelně aktualizovaný zdroj elektronického zpravodajství.

Výhoda zpravodajským webů je ta, že jsou za krátkou dobu (v řádu měsíců až jednotek let) stylisticky stabilní a jsou psány podobným jazykem, zároveň ale mají jasně danou časovou posloupnost. To nám daný úkol značně zjednodušuje.

Zpravodajství na jednotlivých serverech lze jednoduše rozdělit na jednotlivé dokumenty neboli články. Článek na zpravodajském webu má jasně daný obsah, kromě toho je důležité, že nese i informaci o čase vydání (jak datem, kdy se na webu objeví, tak datem, uvedeným v popisu článku přímo na webu). Toho lze využít a informaci o čase dále zpracovávat.

### Témata

Jádrem bakalářské práce je na této množině článků najít „témata“ a jejich četnost případně vynést na časovou osu. Přesnou definici tématu v bakalářské práci nicméně teprve hledám. Je nutné se zatím spokojit s tím, že téma nejasně definuji jako „něco, o čem článek je“. Tuto definici jsem si nazval *naivní* definice tématu.

V první kapitole této práce se budu snažit definici tématu nalézt a zpřesnit; několik nalezených definic pak ve třetí kapitole otestuji, jak „opticky“, tak jasně definovanou metrikou.

# 1. Detekce témat

V této části uvedu několik postupů, jak je možné detekovat témata článku. *Naivní* definici z úvodu jsem si ponechal jako „skutečnou“ definici tématu; pro jednotlivé postupy potom platí, že v každém z nich lze nakonec za definici tématu považovat *ten* formální prvek, které dané postupy vyrábějí.

Některé z těchto postupů jsem neimplementoval; u těch, které jsem implementoval, hodnotím i to, jak moc se výsledky blíží naivní definici (která je samozřejmě už ze své podstaty *velmi* vágní a nepřesná); tj. jak dobře lze říci, že témata, nalezena touto metodou, jsou opravdu to, „o čem je tento článek“. Výsledky samotné včetně výsledků evaluace, kterou také popisuji dále v této části, podrobněji zkoumám v Kapitole 3.

## 1.1 Frekvence a stopslova

První nápad na detekci témat je jednoduché počítání frekvencí slov ve článcích s tím, že častější slova prohlásím automaticky za témata — případně s tím, že budu ignorovat tzv. stopslova.

Termín *stopslova* (*stopword*, případně *stoplist* ve významu seznam stopslov) popisuje např. [Salton – McGill (1983)] a další. Jde o poměrně malou množinu typů slov, která ale představuje velkou část tokenů. (V angličtině je to např. podle [Salton – McGill (1983)] 40-50 procent; v části 3.3.2 ukazuji statistiky na vlastní databázi zpravodajských článků.) Jde převážně o spojky, částice a základní slovesa, která sama o sobě nenesou žádný obsah textu. Je nutno je buď definovat ručně, což je ale jaksi „nesystémové“ řešení; nebo se stopslova definují jako nejčastěji používaná slova (jak je definuji i já), což s sebou nese riziko, že se velmi častá slova, která ale *nejsou* bezvýznamová či pomocná, úplně ignorují.

Pro nalezení množiny témat pro daný článek potom vezmu prvních $k$ nejčastějších slov — buď bez ohledu na stopslova, nebo s jejich vyřazením, kde $k$ je konstantní pro celou databázi.

Tento přístup je sice velmi jednoduchý a dobře definovaný, ale jak ukazuji v části 3.3.1, nedává příliš dobré výsledky. Je to proto, že četnost slova je brána jako *jediné* kritérium pro určení daného slova jako tématu — přičemž dané slovo může sice být ve článku četné, ale je podobně četné i v ostatních článcích, takže velká četnost tohoto slova článek nijak nevymezuje vůči dalším článkům. V podstatě jde o tentýž problém, který se snaží řešit stopslova, pouze v menší míře — nejedná se *přímo* o stopslova, ale přesto tato slova nejsou pro článek natolik důležitá, aby byla jeho tématem.

Dále mi nastává přesně opačný problém, než ten, který jsem nastínil — protože stopslova beru pouze jako *právě ta* nejčetnější slova v celé databázi, některá slova, která by se určitě dala chápat jako témata článku podle naivní definice, se kompletně ignorují (konkrétně se jedná například o jména politických stran).

S touto metodou, stejně jako s následující, je spojen také další, významný problém — tyto metody nacházejí pouze ta slova, která jsou v samotném článku. Jak bude vidět na skutečných datech dále, je nutno je chápat spíše jako jakási klíčová slova — je tedy problém správnost nalezení těchto témat měřit, přičemž návrh metriky kvality zatřídění je jedním z úkolů bakalářské práce.

Témata, nalezená ve článku pouze pomocí této metody počítání frekvencí, jsem si nazval *frekvenční témata*; témata, nalezená za pomoci stopslov jsem si nazval *stop témata*.

## 1.2 TF-IDF

Dalším nápadem je použít váhovou funkci *TF-IDF*, kterou popisuje dobře např. [Salton – Buckley (1988)] a která řeší jednu z mých výtek k minulému přístupu — tj. to, že slova častá v jednom dokumentu jsou častá i ve všech ostatních.

Je zavedena *inverzní frekvence dokumentů* (*inverse document frequency*, *IDF*), která dává větší hodnotu těm slovům, která se vyskytnou ve více dokumentech. Výsledná váha *TF-IDF* (někdy také *tf*idf* či *tf*×*idf*) je poté pro každé slovo ve článku násobkem *TF* (jejich „obyčejná“ frekvence v článku) a *IDF*.

Přesněji definuje [Robertson (2004)] $tf_{i,j}$ jako počet výskytů slova $i$ ve článku $j$; tu můžeme ještě normalizovat vydělením velikostí dokumentu, čímž se dostaneme k podobnému vzorci, jaký uvádí [Wikipedia (2011)]

$$tf_{i,j} = \frac{n_{i,j}}{dl_j},$$

kde $dl_j$ značí délku dokumentu $j$ a $n_{i,j}$ počet výskytů slova $i$ ve článku $j$.

$idf_{i,j}$ je pak (taktéž např. podle [Robertson (2004)]) definována jako $\log \frac{N}{n_i}$, kdy $N$ je velikost korpusu a $n_i$ počet dokumentů, ve kterých se vyskytne slovo $i$ — [Wikipedia (2011)] uvádí ekvivalentní vzorec

$$idf_{i,j} = \log \frac{N}{|\{j : i \in j\}|},$$

kde $j$ je vnímáno jako množina slov ve článku.

[Salton – McGill (1983)] uvádí vzorec pro *idf* explicitně s binárním logaritmem, což ale vlastnosti nijak nemění. V případě, že slovo nemusí být v korpusu, je možné jmenovatel upravit na $|\{j : i \in j\}| + 1$, abychom nedělili nulou.

Potom platí jednoduché

$$\text{TF-IDF}_{i,j} = tf_{i,j} \cdot idf_{i,j}.$$

TF-IDF je opět pouze váhová funkce, tj. je třeba určit nějaké $k$, konstantní pro všechny články, a jako témata vzít pouze $k$ slov s nejvyšším TF-IDF. Těmto tématům říkám *tf-idf témata*.

Pro bigramy, trigramy apod. nedává TF-IDF dobré výsledky, protože IDF je pro bigramy ve většině případů a pro trigramy téměř ve všech případech 1. Stále také platí, že tato metoda najde pouze slova, která se ve článku vyskytují, tj. jde pouze o jakási klíčová slova — platí tedy totéž, co jsem psal v závěru části 1.1.

## 1.3 Klasifikace

Další možností je úlohu definovat jinak — zatřiďovat články do kategorií, které nazveme tématy a které můžeme nějak pojmenovat, aniž by se ale tyto názvy musely jako slova ve článcích vyskytovat. Články jsou na podmnožině článků

do těchto kategorií zatříděny uživatelem, na základě tohoto ručního zatřídění potom zatřiďuje do stejných kategorií stroj. Tato metoda se nazývá *kategorizace* (*categorization*) nebo *klasifikace* (*classification*); témata, nalezená touto metodou, jsem si nazval *kategorická témata*.

### 1.3.1 Definice

Např. podle [Sebastiani (2002)] lze úlohu klasifikace definovat takto: existuje funkce $\bar{\Phi} : D \times C \to \{Ano, Ne\}$, kde $D$ je množina všech dokumentů, $C$ množina všech kategorií, $\{Ano, Ne\}$ pravdivostní hodnota. $\bar{\Phi}$ lze potom interpretovat jako odpověď na otázku „obsahuje tato kategorie tento článek?“.

Funkce $\bar{\Phi}$ je tedy „správné“ zatřídění, nazývaná také *cílová funkce* (*target function*). Úlohou klasifikace je potom tuto funkci $\bar{\Phi}$ nějak aproximovat funkcí $\Phi : D \times C \to \{Ano, Ne\}$, která se nazývá *klasifikátor* (*classifier*).

Jednou možností klasifikace jsou automaty, které mají programátorem daná pravidla pro zatřiďování — tzv. *pravidlové klasifikátory* (*rule-based classifier*). Touto možností jsem se příliš nezabýval, protože ruční sestavení pravidel je příliš pracné; spíše jsem se zaměřil na klasifikace se strojovým učením.

Algoritmy klasifikace se *strojovým učením* (*machine learning*) jsou založeny na tom, že klasifikátor se nejdříve naučí zatřiďovat na množině předem daných článků s *ručně zatříděnými kategoriemi*.

Přesněji popisuje opět [Sebastiani (2002)]: Vezmeme $\Omega \subset D$ podmnožinu všech dokumentů, na níž *známe* výsledky cílové funkce $\bar{\Phi}$, a na základě této znalosti tvoříme funkci $\Phi$ (která se na $\Omega \times C$ může a nemusí shodovat s $\bar{\Phi}$). Ono „známe výsledky funkce“ je právě ruční zatřiďování na podmnožině $\Omega$.

V [Sebastiani (2002)] jsou dobře popsány mnou použité algoritmy pro klasifikaci. Jelikož jsem tyto algoritmy neimplementoval přímo já, ale využil jsem perl modul `AI::Categorizer`, nebudu se zde o nich tolik zmiňovat a případně mohu odkázat na dokumentaci tohoto modulu, viz [Williams (2000–2003)]. V rámci tohoto modulu jsem použil algoritmy Naive Bayes, SVM a Decision Tree.

### 1.3.2 Rysy

Článek je pro všechny tyto algoritmy třeba nějak formálně reprezentovat. Pro účely klasifikace se používají takzvané *rysy* (*features*). Je to množina, která charakterizuje dokument. Formálně, opět podle [Sebastiani (2002)], je možné reprezentaci pomocí rysů popsat takto: $T$ je množina všech rysů, a každý dokument je poté definován jako vektor $d_j = \left\langle w_{1_j}, \ldots w_{|T|_j} \right\rangle$.

Já jsem k definování rysů použil už hotovou funkci TF-IDF, a to dvěma různými způsoby:

1. první varianta je vzít $k$ slov s největší TF-IDF hodnotou a každému z nich dát $w$ stejné, rovné 1,

2. druhá varianta je vzít všechna různá slova ve článku, ohodnocená právě jako výsledek váhové funkce.

### 1.3.3 Kategorie

Čeho jsem se zatím nedotkl, je výběr samotných kategorií a potom také otázky, jestli články mohou být ve více než jedné kategorii.

#### Předem dané kategorie

Jedna z možností je kategorie zadat explicitně předem a zatřiďovat do nich. Následuje otázka — jaké kategorie navrhnout?

Můžeme vybrat například kategorie velmi hrubě podle tématických kategorií, které budou přibližně zrcadlit sekce na zpravodajských serverech. Tuto možnost jsem vyzkoušel s těmito konkrétními kategoriemi: `Politika`, `Ekonomika`, `Krimi`, `Bulvár`, `Kultura`, `Studium`, `Věda`, `Technika`, `Počasí`, `Sport` a `Další`; každá z nich je ještě rozdělena na `domácí` a `svět`.

Kategorie můžeme pojmout i šířeji a vypůjčit si kategorie například z českých WikiNews. Tuto možnost jsem zvažoval, ale nakonec nestihl otestovat.

#### Neomezená množina kategorií

Další možností je nedefinovat kategorie explicitně předem, ale vytvářet je až v průběhu ručního zatřiďování, kde „definice“ kategorie kopíruje naivní definici tématu z úvodu. Pokud jsou kategorie vytvářeny až člověkem při ručním zatřiďování, znamená to, že teoreticky není množina kategorií nijak omezená, a tedy jakýkoliv výraz, který nese nějaký význam, může být tématem článku.

Tato možnost je lepší v tom, že se není třeba dopředu omezovat na seznam kategorií, není seznam kategorií třeba předem vytvářet a hlavně je mnohem jednodušší úloha ručního zatřiďování — v podstatě je redukovaná pouze na otázku „O čem je tento článek?“. Tuto možnost jsem implementoval jako první.

Tato možnost má ale také své nevýhody — vektor kategorií je, jak již naznačuji výše, potenciálně nekonečný. Pokud není v programu implementována sémantická analýza nebo alespoň rozlišování synonymie, polysémie apod., neúplné překryvy témat jsou příliš časté pro to, aby bylo zatřiďování do témat praktické, protože mnoho článků má téma, které se znovu nevyskytne.

Např. článků o pražském Tančícím domě je velmi málo, kategorii `Tančící dům` bude tedy mít pouze malé množství článků. S nějakou úrovní sémantické analýzy by se článek mohl taktéž „automaticky“ zatřídit do kategorie `architektura` nebo `Praha`.

Na druhou stranu, sémantická analýza je příliš složitá, nejednoznačná (jaký je např. vztah témat `premiér ČR` a `Petr Nečas`?) a rozhodně jde za rámec této bakalářské práce.

Je nutno dodat, že ve skutečnosti klasifikátor nezatřiďuje do nekonečně mnoha kategorií, ale opět pouze do konečného množství — a to právě do těch kategorií, které byly označeny v trénovacích datech. Dále je nutné v evaluaci, kterou popisuji v části 1.5, brát v úvahu také pouze existující kategorie.

### 1.3.4 Počet kategorií na článek

Důležitou otázkou je také to, jestli má každý článek pouze jedinou tématickou kategorii, nebo jich může mít více.

Jedna kategorie na každý článek má výhodu kvůli jednodušší evaluaci, na druhou stranu je *složitější* fáze lidského zatřiďování.

Ve skutečnosti se totiž málokdy články týkají pouze jednoho tématu. Podstata novinového článku sama o sobě je spojovat dohromady více skutečností, informovat o dění v určitém místě, spojovat dohromady více osob apod. Proto se článek ze *samé své podstaty* dotýká více než jednoho tématu.

U varianty s neomezenou množinou kategorií jsem nechal možných kategorií na jeden článek více, u varianty s omezenými kategoriemi pouze jednu na článek.

## 1.4 Další přístupy

V této části shrnuji další přístupy, které jsem ovšem neimplementoval, takže jejich případné výhody či nevýhody mohu posuzovat pouze teoreticky.

### 1.4.1 Shlukování

Další variantou je jít na problém jakoby z druhé strany a použít metodu tzv. *shlukování* (*clustering*) — tuto dobře popisuje například [Andrews – Fox (2007)].

Zjednodušeně řečeno jde o to, že na rozdíl od kategorizace se dokumentům nepřiřazují kategorie, ale dokumenty jsou rozděleny do množin na základě vzájemné podobnosti a případné názvy jsou těmto množinám — tzv. *shlukům* (*clusterům*) přiřazovány až ex-post.

Pro úlohu clusteringu je třeba mimo jiné mít nějak definovanou míru podobnosti mezi dvěma různými dokumenty. V našem případě by možná bylo možno tuto „vzdálenost“ definovat jako velikost průniku klíčových slov.

Clustering jsem i kvůli nedostatku času blíže nezkoumal a zde ho uvádím spíše pro srovnání.

### 1.4.2 Latent semantic indexing

*Latent semantic indexing* (zkráceně *LSI*) je mírně rozdílný způsob indexování dokumentů. Postup dobře popisuje [Deerwester et al. (1990)]. LSI v podstatě pouze pomocí jednoduché lineární algebry řeší problémy se synonymií termínů (nikoliv však polysémií); na druhou stranu je ve své původní formě navržen pro problém *text retrieval* — tj. pro mírně odlišný problém, než je ten náš.

Podstata LSI je v tom, že matici $\langle$dokumenty $\times$ termíny$\rangle$ rozloží pomocí techniky *singular value decomposition* (zkráceně *SVD*) a poté zmenší dimenzi prostoru dokumentů a termínů; jednotlivé dimenze pak [Deerwester et al. (1990)] interpretuje jako „koncepty“.

Přesněji je matice $X$ definovaná jako matice o velikosti $t \times d$, $t$ je počet termínů, $d$ je počet dokumentů. V matici se na pozici $x_{termín,dokument}$ nachází počet výskytů termínu *termín* v dokumentu *dokument*.

Metoda SVD převede tuto matici na násobek tří matic $T \times S \times D'$, kde matice $S$ je čtvercová diagonální matice o velikosti $m \times m$ (kde $m \leq \min(t, d)$), matice $T$ má ortonormální sloupce, stejně jako matice $D$, ke které je $D'$ komplexně

sdruženou maticí — matice $T$ je velká $t \times m$, matice $D'$ $m \times d$. [1] Tento rozklad existuje pro *každou* matici.

Můžeme si tedy představit, že prostor termínů a prostor dokumentů má dimenzi $m$ [2] — myšleno tak, že pro daný termín a dokument je pozice v matici $X_{termín,dokument}$ skalárním součinem řádku odpovídajícímu danému termínu v matici $TS^{\frac{1}{2}}$ a řádku odpovídajícímu danému dokumentu v matici $DS^{\frac{1}{2}}$, které jsou oba velké $m$.

Pro zjištění podobnosti dvou dokumentů či dvou termínů je možné si také reprezentovat termíny či body v $m$-rozměrném prostoru, ale tyto body jsou mírně jiné, viz [Deerwester et al. (1990)], zvláště kapitola 4.2.3.

Hlavním krokem LSI je to, že prvky v diagonální matici S lze seřadit podle velikosti, nechat pouze prvních $k$ největších prvků na diagonále a zbytek, tj. naopak $m - k$ nejmenších prvků na diagonále, nahradit nulami. Vznikne poté aproximace matice $X$, nazveme ji $\bar{X}$, je daná součinem $T\bar{S}D$. Jelikož má $S$ na diagonále nuly, je možné[3] řádky s nulami úplně vynechat, vymazat odpovídající řádky a sloupce z matic $T$ a $D$ a mít tedy matice o rozměrech $t \times k$, $k \times k$ a $k \times d$.

Prostor termínů a prostor dokumentů má teď zmenšenou dimenzi $k$ (stejnou úvahou, jako výše). Dokumenty i termíny jsou teď určeny řádky matic $\bar{D}\bar{S}^{\frac{1}{2}}$ a $\bar{T}\bar{S}^{\frac{1}{2}}$.

Podle [Deerwester et al. (1990)] je možné interpretovat tyto řádky jako příslušnost termínů k jednotlivým konceptům.

Je otázkou, jak přesně by k nalezení našich „témat“ LSI přispěl, jestli vůbec; jestli by bylo například možné už tyto koncepty nějak prohlásit za témata, případně použít tuto metodu například ke shlukování dokumentů, slov nebo obojího.

Nad čím by bylo také potřeba uvažovat je způsob SVD rozkladu matice $X$, která se celá nevejde do běžné paměti (pokud bychom opravdu vzali všechny články a termíny) a je velmi řídká.

### 1.4.3 Latent Dirichlet allocation

*Latent Dirichlet allocation* (zkráceně *LDA*) je pravděpodobnostní model, který zajímavě řeší problematiku témat.

LDA byl poprvé popsán v [Blei et al. (2003)]. Je to generativní pravděpodobnostní model, který pracuje se zjednodušením, že u slov ve článku nezáleží na pořadí. Při generování článku v korpusu se postupuje následovně:

- Předem je dáno pravděpodobnostní rozdělení, generující délku článku. Jeho podoba není příliš důležitá. [4]
- Předem je dán také počet témat $k$.
- Dále je předem znám vektor $\alpha$ o velikosti $k$, který je použit jako parametr Dirichletova rozdělení.

[1] Konvence pro velikosti matic u SVD se liší zdroj od zdroje; zde jsem zvolil konvenci od [Deerwester et al. (1990)].

[2] V [Deerwester et al. (1990)] je toto vysvětleno až u redukované matice, ale platí to analogicky i před redukcí

[3] Z principu násobení matic, více opět [Deerwester et al. (1990)]

[4] Autoři [Blei et al. (2003)] uvádějí Poissonovo rozdělení, ale vzápětí také dodají, že toto rozdělení není dále příliš relevantní.

- Dále je předem známá matice $\beta$ velká $k \times v$ (kde $v$ je počet všech slov), která na $\beta_{i,j}$ má pravděpodobnost $p(slovo\, j | téma\, i)$
- Pro každý dokument se nejdříve vygeneruje jeho velikost.
- Poté se vybere parametr $\theta \sim \mathrm{Dir}(\alpha)$, pro **každý dokument zvlášť**.
- Pro **každé slovo zvlášť** se pak podle pravděpodobností v $\theta$ vybere téma $z$ (tedy $z \sim \mathrm{Categorical}(\theta)$[5]).
- Podle pravděpodobnosti v matici $\beta$ se poté vygeneruje slovo.

Znamená to tedy následující:

- Článek nemá jedno téma, ale pravděpodobnost témat $\theta$ pro svoje slova.
- Každý token má naopak svoje téma.
- Jeden typ může být (a často je) ale generován více tématy.
- Těchto slovních témat je omezený počet ($k$) pro celý korpus.

Pokud máme naopak pouze korpus, potřebujeme přiřadit jednoznačně tokeny k tématům, odhadnout $\theta$ pro každý dokument a také parametry $\alpha$ a $\beta$. Jeden z možných algoritmů ke stavbě LDA modelu je popsán opět v [Blei et al. (2003)].

Z uváděných neimplementovaných metod bych nejspíše LDA implementoval jako první a v případné další práci by bylo dobré ji prozkoumat podrobněji, jelikož automaticky řeší velké množství problémů. Je možné, že $k$ takto nalezených témat už by bylo automaticky možné prohlásit za „témata“ — na druhou stranu, je třeba si uvědomit, že $\theta$ neříká to, jaké jsou pravděpodobnosti, že článek patří do tématu, ale opravdu pouze pravděpodobnosti témat jednotlivých slov.

Pro doplnění je dobré dodat, že tzv. online LDA algoritmus na stavbu LDA modelu byl implementován například v projektu GenSim Radimem Řehůrkem z Centra zpracovávání přirozeného jazyka na Fakultě informatiky Masarykovy univerzity, viz [Řehůřek – Sojka (2010)] či přímo stránky projektu: `http://nlp.fi.muni.cz/projekty/gensim/index.html`; autoři popisují průběh algoritmu na korpusu celé anglické wikipedie zde: `http://nlp.fi.muni.cz/projekty/gensim/wiki.html#latent-dirichlet-allocation`.

## 1.5 Evaluace

Nalezení metriky kvality detekce témat bylo jedno ze zadání bakalářské práce. Zatímco klasifikace má sama o sobě dobře definovanou metriku, u dalších metod je evaluace problematická.

[5] [Blei et al. (2003)] uvádí multinomické rozdělení, ale jedná se pouze o jeden výběr, tedy nám stačí kategorické.

### 1.5.1 Klasifikace

Evaluaci u klasifikátorů jsem provedl metodou *cross-validace* (*cross-validation*), popsanou například opět v [Sebastiani (2002)], a jednotlivé testy v cross-validaci jsem provedl pomocí *makroprůměru* (*macroaveraging*) i *mikroprůměru* (*microaveraging*), popsaných tamtéž.

Cross-validace funguje tak, že z množiny všech článků se vybere podmnožina $\Omega$, která je ručně zatříděna do kategorií. Tato množina je rozdělena do $k$ disjunktních množin $Te_1, Te_2, \ldots Te_k$, a poté je na dvojicích $\langle Tv_i = \Omega - Te_i, Te_i \rangle$ provedena klasická evaluace (pokaždé ale s *novým* klasifikátorem), nazývaná v [Sebastiani (2002)] *train-and-test*.

V metodě *train-and-test* je klasifikátor natrénován testovaným algoritmem na množině $Tv$ a menší množinu článků $Te$ poté zatřiďuje do kategorií. Klasifikátor ruční zatřídění na množině $Te$ nezná a znát *nesmí*.

Po každém z těchto dílčích testů je na základě porovnání ručního a strojového zatřídění na množině $Te$ spočítána *přesnost* (*precision*) a *úplnost*[6] (*recall*) — tyto přesněji definuji níže; z nich je poté spočítáno pro každý dílčí test $F_1$*-skóre* ($F_1$*-score*).

Tato skóre jsou poté ze všech $k$ dílčích testů dohromady zprůměrována; pro ilustraci ve výsledcích uvádím i průměrnou přesnost a úplnost.

Jelikož při každém dílčím testu je použita pouze množina, velká $\frac{k-1}{k} |\Omega|$, znamená to, že klasifikátor nikdy nefunguje tak dobře, jak by „mohl“ — ale testování na stejných datech, na kterých byl klasifikátor natrénován, by dávalo nerealisticky dobré výsledky.

Pro dílčí evaluace počítám přesnost a úplnost *vzhledek ke kategoriím* (*with respect to category*), průměrované pomocí *mikro-* a *makroprůměru* (*micro* a *macro averaging*). Jde o to, že pro každou kategorii $c_i$ jsou spočítány hodnoty:

- $TP_i$: *true positive* — počet dokumentů, které měly být a jsou v kategorii $c_i$,
- $FP_i$: *false positive* — počet dokumentů, které neměly být v kategorii $c_i$, ale jsou,
- $FN_i$: *false negative* — počet dokumentů, které měly být v dané kategorii, ale nejsou.

Spolu s $TN_i$ (*true negative* — počet dokumentů, které neměly být v dané kategorii a nejsou) platí pro všechny kategorie $c_i$:

$$(FP_i + TP_i) + (FN_i + TN_i) = |c_i| + |Te \setminus c_i| = |Te| .$$

Poté je přesnost ($\pi$) a úplnost ($\rho$) spočítána dvěma způsoby. Jeden (tzv. mikroprůměr, s indexem $\mu$) je spočítán přes všechna rozhodnutí o kategorizaci, druhý (makroprůměr, s indexem $M$) přes všechny kategorie, tj.

$$\pi_\mu = \frac{TP}{TP + FP} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FP_i)},$$

[6] Český překlad slova *recall* není ustálen — velká část odborné literatury používá přímo anglické názvy, kromě termínu *úplnost* je někdy také používáno termínu *pokrytí*.

$$\rho_\mu = \frac{TP}{TP + FN} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FN_i)},$$

$$\pi_M = \frac{\sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FP_i}}{|C|},$$

$$\rho_M = \frac{\sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FN_i}}{|C|}.$$

V obou případech lze říci, že přesnost znázorňuje podíl toho, že pokud je dokument klasifikátorem zatříděn do kategorie $c$, tak tam má být zatříděn; zatímco úplnost znázorňuje podíl toho, že pokud má být dokument do kategorie zatříděn, tak tam klasifikátorem je zatříděn. To odpovídá českým názvům.

Technicky je nutné dořešit jeden problém, kterému se [Sebastiani (2002)] nevěnuje. Jak přesně spočítat přesnost, pokud jsou $TP$ i $FP$ rovné nule, tj. klasifikátor nevydal ani jedno pozitivní rozhodnutí? Já tento problém vyřešil takto — pokud jsou při výpočtu mikroprůměru $TP$ i $FP$ rovny nule — tj. klasifikátor nevydal *ani jedno* rozhodnutí, je $\pi_\mu$ rovno jedné, protože jsou jakoby všechna rozhodnutí správně. Stejně tak uvažuji ve výpočtu makroprůměru v počítání „malých“ přesností pro každou kategorii.

Ani přesnost, ani úplnost není možno brát jako samostatnou metriku — například $\rho = 1$ lze docílit triviálním klasifikátorem, který zatřídí každý článek do všech kategorií. K průměrování těchto dvou hodnot k dosažení jediné metriky se používá $F_1$-skóre, které je (například opět podle [Sebastiani (2002)]) definováno jako

$$F_1 = \frac{2\pi\rho}{\pi + \rho}.$$

Jelikož mám dvojí přesnost a úplnost, mám i dvojí $F_1$ skóre, jedno pro mikro- a jedno pro makroprůměr; označuji je $F_\mu$ a $F_M$. Tato dvě skóre tedy beru jako finální metriky kvality detekce kategorií.

### 1.5.2 Další metody

Jak jsem již uvedl v sekcích 1.1 a 1.2, tyto metody vrací spíše klíčová slova než témata. Je velmi obtížné pro kvalitu určování těchto klíčových slov navrhnout metriku, která by dávala smysluplné výsledky.

Zkusil jsem pro evaluaci těchto metod využít stejnou metriku, jako pro úlohu klasifikace — tj. použít ručně zatříděná data a výsledky srovnat s nimi. Úlohu evaluace na těchto metodách jsem si pro tyto účely upravil takto:

1. Slova, která tyto metody vrací, jsou názvy kategorií, do kterých je článek strojem zatříděn; tj. z metod, které nefungují na principu klasifikace, vyrobím svého druhu klasifikátory.

2. Naopak názvy ručně určených kategorií jsou často víceslovné; proto chápu pro jednoduchost shodu mezi ruční kategorií, skládající se ze slov $a_1, a_2 \ldots a_n$ a strojem určenou kategorií, skládající se ze slova $b$ tak, že *některé* ze slov $a_1$ až $a_n$ je rovno $b$.

3. Protože takto definované klasifikátory se nijak neučí, není potřeba dělit množinu článků na testovací a trénovací část a provádět cross-evaluaci — je možné *rovnou* spočítat $\pi$, $\rho$ a $F_1$, vše podle definic, uvedených výše.

Tento způsob evaluace má o něco volnější definici *true positive* a dále probíhá evaluace vlastně na mírně jiných datech, než cross-evaluace v případě klasifikátorů; proto nejsou výsledky této evaluace přímo srovnatelné s výsledkami evaluace a raději je tedy uvádím v části 3.4 zvlášť.

# 2. Problémy implementace

V této kapitole se pokusím rozebrat jednotlivé problémy, které úkol sledování témat v elektronickém zpravodajství má, a ukázat různé varianty jejich řešení. Zároveň s tím zmíním některé implementační problémy samotného programu.

## 2.1 Výběr serverů

Seznam webů ke sledování jsem se rozhodl vybrat ručně, čistě podle toho, jak „důvěryhodné“ mi připadaly — tj. převážně ty, které mají za sebou tištěné deníky, a navíc server `http://aktualne.cz`, který je v ostatních médiích také často citován.

Kromě „seriozních“ deníků jsem přidal pro úplnost také bulvární servery `blesk.cz` a `bleskove.cz`.

V souvislosti s dalším problémem jsem vybíral pouze takové servery, které seznam svých článků vydávají ve formátu RSS. Úplný seznam je uveden v Příloze A.

## 2.2 Získávání článků

Získávat články lze několika způsoby. Já zvolil následující dva.

### RSS

RSS je technologie, umožňující serverům vydávajícím pravidelný obsah oznamovat vydávání nových článků přes takzvané RSS kanály a čtenářům naopak umožňující tyto nové články sledovat.

Jelikož drtivá většina českých zpravodajských serverů RSS kanály podporuje, rozhodl jsem se tuto technologii použít ke stahování nových článků.

### Archiv

Z technických důvodů mi ovšem RSS kanály určitou dobu nefungovaly. Proto jsem do programu přidal další funkci — stahování článků přes archiv zpráv.

Všechny mnou sledované weby (kromě, bohužel, serveru iHNed.cz[1]) mají nějakou formu archivu, který je zdarma přístupný. Každý z těchto archivů je možno velmi jednoduše projít a články za období výpadku RSS stáhnout.

Je důležité si ale uvědomit, že články v archivu nejsou vždy shodné se články v RSS, což bude více vidět v kapitole 3.1. Proto se počet a struktura článků ve dnech, kde byly články stahovány první metodou, liší od dnů, kdy byly stahovány druhou metodou.

[1] Server iHNed.cz má sice starší články na webu zdarma přístupné, ale nemá přístupný seznam více než 300 článků v každé kategorii, což odpovídá přibližně dvěma měsícům.

## 2.3 Vyčištění článku

Jak dobře popisuje například [Yi et al. (2003)] či [Lin – Ho (2002)], kolem „skutečného“ obsahu článku, který nese sémantickou informaci, je také velké množství dalšího textu, například odkazy na jiné články, navigační prvky apod. Tyto HTML elementy sice uživateli pomáhají lépe se orientovat na stránce a v nejlepším případě jsou pro člověka jasně vizuálně oddělitelné od obsahu článku, stroj má ale k dispozici pouze zdrojový HTML kód a oddělení „skutečného“ textu článku od „zbytku“ je netriviální problém. Na tento problém existuje několik možností řešení.

### Vyčištění vynechat

První řešení se nabízí vyčištění článku úplně vynechat a (například pomocí programu `links`) jako obsah článku vzít vše, co je zobrazeno.

Jak správně podotýká například [Lin – Ho (2002)], článek je ale potom pro strojové zpracování nepoužitelný, jednak z toho důvodu, že se některé výrazy často opakují, a poté i proto, že délka samotného textu článku je často kratší, než délka okolních navigačních prvků.

### Šablony

Dalším řešením je nějak *a priori* popsat, které HTML prvky jsou opravdu obsahové, které naopak nejsou a tyto seznamy vytvořit pro každý zpravodajský zdroj zvlášť. Tomuto přístupu se někdy říká *používání šablon* (*templating*).

Jeho hlavní nevýhoda je zřejmá — je potřeba předem znát HTML strukturu stránek, které budeme čistit. Vzhledem k tomu, že můj projekt běžel delší dobu, po které se designy a názvy HTML elementů různě měnily, nebyl tento přístup příliš použitelný.

Lepším přístupem je řešit vyčištění článků až podle obsahu jednotlivých HTML DOM elementů bez předchozí znalosti struktury stránky (případně s použitím názvů tagů pouze jako doporučení pro automatické čištění). Z mnoha možností jsem nakonec zvolil následující.

### Vyčištění na základě hustoty odkazů

Rozhodl jsem se využít utilitu `Readability`, která je původně psána jako JavaScriptový browser plugin k pohodlnějšímu uživatelskému čtení internetových článků (a jako takovou ji používá například prohlížeč Safari pod názvem Safari Reader).

Utilita je o něco složitější, ale základní způsob, jakým určuje obsahové části od neobsahových, je poměr odkazů k běžnému textu. Čím větší je podíl odkazů, tím menší je pravděpodobnost, že se jedná o část s obsahem.

`Readability` bohužel nemá žádnou další dokumentaci, kterou bych lze mohl citovat, kromě komentářů ve zdrojovém kódu.[2]

[2]Zdrojový kód lze nalézt na adrese projektu `http://code.google.com/p/arc90labs-readability/`.

## 2.4 Kontinuální zpracování

V zadání práce je nepřímo zmíněno, že sledování zpravodajství a detekce témat bude probíhat dlouhodobě a kontinuálně.

To představuje mírný problém. Například TF-IDF váhová funkce se kontinuálně zlepšuje s tím, jak do databáze přibývají nové a nové články. Vyvstává otázka, zda je třeba s novými články, a tedy s novými inverzními frekvencemi, měnit zpětně články s už nalezenými tf-idf tématy. Algoritmus LSI pro SVD také potřebuje mít matici kompletní.

Nakonec jsem kontinuální zpracování vůbec neřešil — jak bude vidět v části 3.1, sbírání článků jsem v určitém momentě zastavil a úlohu detekce témat řešil pouze na dané množině. Tím jsem se mírně odchýlil od zadání.

## 2.5 Lingvistické předzpracování

### 2.5.1 Lemmatizace

V kapitole 1 jsem mluvil obecně pouze o slovech. Pro lepší výsledky při počítání frekvencí je dobré převést různé tvary stejného slova na kanonickou formu — buď prostřednictvím lemmatizace, nebo pomocí tzv. stemmingu. *Lemmatizace* (*lemmatization*) je převedení slova do základního tvaru, tzv. *lemmatu* — např. u podstatných jmen jde o nominativ singuláru.

V rámci Ústavu formální a aplikované lingvistiky existuje projekt TectoMT, v němž jsou už předinstalované dva lemmatizátory, pojmenované `TagHajic` a `MorCe`. Více informací lze zjistit na domovské stránce projektu, `http://ufal.mff.cuni.cz/tectomt/`.[3]

Pro lemmatizaci používám lemmatizátor `TagHajic` — sice vrací o něco horší výsledky, ale má menší paměťové nároky a pracuje o něco rychleji.

### 2.5.2 Pojmenované entity

Při detekci klíčových slov si ještě mírně pomáhám detekcí tzv. *pojmenovaných entit* (*named entities*), taktéž pomocí vestavěných modulů v `TectoMT`. Pojmenované entity jsou jména osob, míst, organizací a podobně.

Předpokládám, že pojmenované entity budou mít vyšší důležitost, než ostatní slova. Proto jim v rámci TF-IDF vážící funkce dám určitý „náskok“ — tj. že počet výskytů $n_{i,j}$, tedy i celou váhovou funkci, vynásobím koeficientem 2.

Pohlíženo zpět si nejsem zcela jist, jestli tento krok byl vhodný, protože je možné, že tyto entity by tak jako tak dostaly vyšší TF-IDF skóre. Na druhou stranu, jména osob a míst *jsou* často klíčová slova článku, a protože automatický pojmenovávač entit funguje na jiné bázi, než TD-IDF, je dobré tyto dva komplementární přístupy kombinovat.

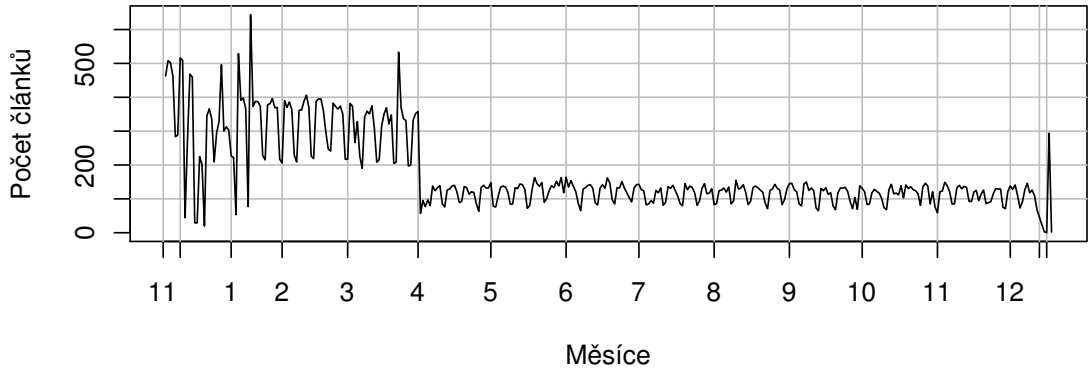
Každopádně jde pouze o teoretickou otázku, kterou jsem přímo neověřoval, pojmenovávač entit jsem nechával jako součást programu a ani koeficient jsem neupravoval.

[3] Pro úplnost dodám, že projekt TectoMT se nově jmenuje TreeX a bude ke stažení z archivu CPAN (`http://cpan.org`).

# 3. Experimenty

## 3.1 Sběr dat

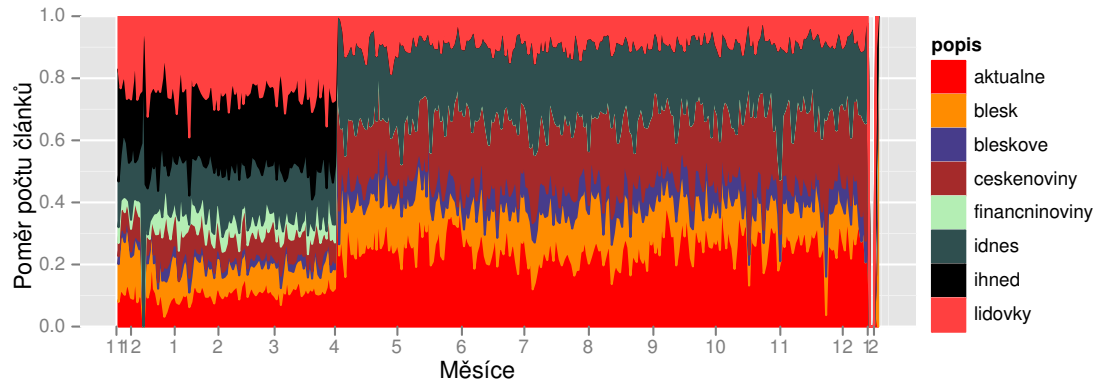
Články byly sbírány mezi prosincem 2009 a lednem 2011. Jak jsem již naznačil v sekci 2.2, část dat jsem sbíral přes RSS, druhou část přes archiv zpráv. Jasný časový předěl mezi těmito dvěma částmi byl 1. duben 2010 — do tohoto data se články sbíraly přes RSS, od tohoto data dále se články sbíraly přes archiv.




Graf 1: Počty článků v čase

První zajímavou veličinou je samotný počet článků, který jsem vynesl do Grafu 1. Na horizontální ose je vynesen čas, na vertikální počet článků. Na horizontální ose jsou pouze dny s nenulovým počtem článků; počáteční měsíc, kdy ještě systém zcela nefungoval, je tedy o něco kratší. Na horizontální ose jsou také vyznačeny měsíce (pro jednoduchost je vynecháno označení roků — většina grafu je rok 2010).

Kromě počátečních a závěrečných výkyvů, způsobených drobnými chybami v experimentu, stojí za pozornost jednak „pád“ v dubnu 2010, způsobený již popisovanou změnou způsobu sběru článků, a potom pravidelné týdenní fluktuace, způsobené menším množstvím zpráv o víkendech.




Graf 2: Poměr článků z jednotlivých zdrojů

Složení jednotlivých zdrojů v databázi jsem vynesl do Grafu 2. Co stojí za zmínku je fakt, že server iHNed jsem po 1. dubnu 2010 do databáze nestahoval, protože na severu iHNed.cz není k dispozici archiv; dále splynuly servery

| | | | | |
|---|---|---|---|---|
| Factum Invenio | průzkum | volby | ODS | ČSSD |
| KSČM | TOP 09 | KDU-ČSL | Věci Veřejné | domácí politika |

Tabulka 1: Moje kategorická témata pro Článek 1

| | | |
|---|---|---|
| Barack Obama | Indie | opice |

Tabulka 2: Moje kategorická témata pro Článek 2

České noviny a Finanční noviny (v grafu to není úplně zřetelné), protože jejich archivy jsou totožné.

## 3.2 Ruční zatřiďování

Pro evaluaci různých metod potřebuji množinu ručně zatříděných článků, jak popisuji v části 1.5. Jak popisuji v části 1.3.3, rozhodl jsem se pro dvě možnosti — u první může být článek ve více kategoriích, které vytvářím až při ručním třídění; u druhé jsem si předem zadal několik obecných kategorií, kde každý článek má tuto kategorii pouze jednu.

Proti kategoriím vzniklým první možností tedy testuji všechny metody (včetně těch, které nejsou přímo založeny na principu klasifikace); klasifikátory poté testuji i proti omezeným, obecným kategoriím, které popisuji v části 1.3.3.

Z celé databáze článků jsem vybral podmnožinu 200 článků, které jsem takto zatřídil. Jak se ukazuje, je ruční kategorizace do témat obtížná sama o sobě (možná *právě* i kvůli neomezenosti množiny témat).

Pro ilustraci jsem z těchto 200 článků vybral náhodně dva, na kterých ukážu problémy jak ručního zatřiďování v této části, tak konkrétní výsledky jednotlivých metod v části další. Jde o článek *Volby skončí patem, Zeman zůstane mimo* ze serveru Týden.cz ze dne 12. 5. 2010 a článek *Indie se chystá na Obamovu návštěvu: chytá opice a očesává kokosy* ze serveru iDnes.cz ze dne 3. 11. 2010; tyto články jsou na další straně.

Na Článku 2 lze demonstrovat, proč jsem zavrhl u neomezených tématických kategorií možnost, že by byla na článek pouze jedna. Jak zvolit jedinou kategorii, ve které by článek byl — byla by to `Indie`, `Obama`, `cesty prezidentů`, `opice`, `Taj Mahal`...?

I přes povolení více kategorií je ale úloha ruční kategorizace poměrně složitá. Úlohu „zvol jedno téma“ jsem změnil na úlohu „najdi všechna témata“ — o jakých všech tématech je ale Článek 2? Tím, že jsem si nedefinoval žádný přesnější rámec, může být tématem cokoliv, co je se článkem sémanticky spojeno, tj. `opice`, ale také např. `kabely`, `chystání státníků na zahraniční cesty`, `útoky přírody na člověka`,...

Ve skutečnosti jde s kategoriemi jít až absurdně „vysoko“ a abstraktně; nebo naopak velmi „nízko“ a konkrétně. V tom spočívá hlavní slabina mé snahy o ruční nalezení všech kategorických témat — je jich prostě až příliš mnoho.

V Tabulkách 1 a 2 jsou moje kategorie pro ilustrační články. Na nich lze demostrovat, že jsem se snažil být poměrně střízlivý; i přesto vyšlo v rámci celého ručního zatřiďování na 200 článků 324 kategorií, většina z nich (přesně 236) obsahuje ale pouze jeden článek!

Volby do Poslanecké sněmovny by podle volebního modelu agentury Factum Invenio z konce dubna vyhráli sociální demokraté s 27,5 procenta hlasů. ODS by volilo 21,7 procenta lidí. Ve sněmovně by zasedli i zástupci KSČM, TOP 09, Věcí veřejných a KDU-ČSL. Sociální demokraté by podle Factum Invenio sice byli vítězi voleb, jejich podpora ale podle agentury poklesla a k sestavení menšinové vlády by jim nadále nestačila podpora KSČM. ČSSD by nedosáhla na většinu ani ve spojení s KDU-ČSL a Věcmi veřejnými. Průzkum naznačuje patovou situaci, protože většinu by neměla ani pravicová koalice ODS, TOP 09 a Věcí veřejných. Podle prognózy posílila Strana práv občanů — zemanovci. Se ziskem 3,2 procenta hlasů by se ale podle agentury do sněmovny nedostala. Za branami parlamentu by zůstali i zelení s 2,9 procenta. Průzkum se odehrál mezi 23. a 28. dubnem na vzorku asi tisíc lidí. KSČM by podle modelu skončila třetí s 13,9 procenta hlasů. Těsně za komunisty by se umístila TOP 09 s 11,1 procenta a téměř stejného zisku by dosáhly Věci Veřejné s 11 procenty. Lidovci by těsně překonali hranici nutnou pro vstup do parlamentu s 5,2 procenta. ČTĚTE TAKÉ: Volební spoty: lepší dobře ukrást než špatně vymyslet V přepočtu na mandáty by ČSSD získala 66 sněmovních křesel, ODS 51, KSČM 30, TOP 09 22, Věci veřejné 23 a KDU-ČSL by disponovala osmi mandáty. Pro sestavení většinové vlády bez podpory komunistů by tak ČSSD i ODS potřebovaly účast všech tří menších stran — TOP 09, Věcí veřejných i KDU-ČSL. Výpočet mandátů je ale podle Factum Invenio zatížen výběrovou chybou, která může způsobit odchylku až dvou poslaneckých křesel. Koalici ČSSD, KDU-ČSL a Věcí veřejných přitom chybí k většině jen tři křesla a pravicovému uskupení čtyři mandáty. Jen čtyři poslanci by také chyběli k většině spojenectví ČSSD a KSČM. Foto: Jan Schejbal, ČTK

Článek 1: Volby skončí patem, Zeman zůstane mimo

Džungle naproti prezidentskému apartmá v hotelu Taj Mahal v Bombaji bude o víkendu jedním z nejstřeženějších míst na světě. Zvláštní komanda vybavená dalekohledy s nočním viděním budou dávat pozor na každé šustnutí. A to také kvůli opicím, které město v poslední době doslova terorizují. Podle deníku The Daily Telegraph makakové neustále přebíhají přes vládní pozemky, překusují kabely, napadají lidi nesoucí jídlo a celkově vyvolávají paniku. Přestože místní noviny pravidelně píší o škodách, které opice napáchaly, a o obětech jejich hladových útoků, úřady se dosud neměly k zásahu. Makak je totiž pro hinduisty symbolem opičího boha Hanumana. Kvůli Obamově návštěvě ale nemohou jinak, než přichystat potřebná opatření. Po zuby ozbrojené, speciálně cvičené protiteroristické jednotky tak doplní chytači opic. „Rozmístíme policejní komanda, odstřelovače a také lidi na chytání makaků, abychom zvýšili Obamovu bezpečnost," řekl jeden z policistů listu The Hindustan Times. Kokosy mizí z ulic Další hrozbu představují pro Obamu bombajské kokosovníky. Padající kokosy v Indii každoročně zraní, ale i zabijí několik lidí. Aby se to nestalo i Obamovi, mizí postupně suché ořechy z palem v místech, kde se bude americký prezident při páteční návštěvě pohybovat. Indové už tak očesali kokosy například u Gándhího muzea, uvedla britská BBC. Dvoupatrový komplex, zasvěcený otci indické nezávislosti Mahátmovi Gándhímu, stojí na jihu Bombaje a cesta k němu je lemována právě kokosovníky. V domě je kromě muzea i pokoj, ve kterém Gándhí žil sedmnáct let. Obama Gándhího otevřeně obdivuje a říká o něm, že pro něj byl inspirací. Jeho portrét si dokonce pověsil do senátorské pracovny. I proto chce jeho muzeum navštívit. Indické úřady budovu na tuto událost připravily, například ji kompletně zrekonstruovaly.

Článek 2: Indie se chystá na Obamovu návštěvu: chytá opice a očesává kokosy

| **lemma** | být | a | s | procento | podle | veřejný | kdu | čsl | 09 | ksčm |
|---|---|---|---|---|---|---|---|---|---|---|
| **četnost** | 19 | 10 | 8 | 8 | 7 | 7 | 5 | 5 | 5 | 5 |

Tabulka 3: Frekvenční témata na Článku 1

| **lemma** | být | a | v | na | se | gándhí | opice | i | o | kokos |
|---|---|---|---|---|---|---|---|---|---|---|
| **četnost** | 7 | 7 | 7 | 6 | 5 | 4 | 4 | 4 | 4 | 4 |

Tabulka 4: Frekvenční témata na Článku 2

| **lemma** | v | být | a | se | na | ten | že | z | on |
|---|---|---|---|---|---|---|---|---|---|
| **četnost** | 50619 | 50617 | 45068 | 44707 | 41691 | 15265 | 13904 | 11874 | 10437 |

| s | který | rok | o | mít | podle | do | za | i | svůj | ale |
|---|---|---|---|---|---|---|---|---|---|---|
| 9937 | 9597 | 7179 | 6930 | 5805 | 5727 | 5355 | 3814 | 3593 | 2452 | 2424 |

Tabulka 5: Nejčetnější frekvenční témata

Zatřiďování v rámci obecných kategoriích je naopak mnohem jednodušší, rozhodování o ručním zatřídění trvá výrazně kratší dobu. Ke Článku 1 jsem přiřadil kategorii `Politika domácí`, ke Článku 2 jsem přiřadil kategorii `Politika svět`.

## 3.3 Detekce témat a trendů

V této kapitole chci ukázat výsledky jednotlivých přístupů k detekci témat ve článcích a ukázat na nich jejich výhody a nevýhody, které zkusím demonstrovat na dvou ukázkových článcích z předchozí kapitoly. Jedním z úkolů bakalářské práce bylo i hledání něčeho, jako je „okurková sezóna“; pokouším se tedy i zjistit, jestli vynesením některých veličin na časovou osu nějaké zajímavé trendy nenaleznu.

### 3.3.1 Frekvenční témata

Prvním pokusem o definování témat bylo jednoduché počítání frekvencí, bez zavedení stop slov. Jak jsem již popsal v části 1.1, takto nalezená témata jsem nazval frekvenčními tématy; počet frekvenčních témat na jeden článek jsem určil jako 10.

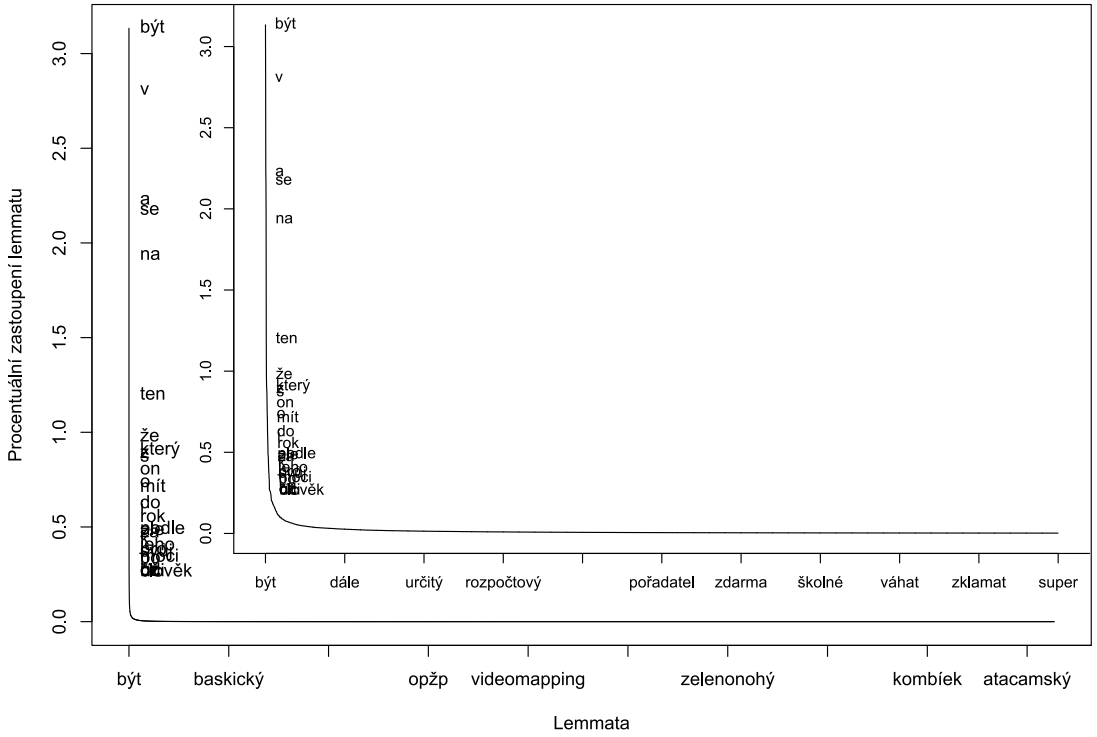
Jak vyjdou frekvenční témata na ukázkových článcích ukazují Tabulky 3 a 4; nejčastější frekvenční témata, posčítaná na celém korpusu, jsou vidět v Tabulce 5.

Jak je z těchto tabulek vidět, nemá toto počítání témat přílišný smysl — nalezenými frekvenční tématy jsou většinou právě stop-slova, tedy pomocná slova bez jakéhokoliv významu.

### 3.3.2 Stop-témata

Stop-tématy jsem si v 1.1 nazval témata, počítána stále „hloupě“ přes frekvence, ale s vynecháváním stop-slov.

Stop-slova jsem neurčil „ručně“, ale rozhodl jsem se je nalézt algoritmicky. K tomu je dobré vědět, která slova se jak často vyskytují a jakou část textu zabírají — proto jsem nejdříve zjistil frekvence všech lemmat v celém textu.




Graf 3: Četnosti všech lemmat a nejčastějších 5000 (výřez)

| **počet lemmat** | 4 | 15 | 49 | 142 | 350 | 752 | 1532 | 3260 | 8666 |
|---|---|---|---|---|---|---|---|---|---|
| **zastoupení** | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% |

Tabulka 6: Počet lemmat pro násobky 10%

Pokud se frekvence všech přibližně 185 tisíc typů lemmat zakreslí do grafu, stoupání je tak rychlé, že připomíná spíše pravý úhel; pokud se ale zakreslí pouze prvních 5 tisíc typů, je stoupání o něco pozvolnější. Oba grafy jsem pro úsporu místa nakreslil přes sebe do Grafu 3 — „vnější“ graf je všech typů, „vnitřní“ pouze nejčetnějších 5 tisíc.

Na svislé ose je vyneseno procentuální zastoupení lemmatu v celé databázi, na vodorovné ose jsou lemmata, seřazena dle četnosti. Pro ilustraci je na vodorovné ose každé dvacetitisícé slovo označeno. Dále je pro lepší ilustraci v grafu označeno prvních 15 nejčetnějších lemmat. Vzhledem k tomu, jak procentuální zastoupení rychle roste, je levá část Grafu 3 v podstatě zcela svislá čára a pravá část naopak zcela vodorovná. Většina této vodorovné části jsou navíc všechno slova s četností 1; tato slova jsou poté seřazena náhodně — příklady, uvedené na pravé části vodorovné osy, tak nemají přílišný smysl, všechna jsou v korpusu právě jednou.

Ke zjištění vhodného počtu lemmat pro množinu stopslov jsem pro $k$ od 1 do 9 zjistil, kolik nejméně typů lemmat dá dohromady $10k$ či více procent textu; tyto hodnoty jsou v Tabulce 6.

Je otázka, jak velkou vzít množinu stopslov — pokud bude příliš velká, můžeme ignorovat i něco, co je pro článek důležité; pokud bude příliš malá, do stop-témat se můžou dostat i nerelevantní lemmata. Nakonec jsem se rozhodl vzít jako stopslova 200 lemmat, která potom zabírají 43.64 procent veškerého textu.

Již zde je ale zjevný problém tohoto přístupu k detekci témat — například slovo `ods` (lemma od názvu strany ODS) je už 44. nejčetnější — dokonce četnější,

| **lemma** | kdu | čsl | 09 | ksčm | mandát | factum | většina | invenio | 2 | hlas |
|---|---|---|---|---|---|---|---|---|---|---|
| **četnost** | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 3 | 3 | 3 |

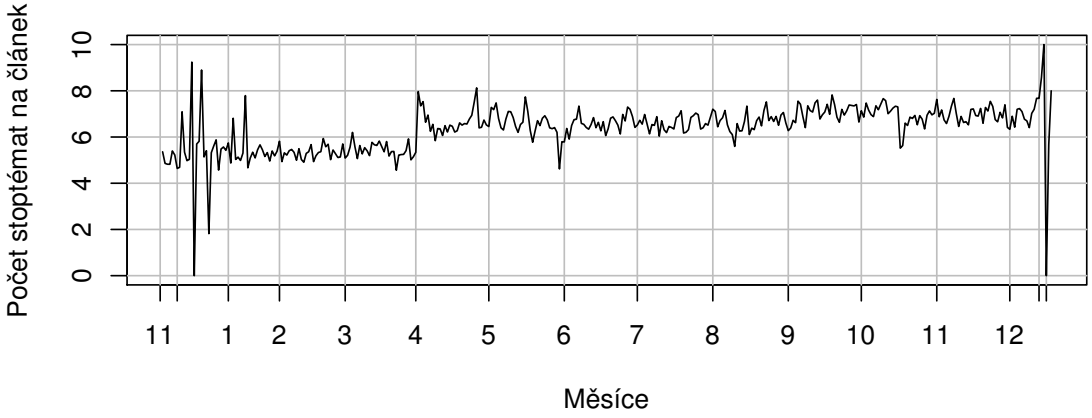
Tabulka 7: Stop-témata pro Článek 1

| gándhí | opice | kokos | makak | návštěva | muzeum | indický | mizet | the | kokosovník |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 4 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 2 |

Tabulka 8: Stop-témata pro Článek 2

| **lemma** | nehoda | nemocnice | řidič | 09 | 1 | dům | * | . | skupina | auto |
|---|---|---|---|---|---|---|---|---|---|---|
| **četnost** | 1721 | 1635 | 1634 | 1566 | 1551 | 1532 | 1496 | 1291 | 1286 | 1269 |

Tabulka 9: Nejčetnější stop-témata




Graf 4: Průměrný počet stop-témat na článek

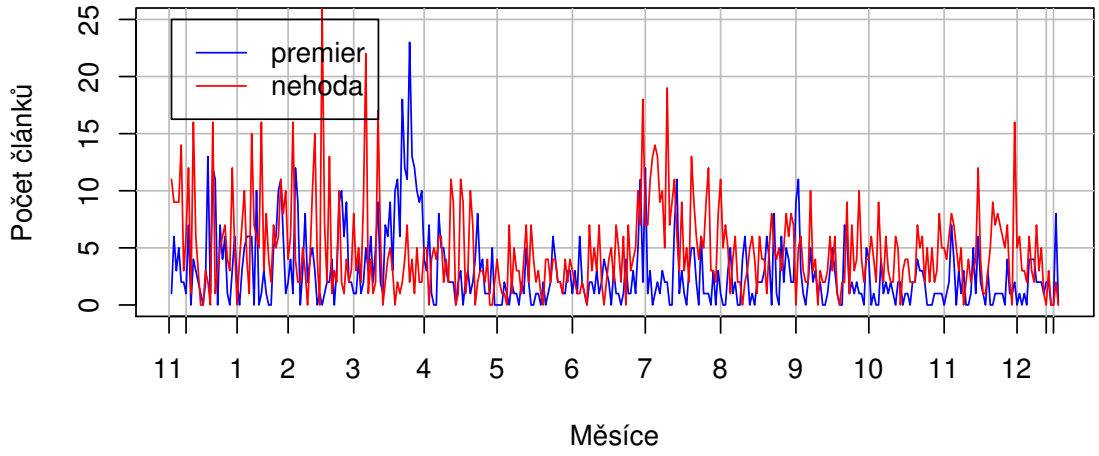
než slova nebo, aby, než, takový apod. Stejně jako ČSSD je už 66. nejčetnější. A v mé množině 200 nejčastějších lemmat je ještě například také slovo nečas, které je na 195. místě.

Velikost počtu stop-témat na článek jsem určil jako 10. Výsledky pro ukázkové články jsou v Tabulkách 7 a 8, posčítaná stop-témata na celé databázi jsou v Tabulce 9. K té ještě uvedu, že lemma 09 nejspíše patří k názvu politické strany TOP09 a lemmata . a * jsou způsobena mírnými chybami v lemmatizátoru; . ovšem neznamená tečku na konci věty, kterou systém úspěšně ignoruje, ale tečku v internetových adresách.

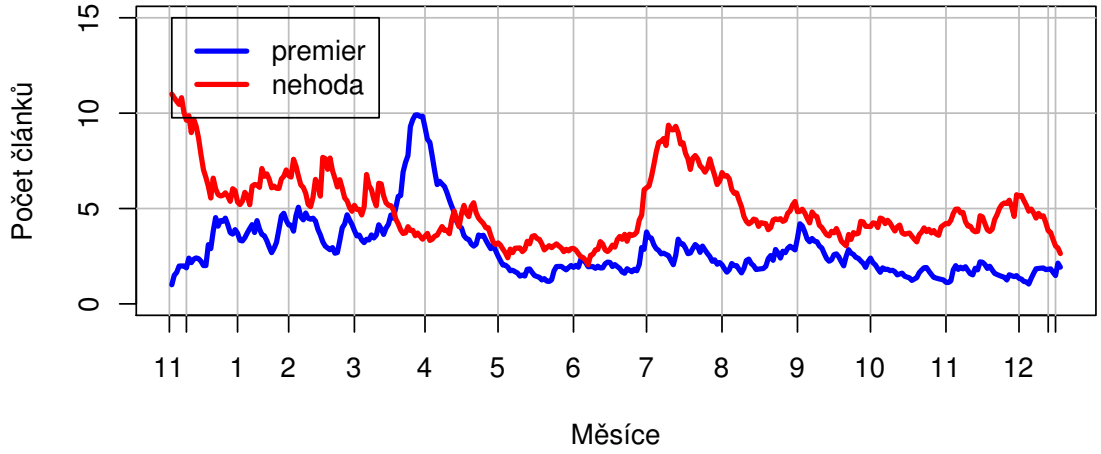
Jak je vidět, stop-témata stále nejsou *tématy* v pravém slova smyslu; je zřejmé, že na ukázkových článcích jde stále spíše o množinu klíčových slov. Například nelze říct, že by Článek 1 byl „o většině“ nebo druhý článek „o Gándhím“; na druhou stranu už lze říci, že druhý článek je „o opicích“, takže stop-témata zcela bez relevance také nejsou.

Zkusím nyní nějak na základě vývoje stop-témat v čase detekovat „okurkovou sezónu“ či nalézt jiné související trendy; prvním nápadem je vyzkoušet průměrný počet stop-témat na jeden článek.

Jeho vývoj je vidět v Grafu 4. Největší „skoky“ zde způsobují jednak nedokonalé ukládání článku na začátku a na závěr a potom dubnová změna způsobu stahování článků. Kromě toho se průměrný počet stop-témat mění, ale jedná se dle mého názoru o změnu spíše náhodnou. Tato veličina nám tedy nic příliš zajímavého neukazuje.




Graf 5: Graf četnosti stop-témat `nehoda` a `premiér`




Graf 6: Vyhlazený graf četnosti stop-témat `nehoda` a `premiér`

Zkusil jsem si do grafů zobrazit vývoj četnosti stop-témat, která se opakují často a u kterých by tento vývoj mohl být zajímavý — nejdříve jsem si vybral slova `nehoda` a `premiér`. Jak je ale vidět na Grafu 5, změny jsou na čistém grafu příliš rychlé a obtížně interpretovatelné.

Pro lepší interpretaci dat jsem v grafech použil tzv. exponenciální vyhlazování, definované například v [NIST/SEMATECH (2010)] takto: pokud je původní sledovaná veličina $y$ závislá na čase, potom vyhlazená $S$ je definována jako

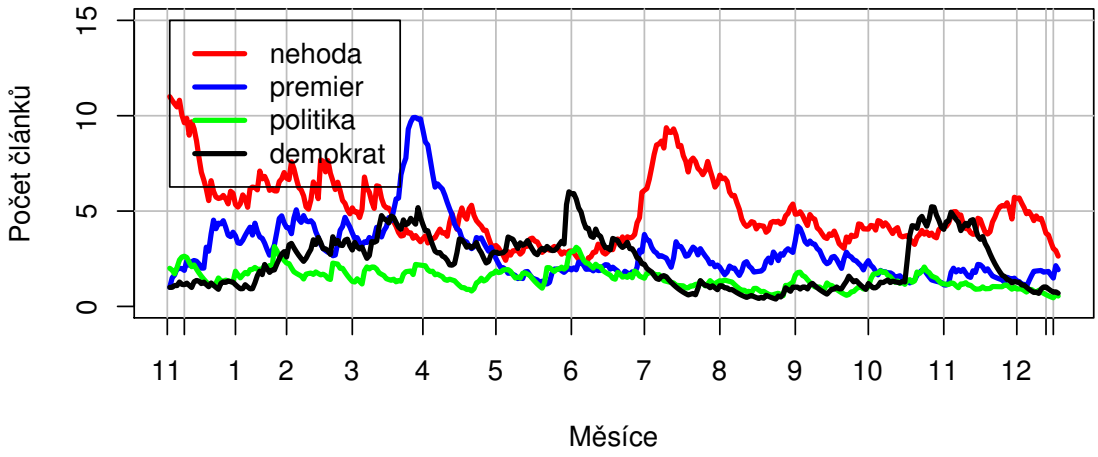
$$S_0 = y_0,$$

$$S_t = \alpha y_t + (1 - \alpha) y_{t-1}.$$

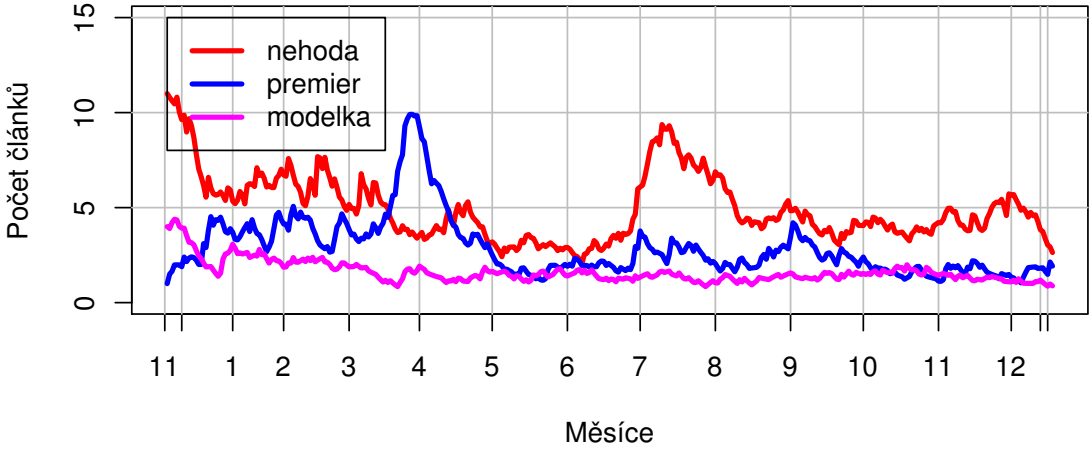
$\alpha$ je konstantní. Vyhlazení má samozřejmě svá rizika a nevýhody, ale dovolil jsem si ho zde použít, jelikož na základě vyhlazených dat k žádným „velkým“ závěrům nedocházím.

Vyhlazená data (s koeficientem $\alpha = 0.1$) jsou vidět v Grafu 6. Je evidentní, že zavedení vyhlazování udělá výsledky o něco „čitelnější“, na druhou stranu ale například zmenšuje maxima.

Je ale například vidět, že v březnu 2010 se stop-téma `premiér` stává častějším. Při prozkoumání archivu zjistíme, že v této době pronesl předseda ODS Mirek




Graf 7: Vyhlazený graf různých stop-témat




Graf 8: Vyhlazený graf různých stop-témat

Topolánek urážlivé výroky na adresu premiéra Fišera, tedy zvýšení četnosti tohoto stop-tématu nejspíše nebylo náhodné. Je ale otázkou, jestli bylo pro podobné zjištění nalézání stop-témat vůbec potřeba; možná by stačilo uvažovat počty *jakýchkoliv* výskytů slova `premiér` ve článcích a získali bychom podobné, možná lepší křivky.

Stejně tak nemůžeme četnosti článků s tématem `premiér` nějak zobecnit a tvrdit, že v té době se píše víc o politice a naopak, že `nehoda` znamená, že se nemá „o čem psát“ a jedná se pouze o „výplň“. Pokud si totiž přidáme do grafu ještě například četnosti článků se stop-tématem `politika` a `demokrat`, které by v případě obecnějšího trendu mělo nějak téma `premiér` následovat (Graf 7, už pouze vyhlazená data), a naopak pokud si přidáme ještě téma `modelka`, které má zase bulvární nádech (Graf 8, také pouze vyhlazená data), tak zjistíme, že trendy, které by nějak určovaly „okurkovou sezónu“, na těchto dílčích stop-tématech vysledovatelné nejsou.

### 3.3.3 tf-idf-témata

Jak jsem popsal výše, tf-idf-témata hledám přes váhovou funkci TF-IDF.

| slovo | TF-IDF |
|---|---|
| factum | 0.08 |
| invenio | 0.06 |
| čsl | 0.06 |
| kdu | 0.06 |
| ksčm | 0.06 |
| procento | 0.05 |
| veřejný | 0.05 |
| mandát | 0.04 |
| top | 0.04 |
| duben | 0.04 |
| čssd | 0.03 |
| křeslo | 0.03 |
| sestavení | 0.03 |
| věc | 0.03 |
| většina | 0.03 |
| hlas | 0.03 |
| ods | 0.02 |
| podpora | 0.02 |
| pravicový | 0.02 |
| sněmovna | 0.02 |

| slovo | TF-IDF |
|---|---|
| kokos | 0.12 |
| gándhí | 0.11 |
| opice | 0.09 |
| makak | 0.09 |
| kokosovník | 0.08 |
| bombaj | 0.05 |
| obamov | 0.05 |
| muzeum | 0.05 |
| překusovat | 0.04 |
| hanuman | 0.04 |
| očesávat | 0.04 |
| mizet | 0.04 |
| komando | 0.04 |
| návštěva | 0.04 |
| očesat | 0.04 |
| chytač | 0.04 |
| šustnutí | 0.04 |
| hindustan | 0.04 |
| indie | 0.03 |
| indický | 0.03 |

Tabulka 10: Tf-idf-témata na ukázkových článcích

| **lemma** | ods | čssd | soud | procento | volba | blesk | strana | nehoda | * | nečas |
|---|---|---|---|---|---|---|---|---|---|---|
| **četnost** | 2960 | 2606 | 2306 | 1920 | 1893 | 1717 | 1711 | 1471 | 1437 | 1397 |

Tabulka 11: Nejčetnější tf-idf-témata

Výsledek na ukázkových článcích (s počtem témat na článek $k = 20$) je popsán v Tabulce 10. Výsledky jsou opticky ještě relevantnější, než stop-témata (i pokud srovnáme pouze prvních 10 tf-idf-témat — naznačeno dvojitou čarou — s tabulkami 7 a 8), ale stále je lepší je vnímat spíše jako klíčová slova než jako skutečně témata. Jen dodám, že `obamov` je chybně zlemmatizované slovo Obama.

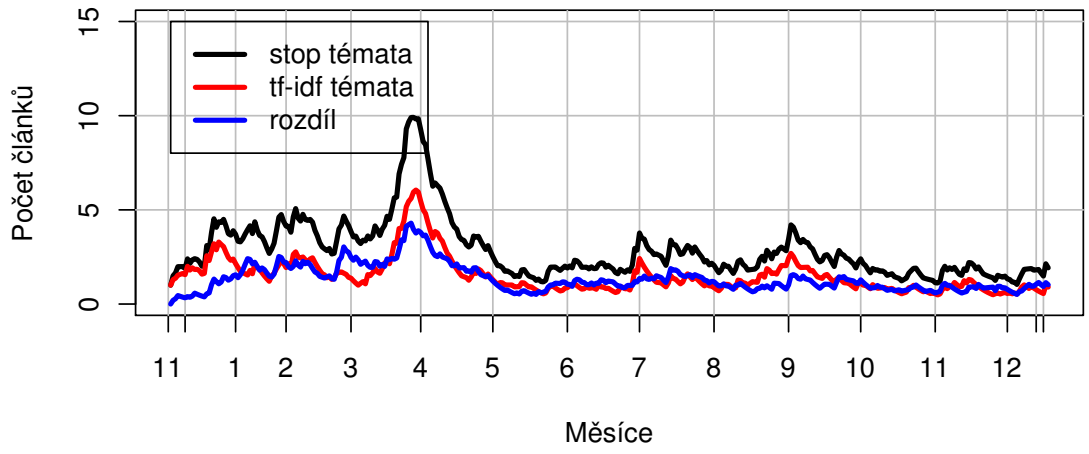
Z výsledků je zatím vidět, že TF-IDF dává sice dobré výsledky, ale dají se spíše chápat jako klíčová slova než jako opravdová témata — například v Tabulce 10 je vidět, že ve Článku 1 jsou témata jako `duben` nebo `většina`.

Nejčetnější tf-idf-témata (stále s počtem $k = 20$) na celé databázi jsou v Tabulce 11 (je nutné dodat, že téma `volba` je zlemmatizovaná forma slova `volby`). Je zajímavé, že do první desítky se dostal i `blesk`, přitom další názvy deníků jsou mnohem méně četné; možná je to proto, že ve článcích Blesku je mnohem častěji zmiňován samotný název novin, než v serioznějších denících.

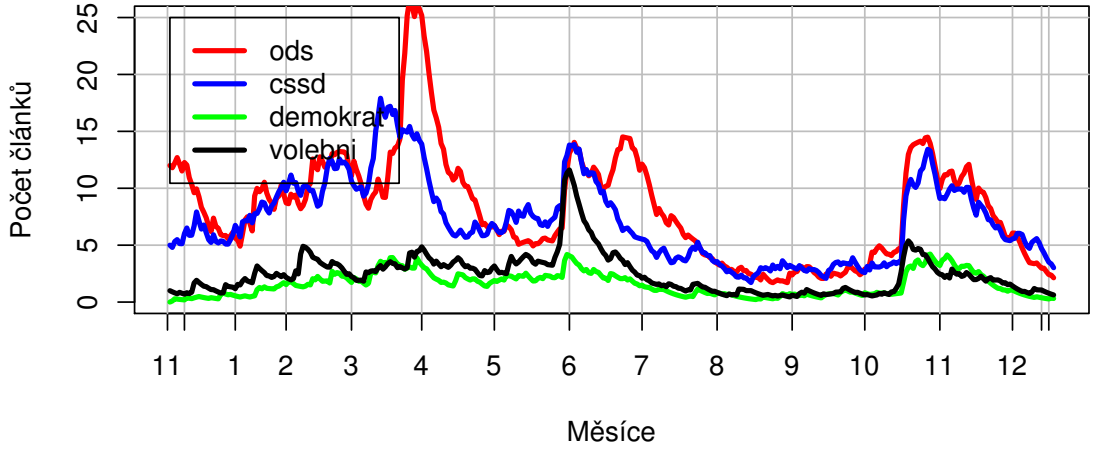
Ač na jednotlivých článcích dává tento algoritmus opticky lepší výsledky, pokud se srovná vývoj četností tf-idf-témat a stop-témat, je zřejmé, že trendy jsou velmi podobné. V Grafu 9 jsem srovnal četnosti tématu `premiér` (četnosti i rozdíl jsou vyhlazené).

Co ale tentokrát můžeme udělat je podívat se „zblízka“ na témata `ods` a `čssd`, která byla jako stoptémata vyřazena, protože jsou to příliš častá slova. Četnost těchto témat jsem zanesl do Grafu 10, spolu s dalšími politickými tématy. Je zajímavé, že témata `ods` a `čssd` mají maxima a minima v podobných místech; další dvě politická témata tato témata částečně kopírují.

Zajímavé je, že všechna tato témata „padají“ právě okolo 1. dubna 2010,




Graf 9: Graf rozdílu tf-idf a stoptématu `premiér`

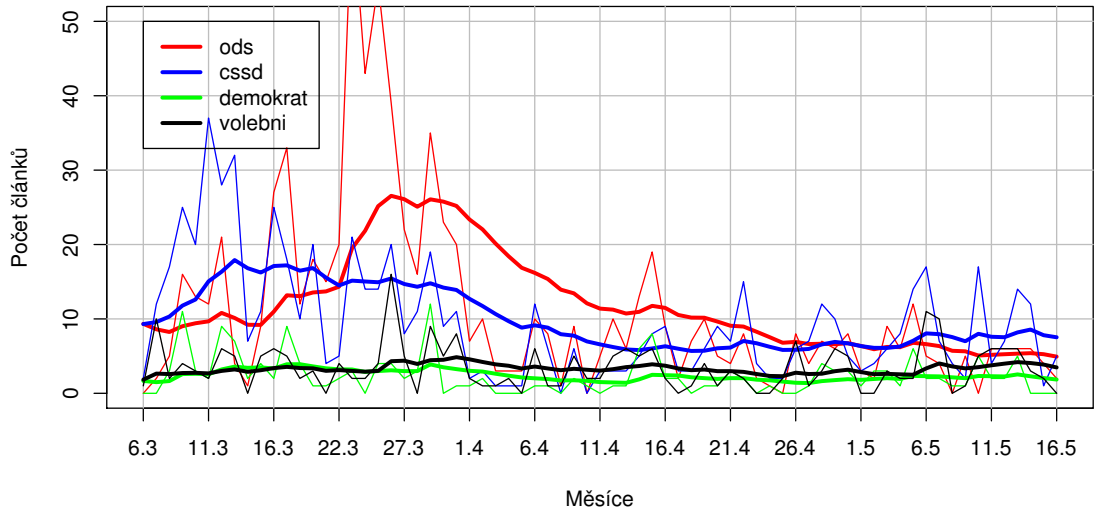



Graf 10: Četnosti politických tf-idf témat

což je datum změny počtu a struktury článků. V Grafu 11 jsem přiblížil blíže na dny kolem tohoto „propadu“ (mimochodem, v přiblížení jsou dobře vidět nepřesnosti exponenciálního vyhlazení). Například heslo `ods` mělo do 1. dubna téměř konstantně nad 10 článků denně, zatímco od 1. dubna naopak nad 10 článků za den téměř nedosáhlo. Spolu s tím, že 1. dubna se žádná významná změna nekonala si spíše myslím, že tento pokles je opravdu způsoben odlišným způsobem stahování článků.

Zajímavý je ale i další pohyb — začátkem (a u `ods` i v průběhu) června byla politická témata velmi častá, od července do poloviny října ale jejich četnost velmi poklesla, aby v polovině října opět prudce stoupla.

Je otázka, jak tyto pohyby interpretovat. Je možné, že opravdu jde o hledanou „okurkovou sezónu“ — je ale možné, že vrchol četnosti na konci května 2010 může být způsoben tím, že 28. a 29. května se konaly volby do Poslanecké sněmovny Parlamentu České republiky a za normálních okolností by četnost politických zpráv v květnu byla nižší. Pro opravdové zjištění, jestli se opravdu jedná o sezónní trend, by asi bylo třeba mít databázi, která pokrývá ještě delší časové období; podobnou databázi ale bohužel nemám k dispozici.




Graf 11: Propad četnosti politických tf-idf témat v dubnu 2010

| **Článek 1** | 10V | *V | 10O | *O |
|---|---|---|---|---|
| NaiveBayes | `domácí politika` | `domácí politika`<br>`volby`<br>`soud`<br>`nehoda` | `politika d` | `politika d` |
| SVM | | | `politika d` | `počasí d` |
| DecisionTree | `ODS`<br>`ČSSD`<br>`průzkum` | | `politika d` | `politika d` |

| **Článek 2** | 10V | *V | 10O | *O |
|---|---|---|---|---|
| NaiveBayes | `nehoda`<br>`domácí politika` | `domácí politika`<br>`nehoda` | `politika d` | `politika d` |
| SVM | | | `počasí d` | `počasí d` |
| DecisionTree | | `Barack Obama`<br>`armáda` | `počasí d` | `počasí d` |

Tabulka 12: Výsledky klasifikátorů na Článcích 1 a 2

### 3.3.4 Klasifikace

Můžeme ještě vyzkoušet tři algoritmy na klasifikaci, jak popisuji v 1.3 — algoritmy NaiveBayes, SVM a DecisionTree. Každý z těchto tří klasifikátorů lze naučit na množině 200 ručně zatříděných článků, jejichž tvorba je popsána v části 3.2.

Protože chci ale demonstrovat klasifikaci na dvou článcích, které *už jsou* v této množině, tak pro tuto konkrétní demonstraci jsem tyto dva články z trénovací množiny vyjmul, klasifikátory jsou tedy natrénovány pouze na 198 článcích.

Klasifikaci jsem, kromě rozdělení podle klasifikačních algoritmů, rozdělil ještě na 4 varianty:

- *10V* nazývám tu variantu, kdy jako rysy článků beru 10 nejcharakterističtějších slov podle TF-IDF, všechny s konstantní vahou 1, a články zatřiďuji do „neomezených“ kategorií, jak popisuji v 1.3.3,
- **V* nazývám tu variantu, kdy jako rysy nechám *všechna* slova s jejich vahou spočítanou podle TF-IDF a články opět zatřiďuji do „neomezených“

kategorií, jak popisuji v 1.3.3,

- *10O* nazývám obměnu varianty 10V s tím, že zatřiďuji do „omezených“, obecnějších kategorií,
- **O* nazývám obměnu varianty *V s tím, že opět zatřiďuji do „omezených“ kategorií.

Konkrétní výsledky jsou uvedeny v Tabulce 12 - pro varianty O jsem části názvů kategorií `domácí` a `svět` zkrátil na `d` a `s`, ale, jak je vidět, klasifikátory vracely pouze kategorie `domácí`. Prázdné políčko tabulky znamená, že klasifikátor nevydal ani jedno rozhodnutí.[1]

Zde je také dobré zmínit, že ze všech klasifikačních algoritmů je suverénně nejrychlejší algoritmus NaiveBayes, a to řádově.

První článek klasifikátory nemají, asi díky vysokému počtu slov, souvisejících s politikou, problémy zatřídit do kategorie `domácí politika` (pokud už nějaké rozhodnutí dají), ačkoliv NaiveBayes *V ještě přidá příliš nesouvisející `soud` a `nehodu`. Na druhém článku mají naopak klasifikátory problém — kromě kategorie `Barack Obama`, kterou DecisionTree *V určí správně, jsou všechny klasifikátory „mimo“ — varianty O ho dokonce z nějakého důvodu přidávají do kategorie `domácí počasí`, která je zcela špatně, naopak ani jednou nebyl článek vložen do kategorií, souvisejících se světovou politikou.

Na těchto dvou článcích se dá demonstrovat to, že klasifikátory mají *někdy* tendenci dávat správné výsledky a zatřiďovat do správných kategorií, ale celkově nejsou výsledky příliš přesvědčivé. Může to být i tím, že 200 článků opravdu není příliš, ale, jak jsem se snažil ukázat v části 3.2, ruční zatřiďování je samo o sobě poměrně náročné.

Rozhodl jsem se, že klasifikátory nebudu testovat na celé databázi a výsledky nebudu vynášet na časovou osu tak, jako v předcházejících částech — jednak kvůli tomu, že trénovací množina je velká pouze 200 článků, kdežto celkový počet článků je 63735, tedy přibližně 300-krát větší; poté kvůli tomu, že (jak bude vidět v části 3.4) klasifikátory nedávají až tak dobré výsledky; a nakonec proto, že, kromě algoritmu NaiveBayes, jsou klasifikátory opravdu velmi pomalé.

## 3.4 Evaluace jednotlivých přístupů

Jednotlivé přístupy jsem porovnal metrikou, popisovanou pro klasifikátory v části 1.5.1 a rozšířenou na další metody (zde jsem si je nazval *neklasifikační*) v části 1.5.2. Protože evaluace je provedena u neklasifikačních metod rozdílně, nedají se oba přístupy přímo porovnávat a proto uvádím výsledky zvlášť.

## 3.5 Neklasifikační metody

Metrikou, popsanou v 1.5.2, jsem porovnal různé způsoby detekce témat; výsledky jsou uvedeny v Tabulce 13.

[1] Žádné rozhodnutí znamená, že všechny kategorie měly příliš malou pravděpodobnost.

| | $\rho_\mu$ | $\pi_\mu$ | $F_\mu$ | $\rho_M$ | $\pi_M$ | $F_M$ |
|---|---|---|---|---|---|---|
| Triviální | 1.67 | 4.50 | 2.44 | 0.31 | 0.01 | 0.03 |
| Frekvenční témata | 28.49 | 7.97 | 12.46 | 13.79 | 13.04 | 13.41 |
| Stop témata | 39.12 | 11.10 | 17.29 | 10.58 | 9.74 | 10.14 |
| TF-IDF, $k = 10$ | 41.40 | 11.83 | 18.40 | 9.80 | 9.28 | 9.54 |
| TF-IDF, $k = 20$ | 51.72 | 7.60 | 13.25 | 6.79 | 5.95 | 6.34 |

Tabulka 13: Přesnosti a úplnosti neklasifikačních metod, v procentech

#### Triviální kategorizace

Pro srovnání s dalšími metodami jsem otestoval i „triviální kategorizaci“, která každému článku přiřadí právě jednu kategorii, a to `ODS`. Výsledky evaluace jsou, dle očekávání, téměř nulové.

#### Frekvenční témata

Frekvenční témata dávají možná lepší výsledky, než by se zdálo ze zběžného prohlížení nalezených témat, jak jsem činil v části 3.3.1. To, že úplnost je téměř 30 %, zatímco přesnost je méně než 10 %, je nejspíš tím, že metoda vrací vždy 10 témat na článek, zatímco průměrný počet témat na článek v ruční kategorizaci je nižší. Zmenšením počtu témat na článek bychom zvýšili přesnost, ale snížili úplnost, jak je vidět i dále na části s TF-IDF.

#### Stop témata

Je zajímavé, že zatímco „opticky“ a v mikroprůměru jsou výsledky stop témat lepší, v makroprůměru se výsledky zhorší.

U tohoto jevu si nejsem schopen vysvětlit příčiny; makroprůměr se ale chová (u $\pi$ i $\rho$) tak, že většina kategorií má nahoře ve zlomku 0, ze zbylých kategorií má zase naopak většina v čitateli i jmenovateli 1, tj. vyskytne se pouze jednou a tehdy je zatřídění úspěšné. Možná zde by mohlo ležet vysvětlení.

#### TF-IDF

Zvýšením počtu témat se zvedla, dle očekávání, úplnost, ale snížila přesnost. Pokud ale necháme počet témat na článek stejný, jako u stop-témat a frekvenčních témat, všechny výsledky v mikroprůměru se zlepší. Naopak se opakuje situace, kdy i přes „opticky“ lepší výsledky a lepší mikroprůměr dostáváme u makroprůměru horší výsledky.

## 3.6 Klasifikační metody

Varianty u klasifikátorů jsou shodné s variantami, definovanými v části 3.3.4.

Lze říci, že výsledky dopadají velmi tristně. Vždy dopadá lépe varianta 10 než varianta *; taktéž je ze všech klasifikátorů nejúspěšnější NaiveBayes (který má také výhodu, že je nejrychlejší).

Taktéž dopadá lépe verze O, kdy kategorie jsou velmi obecné. Je ale otázka, zda by na těchto kategoriích šly sledovat trendy a jestli by naopak nebyl poměr

| | $\rho_\mu$ | $\pi_\mu$ | $F_\mu$ | $\rho_M$ | $\pi_M$ | $F_M$ |
|---|---|---|---|---|---|---|
| NaiveBayes - 10V | 13.18 | 17.13 | 14.85 | 9.01 | 5.85 | 7.04 |
| NaiveBayes - *V | 5.36 | 5.69 | 5.50 | 4.02 | 1.19 | 1.75 |
| SVM - 10V | 2.17 | 90.00 | 4.16 | 1.72 | 1.94 | 1.81 |
| SVM - *V | 0.00 | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| DecisionTree - 10V | 11.57 | 33.55 | 16.89 | 7.79 | 8.21 | 7.93 |
| DecisionTree - *V | 9.00 | 13.93 | 10.77 | 5.34 | 5.25 | 5.25 |
| NaiveBayes - 10O | 48.50 | 48.50 | 48.50 | 30.05 | 25.88 | 27.49 |
| NaiveBayes - *O | 31.50 | 31.50 | 31.50 | 12.69 | 7.92 | 9.40 |
| SVM - 10O | 16.00 | 16.00 | 16.00 | 13.75 | 14.90 | 13.33 |
| SVM - *O | 6.00 | 6.00 | 6.00 | 7.26 | 3.71 | 4.07 |
| DecisionTree - 10O | 24.50 | 24.50 | 24.50 | 16.96 | 18.75 | 17.41 |
| DecisionTree - *O | 33.50 | 33.50 | 33.50 | 22.43 | 23.83 | 22.82 |

Tabulka 14: Přesnosti a úplnosti klasifikátorů, v procentech

těchto kategorií v čase stále konstantní; s jistotou ale nevím a tak můžu pouze spekulovat. Možná, že by bylo lepší se opravdu soustředit pouze na takto definované kategorie, jelikož zatřiďování do nich je rychlé a klasifikátory nemají až *tak* špatné výsledky.

Jen dodám, že $\pi_\mu = 1$ u varianty SVM-*V proto, že klasifikátor nevydá ani jedno rozhodnutí.

# Závěr

V práci jsem v části 1 uvedl několik způsobů detekce zpravodajských témat. Některé z nich jsem vyhodnotil, a to jak pomocí demonstrace na několika ukázkových článcích, tak přesně definovaným porovnáním oproti ručnímu zatřiďování. Konkrétně šlo o algoritmy počítání frekvencí slov, TF-IDF a klasifikátory se strojovým učením.

V části 3.4 jsem prezentoval výsledky porovnání těchto algoritmů s ručním zatřiďováním. Je otázka, jestli je možno tyto výsledky interpretovat jako dobré. Přiznávám se, že si nejsem zcela jist.

Jak jsem se snažil částečně ukázat už v části 3.2, problematickým vnímám hlavně zatřiďování do ručních kategorií, které je nejasně definované, proto je jednak poměrně náročné a také je evaluace proti němu problematická, stejně jako je problematické na těchto kategoriích učit algoritmy, založené na principu klasifikace se strojovým učením.

Pokud bychom trvali na požadavku metriky, určené jako shody s lidským zatřiďováním, bylo by asi nutné zlepšit definici kategorií pro ruční zatřiďování — kromě velmi obecných kategorií, uvedených v části 1.3.3, se mi ale nepodařilo žádnou lepší definici nalézt.

Pokud bychom od požadavku metriky, založené na shodě s lidským zatřiďováním, upustili, bylo by možná dobré dále zkoumat metody, popsané v části 1.4 — osobně mi přišly velmi zajímavé metody, vycházející z Latent Dirichlet allocation.

Dále se mi v části 3.3 podařilo určité pohyby v počtu a složení detekovaných témat nalézt; nejsem si ale zcela jist, zda jde o projevy nějakého hlubšího trendu, zda tyto pohyby nesouvisí spíše s jednotlivými událostmi, nebo zda se nejedná o pohyby čistě náhodné.

# Seznam použité literatury


[Andrews – Fox (2007)] ANDREWS, N. O. – FOX, E. A. Recent Developments in Document Clustering. Technical report, Blacksburg, VA, USA, 2007. Dostupné z: `http://eprints.cs.vt.edu/archive/00001000/01/docclust.pdf`.

[Blei et al. (2003)] BLEI, D. M. – NG, A. Y. – JORDAN, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. Leden 2003, 3, s. 993–1022. Dostupné z: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.80.7000&rep=rep1&type=pdf`.

[Deerwester et al. (1990)] DEERWESTER, S. et al. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*. 1990, 41, 6, s. 391–407. Dostupné z: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.1152&rep=rep1&type=pdf`.

[Lin – Ho (2002)] LIN, S.-H. – HO, J.-M. Discovering informative content blocks from Web documents. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, s. 588–593, New York, NY, USA, 2002. ACM. Dostupné z: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.18.7133&rep=rep1&type=pdf`. ISBN 1-58113-567-X.

[NIST/SEMATECH (2010)] NIST/SEMATECH. *NIST/SEMATECH e-Handbook of Statistical Methods*. 2010. Dostupné z: `http://www.itl.nist.gov/div898/handbook/`.

[Řehůřek – Sojka (2010)] ŘEHŮŘEK, R. – SOJKA, P. Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, s. 45–50, Valletta, Malta, Květen 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[Robertson (2004)] ROBERTSON, S. Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*. 2004, 60, s. 503–520. ISSN 0022-0418. Dostupné z: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.7340&type=pdf&rep=rep1`.

[Salton – McGill (1983)] SALTON, G. – MCGILL, M. J. *Introduction to Modern Information Retrieval*. McGraw-Hill computer science series. New York, NY, USA : McGraw-Hill, Inc., 1983. ISBN 0070544840.

[Salton – Buckley (1988)] SALTON, G. – BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*. 1988, 24, 5, s. 513–523. ISSN 0306-4573. doi: 10.1016/0306-4573(88)90021-0. Dostupné z: `http://dspace.library.cornell.edu/bitstream/1813/6721/2/87-881.ps`.

[Sebastiani (2002)] SEBASTIANI, F. Machine learning in automated text categorization. *ACM Comput. Surv.* March 2002, 34, s. 1–47. ISSN 0360-0300. Dostupné z: `http://arxiv.org/pdf/cs.ir/0110053`.

[Wikipedia (2011)] WIKIPEDIA. Tf–idf — Wikipedia, The Free Encyclopedia, 2011. Dostupné z: http://en.wikipedia.org/w/index.php?title=Tf%E2%80%93idf&oldid=419643129. [Online; accessed 5-April-2011].

[Williams (2000–2003)] WILLIAMS, K. Dokumentace k AI::Categorizer, 2000–2003. Dostupné z: http://search.cpan.org/~kwilliams/AI-Categorizer-0.09/lib/AI/Categorizer.pm.

[Yi et al. (2003)] YI, L. – LIU, B. – LI, X. Eliminating noisy information in Web pages for data mining. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, s. 296–305, New York, NY, USA, 2003. ACM. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.4087&rep=rep1&type=pdf. ISBN 1-58113-737-0.

# Přílohy

# A. Seznam zpravodajských zdrojů

- http://www.blesk.cz
- http://www.idnes.cz
- http://www.lidovky.cz
- http://www.tyden.cz
- http://www.ihned.cz
- http://aktualne.centrum.cz
- http://www.ceskenoviny.cz
- http://www.financninoviny.cz
- http://bleskove.centrum.cz

# B. Obsah přiloženého DVD

Na přiloženém DVD je obsaženo:

- Tato práce ve formátech LaTeX a PDF v adresáři `thesis`
- Stažené zprávy a z nich získaná data data v adresáři `data`
- Grafy, použité v této bakalářské práci, v adresáři `R_graphs`
- Program Zpravostroj, popisovaný v této bakalářské práci, v adresáři `Zpravostroj`
- Krátká dokumentace k programu v adresáři `dokumentace`

Práva ke staženým zprávám vlastní jejich autoři. Doufám, že jejich přiložením k této práci neporušuji autorský zákon.

Obsah DVD kromě stažených zpráv a z nich zjištěných dat je k dispozici též online na adrese `http://github.com/runn1ng/zpravostroj2`. Samotný program Zpravostroj je licencován pod licencí Apache 2.0.