**UNIVERZITA KARLOVA V PRAZE**

FAKULTA SOCIÁLNÍCH VĚD

Institut komunikačních studií a žurnalistiky

Jindřich Libovický

# Automatizace kvantiativní obsahové analýzy

*Bakalářská práce*

Praha, 2014

Autor práce: **Jindřich Libovický**
Vedoucí práce: **PhDr. Jan Křeček, Ph.D.**

Rok obhajoby: 2014

**Bibliografický záznam:**

LIBOVICKÝ, J. *Automatizace kvantitativní obsahové analýzy.* Bakalářská práce (Bc.), Univerzita Karlova v Praze: Praha, 2014. Fakulta sociálních věd, Institut komunikačních studií a žurnalistiky. Katedra mediálních studií. Vedoucí bakalářské práce PhDr. Jan Křeček, Ph.D.

**Anotace:**

Práce je rešerší, která shrnuje základní možnosti, které poskytují dnešní jazykové technologie pro automatizaci kódování při provádění kvantitativní obsahové analýzy. Kromě velice stručného úvodu do kvantiativní obsahové analýzy jako vědecké metody, práce obsahuje přehled dostupných softwarových nástrojů a jejich porovnání s nástoji dostupnými v roce 2002. Další část práce představuje metody, které je možné využít při navrhování systému, který kódování obsahu prováděl. Jedná se jak o postupy, které už patrně jsou součástí existujících softwarových řešení, tak metody, které byly nedávno publikovány a existuji pouze jako výzkumné prototypy. Pro lepší porozumění technologickým pasážím je poslední kapitola věnována vysvětlení některých základních pojmů vyšší matematiky.

**Abstract:**

The thesis is a summary of the state-of-the-art language technologies could be used for automation of the coding phase of the quantitative content analysis pipeline. After a short introduction to the methodology of the quantitative content analysis, an overveiw of the available software solutions is presented. The current offer is compared with what was available in 2002. The biggest part of the thesis introduces readers to the natural language processing methods applicaple while developing a system for automated quantative content analysis. Those are both well known methods which are very likely implemented in the currently existing software and both recently published techniques which exist only as research prototypes. For better understanding of the technologial parts, the last chapter contains a brief explanation of some more complex mathematical concept which are frenquently used in the previous chapters.

**Klíčová slova:**

obsahová analýza, zpracování textů

**Key words:**

content analysis, text processing

**Rozsah práce:** 78 647 znaků

## Prohlášení

1. Prohlašuji, že jsem předkládanou práci zpracoval samostatně a použil jen uvedené prameny a literaturu.
2. Prohlašuji, že práce nebyla využita k získání jiného titulu.
3. Souhlasím s tím, aby práce byla zpřístupněna pro studijní a vědecké účely.

V Praze, dne 15. května 2014 Jindřich Libovický

## Poděkování

Rád bych poděkoval svému vedoucímu za to, že byl ochoten se se mnou věnovat práci, která zcela nespadá do jeho specializace. Mé poděkování patří také všem, které jsem kvůli psaní práce ošidil o svůj čas.

# Obsah

# Úvod

V průběhu studia médií či jiných sociálních věd se každý dříve nebo později setká s kvantitativní obsahovou analýzou. Naučí se, že je to metoda, kterou je možné za jistých okolností měřit určité kvality mediálních obsahů, a začne si osvojovat konkrétní realizaci této metody. Naučí se, že má vzít nějaké předpoklady – co by měl nebo neměl daný mediální obsah splňovat – vymyslet si nějaké statisticky verifikovatelné hypotézy, nalézt kvantitativní indikátory, které za pomocí metod matematické statistiky dokáží něco vypovídat o vyslovených hypotézách. Tyto indikátory nazve proměnné, dá jim jen obtížně srozumitelná jména a začíná mravenčí, téměř strojová práce, na jejímž konci má výzkumník v ruce data, na která konečně může použít statistické nástroje, které se na počátku výzkumu rozhodl použít.

Tuto metodu, jejíž kroky jsem zjednodušeně parafrázoval v předchozím odstavci, lze kriticky napadat z různých pohledů. Stojíme-li na půdě sociálních věd, můžeme říkat, že operacionalizace problému tak, aby bylo možné jej kvantitativně analyzovat, vždy znamená redukci širokého kontinuálního spektra možných alternativ do několika diskrétních škatulek takovým způsobem, že už sotva něco vypoví o zkoumaném.

Stejně můžeme použitou metodu podrobit kritice z matematického pohledu. Pro většinu statistických metod je nutné předpokládat, že sesbíraná data jsou produkována určitými známými pravděpodobnostními rozděleními, obvykle normálním nebo geometrickým rozdělením. Většinou však nelze odhadnout, jaké pravděpodobnostní rozdělení může stát za proměnnou jako je výskyt pozitivních zmínek o nějaké osobnosti v novinách. Zároveň většina statistických postupů má za své předpoklady limitní tvrzení jako je zákon velkých čísel (tedy, že aritmetický průměr je dobrým odhadem střední hodnoty, pokud se množství sesbíraných dat blíží k nekonečnu) nebo centrální

limitní věta (tvrzení, že rozdělení průměrné hodnoty proměnné se blíží normálnímu rozdělení, pokud se počet sesbíraných hodnot dané proměnné blíží k nekonečnu). Podobně jako u předchozího problému jen těžko lze odhadnout, jaké množství dat je potřeba sesbírat pro velmi specifické proměnné, se kterými se v obsahové analýze pracuje, abychom byli dostatečně blízko nekonečnu a příslušná limitní tvrzení alespoň přibližně platila.

Přes tyto slabiny je kvantitativní obsahová analýza intenzivně používanou metodu a je-li dobře provedena, obvykle nelze pochybovat o její reliabilitě a validitě.

V úvodním odstavci jsem sběr dat – kódování obsahu do předem dohodnutého formátu – označil jako mravenčí, téměř strojovou práci. V době, kdy jsme obklopeni technologiemi, které bychom před několika lety nutně museli považovat za součást sci-fi filmů, je téměř s podivem, jak málo se píše o tom, jak tyto technologie spolu s poměrně mladými matematickými nástroji, mají potenciál výzkumníkům usnadnit jejich práci. Matematická lingvistika a umělá inteligence už léta poskytují nástroje, které by mohly být využity i v kvantitativních metodách v sociálních vědách. Matematická lingvistika se zaměřuje především na popis samotného přirozeného jazyka matematickými prostředky a vynalézá při tom způsoby, jak lze jazyk strojově zpracovávat. Umělá inteligence se zabývá nalezením způsobů, jak mohou nějací autonomní agenti vyřešit problémy, aniž by k tomu byly explicitně naprogramováni.

Kromě akademické sféry přicházejí nástroje i z komerčního prostředí. V IT světě se teď často hovoří o fenoménu *Big Data*. Firmy se snaží z velkého množství dat, která z různých důvodů po léta shromažďují, získat co největší množství informací, použitelných v jejich podnikání. Taková data mohou být třeba účetní archivy, různé obchodní statistiky, ale i data psaná v přirozeném jazyce – různé zprávy, ale také obsahy internetových diskuzí nebo sociálních sítí, z nichž se dají odhadovat nálady na spotřebitelském trhu. I v tomto oboru vznikají nástroje, které by bylo možné při obsahové analýze použít.

V oblasti umělé inteligence a strojového učení, se dnes výsledky základního výzkumu dostávají ve srovnání s jinými vědními obory nebývale rychle do praxe. Technologičtí giganti používají ve svých programech poznatky z matematické statis-

tiky, které jsou často jen rok nebo dva staré, nezřídka se stává, že Google přichází s produkty, které překonají do té doby existující výzkumné prototypy vyvinuté univerzitami.

Přestože je pokrok v těchto oblastech na první pohled patrný, sociální vědci se snad z jakési konzervativnosti drží vyzkoušených postupů a ani neví, že na jiných pracovištích týchž univerzit, kde sami působí, probíhá výzkum, jehož závěry by pro ně mohly být velice užitečné. Podobně jejich kolegové, kteří se pohybují na rozhraní matematiky, přírodních a technických věd často ani netuší, že nástroje, které pracně vyvíjí, a myšlenkový aparát, který kousek po kousku budují, by mohl sloužit i v diametrálně odlišných vědách. Místo toho se často ženou – vedeni cílem co nejdříve dát na odiv praktické výsledky svého teoretického bádání a ospravedlnit se tak v očích utilitaristicky uvažující veřejnosti – za aplikacemi co nejpodobnějšími těm, co přináší komerční svět.

Je lákavé a bylo by jistě zajímavé ptát se po příčinách těchto nepodložených spekulací, které jsem v několika předchozích odstavcích vyslovil. V bakalářské práci není dostatek prostoru k tomu zabývat se takto závažným tématem a zároveň postrádám dostatečnou erudici v oblasti sociální teorie k tomu, abych mohl vyslovovat něco kromě pár nepodložených soudů. Přesto si dovolím dát tento jev do souvislosti s proměnou veřejnosti osobností veřejně používajících rozum na veřejnost konzumentů. Jak uvádí Habermas (2000):

> *Rozpad literární veřejnosti je ještě shrnut v tomto jevu: byla narušena rezonanční půda vzdělanostní vrstvy, vychované pro veřejné užívání rozumu; publikum se rozpadlo na menšiny neveřejně uvažujících odborníků a velkou masu veřejně recipujících konzumentů. Tím ztratilo specifickou komunikační podobu publika.*

Doufám, že se mi v této práci podaří alespoň na papíře překlenout alespoň některé chybějící kontakty mezi izolovaně a neveřejně uvažujícími skupinami odborníků.

# Kapitola 1

# Kvantitativní obsahová analýza

Následující kapitola obsahuje stručné shrnutí kvantitativní obsahové analýzy jako vědecké metody. Kapitola je stručnou kompilací úvodních textů do metody kvantitativní obsahové analýzy (Scherer, 2011; Berger, 1998; Bertrand – Hughes, 2005) a je doplněna o několik málo komentářů.

## 1.1 Kvantitativní obsahová analýza jako vědecká metoda

*Kvantitativní obsahová analýza* je jeden ze způsobů analýzy mediálních obsahů používaný jak v soukromé tak akademické sféře, který obvykle slouží následujícím účelům: politické poradenství, kritické zkoumání účinnosti mediálního systému a ověřování vědeckých teorií.

V porovnání s hermeneutickou textovou a obsahovou analýzou, která je kvalitativní metodou, vychází z literárně vědní textové interpretace a slouží tak spíše k vytváření hypotéz, je kvantitativní obsahová analýza mnohem strukturovanější a nabízí systematický, intersubjektivně ověřitelný způsob zkoumání médií. Vysoká míra strukturovanosti a rigoróznost při popisu jednotlivých výzkumných kroků zajišťuje opakovatelnost a ověřitelnost výsledků.

Velice obecně se dá říci, že to, co výzkumník provádí při kvantitativní obsahové analýze, je měření množství něčeho v nějakém vzorku. Předmětem zkoumání může být prakticky cokoli – v tištěných médiích nás mohou zajímat počty článků, kde je něco pozitivně nebo negativně zmíněno, počty vět s vybranými obraty, počet fotografií určitých osobností nebo jiných jevů, plocha, kterou fotografie zabírají, plocha, kterou zabírají články a tak dále. U vysílacích médií se typicky měří stopáž pořadů nebo jejich úseků, které se věnují určitému tématu nebo zobrazují nějaký konkrétní jev (takto nás může zajímat například měřit stopáž reklam v dětských pořadech, které propagují produkty, jejichž konzumace má negativní dopad na stav chrupu dětí), u internetových médií se pak k už zmíněným měřením může přidat i analýza diskuzních příspěvků pod články. Kvantitativní obsahové analýze je možné podrobit téměř cokoli – projevy politiků, venkovní reklamu, sdělení na graffiti ve vybrané městské čtvrti nebo nápisy na pánských záchodcích.

Při kvantitativní obsahové analýze se zabýváme *přímo a pouze* obsahem určitých, obvykle mediálních sdělení. Zcela stranou zůstává očekávaná interpretace obsahu publikem – vnímání obsahů je u různých příjemců různé a kromě samotného obsahu sdělení závisí na spoustě dalších ukazatelů, které se sdělením jako takovým nesouvisí. Obvykle se uvažují sociodemogafické udaje jako sociální postavení, věk, vzdělání, politická orientace, etnický nebo geografický původ příjemců obsahů. Kvantitativní výzkum, který by měl na takovou otázku odpovědět, by musel kromě práce se samotným obsahem zahrnout do zkoumání i reprezentativní vzorek respondentů, na němž by muselo být provedeno poměrně náročné šetření. Kromě velké náročnosti takového výzkumu by se obtížně zajišťovala i jeho reliabilita a validita. Tomuto problému se při kvantitativní obsahové analýze obvykle vyhýbáme tím, že na začátku výzkumu stojí nějaké normativní předpoklady, podle nichž má smysl měřit množství nějakých jevů v obsahu. Zmínili jsme měření stopáže reklamních sdělení v dětských pořadech, které propagují produkty, jejichž konzumace může mít negativní vliv na stav chrupu dětí. Ten, kdo takový výzkum provádí, nutně předpokládá, že sledování takových reklamních sdělení děti podněcuje k tomu, aby prezentované produkty konzumovaly a v důsledku toho se jim kazily zuby.

Z toho, že se věnujeme pouze obsahu jako takovému, plyne další důležitá vlastnost kvantitativní obsahové analýzy, a to je *její časová i prostorová invariantnost*. Při zkoumání obsahu máme k dispozici obvykle digitalizované záznamy zkoumaných obsahů a nezáleží tedy na tom, kde a kdy daný jev zkoumáme. Naopak při práci s respondenty by bylo nutné vyhledat respondenty ve správný čas na správném místě. Kvantitativní obsahová analýza se tedy dá používat i jako srovnávací nástroj. Kromě porovnání mezi jednotlivými médii ji můžeme použít i k porovnání v čase nebo ke geografickému srovnání.

To, že metoda vyžaduje rigorózní definici jednotlivých výzkumných kroků s sebou nese i určité problémy. Často nemusí být na první pohled jasné, jak definovat, co to vlastně měříme – třeba když si řekneme, že budeme měřit množství násilí v novinových článcích, musíme přesně stanovit, co považujeme za násilí – přestože u různých tvůrců obsahu a konzumentů obsahu může být vnímání toho, co je násilí, zcela odlišné. Exaktní konceptualizace problému úzce souvisí s normativními předpoklady, jež na začátku do výzkumu vnášíme a může mít za následek, že výsledky určitých analýz, byť by byly velice precizně provedeny, budou jen obtížně použitelné v jiných obdobných výzkumech, kde je konceptualizace problému odlišná. Podobně náročným problémem je také stanovit jednotku, kterou budeme počítat nebo měřit – kromě jemnosti měření (např. zda budeme počítat počet vydání, počet stran, počet článků nebo počet vět) můžeme počítat i ukazatele, které v sobě mohou nést další arbitrárně volené předpoklady – pokud například počítáme, kolikrát v měsíci je na titulní straně bulvárního deníku určitá osobnost, implicitně vnášíme do výzkumu předpoklad, že titulní strana má v deníku nějaké speciální postavení.

Silnou stránkou kvantitativní obsahové analýzy je její exaktnost a opakovatelnost a to že žádným způsobem neovlivňuje to, co je zkoumáno, pracuje se s texty a elektronickými záznamy. Nevýhodou kvantitativní obsahové analýzy může být její jistá hrubost a necitlivost a to, že nevypovídá o ničem jiném, než o obsahu jako takovém v kategoriích, které byly pro výzkum zvláště definovány – nemůže vypovědět nic o tom, co si výzkumník nevzal za cíl, nevypovídá o tom, jak je obsah vnímán jeho příjemci.

## 1.2 Stručný popis struktury kvantitativní obsahové analýzy

Výzkumný proces při provádění kvantitativní obsahové analýzy se dá rozdělit do několika fází:

1. *výzkumné téma*
2. *operacionalizace* – vypracování metody, která umožní odpovědět na zadání tématu
3. *plánování a organizace* – organizační příprava výzkumu
4. *přípravná o ověřovací fáze* – ověření, jak je v praxi proveditelné to, co bylo naplánováno
5. *sběr dat*
6. *vyhodnocení dat*

Zadání výzkumu je obvykle velice abstraktní, zároveň je to, co vnáší do výzkumu normativní předpoklady. Pokud se například rozhodneme zjistit, zda některé médium protežuje určitou politickou stranu, často zároveň s tím vyslovujeme předpoklad, že by média měla jednotlivým politickým stranám poskytovat stejný prostor nebo například prostor úměrný jejich volebnímu zisku. Zároveň stojíme před nutností nějakým způsobem exaktně popsat, jak se protežování jednoho politického subjektu kvantitativně projeví na zkoumaných mediálních obsazích. Takovým ukazatelem může být pochopitelně frekvence zmínek o dané straně, popřípadě porovnání frekvence zmínek se zmínkami o jiných stranách nebo porovnání s jinými médií.

Právě rozhodnutí, jaké ukazatele zvolit a jak je měřit, je to, co je nutné provést ve fázi *operacionalizace*. V této fázi výzkumník provádí designová rozhodnutí, která jsou stěžejní pro validitu výzkumu (tedy aby skutečně měřil to, na co se ptá zadání) a reliabilitu (aby byl výzkum opakovatelný a nezávislý na místě a čase provedení). Konkrétně se jedná o to, jaká média sledovat, jaké obsahy v nich sledovat, v jakém časovém období a stanovit takzvanou *kódovací jednotku*, tedy takovou část obsahu,

kterou pokud obsahuje sledovanou kvalitu, započítáme jako jednotkový přírůstek absolutní frekvence či jako absolutní přírůstek měřené kvality.

Důležitá je také volba časového období, kdy budeme vybraná média sledovat a části médií, na které se hodláme zaměřit. Základní literatura poskytuje poměrně podrobné konkrétní návody, většinu z nich lze ale odvodit kritickým zamyšlením nad problémem. Jedná se o doporučení následujícího charakteru: Pokud analyzujeme nenadálou událost jako je např. zemětřesení, není potřeba zjišťovat, jak se o zemětřesení psalo před událostí. Pokud se zabýváme výhradně zahraničním zpravodajstvím v deníku, je v pořádku dívat se pouze na strany, o kterých víme, že se věnují zahraničnímu zpravodajství potom, co takovou skutečnost odvodíme ze struktury několika prolistovaných vydání.

Mnohem zajímavějším problémem je určení kódovací jednotky. Oproti výše zmíněné matematicky postavené definici, můžeme postavit vágnější popis kódovací jednotky – jedná se o individuum, jehož měřené kvality nazýváme proměnnými. V případě tištěných médií může být jednotkou strana, článek, nadepsaná sekce článku, fotografie, jedna věta. V případě vysílacích médií to může být pořad, reportáž nebo nějaká část pořadu, např. závěrečné titulky. Definici kódovací jednotky musí být věnována dostatečná pozornost. Při její volbě nenastavujeme pouze granularitu měření, ale výraznou měrou zde ovlivňujeme reliabilitu výzkumu. Například zmíněná segmentace obsahu tištěných médií, která by se mohla zdát být intuitivně zcela zřejmá, může narazit na rozdílné pojetí různých pojmů u jednotlivých výzkumníků. Například nemusí být jasné, zda pokračování článku z titulní strany, které je uvnitř deníku, má být považována za tentýž článek a nebo už jiný článek.

V literatuře nalezneme také technické návody, jak provádět, resp. zapisovat samotné kódování. Doporučení zahrnují, jak volit názvy proměnných a často i takové detaily, které nemají žádný vliv na prováděný výzkum a výzkumník je může volit arbitrárně. Příkladem takového doporučení mohou být názvy proměnných složené ze zřetězených obtížně srozumitelných zkratek nebo zavést číselnou substituci pro hodnoty kategoriálních proměnných. Doporučení nejspíše plynou z obtížně přenositelných

zkušeností autorů knih, poznamenaných možnostmi výpočetní techniky, se kterou přišli autoři knih do styku.

Takto sesbíraná data musí projít následným zpracováním. Pro zajištění větší spolehlivosti výsledků je například možné nechat některé nebo všechny jednotky zpracovat více výzkumníky a jejich, potenciálně rozdílné závěry nějakým způsobem zpracovat. Velice propracovanou metodologii si lze v tomto případě vypůjčit z oblasti korpusové lingvistiky, která používá propracované metody pro zajištění reliability jazykových dat. V případě více výzkumníků je možné v rané fázi výzkumu sledovat rozptyl jejich výsledků na stejných jednotkách popřípadě, tzv. mezianotátorskou shodu, jejichž vysoké hodnoty mohou naznačit nejasnosti v operacionalizaci problému.

Ke zpracování už definitivních nasbíraných dat se používají statistické metody. V nejjednodušším případě si lze vystačit s aglometrativními statistikami jako je aritmetický, geometrický nebo harmonický průměr, medián, modus nebo směrodatná odchylka. Pro zajištění vyšší spolehlivosti výzkumu je potom vhodné použít pokročilejší statistické metody, které vykazují vyšší robustnost vůči chybám v datech. Jako základní metodu můžeme použít statistické testování hypotéz (metody jako $t$-test, $\chi^2$-test), nebo například pokročilejší metody Bayesovské statistiky, které nám mohou pomoci odhalit i komplexní statistické závislosti mezi jednotlivými proměnnými.

Základní literatura ke kvantitativní obsahové analýze obvykle obsahuje téměř výhradně metodologická doporučení, které člověk znalý základních principů „Popperovské vědy“ musí považovat za samozřejmost a k jejich odvození nepotřebuje více než zdravý rozum. Naopak chybí byť letmé vysvětlení základních metod matematické statistiky (např. testování hypotéz, intervaly spolehlivosti), které mohou dát výsledkům zpracování dat mnohem hlubší význam než pouhé porovnání číselných výsledků aglomerativních statistik.

# Kapitola 2

# Automatická kvantitativní obsahová analýza

Provedení kvantitativní obsahové analýzy zahrnuje spoustu rutinních, až mechanických kroků a výpočtů, které je možné pomocí výpočetních nástrojů výrazně usnadnit. Často zůstáváme u toho, že výsledky v podstatě ručně provedené analýzy zanášíme do počítače a takto získaná data statisticky zpracujeme buď tabulkovým procesorem (Excel, OpenOffice Calc) nebo nějakým pokročilejším statistickým nástrojem (SPSS, R, Matlab). Existují ale i nástroje, které jsou vytvořeny speciálně pro provádění kvantitativní obsahové analýzy a jejichž úkolem je ulehčit výzkumníkovu práci. Matematická informatika (především pak umělá inteligence, strojové učení a matematická lingvistika) zná nástroje, které mohou posunout možnosti takového softwaru ještě dále a pomalu se již stává součástí technologické praxe.

V této kapitole nejprve shrneme, na jakých principech fungují běžně dostupné nástroje a porovnáme na několika v současnosti dostupných nástrojích se situací v roce 2002. Druhá část kapitoly obsahuje zjednodušený přehled pokročilých statistických metod zpracování textu, které jsou využitelné v kvantitativní analýze textu. Soustředíme se především na *kódování obsahu*, které bývá nejnáročnější částí provádění výzkumu. Následnou statistickou analýzou kódovaných obsahů se nezabýváme, především protože příliš nezávisí na tom, jakou metodou byl obsah kódován. Je nutné pouze mít na paměti, že použití automatických metod vždy zvyšuje

statistickou chybu oproti ručně prováděnému kódování. Pokud provádíme testování statistických hypotéz, je vhodné použít testy, které jsou robustnější vzhledem k chybám měření.

## 2.1 Software pro automatickou obsahovou analýzu

### 2.1.1 Základní principy fungování

Software pro kvantitativní obsahovou analýzu můžeme rozdělit do dvou hlavních skupin. Jednou velkou kategorií je software, který pomáhá výzkumníkům s kódováním obsahu – zobrazuje mu zkoumaný obsah a pokládá mu příslušné otázky. Při tom také kontroluje, jestli jeho odpovědi dávají smysl – například, jestli je možné, aby zadaný řetězec byl číslem strany v deníku, aby hodnoty kategoriálních veličin byly vybírány skutečně pouze z povolené množiny apod. Program jako takový žádnou analýzu neprovádí.

Z hlediska této práce jsou mnohem zajímavější programy, které provedou automaticky alespoň část kódování obsahu, především veličin, které vycházejí z textu jako takového. Základní použitou metodou je počítání četností slov nebo víceslovných pojmů, popřípadě slovních lemmat a následně testování statistických hypotéz o těchto počtech. Můžeme zjistit takové statistiky, jako například jak se mění četnost novinových článků, kde se vyskytuje slovo „Havel“, v čase, nebo třeba jak se liší mezi jednotlivými deníky. Můžeme také testovat hypotézy, zda je v určitém čase signifikantně článků s daným slovem než v jiných časech a podobně.

Velice jednoduchou metodou, jak rozšířit prosté počítání slov, je předpřipravený slovník, který obsahuje pojmy, jejichž četnost nás zajímá a k nim množiny řetězce, které je kódují. Takovým slovníkovým záznamem může být např.

```
Václav Havel = { Václav Havel, Havel, Václava Havla,
                 Havlovi... }
```

Výskyt jakéhokoli řetězce z množiny je potom programem interpretován jako výskyt příslušného pojmu. Nevýhodou této jednoduché metody může být pracnost přípravy takového slovníku. V sekci 2.2 představujeme postupy, jak i tuto část zautomatizovat.

Jedná se o velice primitivní metodu. Její výhodou je, že takový program lze napsat během několika minut, a pokud nevytváříme velké slovníky pojmů, během několika dalších minut je možné mít výsledky z textů, které mohou mít miliony slov. Takové počítání řetězců je ovšem metodologicky zpochybnitelné a spíše než ke skutečnému kódování obsahu je vhodné spíše k předběžné analýze zkoumaných textů. Zajímavou ukázkou toho, co je možné s takovým jednoduchým řetězců zístkat za informace je projekt Google N-grams[1], kde je možné získat časové řady výskytu zadaných řetězců v korpusech tvořených z různých textů v různých jazycích.

Analytické nástroje, jejichž úkolem je vytěžovat informace ze strukturovaného nebo i nestrukturovaného textu, prochází v současnosti intenzivním vývojem. Systémy, které denně používáme k různým úkolům, za běhu zaznamenávají velké množství dat. Mnoho konverzací, které dříve probíhaly osobně, jsou dnes vedeny elektronicky v psané formě – prostřednictvím emailu, instant messagingu nebo sociálních sítích. Vlastnictví takových dat dává firmám možnost vytěžit užitečné informace ať už o vlastních zaměstnancích nebo o zákaznících (využití v marketingu, pojistné matematice, atd.). Existence výpočetních cloudů, tedy možnost krátkodobého pronájmu velké výpočetní kapacity, která byla ještě nedávno dostupná jen ve výpočetních centrech univerzit a velkých firem, umožňuje analyzovat data ve velkém rozsahu téměř každému. Uvádí se, že každý dolar investovaný do analýzy velkých dat se vrátí třicetinásobně (Asay, 2013). Tento technologický posun přináší do jisté míry paradigmatickou změnu ve vnímání zpracování dat. Hovoří se i tom, že současná znalostní ekonomika se mění na daty řízenou *data-driven knowledge economy* (Cuong et al., 2014). Příkladem toho, co zpracování velkých dát může přinést, je superpočítač IBM Watson, který dokázal porazit šampiony v americké televizní hře Jeopardy (Jackson, 2011) pouze na základě informací z nestrukturovaných textů, které nalezl na Internetu. Ve světle těchto utilitaristicky chápaných úspěchů výzkumu umělé inteligence se nelze divit, že někteří

[1] `https://books.google.com/ngrams`

kriticky orientovaní myslitelé tento vývoj vnímají negativně, například Šerý (2009) jej považuje za „definitivní konec lidského člověka“, byť samotná média povauje přinejmenším za rovnocenného spoluviníka tohoto jevu.

V akademickém světě se pak můžeme setkat s přívlastkem *komputační* u názvů tradičních přírodních, ale i společenských a humanitních věd. Kromě už etablovaných oborů jako je komputační lingvistika nebo komputační biologie, se můžeme setkat i s pojmy jako komputační historie (Turkle et al., 2009), kdy se vědci snaží získat informace z velkého množství naskenovaných textů.

Jednoduché počítání slov zmíněné v úvodu této části pochopitelně není ekvivalentní s tím, když člověk zjišťuje, zda je něco zmíněno v článcích. Počítání výskytu slovníkových záznamů tedy přináší po operacionalizaci výzkumného tématu další intenzivní zjednodušení. Jak do jisté míry oslabit vliv tohoto dodatečného předpokladu ukážeme v části 2.2.

### 2.1.2 Přehled nástrojů z roku 2002

Lowe (2002) uvádí 21 softwarových nástrojů dostupných v roce 2002. Patnáct z nich jsou nástroje určené k analýze samotné, sedm z nich tehdy dostupných zdarma. Všechny zmíněné programy používají pouze počítání frekvencí slov, popřípadě slovních spojení uvedených ve slovníku a testují hypotézy spojené s těmito počty. Často je připojena i uživatelsky přívětivá vizualizace. Vzhledem k tomu o jak programátorsky jednoduchý software se jedná (odhadem několik dní práce jednoho člověka), jsou ceny, za které se programy prodávaly (okolo 500 $), poněkud zarážející.

Dále jsou zmíněné tři vývojářské knihovny, které umožňují připravit vlastní slovníky a modifikovat algoritmy, kterými se texty zpracovávají. Spíše ale než k obsahové analýze z pohledu společenských věd jsou zaměřeny na úkol, který se v informatice označuje jako information retrieval – tedy velmi povrchové sémantické předzpracování textů pro účely pozdějšího vyhledávání.

Nástroje pro automatickou morfologickou analýzu nebyly v té době na takové úrovni a tak snadno dostupné, aby bylo možné provádět lemmatizaci – automaticky

uvádět slova do základního tvaru, přestože metody, které se dnes používají byly známy a v té době publikovány i v monografiích (Manning – Schütze, 1999; Jurafsky et al., 2000).

### 2.1.3 Současné nástroje

Dnes je k dispozici poměrně velké množství nástrojů, se kterými je možné analyzovat texty, obvykle dostupných komerčně. Stačí do internetového vyhledávače zadat příslušné heslo a uživatel ihned může procházet nabídku dostupných programů. Bohužel se nejedená o nástroje, které by byky určeny k provádění metodologicky korektní kvantitativní obsahové analýzy, nicméně v oblasti public relations a marketingu se využívají stejně jako kdyby se jednalo o rigorózně provedenou obsahovou analýzu. V následujících odstavcích zmíním ty, které jsou při vyhledávání na Internetu nejčastěji zmiňované.

Pravděpodobně největší funkcionalitu poskytuje *IBM Content Analytics*[2], který je prezentován jako nástroj pro podporu rozhodování v soukromé i státní sféře na základě analýzy strukturovaných dat i nestrukturovaných textů. Ve svém typickém nasazení automaticky z Internetu stahuje požadované obsahy – články na zpravodajských serverech, blogy, diskuzní fóra, příspěvky na sociálních sítích – a provádí analýzu tak, jak si ji uživatel zadal. Typicky se jedná o sledování určitých jmen nebo témat charakterizovaných seznamem klíčových slov nebo frází, zároveň je možné provádět základní analýzu sentimentu a zjistit, jestli je text spíše negativní nebo spíše pozitivní. Takto sesbíraná data potom lze použít k počítání statistik, například jak moc koreluje výskyt určitého jména s určitým tématem a zda je určité téma nebo třeba spotřebitelský produkt vnímán pozitivně nebo negativně.

Velmi podobným řešením je *LEXIALYTICS*[3] nebo *Provalis*[4]. Na monitoring a automatickou analýzu sociálních sítí se zaměřuje český software *YESETER*[5].

[2] `http://www-03.ibm.com/software/products/en/category/content-analytics`
[3] `http://www.lexalytics.com/`
[4] `http://provalisresearch.com/`
[5] `http://www.yeseter.com/`

## 2.2 Pokročilé statistické metody

Grimmer – Stewart (2013) přináší poměrně obsáhlý popis, jaké prostředky může v současnosti matematická lingvistika nabídnout těm, kteří se rozhodnout požívat automatické nástroje k analýze textu v anglickém jazyce. Na následujících stránkách přinášíme komentované shrnutí této práce s ohledem na specifika českého jazyka a tam, kde je to možné i doporučení, jaké, především programátorské nástroje by bylo možné použít, pokud bychom chtěli využívat takové metody.

Pro čtenáře, kteří mají méně zkušeností s pojmy z vyšší matematiky doporučujeme nejprve přečíst kapitolu 3, která poskytuje velice stručný úvod v této oblasti.

### 2.2.1 Předzpracování a reprezentace textu

Jakýkoli text v přirozeném jazyce je velice složitá struktura a předtím, než je automaticky zpracován, musí zpravidla dojít k nějakému velice silnému zjednodušení. Pro účely takovéto analýzy se obvykle pracuje s dokumenty (nebo jinými dostatečně dlouhými jednotkami, můžeme použít například odstavce) jako vektory. Každé slovo, které se v dokumentu vyskytuje, dostane číslo, řekněme $i$. Potom ve vektoru, který reprezentuje dokument, bude na $i$-tém místě počet výskytů daného slova v dokumentu.

V dokumentech se mohou vyskytovat slova, která o něm nic nevypovídají – obyčejná slova, která jsou přítomna ve všech textech – pomocná slovesa, zájmena apod. Ta bývají označována jako *stop words* a pro účely takovéto analýzy odstraňována. Kromě takovýchto slov to mohou být i jiná slova, která v dané doméně textů mají jen malou váhu. Pokud bychom například zpracovávali texty o politických stranách, slovo „strana“ bude pro potřebu analýzy nadbytečné. Může tedy být užitečné pro potřeby analýzy vytvořit slovník takových pojmů.

Pokud se vyskytují v textu různé formy téhož slova, obvykle se předpokládá, že téměř jistě odkazují k témuž pojmu. Sémantická disambiguace homonymních termínů je sice zajímavým výzkumným tématem[6], ale v praktických aplikacích se obvykle za-

[6]Podobně jako v jiných otevřených informatických výzkumných tématech se v řešení této úlohy soutěží. Soutěž se nazývá SENSEVAL (`http://www.senseval.org/`) a probíhá od roku 1998 každé tři roky.

nedbává. Není to tak častý jazykový jev, aby signifikantně ovlivnil výpočty. Existují pochopitelně i příklady, kdy by mohlo být důležité se tímto jevem zabývat. Pokud bychom například chtěli znát, jak často se v nějakých textech mluví o hráčích na violoncello, tedy o *čelistech*, rádi bychom se vyhnuli *čelistem*, které odkazují na část těla obratlovců.

Častým modelem zkoumaného dokumentu je vektor četností jednotlivých slov, která se vyskytují v dokumentu. (Pojem vektoru je vysvětlen v sekci 3.1.) Takový model si můžeme představit například jako tabulku, kde jsou jednotlivé sloupce nadepsány slovy, která se v dokumentech vyskytují, a na jednotlivých řádcích této tabulky máme četnosti těchto slov v jednotlivých dokumentech. Zjevně se jedná se o ztrátovou reprezentaci textu – z takovéto reprezentace není možné text zpětně rekonstruovat.

Už samotná taková tabulka nám může přinést zajímavé informace o textech. Kromě přímé interpretace úsudkem člověka, který tabulku čte, může být takový jednoduchý model použit například pro různé metody strojového učení (viz sekce 3.3). V následujících odstavcích si navíc ukážeme, jak udělat takovouto vektorovou reprezentaci ještě informativnější, které Grimmer – Stewart (2013) neuvádí, z části protože při analýze anglických textů nejsou zdaleka tak užitečné jako při analýze textů ve flektivních jazycích.

Jedním ze způsobů je použít jinou metriku, než je prostá četnost slov. Nejčastěji se používá v počítačové vědě dlouho známé *tf-idf* skóre, které říká, jak moc je daný pojem pro dokument typický (Salton – Waldstein, 1978). Formálně jej definujeme následovně. Mějme množinu dokumentů $D$ a zajímá nás skóre pro slovo $t$ v dokumentu $d \in D$, potom jeho skóre spočítáme jako:

$$\mathrm{tf-idf}(d,t) = \frac{\text{počet slov } t \text{ v dokumentu } d}{\text{celkový počet slov v dokumentu } d} \cdot \log \frac{\text{počet dokumentů}}{\text{počet dokumentů obsahující slovo } t}.$$

Druhý zlomek v rovnici (*idf* – inverse document frequency) popisuje jak moc příslušné slovo charakterizuje dokument, pokud se v něm vyskytuje – jinými slovy, jak moc je

slovo $t$ informativní v množině $D$. Pokud bychom měli 100 dokumentů, slovo, které by se vyskytovalo pouze v jednom z nich, obdrželo by skóre $\log 100 \approx 4.6$, zatímco slovo, které se vyskytuje ve všech by mělo váhu pouze $\log 1 = 0$. Volíme logaritmickou škálu, protože eliminuje velice častá slova a stírá násobné rozdíly u velmi řídkých slov. Pokud bychom použili lineární stupnici, slovo, které se vyskytuje ve dvou dokumentech by obdrželo poloviční skóre než slovo, které se vyskytuje pouze v jednom. Tento člen vážíme prvním zlomkem, který je relativní četnosti slova v textu. Pro dokument tedy získáme vektor, který přibližně říká, jak moc dobře které slovo vystihuje příslušný dokument. Zároveň je možné z dalšího zpracování odstranit slova, která nepřináší příliš informace a zjednodušit tak další výpočty. Není ani nutné dopředu odstraňovat stop words, protože z vlastnosti metody víme, že tato slova nutně dostanou velmi nízké skóre.

Před počítáním takových statistik je vhodné použít *stemming* – nahradit slova jejich kořeny – nebo *lemmatizaci* – nahradit slova jejich základními tvary, abychom se vyhnuli tomu, že různé tvary téhož slova budou považovány za různá slova. Pro angličtinu je možné oba tyto postupy naprogramovat na několik desítek řádků a existuje mnoho volně použitelných řešení, například *Porter Stemmer* (Porter, 2001). V češtině a jiných morfologicky bohatých jazycích je situace výrazně složitější. Je to problém, kterým se matematická lingvistika dlouhodobě zabývá a existují dobře použitelná řešení. Pro češtinu lze použít tzv. Hajičovu morfologii (Hajič, 2004) [7], která ovšem v současnosti není volně šiřitelná nebo program *Morče* (morfologie češtiny)[8] vyvinutý na Matematicko-fyzikální fakultě UK.

Velmi užitečné mohou být i další způsoby předzpracování textu, které (Grimmer – Stewart, 2013) neuvádí. V jazyce se často setkáváme s víceslovnými pojmy (kliková hřídel, Miloš Zeman, Česká televize), které by nám statistika počítaná po slovech neumožňovala analyzovat. Chceme-li s takovými pojmy pracovat, musíme je v textech identifikovat a zacházet s nimi jako s lexikální jednotkou.

[7] Volně dostupnou verzi, která používá omezený slovník lze nalézt na `http://ufal.mff.cuni.cz/pdt1/Morphology_and_Tagging/Morphology/index.html`.

[8] Program lze stáhnout na adrese `http://ufal.mff.cuni.cz/morce/index.php`. Je volně šiřitelný pod restriktivní svobodnou licencí GPL.

Jak už bylo zmíněno v předchozí sekci, triviálním způsobem, jak toho docílit, je předem si připravit seznam víceslovných pojmů, které se mohou vyskytnout. Existují i sofistikovanější statistické metody, jak tyto pojmy vyhledat. Jednou z nich je statistické vyhledávání kolokací (Manning – Schütze, 1999)[s. 152]. Principiálně se dá postup popsat následovně: nalezneme v textu slova, která se spolu vyskytují signifikantně častěji, než by se vyskytovala náhodou. Obvykle uvažujeme, že slova jsou již lemmatizovaná.

Pro představu popíšeme jeden jednoduchý způsob, jak takový výpočet provést. Pravděpodobnost, že se slovo vyskytne v textu odhadneme jednoduše jako poměr četnosti daného slova ku počtu slov v textech. Pravděpodobnost výskytu po sobě jdoucí dvojice slov odhadneme jako poměr četnosti výskytu této dvojice ku všem dvojicím slov v textech. Pokud by mezi dvěma slovy $w_1$ a $w_2$ nebyla žádná souvislost – byly to tedy dva nezávislé pravděpodobnostní jevy – byla by pravděpodobnost jejich společného výskytu rovna součinu jejich individuálních pravděpodobností. Pokud je ale empirická pravděpodobnost jejich sdruženého výskytu odhadnutá z dat výrazně vyšší, říká nám to, že slova se spolu nevyskytují náhodou a považujeme jejich společné výskyty za kolokaci. Formálněji můžeme říci, že testujeme statistickou hypotézu (viz sekce 3.2):

$$\mathsf{P}(w_1, w_2) > \mathsf{P}(w_1) \cdot \mathsf{P}(w_2).$$

Výhodou tohoto přístupu je, že nepotřebuje žádná trénovací data kromě zpracovávaného textu samotného. Více informací o víceslovných pojmech můžeme zásjat metodami *Named Entity Recognition* (Tjong Kim Sang – De Meulder, 2003). S pomocí těchto metod dokážeme v textech označit řetězce, které jsou pojmy a navíc dokážeme tyto pojmy zařadit do předem připravených kategorií. Původně byla tato metoda použita pro detekci vlastních jmen a názvů firem v textech, lze ji ale použít pro libovolné kategorie pojmů. Její nevýhodou je, že vyžaduje trénovací data – poměrně velký objem textu, kde jsou lidmi ručně anotovány příklady toho, co se má v textu vyhledávat. Pro obecné texty a obecné kategorie existují trénovací korpusy[9]. Máme-li

[9]Pro češtinu je volně dostupný Czech Named Entity Corpus dostupný na `http://ufal.mff.cuni.cz/cnec/`

taková data k dispozici, můžeme natrénovat některý z dostupných nástrojů a ten dále používat nebo použít už předtrénované modely.[10] Metody Named Entity Recognition obvykle používají velice složité metody diskrétní statistiky a strojového učení a i jejich stručný popis by byl nad rámec této práce.

Další technikou, která by mohla být užitečná, je řešení koreferencí v textu – tedy zjištění, na jaké pojmy odkazují zájmena nebo jiná slova v textech. Pokud bychom takovou informaci měli, mohli bychom dosáhnout přesnějšího odhadu, jak často jsou pojmy v textech zmiňovány. Jedná se o intenzivně studované téma, nicméně publikované práce zatím nedosahují výsledků, které by šlo uplatnit v praxi.

Už samotné takovéto statistické zpracování textu, které je popsáno v této sekci, nám může poskytnout cenné informace. Můžeme například počítat statistiky, jak často byla zmiňována určitá jména nebo pojmy v novinových článcích nebo jak podstatná pro daný text byla – zda daný člověk nebo pojem byl tématem daného článku nebo byl zmíněn pouze okrajově. Ve velice krátkém čase můžeme zpracovat množství textu, které by šlo ručně zpracovat jen velice obtížně, a získat časové řady pro významnost určitých pojmů ve zkoumaných textech.

### 2.2.2 Klasifikace dokumentů do předem známých kategorií

Klasifikace dokumentů do kategorií je typickou úlohou při provádění analýzy. Kromě třídění dokumentů do nějakých triviálních skupin – např. zda se jedná o článek o politice nebo o sportu – můžeme tento úkol zobecnit na přiřazování libovolných kategoriálních příznaků k textům – například, zda článek odpovídá pozitivnímu nebo negativnímu vedení předvolební kampaně. Z algoritmického hlediska se jedná o totožné úkoly, a tak budeme dále hovořit pouze o přiřazování kategorií.

Nejjednodušší metodou, kterou lze použít, je vytvořit slovník, kde budou pojmy, které přispívají určitým kategoriím a pro každou kategorii navíc odhadneme jakou

[10] K dispozici je několik programátorských nástrojů. Zcela volně dostupné je *OpenNLP* (`http://opennlp.apache.org/`). Za nejlepší nástroj je považováno *StanfordNLP* (`http://www-nlp.stanford.edu/`), dostupné pod licencí GPL nebo komerčně.

měrou příslušný pojem přispívá. Na základě tohoto můžeme spočítat skóre, jak moc dokument náleží do určité kategorie.

Triviální model, který uvádí Grimmer – Stewart (2013), představíme v obecnější podobě. Předpokládejme, že pro kategorii $k$ máme $M$ různých slovníkových záznamů. Pro každé slovo $m$ máme na základě výzkumníkovi znalosti problematiky odhadnuté reálné číslo $s_{km}$, které říká, jak moc pojem $m$ přispívá k tomu, že dokument náleží do kategorie $k$. Skóre toho, že $i$-tý dokument náleží do kategorie $k$ potom spočítáme jednoduše jako normovaný součet skóre pro dokument:

$$t_{ki} = \frac{\sum_{m=1}^{M} s_{mk} \cdot f_{im}}{N_i},$$

kde $N_i$ je délka $i$-tého dokumentu a $f_{im}$ je četnost $m$-tého slovníkového záznamu v $i$-tém dokumentu. Místo prosté četnosti je možné použít nějaké vhodnější komplikovanější skóre, například *tf-idf* skóre, viz sekce 2.2.1. Můžeme si také povšimnout, že takto zapsaný model je z hlediska lineární algebry skalárním součinem, což je výhodné pro použití metod strojového učení, které představíme v následujících odstavcích.

Popsaná metoda vyžaduje poměrně hlubokou expertní znalost ze strany výzkumníka. Musí být velice dobře obeznámen s tím, jak zkoumané texty vypadají a zároveň potřebuje alespoň nějakou zkušenost s počítáním ve vícerozměrných vektorových prostorech. Navíc příprava takového slovníku je velice pracná a její výsledek je jen obtížně přenositelný do dalšího výzkumu. Možností, jak se tomuto problému vyhnout, je použít metody řízeného strojového učení.

Strojové učení umožňuje odhadnout příslušná skóre $s_{mk}$ automaticky bez toho, aby je musel výzkumník odhadovat. Na základě trénovacích příkladů – klasickým způsobem, ručně kódovaných obsahů – je možné tyto parametry odhadnout nejenom pro několik slovníkových záznamů, ale pro všechny pojmy, které se v textech vyskytují. Matematicky řečeno, pro vektorovou reprezentaci dokumentu můžeme nalézt vektor vah jednotlivých pojmů, které nám umožní dokumenty kategorizovat porovnáním výsledné hodnoty skalárního součinu s předem danou hodnotou. Tento pohled odpovídá historicky nejstaršímu pohledu na učení z dat na základě příkladů, tzv.

perceptronovému algoritmu (Rosenblatt, 1958), který je inspirován fungováním nervových buněk.

V následujících pěti desetiletích poměrně bouřlivého vývoje počítačové vědy vznikly další modely, které často mají v pozadí zcela jiný princip než skalární součin dvou vektorů. Základní informace o těchto metodách lze nalézt v části 3.3. Z uživatelského hlediska není potřeba vědět, jak tyto metody fungují – existují programy s grafickým uživatelským rozhraní, které umí strojové učení provádět[11]. Nicméně je vhodné zvolit některou z metod, které fungují dobře i při nadbytečných dimenzích ve vektorové reprezentaci, například *náhodný les* (random forest, (Breiman, 2001)). V takovém případě už ale klasifikace není modelována skalárním součinem.

Není také nutné zůstávat u jednoduchého vektoru rysů, jako informativnost jednotlivých v textu. Vektorovou reprezentaci dokumentu můžeme rozšířit o další informace – např. na jaké straně deníku se text nachází, zda je na barevné nebo černobílé stránce, kolikátý odstavec textu to je, jaká je průměrná poslechovost daného pořadu apod. Všechny tyto dodatečné informace dovedou algoritmy strojového učení použít ke svému rozhodování. Jistou nevýhodou může být, že při použití složitějších modelů, jako jsou zmíněné náhodné lesy, a použití široké sady dalších rysů ztrácíme možnost intuitivně interpretovat výsledky učícího algoritmu. Jedinou možností, jak validovat chování algoritmu strojového učení, je vytvořit správně kódovanou testovací sadu a s její pomocí odhadovat, jak velké chyby se algoritmus dopouští při zpracování cílových dat.

Nevýhodou při uplatnění těchto metod je nutnost mít dostatečné množství trénovacích příkladů. Potřebné množství trénovacích dat se liší především podle toho, jak obtížné je na základě pozorovaných dat od sebe kategorie odlišit. Obecně se dá říci, že je potřeba mít stovky trénovacích příkladů, a to jak pozitivních tak negativních, a destíky testovacích příkladů na ověření funkčnosti. V případě kvantitativní obsahové analýzy to může znamenat, že se nevyplatí tyto metody aplikovat, protože objem

[11]Mezi nejpoužívanější nástroje patří *WEKA* z univerzity v novozélanském Waikato (`http://www.cs.waikato.ac.nz/ml/weka`) pod licencí GPL, a komerční nástroj *Rapidminer* (`http://rapidminer.com`), který je v základní verzi volně šiřitelný pod licencí AGPL.)

potřebných trénovacích dat by byl větší než množství zkoumaného obsahu. Na druhou stranu, máme-li natrénovaný model, můžeme analyzovat miliony stran textu do několika minut nebo hodin.

Existují i tzv. *neřízené metody* strojového učení, které nevyžadují připravená trénovací data a automaticky naleznou v textech struktury a vlastnosti. Zjednodušeně se dá říct, že naleznou skupiny popřípadě hierarchické struktury textů s podobnými vektorovými reprezentacemi. Takové metody mohou být vhodné pro prvotní analýzu textu, na základě neřízeně získaných kategorií je možné leccos usoudit. Z hlediska kvantitativní obsahové analýzy se nejedná o metodologicky správný postup, protože hypotézy musí být formulovány předtím, než probíhá kódovaní, ať už ručně nebo automaticky.

### 2.2.3 Analýza sentimentu

Algoritmy popsané v předchozí sekci jsou v obecnosti schopny rozhodovat libovolné otázky o textech. Bohužel množství trénovacích dat, které by bylo potřebné k rozhodování některých otázek, je v některých případech tak obrovské, že dělají tyto metody zcela nepoužitelnými.

Při analýze textu bychom často chtěli znát, i jaké je emocionální zabarvení příslušných textů nebo polaritu názoru, který autor zastává. Bohužel tento z praktického hlediska velmi užitečný úkol patří k těm, pro které modely popsané v předchozí části potřebují velké množství trénovacích dat. Tento problém se v matematické lingvistice obvykle označuje jako *analýza sentimentu* (sentiment analysis) nebo *vytěžování názorů* (opinion mining).

Přestože se jedná vlastně jen o speciální případ úkolu popsáného v předchozí části, je to problém, který si zaslouží zvláštní pozornost. Jedním z důvodů je jeho komerční žádanost, dalším důvodem je sporná reliabilita výsledků této metody.

Výzkum v této oblasti je hnán dopředu také potenciálním využitím v marketingu (de Haaff, 2010), kterému umožňuje získávat velice rychlou zpětnou vazbu o reakcích potenciálních adresátů kampaní na Internetu. Analýzou víceméně volně dostupné elek-

tronické komunikace (diskuze na zpravodajských serverech, diskuzní fóra, sociální sítě) je možné hodnotit mínění čtenářů o reklamních kampaní, postoj lidí ke značce nebo například postoje lidí k politickým otázkám. Kromě toho, že se jedná o poměrně často akademicky zkoumané téma, takovouto funkcionalitu nabízí i všechny komerčně dostupné nástroje pro analýzu textů zmíněné v části 2.1.3.

Obvyklým východiskem pro implementaci analýzy sentimentu je předpoklad, že je možné použít nějaká univerzální trénovací data, která se shromáždí a anotují za tímto účelem. Počítá se tedy s tím, že modely vytvořené na těchto datech se budou dát znovu použít na libovolných jiných textech. Předpokládá se tedy, že ve většině diskurzů jsou prostředky pro vyjadřování pozitivního nebo negativního postoje obdobné. Přestože se obvykle používá velice pokročilé metody automatické lingvistické analýzy textů, obvykle syntaktická analýza, někdy i povrchová sémantická analýza (Nasukawa – Yi, 2003; Choi – Cardie, 2008), jádrem všech metod je použití lidmi vytvořeného slovníku frází značících pozitivní nebo negativní postoj. Význam syntaktické analýzy spočívá v tomto případě v tom, že s její znalostí je možné odhalit modalitu frází ze slovníku a odlišit tak například sentiment vět: „Bylo to dobré.“, „Bývalo to mohlo být dobré“. Mimoto umožňuje rozpoznat víceslovné formulace, které byly ve větě rozděleny funkční slovy nebo větnými modifikátory. Z pohledu strojového učení se dá říci, že příznaky ve vektorové reprezentaci pocházejí ze syntaktické nebo sémantické analýzy věty simulují skutečné porozumění textu člověkem.

Analýzu sentimentu lze metodologicky napadat z několika hledisek. Nejobecnějším argumentem by mohlo být jakýsi kulturální pohled – tedy to, že emoce a postoje vložené do textu jsou různými příjemci interpretovány různě. Tento argument ale obvykle zavrhujeme už s normativními předpoklady, které si klademe na počátku provádění obsahové analýzy. Zároveň pak téměř pro každou metodu můžeme nalézt věcnější protipříklady, kdy metoda selže. Typickým protipříkladem může být článek, který ve velké míře cituje jiné články zastávající opačný názor než autor textu a snaží se je při tom zesměšnit.

## 2.3 Shrnutí

V předchozích odstavcích jsme představili některé výpočetní metody, které by bylo možné uplatnit při obsahové analýze. S pomocí těchto metod jsme schopni:

- bez ručního kódování zjistit přítomnost a četnost výskytu pojmů (i víceslovných) v textech a odhadnout, jak významnou roli hrají pro různé texty
- s pomocí strojového učení a ručně kódovacích trénovacích příkladů rozhodnout o kategoriálních vlastnostech textů
- odhadnout sentiment, který se váže k určitým pojmům v textech

Jakékoli automatické zpracování jazyka po slovech, trpí nejednoznačností. Zmíněný řetězec „Havlovi“ může odkazovat k Václavu Havlovi, ale může být součástí slovního spojení „bratři Havlovi“. Samotný pojem „Václav Havel“ je také homonymní – může odkazovat k bývalému prezidentovi a dramatikovi, ale také k jeho dědečkovi.

Je nutné mít na paměti, že automatické metody fungují pouze přibližně, s určitou chybou. Je tedy možné je uspokojivě použít tehdy, pokud nám nezáleží na přesných číslech a je důležitější uhádnout nějaký trend. Musíme také zdůraznit, že popsané metody jsou vhodné pouze, pokud zkoumáme texty a pokud pozorované jevy nejsou příliš založeny na lidském úsudku a „čtení mezi řádky“. Například při odhalování jazykových jevů na pragmatické úrovni, jako je ironie, může být obtížně posouditelné i pro člověka. V takovém případě se automatické metody stávají téměř bezzubé – na druhou stranu i při výzkumech prováděných lidmi je snaha minimalizovat prostor pro osobitou interpretaci výzkumníka, aby nebyla ohrožena reliabilita výzkumu.

Statistické metody jsou často velice citlivé na nastavení parametrů[12], navíc s sebou přinášejí problémy, které v obsahové analýze dříve nebyly. Může to být problém přeučení (*overfitting*) známý ze strojového učení, kdy se model naučí specifika trénovacích dat a není dostatečně obecný pro dosud neviděná data.

Zcela jsme také opomenuly analýzu obrazového materiálu, který je nedílnou součástí mediálních obsahů. Interpretace obrázků a fotografií počítači zůstává i přes

[12]Ve strojovém učení se hovoří o tzv. „prokletí dimenzionality“ – paradoxu, že čím více různých vstupních informací použijeme, tím větší je šance, že model špatně odhadne parametry. Často se totiž stane, že přínos další informace je převáží komplexita způsobená odhadováním dalšího parametru.

dlouhá léta intenzivního výzkumu na velmi nízké úrovni a strojové vidění (machine vision), podobor umělé inteligence, nám nedává příliš spolehlivých nástrojů, které by se daly použít, přestože takové existují – například pro rozpoznávání tváří nebo rozpoznávání objektů na fotografiích. Obvykle ale vyžadují velké množství ručně zpracovaných trénovacích dat, které přesahují obvykle i objem dat zpracovávaný při obsahových analýzách a provádět nějaké komplikovanější úsudky blížící se například sémiotické analýze je takřka nemožné. Pro představu, úspěšnost současných systémů rozpoznávání textu v obrázku je okolo 40 % (Karatzas et al., 2013), úspěšnost rozpoznávání objektů 100 000 typů objektů z každodenního života je okolo 25 % (Dean et al., 2013).

Automatické metody obsahové analýzy by tu neměly být od toho, aby nahradily lidské schopnosti, ale aby naopak lidské schopnosti rozšířily. Je potřeba mít na paměti, že všechny statistické metody fungují s nějakou více či méně předem odhadnutelnou chybou. V úplné obecnosti je důvodem samotná definice statistiky jako inverzního výpočetního procesu k pravděpodobnostním jevům. Při provádění libovolné statistiky předpokládáme, že pozorování, která máme k dispozici a která jsou nejspíše důsledkem nějakého *deterministického procesu* – i tvorba mediálních obsahů je deterministickým procesem – jsou výsledkem *náhodného procesu*, který s určitými pravděpodobnostmi generuje určitá pozorování. Při statistickém výpočtu pak odhadujeme parametry pravděpodobnostního rozdělení, které tyto jevy vygenerovalo.

# Kapitola 3

# Minimální úvod do vyšší matematiky

V této kapitole čtenář nalezne vysvětlení pojmů z vysokoškolské matematiky, které byly používány v předchozím textu práce. Následující stránky jsou zaměřeny na ty, kteří skončili s matematikou na střední škole a rádi by porozuměli pojmům, které se používají při složitějších kvantitativních výpočtech, které jsou součástí dnešních technologických řešení.

Veškeré pojmy jsou vysvětleny velice neformálně, bez jakýchkoli důkazů nebo odvození. Příklady, na kterých jsou pojmy demonstrovány jsou záměrně vybrány tak, aby souvisely s problémem analýzy textů.

Text je členěn následovně – nejprve jsou představeny pojmy z lineární algebry, která poskytuje základní formalizmus pro práci s texty jako číselnými objekty. V další části jsou představeny základy pravděpodobnosti a statistiky. V závěru kapitoly je letmý úvod do strojového učení, které stojí na poznatcích lineární algebry a pravděpodobnosti a statistiky a je východiskem pro všechny dnešní aplikace v umělé inteligenci a matematické lingvistice. Zakládají se na něm všechny metody zpracování velkého množství dat.

## 3.1 Lineární algebra

Lineární algebra je odvětví matematiky, které se zabývá studiem vektorů a vektorových prostorů. Svůj původ má v analytickém zkoumání dvourozměrného a trojrozměrného Eukleidovského prostoru, známého ze středoškolské matematiky jako analytická geometrie. Vektory jsou také většině lidí známé i z fyziky, kde se používají k popisu veličin, které mají kromě své velikosti i nějaký směr.

Vektor o $n$ dimenzích si lze nejlépe popsat jako $n$-tici reálných čísel. V případě dvou nebo tří dimenzí si lze takové vektory představit jako souřadnice bodů nebo definice směrů v ploše nebo prostoru s různým směrem a různou velikostí. Při přechodu k vícedemenzionálním prostorům už taková intuitivní představa možná není. Matematici při práci s mnohodimenzionálními vektorovými prostory obvykle ve svých úvahách používají analogie ke dvou a trojrozměrnému prostoru, protože tyto prostory se „chovají“ podobně, pouze se pracuje s větším množstvím čísel. Při budování této analogie je také užitečné si povšimnout, o kolik náročnější je trojrozměrná geometrie oproti dvourozměrné. Stejný nárust komplexity se děje při přidání každé další složky do vektoru. To je důležité především při strojovém učení, kde hrozí, že zvýšení komplexity prostoru přidáním dalšíc rysů převáží informativní přínos přidaných rysů.

Při analýze textů se obvykle pracuje s mnohatisícirozměrnými vektory, kde každá složka vektoru reprezentuje četnost nebo význam nějakého slova. Pro účely strojového učení je možné do vektoru dodat i další veličiny (viz část 2.2.2). K používání vektorů nás vede především to, že dávno předtím, než vznikla potřeba strojově analyzovat texty, znala lineární algebra prostředky k tomu, jak efektivně pracovat s $n$-ticemi čísel. Například to, že jsou dva vektory podobné (reprezentují body, které jsou v prostoru blízko u sebe), se pozná jednoduše tak, že mají malou vzájemnou vzdálenost. Ta se dá snadno spočítat pouze za použití Pythagorovy věty.

Tento způsob měření podobnosti vektorů může často selhat v případě textů, které mají výrazně rozdílnou délku. Vzdálenost jejich vektorů v prostoru totiž nebude zachycovat pouze rozdílnou váhu různých slov (popřípadě termínu, viz 2.2.1), ale také jejich rozdílnou délku v prostoru. Je tedy praktičtější počítat velikost úhlu těchto vektorů. Ge-

ometricky si to lze přestavit jako měření úhlu úseček, které vedou od počátku soustavy souřadné k bodům, které příslušné vektory reprezentují. Tento úhel se spočítá podle známého středoškolského vzorce – kosinus úhlu je roven podílu skalárního součinu vektorů a součinu jejich velikostí.

Dalším stěžejním pojmem lineární algebry je pojem matice. Vektory obvykle považujeme za sloupcové. Matice je potom zřetězení vektorů vedle sebe. S maticemi se většina lidí setká ve středoškolské matematice jako s nástrojem pro řešení soustav lineárních rovnic Gaussovou eliminiační metodou.

Pojem, který nelze opomenout, je skalární součin, který byl již zmíněný v části 2.2.2. Ten je pro dvojici vektorů definován jako součet jejich po složkách vynásobených prvků. Kromě jakési formy váženého součtu, je ale i úzce spojený s úhlem, který vektory svírají, jak už bylo zmíněno v předchozích odstavcích.

Lineární algebra nám nabízí prostředky, jak s pomocí velice jednoduchých principů známých ze střední školy vytvořit model tak složitého objektu jako jsou texty v přirozeném jazyce, byť vytvoření příslušných vektorů může předcházet poměrné komplikované předzpracování (např. zmíněné tf-idf skóre, vyhledání víceslovných termínů). Vysokoškolské kurzy lineární algebry obvykle obsahují hlubší základy maticového počtu, bez kterých se nelze obejít při detailním pochopení mnoha postupů ve statistice a strojovém učení. Zájemcům o hlubší pochopení je možné doporučit učebnici *Pěstujeme lineární algebru* (Motl – Zahradník, 1995).

## 3.2 Statistika a pravděpodobnost

*Teorie pravděpodobnosti* je odvětví matematiky, které se zabývá náhodnými jevy (ideálními jevy, platónskými idejemi takových jevů, nikoli jevy z reálného světa). Nádodným jevem rozumíme takový jev, který když pozorujeme, nejsme schopni deterministicky určit (nebo nás to prostě nezajímá), s jakým skončí výsledkem. Víme však, že pokud bychom takový jev pozorovali dostatečně krát, četnosti různých výsledků by se stále více přibližovali nějaké deterministické funkci. Triviálním případem je házení mincí. Hodíme-li mincí, není možné jednoduše říci, zda dopadne na zem lícem nebo

rubem, víme však, že pokud je mince dobře vyvážená, skončí přibližně ve stejném počtém případů vzhůru rubem i licem. Zároveň, čím déle to budeme zkoušet, tím přesnější odhad získáme.

*Statistika* se jako součást matematiky zabývá matematickými nástroji, které nám umožní, pokud pozorujeme vnější projevy nějakého náhodného jevu, zpětně vypozorovat vlastnosti tohoto náhodného jevu. Dříve, než se pustíme do statistické analýzy nějakého jevu, musíme tedy předpokládat, že jev, který pozorujeme se podobá nějakému *ideálnímu* náhodnému jevu a jediné, co o něm nevíme jsou nějaké parametry, které je potřeba dopočítat.

Jako banální příklad je možné uvést tělesnou výšku lidí v populaci. Je zřejmé, že výška lidí má kauzální příčiny – působí na ni především dědičnost, výživa a fyzická zátěž v určitých stěžejních momentech vývoje jedince. Pokud například vytváříme design sedadla řidiče v osobním vozu, je možné od těchto hledisek odhlédnout a spokojit se s modelem, že velikost lidí je generovaná normálním rozdělením s určitými parametry. Parametry normálního rozdělení jsou aritmetický průměr a rozptyl a ty můžeme snadno empiricky odhadnout měřením na reprezentativním vzorku populace. Zde plní statistika i prediktivní funkci – dokážeme odhadnou, jak velcí budou další lidé, které jsme neměřili a sednou si do auta. Pokud by nás ale zajímalo, jak budou velcí lidé v budoucnosti, bylo takové normální rozdělení jako model naprosto nevyhovující, protože náhodný proces, který předpokládám v pozadí nezohledňuje změnu kauzálních příčin, které tělesnou výsku ovlivňují.

Tento příklad ukazuje, že funkce statistiky je dvojí. Jednak zjišťovat, jaké parametry by nejpravděpodobněji muselo mít předpokládané rozdělení, které vygenerovalo data, která pozorujeme, a jednak odhalovat statistické zákonitosti, které by za určitých okolností mohly pomoci predikovat vývoj nějakých dat do budoucnosti.

V prvním případě se jedná o *testování hypotéz*. Při něm se ptáme, jestli náhodný proces, který vygeneroval naše pozorování mohl mít nějaké vlastnosti. Často je rozumné předpokládat, s okdazem na cetrální limitní větu, že veličina má normální rozdělení. Tato věta říká, že pokud agregujeme dostatečné množství hodnot, má jejich součet (a tedy i jemu přímo uměrný průměr) rozdělení blízké normálnímu rozdělení.

Pokud bychom například počítali relativní prostor, který dávají dva různé deníky dvěma různým politickým stranám a nasbírali bychom dostatečné množství pozorování (minimálně desítky), mohli bychom prohlásit, že rozdělení získaných hodnot se blíží normálnímu rozdělení a porovnat, zda se taková rozdělení liší statisticky signifikatně nebo se z větší části překrývají.

Testování hypotéz je hlavní nástroj pro vyhodnocení obsahových analýz, nicméně v této práci se věnujeme především automatizaci kódování při obsahové analýze, kde je možné využít statistiku spíše v její prediktivní funkci jako nejdůležitější nástroj strojového učení.

O matematické statistce bylo napsáno mnoho monografií i různě pokročilých učebnic. Čtenářům, kteří nemají příliš v oblibě zcela formální texty a přesto chtějí porozumět i princpům důkazů hlavních tvrzení lze doporučit knihu *Cartoon Guide to Statistics* (Gonick – Smith, 1994), pro hlubší pochopení teorie například vysokoškolskou učebnici *Pravděpodobnost a statistika* (Zvára – Štěpán, 2001).

## 3.3 Strojové učení

*Strojové učení* je poddobor informatiky a matematiky, který se zabývá tím, jak se na základě ukázkových dat naučit rozhodovat podle určité množiny rysů příslušnost objektů do nějakých tříd. Rozhodování se učí automaticky z dostupných trénovacích dat. Učí na základě vybraných rysů, aniž by bylo potřeba znát zákony (nějakou obdobu fyzikálních zákonů), které stojí v pozadí, a často aniž by bylo možné z výsledků učení o takových zákonech něco usuzovat. Jak už bylo zmíněno v části 2.2.2, při kvantitativní obsahové analýze se dá použít k určování kategoriálních proměnných při kódování obsahů. V takovém případě je pro strojové učení instancí jednotka mediálního obsahu, kterou chceme kódovat.

Pro objekty, o kterých se chceme učit, musíme umět vytvořit seznam rysů, vlastnostní, který daný objekt charakterizují. Na takový seznam převedený na seznam reálných čísel (kategoriální rysy se jednoduše nahradíme na seznamem binárních rysů) se potom nahlíží jako na vektor. Na jednotlivé trénovací a potom i testovací instance

se tedy lze dívat jako na body (v geometrické interpretaci vektorů) v nějakém mnoharozměrném prostoru, jak jsme již rozebrali v části 3.1.

Jednou ze základních možností, jak lze takové rozhodování učit, je najít váhový vektor, který by při skalárním násobení s vektorem dával číslo menší nebo větší než předem daná konstanta. Součin lze interpretovat zaprvé jako jakýsi vážený součet – složky váhového vektoru odpovídají tomu, jak moc přispívají jednotlivé rysy k tomu, aby objekt patřil do dané třídy. Pokud je tento součet dostatečně velký, objekt do třídy patří a můžeme mu přisoudit hledaný kategoriální rys. Jiný pohled vede přes geometrickou interpretaci skalárního součinu. Chceme rozlišit ty objekty, jejichž váhové vektory mají velký skalární součin s vektorem rysů, tedy nazvájem co nejrovnoběžnější, body ležící od středu soustavy souřadné co nejpodobnějším směrem. Vlastně tak chceme ve vektrovém prostoru rysů vymezit poloprostor, ve které objekty mají hledanou kategoriální vlastnost.

Existují různé způsoby, jak nalézt optimální váhový vektor. Historicky nejstraší je takzvaný *perceptronový algoritmus* (Rosenblatt, 1958). Ten prochází trénovací data a pokaždé, když narazí na chybu, upraví váhový vektor tak, aby byl rovnoběžnější s trénovacími příklady. Tento algoritmus je zajímavý především tím, že je implementací hypotézy, jak nervové buňky provádějí zpětnovazbné učení. Nutno dodat, že tato hypotéza stojí na zjednodušeném předpokladu, že neuron pošle signál po neuritu tehdy, když vážený součet (tj. skalární součin) velikosti napětí na dendritech překročí určitou mez. Experimenty s perceptronovým algoritmem ukazují, že takové zpětnovazebné učení teoreticky možné je. Nevysvětluje však, jak se stane, že na dendritech vždy vystupuje sada dobře vybraných rysů. Dnes tedy už neurověda pracuje s jinými hypotézami. Ve strojovém učení se místo tohoto algoritmu používá *Support Vector Machine* (Cortes – Vapnik, 1995), což je algoritmus, který optimalizuje váhy analytick tak, že jednoduchými, ale důmyslnými operacemi převede problém na řešení obrovské soustavy lineárních rovnic.

Výhodou klasifikace pomocí prostého skalárního součinu je, že nám dává i jednoduchou interpretaci, jakou váhu mají jednotlivé složky váhového vektoru (např. výskyt určitého termínu, poloha na stránce apod.). Zde je ale na místě jistá opatrnost – rysy,

které jsou velice korelované si obvykle svůj přínos správné klasifikaci mezi sebou rozdělí. Obykle se této vlastnosti nevyužívá, protože lepších výsledků se dosahuje, když se vektory rysů nejprve zobrazí do komplikovanějšího prostoru.

Existují i diamterálně odlišné přístupy, které seasto dávají pro zpracování textů lepší výsledky. Jedná se o takzvané *rozhodovací stromy*. Rozhodovací strom je formalizmus, který zachycuje posloupnost kroků, které nakonec vedou k nějakému rozhodnutí. Obvykle se jako příklad uvádí lékařská diagnostika, kdy lékař popořadě provádí určité kroky a na základě jejich výsledků volí další kroky. Takové stromy se dají učit automaticky z dat různými algoritmy, například tak, že se vždy volí rozhodnutí podle rysu, tak aby co nejvíce zmenšilo entropii chybně provedených klasifikací, které by se provedly, pokud by bylo toto rozohdnutí tím posledním. Výhodou rozhodovacíh stromů je, že na rozdíl od vektorových součinů nesnižuje účinnost učícího algoritmu přítomnost většího počtu rysů. Navíc je možné provádět přímo klasifikace do více tříd, kdežto při použití skalárního součinu je nutné více tříd simulovat poslouností binárních klasifikací.

Dalšími možnostmi jsou baysovské sítě nebo umělé neuronové sítě, které lépe zachycují to, že k náležení objektu do určité třídy mohou přispívat různou měrou i různé kombinace rysů, ne pouze rysy jednotlivě.

Tímto základním přehledem použitelných metod opouštíme bouřlivý vývoj této vědní disciplíny na začátku devadesátých let. Modely publikované od té doby obvykle nejsou v ničem principiálně odlišné, pouze používají mnohem sofistikovanější metody, jak dosáhnout svých cílů. Pro zájemce o strojové učení je možné doporučit online kurz ze Standfordské univerzity na serveru *Coursera* [1], náročnějším čtenářům potom některou z monografií o strojovém učení (Mitchell, 1997; Alpaydin, 2004).

[1] `http://www.coursera.org`

# Závěr

Práce přináší rešerši – obsáhlejší referát – shrnující jaké jsou v současnosti možnosti automatizace kvantitativní obsahové analýzy. Přestože bakalářské závěrečné práce nejsou obecně příliš čtenými texty, doufám, že se mi podařilo na tomto prostoru přiblížit čtenářům, kteří mají jen minimální znalosti matematiky a informatiky, jaké možnosti jim tyto vědy poskytují.

Jako člověka s původně přírodovědným vzděláním mě překvapilo, jak velké množství „ruční práce“ musí být vykonáno při provádění kvantitativních šetřeních v sociálních vědách. Připadalo mi, že velká část této práce by mohla být s většími či menšími obtížemi přenechána strojům a výzkumníci by se mohli věnovat práci, kde je lidský úsudek nenahraditelný.

Při studiu zdrojů, obvykle odborných informatických článků a monografií, jsem si uvědomil, že se zdaleka nemusí jednat o tak snadný úkol, jak by se na první pohled mohlo zdát. Naprogramovat takový software bz bylo technicky i časově náročné. Jedním z důvodů je to, že velká část potřebných nástrojů neexistuje v produkční kvalitě, ale jsou k dispozici pouze výzkumné prototypy, které obvykle nejsou stabilní a obtížně se začleňují do softwarových produktů. Vývoj takových produktů klade i velké nároky na tým lidí, který by takový software vyvinul. Svědčí o tom i pracovní nabídky firem, které se vývojem softwaru pro analýzu textů zabývají – kromě lidí s technickým a matematickým vzděláním mají zájem i o lingvisty a mediální odborníky, často vyžadují doktorský stupeň vzdělání a zkušenosti s akademickým výzkumem. Další velkou překážkou pro masové uplatnění automatických metod jsou omezení, která kladou současné metody strojového učení. Jsou potřeba trénovací data specificky

vytvořená pro daný problém a jejich příprava se vyplatí jenom v případě zpracování opravdu velkého množství dat.

V této rozsahem poměrně stručné práci jsem se nestihl věnovat hlouběji postavení kvantitativní obsahové analýzy jako vědecké metody. V případě automatizace analýzy je část práce původně vykonávané lidmi, o nichž předpokládáme, že chybují jen minimálně, ale zároveň nedokážeme pravděpodobnost jejich chyb dobře kvantifikovat, přenechána počítačům, o nichž víme, že chybují a zároveň máme velice přesné odhady, jakým způsobem a jak moc chybují. To může vést k potřebě odlišně interpretovat výsledky strojově a ručně provedených analýz. Nejspíš je nutné i zvážit, kdy za jakých okolností zůstává taková analýza nástrojem k verifikaci nebo falzifikaci vědeckých hypotéz. Zajímavý by byl především pohled z hlediska práce Karla Poppera nebo některých pozdějších teoretiků vědy, například Thomase Kuhna.

Aby práce mohla lépe plnit částečně zamýšlenou didaktickou roli, bylo by možné doplnit části detailně popisující některé z metod výsledky na ukázkových datech, popřípadě zjednodušeným kódem, který by čtenáři naznačil, jak je možné něco takového naprogramovat. Užitečné by také mohlo být přidání jakési ukázky potenciálu metod, v technologické praxi nazývané *proof of concept* – demonstrace toho, co lze relativně snadno naprogramovat a jakých výsledků lze dosáhnout v porovnání s klasickým způsobem provedenou analýzou.

# Summary

Quantitative content analysis is one of the most widely used method for researching the media content. Despite often being criticized as insensitive because of the necessity to reduce broad spectra of real world phenomena into a few discrete categories, it used for its other qualities. It is time and space invariant and if it is well designed there cannot be any doubts about its reliability and validity. However, it is strongly dependent on the initial normative assumption which strongly limits the re-usability of data produced by previously made analyses.

The analysis is done in few distinct steps: assignment of the research topic, operationalization of the topic, planning the data collection and testing whether it is feasible, then the data collection and coding follows and finally the collected data are evaluated.

The coding phase of the content analysis is a particularly tedious process with many mechanical steps that needs to be done repeatedly. The recent research in artificial intelligence and computational linguistics offers many tool that could be possibly used for automation of this process, however there are rarely used in production software and the technologies being developed still remain available only as research prototypes. If the content we want to analyze is purely textual, a wide range of tools can be applied.

At the moment, there are few software tools available for analyzing the textual media content which are designed as a end user software and therefore does not require the user to do any programming. Their functionality is usually limited to finding occurrences of certain words or phrases, finding topics of texts and detect sentiment related to particular topics. However, the software tools are not supposed to be used for scientific research of media content, but are usually intended for use in commercial or political marketing.

Artificial intelligence and computational linguistics offer a range of methods that can be used to automate some work done during the coding phase. There exist methods for finding which words or multi-word terms are important for the texts. Sentiment analysis could be used to detect positive and negative judgments associated with the terms. Handcrafted example coding can be used to learn machine learning classifiers that can be used for automatic assignment of values of categorical variables. After having been trained the machine learning classifiers can be applied within a relatively short on a large amounts of texts.

# Literatura


ALPAYDIN, E. *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. Cambridge, USA : The MIT Press, 2004.

ASAY, M. Big Data Investments Currently Earn 50 Cents For Every Dollar Invested. *ReadWrite*. 2013. Dostupné z: `http://readwrite.com/2013/09/19/big-data-investments-currently-earn-50-cents-for-every\\-dollar-invested`.

BERGER, A. A. *Media research techniques*. London : Sage, 1998.

BERTRAND, I. – HUGHES, P. *Media Research: Audiences, Institutions, Texts*. Palgrave He, Print UK, 2005. Dostupné z: `http://books.google.cz/books?id=nT5iQgAACAAJ`.

BREIMAN, L. Random forests. *Machine learning*. 2001, 45, 1, s. 5–32.

CHOI, Y. – CARDIE, C. Learning with Compositional Semantics As Structural Inference for Subsentential Sentiment Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, s. 793–801, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

CORTES, C. – VAPNIK, V. Support vector machine. *Machine learning*. 1995, 20, 3, s. 273–297.

CUONG, E. T. – CAVARRETTA, F. – LEFEBVRE, F. Data-Driven Knowledge Economy: 5 Big Changes for Companies. February 2014. Dostupné z: `http://knowledge.essec.edu/points-of-view/data-driven-knowledge-economy-5-big-changes-for-companies.html`.

HAAFF, M. Sentiment Analysis, Hard But Worth It! *CustomerThink*. 2010. Dostupné z: `http://customerthink.com/sentiment_analysis_hard_but_worth_it`.

DEAN, T. et al. Fast, Accurate Detection of 100,000 Object Classes on a Single Machine. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

GONICK, L. – SMITH, W. *The Cartoon Guide to Statistics*. HarperResource, 1994.

GRIMMER, J. – STEWART, B. M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*. 2013, 21, 3, s. 267–297. doi: 10.1093/pan/mps028. Dostupné z: `http://pan.oxfordjournals.org/content/21/3/267.abstract`.

HABERMAS, J. *Strukturální přeměna veřejnosti.* Praha : Filosofia, 2000.

HAJIČ, J. *Disambiguation of rich inflection: computational morphology of Czech*. Praha : Karolinum, 2004.

JACKSON, J. IBM Watson Vanquishes Human Jeopardy Foes. *PC World*. 2011. Dostupné z: `http://www.pcworld.com/article/219893/ibm_watson_vanquishes_human_jeopardy_foes.html`.

JURAFSKY, D. et al. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Cambridge, USA : MIT Press, 2000.

KARATZAS, D. et al. ICDAR 2013 Robust Reading Competition. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, s. 1484–1493. IEEE, 2013.

LOWE, W. Software for content analysis – A review. *Cambridge: Weatherhead Center for International Affairs and the Harvard Identity Project*. 2002.

MANNING, C. D. – SCHÜTZE, H. *Foundations of statistical natural language processing*. Cambridge, USA : MIT Press, 1999.

MITCHELL, T. M. *Machine learning*. New York : McGraw-Hill, 1997.

MOTL, L. – ZAHRADNÍK, M. *Pěstujeme lineární algebru*. Praha : Karolinum, 1995.

NASUKAWA, T. – YI, J. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. In *Proceedings of the 2Nd International Conference on Knowledge Capture*, K-CAP '03, s. 70–77, New York, 2003. ACM.

PORTER, M. F. Snowball: A language for stemming algorithms, 2001.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*. 1958, 65, 6, s. 386.

SALTON, G. – WALDSTEIN, R. K. Term relevance weights in on-line information retrieval. *Information Processing & Management*. 1978, 14, 1, s. 29–35.

SCHERER, H. Úvod do metody obsahové analýzy. In REIFOVá, I. e. a. (Ed.) *Analýza obsahu mediálních sdělení*. Praha: Univerzita Karlova, 2011.

ŠERý, L. *Laserová romance*. Analogon: Malá řada. Sdružení Analogonu, 2009.

TJONG KIM SANG, E. F. – DE MEULDER, F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning*, s. 142–147. Association for Computational Linguistics, 2003.

TURKLE, W. – CRYMBLE, A. – MACEACHERN, A. The programing historian, 2009.

ZVÁRA, K. – ŠTĚPÁN, J. *Pravděpodobnost a matematická statistika*. Praha : Matfyzpress, 2001.