VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
FACULTY OF MECHANICAL ENGINEERING

ÚSTAV MATEMATIKY
INSTITUTE OF MATHEMATICS

TEXT MINING SE ZAMĚŘENÍM NA SHLUKOVACÍ A FUZZY SHLUKOVACÍ METODY
TEXT MINING FOCUSED ON CLUSTERING AND FUZZY CLUSTERING METHODS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE Bc. Kateřina Zubková
AUTHOR

VEDOUCÍ PRÁCE doc. RNDr. Libor Žák, Ph.D.
SUPERVISOR

BRNO 2018

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadání diplomové práce

Ústav:	Ústav matematiky
Studentka:	**Bc. Kateřina Zubková**
Studijní program:	Aplikované vědy v inženýrství
Studijní obor:	Matematické inženýrství
Vedoucí práce:	**doc. RNDr. Libor Žák, Ph.D.**
Akademický rok:	2017/18

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Text mining se zaměřením na shlukovací a fuzzy shlukovací metody

Stručná charakteristika problematiky úkolu:

Rozpoznávání textu je v poslední době důležitou součástí rozhodování v různých oblastech hospodářství, např. bankovnictví, pojišťovnictví apod. K získávání informací z textu se využívají metody založené na text miningu a také metody založené na shlukování. Diplomová práce by se měla zaměřit na tuto oblast problematiky.

Cíle diplomové práce:

– nastudování text miningu a shlukovacích metod
– hledání vhodných metod pro reálná data z oblasti text miningu
– implementace nalezených metod ve vhodném programovém prostředí
– použití naprogramovaných metod na reálná data
– srovnání výsledků získaných rozdílnými metodami a jejich uplatnění v praxi

Seznam doporučené literatury:

LUKÁŠOVÁ, A. a J. ŠARMANOVÁ. Metody shlukové analýzy. Praha: SNTL, 1985, ISBN 04-014-85.

BEZDEK, J. C. a S. K. PAL. Fuzzy Models for Pattern Recognition. Methods that Search for Pattern in Data. New York: IEEE Press, 1992. internet

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2017/18

V Brně, dne

L. S.

prof. RNDr. Josef Šlapal, CSc. ředitel ústavu	doc. Ing. Jaroslav Katolický, Ph.D. děkan fakulty

ABSTRAKT

Tato práce se zabývá shlukovou analýzou v oblasti text miningu (dolování v textech) a její aplikací na reálná data. Cílem práce je najít vhodné kategorie (shluky) v přepsaných hovorech zaznamenaných v kontaktním centru České pojišťovny a.s., a to převedením těchto textových dokumentů do vektorového prostoru za použití základních metod text miningu a následným využitím implementovaných algoritmů shlukové analýzy. Z formálního hlediska práce obsahuje popis předzpracování a reprezentace textových dat, dále popis několika běžných shlukovacích metod, validace shluků a samotnou aplikaci.

KLÍČOVÁ SLOVA

dolování v textech, shluková analýza, k-means, fuzzy c-means, hierarchické shlukování, validace shluků

ABSTRACT

This thesis is focused on cluster analysis in the field of text mining and its application to real data. The aim of the thesis is to find suitable categories (clusters) in the transcribed calls recorded in the contact center of Česká pojišťovna a.s. by transferring these textual documents into the vector space using basic text mining methods and the implemented clustering algorithms. From the formal point of view, the thesis contains a description of preprocessing and representation of textual data, a description of several common clustering methods, cluster validation, and the application itself.

KEYWORDS

text mining, cluster analysis, k-means, fuzzy c-means, hierarchical clustering, cluster validation

ZUBKOVÁ, Kateřina. *Text mining se zaměřením na shlukovací a fuzzy shlukovací metody*. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2018. 77 s. Vedoucí diplomové práce doc. RNDr. Libor Žák, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Text mining se zaměřením na shlukovací a fuzzy shlukovací metody“ jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autora

PODĚKOVÁNÍ

Ráda bych poděkovala svému vedoucímu práce doc. RNDr. Liborovi Žákovi, Ph.D. za jeho odborné vedení, Ing. Jiřímu Horovi za poskytnutá data a příležitost vyzkoušet si text mining v praxi, Ing. Denisovi Podborskému za pomoc při nahrávání dat do systému EMMA a všem kolegům z TKŘ za příjemnou přátelskou atmosféru. V neposlední řadě bych ráda poděkovala své rodině a blízkým přátelům za veškerou podporu během studia i mimo něj.

Brno . . . . . .

. . . . . .

podpis autora

OBSAH

1 ÚVOD

V dnešním světě moderních technologií vzniká denně obrovské množství dat – jen na sociální síti *Twitter*[1] je to 656 milionů tweetů! (Více zajímavých čísel lze nalézt v [16].) E-maily, sociální sítě, webové stránky, jednotlivé organizace a mnoho dalšího přispívá ke vzniku dat, ze kterých je možno vytěžit důležité informace, např. jak zvýšit zisk, zlepšit kvalitu služeb dané organizace, responzi na reklamní kampaň apod. Získáváním těchto informací z takového kvanta dat (označováno jako *Big Data*) se věnuje disciplína *dolování v datech (DM – Data Mining)* a *dolování v textech (TM – Text Mining)*. Hlavní rozdíl mezi těmito dvěma disciplínami spočívá v typu dat, se kterými pracuje. TM hledá informace nebo zajímavé vzory v nestrukturovaných nebo semistrukturovaných datech pomocí vědního oboru *zpracování přirozeného jazyka (NLP – Natural Language Processing)*, kdežto DM pracuje se strukturovanými daty. Tato práce se bude zabývat kombinací obou přístupů – nejprve jsou pomocí metod text miningu nestrukturovaná data převedena na strukturovaná a následně jsou použity metody data miningu, konkrétně techniky *strojového učení (ML – Machine Learning) bez učitele (unsupervised learning)*.

Struktura této práce je následující: V kapitole druhé jsou popsány základní pojmy a úlohy text miningu. Hlavní částí této kapitoly je popis převedení textových dat do strukturovaných a jsou zde uvedeny možnosti snížení dimenze vektorového prostoru pomocí kterého jsou data reprezentována.

Kapitola třetí se věnuje jedné z úloh text miningu, a to shlukové analýze. Jsou zde uvedeny základní metody shlukování, které jsou rozděleny na hierarchické, nehierarchické a fuzzy. Poslední část této kapitoly obsahuje popis validace shluků čili kvality vytvořeného rozkladu původních dat. Jsou zde také uvedeny metody, jak určit správný počet shluků.

Kapitola čtvrtá se zabývá aplikací metod z předchozích kapitol na reálná data, konkrétně se jedná o nahrané hovory z kontaktního centra České pojišťovny a.s. (ČP) přepsaných do textu pomocí umělé inteligence. Výsledky shlukové analýzy jsou také srovnány se zařazením dokumentů do kategorií současného systému ČP a jsou zde uvedeny výhody a nevýhody obou přístupů.

Poslední kapitola shrnuje získané poznatky této diplomové práce.

[1] https://twitter.com/

2 TEXT MINING (DOLOVÁNÍ V TEXTECH)

V této části si nejprve zavedeme několik základních pojmů podle [6], které budeme v práci používat. Dále definujeme samotný *text mining* podle [14] a uvedeme si příklady jeho využití ([5, 14]). Nakonec si popíšeme jednotlivé kroky této úlohy převzaté z [2, 7, 8, 10, 15, 17, 20].

2.1 Základní pojmy a definice

Text rozdělujeme na základní jednotky zvané *tokeny*. Můžou to být např. čísla, interpunkční znaménka či *termy*. Pod pojmem *term* rozumíme samostatné slovo nebo víceslovné spojení jako je např. „rodné číslo“, „datum narození“ apod.

Dokument je další základní jednotka v text miningu. Jedná se o část textových dat reprezentující např. vědecký či novinový článek, e-mail, zprávu apod. Upozorněme, že jeden dokument se může vyskytovat ve více kolekcích a jednotlivé kolekce můžou obsahovat různé typy dokumentů.

Kolekcí dokumentů nebo *korpusem* se obecně rozumí jakákoli skupina textových dokumentů. Prakticky se však metody text miningu zaměřují na velké kolekce, kde se počet dokumentů pohybuje v rozmezí několika stovek až miliónů. Tyto kolekce můžou být *statické*, tedy původní soubor dokumentů zůstává nezměněn, nebo *dynamické*, kdy jsou dokumenty časem aktualizovány nebo je daná kolekce doplňována novými dokumenty.

Text mining (dolování v textech) je netriviální extrakce implicitních, předem neznámých a potencionálně užitečných informací z (velkého množství) textových dat.

2.2 Příklady využití text miningu

Vyhledávání informací (information retrieval)

Základní úlohou je vyhledání *odpovědi* na uživatelský *dotaz*. *Dotazem* se přitom rozumí posloupnost klíčových slov, které mohou být spojeny vyhledávacími operátory umožňující přesnější formulaci dotazu. *Odpovědí* je skupina dokumentů odpovídající požadovanému tématu. Klasickým příkladem této úlohy je webový vyhledávač.

Extrakce informací (information extraction)

Získávání strukturovaných informací z nestrukturovaných dokumentů. Úkolem je *rozpoznávání pojmenovaných entit (named entity recognition)*, jako jsou osoby, organizace, místa, produkty apod., a *vztahů (relací)* mezi nimi (např. sídla organizací, data narození lidí apod.).

Sumarizace textů (text summarisation)

Vytvoření souhrnu obsahu dokumentu nebo skupiny dokumentů. Jedná se o nalezení nejdůležitější informace v textu, její zestručnění a prezentace v přirozeném jazyce. Může jít buď o *abstrakt* (zohledňuje gramatiku a návaznost textu), nebo *extrakt* (vybrané, často nesouvislé úseky původního textu).

Klasifikace dokumentů (classification)

Jednotlivé dokumenty jsou zařazovány do předem daných kategorií. Možná aplikace je eliminace spamu, organizování záznamů, indexované knihovny dokumentů apod.

Shlukování dokumentů (clustering)

Jedná se o úlohu podobnou klasifikaci, rozdíl je však v tom, že kategorie nejsou dopředu známy. Úkolem je tedy nalézt takové *skupiny (shluky)* dokumentů, aby si dokumenty obsažené v daném *shluku* byly v určitém smyslu podobné. Je to možný způsob, jak zjistit, co daný soubor dokumentů obsahuje. Této úloze se více budeme věnovat v kapitole 3.

2.3 Předzpracování dat

Prvním důležitým krokem dolování v textech je jejich příprava neboli předzpracování. To se skládá z několika částí, které si popíšeme v následujících odstavcích.

2.3.1 Normalizace textu

V případě internetových stránek je třeba odstranit HTML značky, URL či reklamy. Dále je nutné vypořádat se s obrázky, tabulkami, matematickými vzorci apod. Odstraníme také interpunkci a převedeme velká písmena na malá.

2.3.2 Stematizace nebo lemmatizace

V českém jazyce může jedno slovo nabývat různých tvarů – sloveso může být časováno pro vyjádření např. osoby, čísla, času atd. Další slovní druhy můžou být skloňovány pro vyjádření různých mluvnických kategorií, jako je číslo, pád, rod atd. Vzniká tedy problém, kdy je jedno slovo reprezentováno několika termy. To navyšuje dimenzi úlohy (viz dále), a proto je třeba tyto různé tvary najít a sjednotit. Existují dva přístupy:

1. **Stematizace** – jde o nalezení *kmene* slova, tedy části, která se jeho ohýbáním nemění. Někdy *kmen* bývá totožný s *kořenem*. V tomto procesu se hledají koncovky a předpony slova, které se následně odstraní. Výhodou je relativní jednoduchost provedení. Jednou z nevýhod tohoto přístupu je možnost, že podobná slova budou zkrácena na stejný tvar, např. slova „pas“, „past“, „pasta“ jsou převedena na tvar „pas“. Další nevýhodou je např. vznik slov, která v daném jazyce neexistují.

2. **Lemmatizace** – jednotlivá slova převede na jejich základní tvar, tzv. *lemma*. V text miningu je lemmatizace upřednostňována před stematizací kvůli zachování významu slova. Pokud však není znám kontext, může nastat problém s mnohoznačnými slovy. Např. slovo „tancích“ může znamenat „tanec“ nebo „tank“. Tento přístup navíc vyžaduje morfologickou analýzu, a je tedy mnohem náročnější a pomalejší než stematizace.

2.4 Reprezentace kolekce textových dokumentů

V této části si uvedeme, jak reprezentovat kolekci textových dokumentů pomocí vektorů. S tím souvisí přiřazování váhy neboli důležitosti jednotlivým termům. Nakonec si ukážeme, jak vyjádřit (ne)podobnost či vzdálenost dvou dokumentů. Tyto informace byly čerpány z [7, 8, 17].

2.4.1 Bag-of-words model

Jedná se o nejjednodušší model reprezentace textu označovaný také jako *BOW*. Danou kolekci dokumentů převede na tzv. *slovník*, který každému termu t z kolekce přiřazuje číslo vyjadřující jeho počet výskytů v celé kolekci. Následující příklad ukazuje, jak je použitím BOW modelu reprezentována kolekce $\mathcal{K}$, která obsahuje dva dokumenty o jedné větě (vyjmuty z [18]), pomocí slovníku $\mathcal{S}$.

Příklad použití modelu bag-of-words

$\mathcal{K}$ = {„Muší to teď dát hádanku nám, můj milášku, jiště, jiště.“, „Muší náš to nechat hádat třikrát, můj milášku, třikrát!“}

$\mathcal{S}$ = {muší: 2, to: 2, teď: 1, dát: 1, hádanku: 1, nám: 1, můj: 2, milášku: 2, jiště: 2, náš: 1, nechat: 1, hádat: 1, třikrát: 2}

2.4.2 Vektorový model

Tento model označovaný také jako *VSM (Vector Space Model)* je zobecněním modelu BOW. Textový dokument převede na množinu termů z celé kolekce s přiřazeným číslem, tzv. *váhou* daného termu v dokumentu. Každý dokument tedy můžeme zapsat jako vektor

$$\mathbf{d}_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,t}), \qquad i = 1, 2, \ldots, n, \tag{2.1}$$

kde n je počet dokumentů v kolekci, t je počet různých termů v kolekci (vytvářející slovník $\mathcal{S}$, $|\mathcal{S}| = t$) a $w_{i,j}$ je váha j-tého termu v i-tém dokumentu $\mathbf{d}_i$, $j = 1, 2, \ldots, t$. Kolekci textů pak můžeme reprezentovat tzv. *dokument-term maticí*

$$\mathbf{DTM} = (\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n)^T = (w_{i,j}), \tag{2.2}$$

jejíž velikost je $n \times t$. Její řádky tak představují jednotlivé dokumenty v korpusu a sloupce jednotlivé termy.

Poznamenejme, že pokud jako váhu termu bereme jeho počet výskytů v daném dokumentu, pak vektor $(\sum_i w_{i,1}, \sum_i w_{i,2}, \ldots, \sum_i w_{i,t})$ představuje bag-of-words model. Ukažme si použití takového modelu na kolekci z předchozího příkladu.

Příklad použití vektorového modelu

$\mathcal{K}$ = {„Muší to teď dát hádanku nám, můj milášku, jiště, jiště.“, „Muší náš to nechat hádat třikrát, můj milášku, třikrát!“}

$\mathcal{S}$ = {muší: 2, to: 2, teď: 1, dát: 1, hádanku: 1, nám: 1, můj: 2, milášku: 2, jiště: 2, náš: 1, nechat: 1, hádat: 1, třikrát: 2}

	muší	to	teď	dát	hádanku	nám	můj	milášku	jiště	náš	nechat	hádat	třikrát
DTM =	1	1	1	1	1	1	1	1	2	0	0	0	0
	1	1	0	0	0	0	1	1	0	1	1	1	2
$\sum$	2	2	1	1	1	1	2	2	2	1	1	1	2

2.4.3 Váha termu

Váha termu představuje významnost či důležitost termu v daném dokumentu z kolekce. Existuje několik přístupů jak tuto váhu určit, uvedeme si ty nejznámější (více metod lze nalézt například v [8]).

V následující části budeme používat označení $f_{i,j}$ pro počet výskytů termu t_j v dokumentu $\mathbf{d}_i$.

Booleovské vážení

Nejjednodušší přístup je nastavit váhu na 1 pokud se term v daném dokumentu vyskytuje a na 0 v opačném případě:

$$w_{i,j} = \begin{cases} 1 & \text{pokud } f_{i,j} > 0 \\ 0 & \text{jinak.} \end{cases}$$

TF (term frequency)

Term frequency (TF) vyjadřuje četnost termu v dokumentu. Vychází se z předpokladu, že čím více se term vyskytuje v dokumentu, tím více odpovídá tématu tohoto dokumentu. Váhu daného termu t_j v dokumentu $\mathbf{d}_i$ pomocí TF určujeme jako

$$w_{i,j} = tf_{i,j},$$

kde $tf_{i,j}$ počítáme např. podle následujících vztahů:

$$tf_{i,j} = f_{i,j}$$

nebo

$$tf_{i,j} = \frac{f_{i,j}}{\sum_j f_{i,j}},$$

kde jmenovatel odpovídá počtu slov v dokumentu $\mathbf{d}_i$. Jedná se o normalizaci, která zohledňuje délku dokumentu (v delším dokumentu se daný term může vyskytovat víckrát než v kratším a zároveň může být pro kratší dokument relevantnější).

IDF (inverse document frequency)

Inverse document frequency (IDF) vyjadřuje převrácenou četnost dokumentů obsahujících daný term v celé kolekci. Předpoklad je takový, že čím častěji se term vyskytuje ve všech dokumentech v kolekci, tím méně je důležitý (tím méně přispívá k rozlišení jednotlivých dokumentů). Nejčastěji se počítá podle následujícího vztahu

$$idf_{j,\mathcal{K}} = \log\left(\frac{n}{DF_{j,\mathcal{K}}}\right),$$

kde n je počet dokumentů v kolekci $\mathcal{K}$, tedy $n = |\mathcal{K}|$, a $DF_{j,\mathcal{K}}$ (document frequency) je počet dokumentů v kolekci $\mathcal{K}$, ve kterých se vyskytuje term t_j, tedy $DF_{j,\mathcal{K}} = |\{\mathbf{d} \in \mathcal{K} : t_j \in \mathbf{d}\}|$.

TF-IDF vážení

Předpoklady z TF a IDF můžeme zkombinovat a dostat tak známou metodu *TF-IDF vážení*. Váhu termu t_j v dokumentu $\mathbf{d}_i$ z kolekce $\mathcal{K}$ pak určujeme předpisem

$$w_{i,j} = tf_{i,j} \cdot idf_{j,\mathcal{K}}.$$

2.4.4 Kosinová podobnost

Nyní, když máme kolekci textových dokumentů vyjádřenou pomocí matice, je snadné měřit podobnost jednotlivých dokumentů. Nejčastěji používanou mírou podobnosti je v tomto případě *kosinová podobnost* vyjadřující kosinus úhlu, který svírají dva vektory (dokumenty). Tato míra neměří velikost jednotlivých vektorů, ale jejich orientaci v normovaném prostoru. Předpokládejme, že máme dokument, ve kterém se 100× vyskytuje slovo „hobit“ a v dalším dokumentu se toto slovo vyskytuje 10×. Euklidovská vzdálenost je mezi těmito dokumenty větší, ale úhel je pořád malý, protože míří stejným směrem, a právě o to v porovnávání dokumentů jde.

Tato míra vychází ze skalárního součinu a pro dva vektory $\mathbf{a}$ a $\mathbf{b}$ délky n je definována následovně:

$$S_{\cos}(\mathbf{a}, \mathbf{b}) = \cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}. \tag{2.3}$$

Obecně tato míra může nabývat hodnot z intervalu $\langle -1, 1\rangle$, v případě měření podobnosti dokumentů však nabývá pouze nezáporných hodnot (protože hodnoty v dokument-term matici nejsou záporné). Pokud $S_{\cos}(\mathbf{a}, \mathbf{b}) = 0$, pak jsou tyto vektory na sebe kolmé a znamená to, že jsou od sebe odlišné v tom nejširším slova smyslu. V případě $S_{\cos}(\mathbf{a}, \mathbf{b}) = 1$ jsou vektory totožné (svírají úhel 0°).

V takto nezáporném prostoru pak můžeme definovat *kosinovou vzdálenost* nebo *nepodobnost* jako

$$D_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - S_{\cos}(\mathbf{a}, \mathbf{b}). \tag{2.4}$$

Poznamenejme, že se nejedná o formální *metriku*, neboť není splněna podmínka *trojúhelníkové nerovnosti*.

Měření vzdálenosti pomocí této definice je velice efektivní v případě řídkých vektorů, neboť počítá pouze s nenulovými složkami. Proto je tato míra v text miningu (kde je kolekce dokumentů reprezentována řídkou maticí) tak oblíbená.

Uveďme si nyní jednoduchý příklad, kde vyjádříme kosinovou podobnost mezi čtyřmi dokumenty.

Příklad na měření kosinové podobnosti

$\mathcal{K}$ = {„Venku prší a je zima.“, „Je mi zima.“, „Je mi teplo.“, „Dám si čaj.“}

$$\mathbf{DTM} = \begin{array}{c|cccccccccc|c} & \text{venku} & \text{prší} & \text{a} & \text{je} & \text{zima} & \text{mi} & \text{teplo} & \text{dám} & \text{si} & \text{čaj} & \sum \\ \hline & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 5 \\ & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 3 \\ & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 3 \\ & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 3 \end{array}$$

$$\mathbf{S}_{\cos}(\mathcal{K}) = \begin{array}{cccc|l} 1 & 0{,}516 & 0{,}258 & 0 & \text{Venku prší a je zima.} \\ 0{,}516 & 1 & 0{,}667 & 0 & \text{Je mi zima.} \\ 0{,}258 & 0{,}667 & 1 & 0 & \text{Je mi teplo.} \\ 0 & 0 & 0 & 1 & \text{Dám si čaj.} \end{array}$$

Vidíme, že matice podobnosti $\mathbf{S}_{\cos}(\mathcal{K})$ je symetrická a má jednotkovou diagonálu. To je logické, neboť podobnost prvního a druhého dokumentu je stejná jako podobnost druhého a prvního dokumentu. Podobnost dokumentu se sebou samým je samozřejmě rovna jedné (jedná se o identický dokument). V tomto příkladě si poslední dokument není s žádným jiným z kolekce podobný. Nejpodobnější si pak jsou druhý a třetí dokument s podobností rovné 0,667.

2.5 Redukce dimenze

Velkým problémem v text miningu je vysoká dimenzionalita vektorového prostoru, neboť každý term vyskytující se v kolekci představuje jednu dimenzi. Obvykle jich bývá několik tisíc až stovky tisíců. Kromě vysoké výpočetní náročnosti to představuje také problém v počtu potřebných dokumentů - pro klasické metody strojového učení je nutné, aby byl počet dokumentů výrazně větší než počet atributů (v tomto případě termů). Proto jedním z dalších důležitých kroků v dolování v textech je redukce dimenze. V následující části ukážeme, jak lze této redukce dosáhnout. (Čerpáno z [8, 10, 15].)

2.5.1 Stopslova

Stopslova (stopwords) jsou slova, která nenesou žádnou informaci a přitom se v textech vyskytují velice často. Mezi česká stopslova tak patří například spojky, předložky, zájmena a některá slovesa jako třeba „být“ nebo „mít“. Tento krok můžeme zařadit do sekce předzpracování dat, jelikož existují již vytvořené seznamy stopslov pro

několik jazyků. Tyto seznamy však můžeme doplnit o další slova, která nás v daném korpusu nezajímají. Ta můžeme dostat například následující metodou.

Prahování DF

Document frequency (DF) je, jak už bylo výše zmíněno, četnost dokumentů obsahujících určitý term v celé kolekci. Při *prahování DF* se spočítá DF pro každý term a následně se odstraní všechny termy, jejichž DF je menší nebo větší než nějaké předem definované hodnoty, tzv. prahy. Základním předpokladem je, že tyto termy buď nenesou žádnou informaci, nebo nemají v globálním měřítku význam.

2.5.2 Latentní sémantická analýza

Latentní sémentická analýza (LSA) (v disciplíně vyhledávání informací označována jako *latentní sémantická indexace (LSI)*) je způsob, jak výrazně snížit dimenzi dokument-term matice. Podstatou je nalézt jakési *koncepty* v kolekci dokumentů. Například slova „fantasy“, „sci-fi“, „detektivka“, „kuchařka“, „komiks“ můžou představovat koncept „kniha“. Tato technika tak zobrazuje dokumenty do prostoru sémantických dimenzí, kdy do stejné dimenze jsou zobrazena slova, která jsou sémanticky podobná a ta, která jsou sémanticky odlišná jsou zobrazena do jiných dimenzí. Dva dokumenty neobsahující stejná slova tak stále můžou být podobné. Tato metoda kromě samotných konceptů poskytuje i jejich váhu, tudíž zanedbáním těch nejméně významných můžeme docílit potřebné redukce dimenze. LSA je založena na velice známé úloze lineární algebry, a to na *singulárním rozkladu matice (Singular Value Decomposition - SVD)*, kterou si nyní stručně popíšeme.

Singulární rozklad matice

Singulární rozklad matice je založen na větě z lineární algebry, která říká, že matice $\mathbf{A} \in \mathbb{C}^{m \times n}$ se dá vyjádřit jako součin tří matic:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,$$

kde $\mathbf{U}$ je $m \times m$ unitární matice, $\mathbf{V}$ je $n \times n$ unitární matice a $\mathbf{\Sigma}$ je diagonální $m \times n$ matice, jejíž prvky jsou *singulární čísla* $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(m,n)} \geq 0$. Sloupce matice $\mathbf{U}$ se nazývají *levé singulární vektory* a sloupce matice $\mathbf{V}$ se nazývají *pravé singulární vektory*.

Pro snížení dimenze se pak využívá *k-redukovaný singulární rozklad*:

$$\mathbf{A}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T,$$

kde $\mathbf{U}_k$ je $m \times k$ matice získaná z prvních k sloupců matice $\mathbf{U}$, $\mathbf{V}_k$ je $n \times k$ matice získaná z prvních k sloupců matice $\mathbf{V}$ a $\mathbf{\Sigma}_k$ je diagonální matice $k \times k$ obsahující prvních k singulárních čísel matice $\mathbf{\Sigma}$.

Použitím vztahu

$$\mathbf{D}_k = \mathbf{U}_k\mathbf{\Sigma}_k$$

získáme redukovanou $m \times k$ matici, použitím

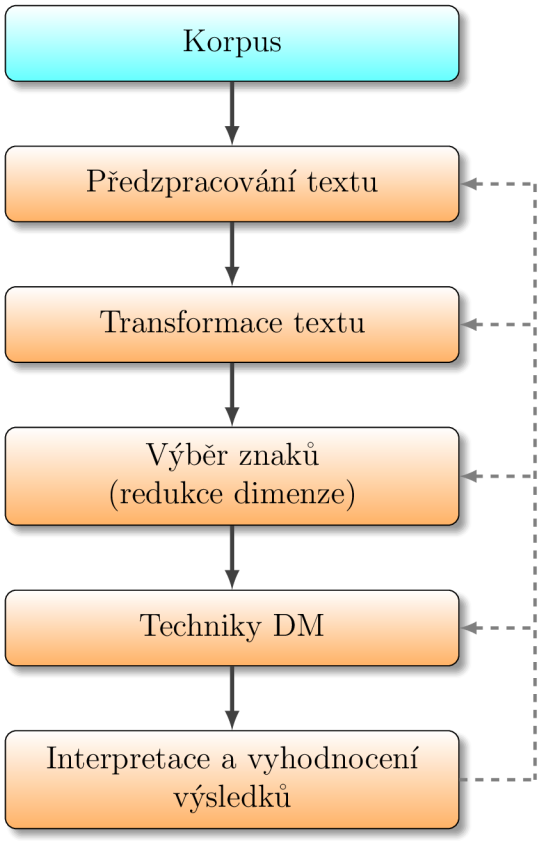
$$\mathbf{T}_k = \mathbf{\Sigma}_k\mathbf{V}_k^T$$

pak získáme redukovanou $k \times n$ matici.

2.6 Další kroky text miningu

Nyní, když máme redukovanou matici reprezentující daný korpus, můžeme aplikovat libovolnou klasickou metodu *data miningu (dolování v datech)*. Příkladem můžou být *rozhodovací stromy*, *asociační pravidla*, *regresní analýza* či *shluková analýza* (kterou se tato práce bude dále zabývat).

Dalším krokem procesu dolování v textech je *interpretace a vyhodnocení výsledků*. Tímto však úloha text miningu nekončí - jedná se o iterační proces, tedy na základě dosažených výsledků se vracíme k libovolnému z předchozích kroků, ve kterém opět vhodně upravíme data a zbytek úlohy opakujeme. Celý proces text miningu je zachycen v diagramu na Obr. 2.1.



Obr. 2.1: Diagram procesu text miningu.

3 SHLUKOVÁ ANALÝZA

Nyní se budeme věnovat shlukové analýze, která může být součástí procesu dolování v datech. Nejprve zavedeme základní pojmy které budeme dále potřebovat a popíšeme shlukovou analýzu jako takovou. (Informace v této části jsou převzaty z [11, 13].)

3.1 Základní pojmy

Na začátku máme množinu $\mathcal{O}$, která obsahuje n objektů $\mathbf{o}$, přičemž každý tento objekt je popsán m znaky:

$$\begin{aligned} \mathbf{o}_i &= (z_{i,1}, z_{i,2}, \ldots, z_{i,m}), \qquad i = 1, 2, \ldots, n \\ \mathcal{O} &= \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_n\}. \end{aligned}$$

Tuto množinu je pak možno reprezentovat pomocí matice podobně jako (2.1) a (2.2), kde dokument $\mathbf{d}_i$ odpovídá objektu $\mathbf{o}_i$ a váhy $w_{i,j}$ odpovídají znakům $z_{i,j}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, t = m$. Označme tuto matici $\mathbf{Z} = (z_{i,j})_{n,m}$.

Obecně každý znak popisující objekt může být měřen v různých jednotkách. Pak se může stát, že některé znaky se jeví jako dominující. Proto je vhodné data nejprve standardizovat. Pro každý sloupec (znak) z_j vypočteme střední hodnotu $\bar{z_j}$ a směrodatnou odchylku s_j podle vzorců

$$\begin{aligned} \bar{z_j} &= \frac{1}{n} \sum_{i=1}^{n} z_{i,j}, \\ s_j &= \sqrt{\frac{1}{n} \sum_{i=1}^{n} (z_{i,j} - \bar{z_j})^2}. \end{aligned}$$

Původně naměřené hodnoty $z_{i,j}$ pak přepočteme na standardizované hodnoty

$$y_{i,j} = \frac{z_{i,j} - \bar{z_j}}{s_j}.$$

Dostaneme tak novou matici $\mathbf{Y} = (y_{i,j})_{n,m}$, jejíž znaky mají nulovou střední hodnotu, jednotkový rozptyl a směrodatnou odchylku.

Další vhodnou úpravou může být normalizace objektů. Každou složku objektu (znak řádku matice $\mathbf{Y}$) podělíme normou tohoto objektu. Vektorově tedy

$$\mathbf{x}_i = \frac{\mathbf{y}_i}{\|\mathbf{y}_i\|}.$$

Po všech předchozích úpravách tedy dostáváme množinu $\mathcal{X}$, se kterou budeme dále pracovat. Konkrétní vlastnosti této množiny a jejich prvků udávají vztahy (3.1).

$$\begin{aligned} \mathcal{X} &= \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}, \\ \mathbf{X} &= (x_{i,j})_{n,m}, \\ \bar{x_j} = 0,\ s_j &= 1,\ \|\mathbf{x}_i\| = 1, \quad i = 1, \ldots, n,\ j = 1, \ldots, m. \end{aligned} \tag{3.1}$$

3.1.1 Podobnostní vztahy

Důležitou součástí shlukové analýzy je vyjádření *podobnosti objektů*. Jedná se o předpis $\pi : \mathcal{X} \times \mathcal{X} \longmapsto \mathbb{R}$, který každé dvojici objektů $(\mathbf{x}_i, \mathbf{x}_j)$ přiřadí reálné nezáporné číslo $\pi(\mathbf{x}_i, \mathbf{x}_j)$ tak, aby byly splněny následující požadavky:

$$\begin{aligned} \pi(\mathbf{x}_i, \mathbf{x}_j) &\geq 0, \\ \pi(\mathbf{x}_i, \mathbf{x}_j) &= \pi(\mathbf{x}_j, \mathbf{x}_i). \end{aligned}$$

Dalším vhodným požadavkem je, aby v případě $\mathbf{x}_i = \mathbf{x}_j$ hodnota $\pi(\mathbf{x}_i, \mathbf{x}_j)$ nabývala maximální hodnoty z oboru hodnot π. Tento předpis nám tedy říká, že čím větší je hodnota $\pi(\mathbf{x}_i, \mathbf{x}_j)$, tím větší je vzájemná podobnost objektů $\mathbf{x}_i, \mathbf{x}_j$. Jeden z takových předpisů jsme již definovali dříve jako *kosinovou podobnost* (viz (2.3)).

Ve většině případů se však používá spíše duálního pojmu. *Nepodobnost objektů* $d : \mathcal{X} \times \mathcal{X} \longmapsto \mathbb{R}$ splňuje následující podmínky:

$$\begin{aligned} d(\mathbf{x}_i, \mathbf{x}_j) &= 0 \Leftrightarrow \mathbf{x}_i = \mathbf{x}_j, \\ d(\mathbf{x}_i, \mathbf{x}_j) &\geq 0, \\ d(\mathbf{x}_i, \mathbf{x}_j) &= d(\mathbf{x}_j, \mathbf{x}_i). \end{aligned}$$

Příkladem může být známá *Euklidovská metrika* nebo *kosinová nepodobnost* (viz (2.4)).

3.1.2 Shlukování a pojem shluku

Existuje několik pokusů charakterizovat shlukovou analýzu. Příkladem může být od Tryona [19]:

> „Shluková analýza je obecný logický postup formulovaný jako procedura, pomocí níž seskupujeme objektivně jedince do skupin na základě jejich podobností a rozdílností.“

Bonner [1] tento proces charakterizoval takto:

> „Je dána množina objektů, z nichž každý je definován pomocí množiny znaků s ním souvisejících. Tato množina znaků je pro každý objekt stejná. Máme nalézt shluky objektů (podmnožiny původní množiny objektů) tak, aby si členové shluku byli vzájemně podobní, ale nebyli si příliš podobní s objekty mimo tento shluk.“

Shluková analýza či *shlukování* se snaží nalézt jisté vzory dané množinou objektů a tyto objekty pak na základě toho rozdělit do jakýchsi „přirozených“ kategorií či skupin, přičemž jejich počet musíme buďto určit předem, nebo se tento počet zjistí v průběhu shlukování. Předpokládejme, že chceme získat k shluků, kde $1 < k < n$. Pak můžeme říci, že cílem shlukovacího procesu je definovat zobrazení $\varphi : \mathcal{X} \longmapsto \mathcal{C}_0$, jež každý objekt $\mathbf{x}_i \in \mathcal{X}$, $i = 1, 2, \ldots, n$ přiřadí do shluku $\mathcal{C}_j \in \mathcal{C}_0$, $\mathcal{C}_0 = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$, $\bigcup_{j=1}^{k} \mathcal{C}_j = \mathcal{X}$, přičemž pojem shluku můžeme definovat pomocí nepodobnosti objektů následovně:

Shlukem nazveme takovou podmnožinu $\mathcal{C}$ množiny objektů $\mathcal{X}$, pro niž platí

$$\max_{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{C}} d(\mathbf{x}_i,\mathbf{x}_j) < \min_{\substack{\mathbf{x}_k\in\mathcal{C},\\ \mathbf{x}_l\notin\mathcal{C}}} d(\mathbf{x}_k,\mathbf{x}_l).$$

Shluk můžeme také definovat pomocí *T-souvislých objektů* jako *T-shluk*: Objekt $\mathbf{x}_p$ je *T-souvislý* s objektem $\mathbf{x}_q$ pro daný práh T, existuje-li řetěz objektů $\mathbf{x}_p = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m = \mathbf{x}_q, m > 1$, že $d(\mathbf{x}_i, \mathbf{x}_{i+1}) \leq T$ pro $i = 1, 2, \ldots, m-1$, kde $d(\mathbf{x}_i, \mathbf{x}_j)$ představuje koeficient nepodobnosti objektů $\mathbf{x}_i, \mathbf{x}_j$. *T-souvislý shluk (T-shluk)* je podmnožina $\mathcal{C} \subset \mathcal{X}$, která splňuje podmínky:

1. Každý pár objektů z $\mathcal{C}$ je T-souvislý.
2. Žádný objekt z $\mathcal{X} - \mathcal{C}$ není T-souvislý s žádným objektem z $\mathcal{C}$.

3.2 Hierarchické shlukování

Hierarchické shlukování je jedna z nejstarších metod shlukové analýzy. Existují dva základní přístupy:

- **Aglomerativní:** Na začátku každý objekt vytváří jeden shluk a v každém dalším kroku se dva nejbližší (nejpodobnější) shluky spojí v jeden. Proces se opakuje tak dlouho, dokud nezbude jeden shluk obsahující všechny objekty. Toto vyžaduje definici podobnosti shluků.
- **Divizivní:** Na začátku je jeden shluk obsahující všechny objekty a v každém dalším kroku se shluk (nebo shluky) rozdělí. Dělení probíhá tak dlouho, dokud nezbudou pouze shluky tvořené samostatnými objekty. V tomto případě musíme v každém kroku určit, který shluk (nebo které shluky) rozdělit a jak to provést.

Pokud bychom chtěli rozdělit jeden shluk na dva, existuje $2^{n-1} - 1$ způsobů jak to udělat. Protože je to neporovnatelně více výpočtů (zvláště pro velká n) než v případě aglomerativního shlukování (kde je potřeba v prvním kroku vypočítat $n \cdot (n-1)/2$ koeficientů nepodobnosti a v každém dalším méně), nebudeme se divizivními metodami v této práci zabývat.

3.2.1 Aglomerativní metody

Hierarchická aglomerativní metoda přiřadí množině objektů $\mathcal{X}$ posloupnost jejích rozkladů $\Omega_0, \Omega_1, \ldots, \Omega_{n-1}$ na shluky. Rozklad Ω_0 je tvořen n jednoprvkovými shluky, Ω_{n-1} je tvořen jedním n-prvkovým shlukem. Tato metoda zároveň přiřadí každému shluku $\mathcal{C}$ reálné nezáporné číslo $h(\mathcal{C})$, které nazýváme *shlukovací hladina*. Pro shluk obsahující jediný objekt $\mathcal{C}_{0i} = \{\mathbf{x}_i\}$ definujeme $h(\mathcal{C}_{0i}) = 0$ a pro každý nový shluk $\mathcal{C}_{kj}$, který vznikl sloučením dvou shluků $\mathcal{C}_k, \mathcal{C}_j$ pak $h(\mathcal{C}_{kj}) = h(\mathcal{C}_k \cup \mathcal{C}_j) = D(\mathcal{C}_k, \mathcal{C}_j)$, kde zobrazení D je *koeficient nepodobnosti shluků* (viz kapitola 3.2.2).

Zjednodušený princip jsme si uvedli dříve, formálněji ho popisuje Algoritmus 1.

Algoritmus 1 Základní aglomerativní hierarchické shlukování

1: Spočti matici nepodobnosti shluků.
2: **repeat**
3: Sluč dva shluky s nejmenším koeficientem nepodobnosti.
4: Přepočti matici nepodobnosti shluků.
5: **until** Zbývá jediný shluk.

3.2.2 Koeficient nepodobnosti shluků

Klíčovou roli v hierarchickém shlukování hraje výpočet nepodobnosti shluků. Pro shluky $\mathcal{C}_i, \mathcal{C}_j$ z rozkladu Ω definujeme *koeficient nepodobnosti shluků* jako zobrazení $D : \Omega \times \Omega \longmapsto \mathbb{R}$, které splňuje tyto podmínky:

$$\begin{aligned} D(\mathcal{C}_i, \mathcal{C}_i) &= 0, \\ D(\mathcal{C}_i, \mathcal{C}_j) &\geq 0, \\ D(\mathcal{C}_i, \mathcal{C}_j) &= D(\mathcal{C}_j, \mathcal{C}_i). \end{aligned}$$

Jako *matici nepodobnosti shluků* pak budeme označovat čtvercovou matici $\mathbf{D} = (D_{i,j})_{k,k}$, kde $D_{i,j} = D(\mathcal{C}_i, \mathcal{C}_j), 1 \leq i, j \leq k = |\Omega|$.

Nyní si uvedeme nejpoužívanější způsoby zavedení tohoto koeficientu.

Metoda nejbližšího souseda

Pro shluky $\mathcal{A}, \mathcal{B}$ definujeme jejich nepodobnost jako

$$D(\mathcal{A}, \mathcal{B}) = \min_{\substack{\mathbf{x}_i \in \mathcal{A} \\ \mathbf{x}_j \in \mathcal{B}}} d(\mathbf{x}_i, \mathbf{x}_j),$$

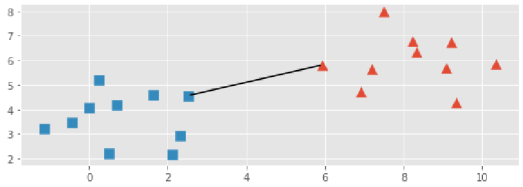
kde d je koeficient nepodobnosti. Vzdálenost dvou shluků $\mathcal{A}, \mathcal{B}$ je tedy vyjádřena jako nepodobnost dvou nejbližších (nejméně nepodobných) objektů ze shluku $\mathcal{A}$ a shluku $\mathcal{B}$. Tato metoda se také označuje jako *single linkage* a je znázorněna na Obr. 3.1a.

Metoda nejvzdálenějšího souseda

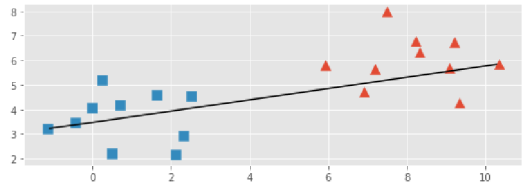
Definuje se obdobně jako předchozí metoda, místo nepodobnosti dvou nejbližších objektů se však vybírá nepodobnost dvou nejvzdálenějších (nejvíce nepodobných) objektů:

$$\begin{aligned} D(\mathcal{A}, \mathcal{B}) &= \max_{\substack{\mathbf{x}_i \in \mathcal{A} \\ \mathbf{x}_j \in \mathcal{B}}} d(\mathbf{x}_i, \mathbf{x}_j) \quad \text{pro } \mathcal{A} \neq \mathcal{B}, \\ D(\mathcal{A}, \mathcal{A}) &= 0. \end{aligned}$$

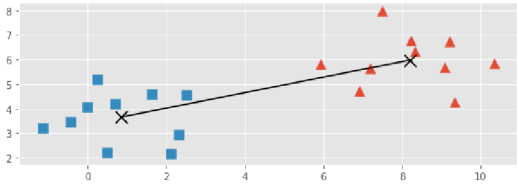
Tato metoda se také označuje jako *complete linkage* a je znázorněna na Obr. 3.1b.



(a) Metoda nejbližšího souseda.



(b) Metoda nejvzdálenějšího souseda.



(c) Centroidní metoda.

Obr. 3.1: Koeficient nepodobnosti shluků.

Centroidní metoda

Tato metoda určuje nepodobnost mezi shluky jako nepodobnost centroidů těchto shluků.

$$\begin{aligned} \bar{\mathbf{a}} &= \frac{1}{|\mathcal{A}|}\sum_{\mathbf{x}\in\mathcal{A}}\mathbf{x}, \\ \bar{\mathbf{b}} &= \frac{1}{|\mathcal{B}|}\sum_{\mathbf{x}\in\mathcal{B}}\mathbf{x}, \\ D(\mathcal{A},\mathcal{B}) &= d(\bar{\mathbf{a}},\bar{\mathbf{b}}). \end{aligned}$$

Tato metoda se také označuje jako *weighted group method* a je znázorněna na Obr. 3.1c.

Metoda průměrné nepodobnosti

Koeficient nepodobnosti se v tomto případě určí jako průměrná nepodobnost všech dvojic objektů z dvou různých shluků:

$$\begin{aligned} D(\mathcal{A},\mathcal{B}) &= \frac{1}{|\mathcal{A}||\mathcal{B}|}\sum_{\mathbf{x}_i\in\mathcal{A}}\sum_{\mathbf{x}_j\in\mathcal{B}} d(\mathbf{x}_i,\mathbf{x}_j), \\ D(\mathcal{A},\mathcal{A}) &= 0. \end{aligned}$$

Mezi další často používané metody patří *mediánová metoda* nebo *Wardova metoda*, jejichž definici můžeme nalézt např. v [11].

Lance-Williamsův vzorec

Výše definované koeficienty nepodobnosti shluků umožňují lépe porozumět daným metodám, z hlediska implementace je však výhodnější použít *Lance-Williamsův vzorec* (3.2), díky kterému není třeba uchovávat původní data. Matice nepodobnosti

shluků rozkladu Ω_i může být přepočítána pomocí matice z předchozího rozkladu Ω_{i-1} následovně:

1. $D(\{\mathbf{x}_i\}, \{\mathbf{x}_j\}) = d(\mathbf{x}_i, \mathbf{x}_j)$ pro $1 \leq i, j \leq n$.
2. Nechť $\mathcal{R} = \mathcal{A} \cup \mathcal{B}$ je shluk rozkladu Ω_i vzniklý sjednocením shluků $\mathcal{A}, \mathcal{B} \in \Omega_{i-1}$. Pak pro všechny shluky $\mathcal{C} \in \Omega_{i-1}$ přecházející do rozkladu Ω_i beze změny platí:

$$D(\mathcal{R}, \mathcal{C}) = \alpha_{\mathcal{A}} D(\mathcal{A}, \mathcal{C}) + \alpha_{\mathcal{B}} D(\mathcal{B}, \mathcal{C}) + \beta D(\mathcal{A}, \mathcal{B}) + \gamma |D(\mathcal{A}, \mathcal{C}) - D(\mathcal{B}, \mathcal{C})|, \quad (3.2)$$

kde koeficienty $\alpha_{\mathcal{A}}, \alpha_{\mathcal{B}}, \beta, \gamma$ jsou pro jednotlivé metody uvedeny v Tabulce 3.1. Tyto koeficienty pro centroidní, mediánovou a Wardovu metodu však byly odvozeny pro čtverec Euklidovské vzdálenosti jako koeficient nepodobnosti objektů.

Tab. 3.1: Koeficienty Lance-Williamsova vzorce pro metody hierarchického shlukování.

Shlukovací metoda	$\alpha_{\mathcal{A}}$	$\alpha_{\mathcal{B}}$	β	γ
Metoda nejbližšího souseda	$1/2$	$1/2$	0	$-1/2$
Metoda nejvzdálenějšího souseda	$1/2$	$1/2$	0	$1/2$
Centroidní metoda	$\frac{\lvert\mathcal{A}\rvert}{\lvert\mathcal{A}\rvert+\lvert\mathcal{B}\rvert}$	$\frac{\lvert\mathcal{B}\rvert}{\lvert\mathcal{A}\rvert+\lvert\mathcal{B}\rvert}$	$\frac{-\lvert\mathcal{A}\rvert\lvert\mathcal{B}\rvert}{(\lvert\mathcal{A}\rvert+\lvert\mathcal{B}\rvert)^2}$	0
Metoda průměrné nepodobnosti	$\frac{\lvert\mathcal{A}\rvert}{\lvert\mathcal{A}\rvert+\lvert\mathcal{B}\rvert}$	$\frac{\lvert\mathcal{B}\rvert}{\lvert\mathcal{A}\rvert+\lvert\mathcal{B}\rvert}$	0	0
Mediánová metoda	$1/2$	$1/2$	$-1/4$	0
Wardova metoda	$\frac{\lvert\mathcal{A}\rvert+\lvert\mathcal{S}\rvert}{\lvert\mathcal{A}\rvert+\lvert\mathcal{B}\rvert+\lvert\mathcal{S}\rvert}$	$\frac{\lvert\mathcal{B}\rvert+\lvert\mathcal{S}\rvert}{\lvert\mathcal{A}\rvert+\lvert\mathcal{B}\rvert+\lvert\mathcal{S}\rvert}$	$\frac{-\lvert\mathcal{S}\rvert}{\lvert\mathcal{A}\rvert+\lvert\mathcal{B}\rvert+\lvert\mathcal{S}\rvert}$	0

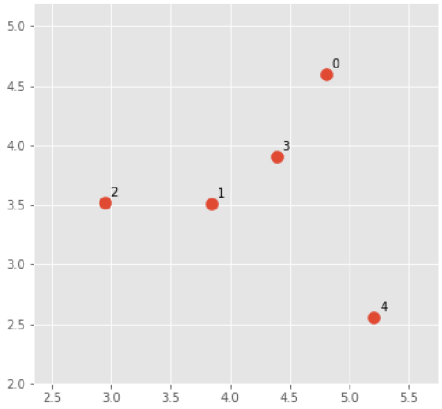
3.2.3 Vizualizace hierarchického shlukování

Hierarchické shlukování bývá graficky znázorněno pomocí stromového diagramu nazývaného *dendrogram*. Zobrazuje vztahy mezi jednotlivými shluky a podshluky a také pořadí, ve kterém byly shluky sloučeny (v případě divizivního shlukování rozděleny). Jedna osa reprezentuje shlukovací hladinu, na druhou osu jsou vyneseny indexy jednotlivých objektů. Řez přes určitou shlukovací hladinu dává příslušný rozklad množiny objektů, jednotlivé shluky jsou dány uzly. Příklad dendrogramu vidíme na Obr. 3.2f.

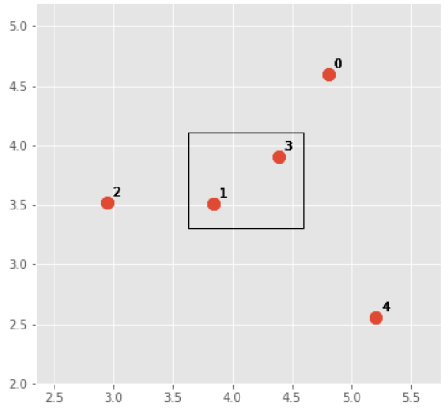
Hierarchické shlukování dvourozměrných objektů může být také znázorněno pomocí *diagramu vnořených shluků*, viz Obrázek 3.2e.

3.2.4 Ilustrační příklad

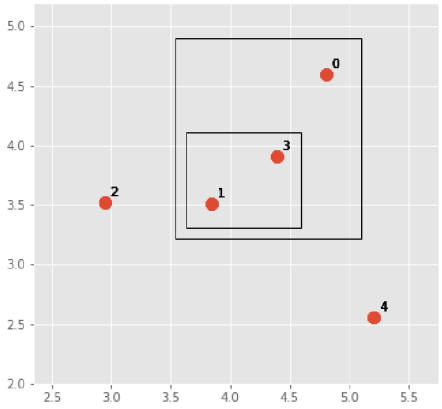
Na ilustračním příkladě ukážeme jednotlivé kroky aglomerativního hierarchického shlukování centroidní metodou. Na Obr. 3.2a vidíme 5 dvourozměrných bodů, které chceme shlukovat. Obrázek 3.2b ukazuje sloučení bodů 1 a 3 do jednoho shluku. Dalším slučováním nakonec dospějeme k jednomu shluku zobrazeném na Obr. 3.2e, který také představuje tzv. diagram vnořených shluků. Obrázek 3.2f pak zobrazuje dendrogram tohoto procesu. Můžeme si všimnout, že např. $D(1,3) \approx 0,7$ a $D(2,4) \approx 1,8$. Dále, řezem shlukovací hladinou $h = 1,25$ obdržíme 3 shluky: $\{4\}, \{0,1,3\}, \{2\}$.



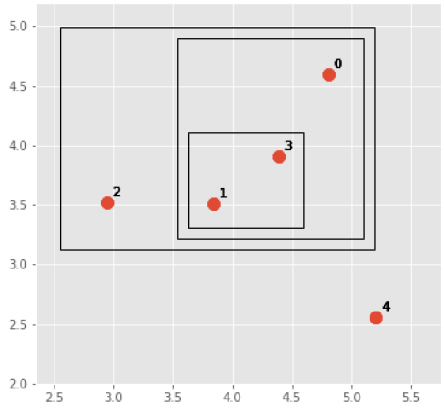
(a)



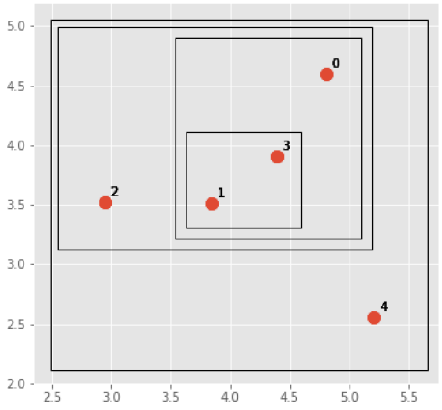
(b)



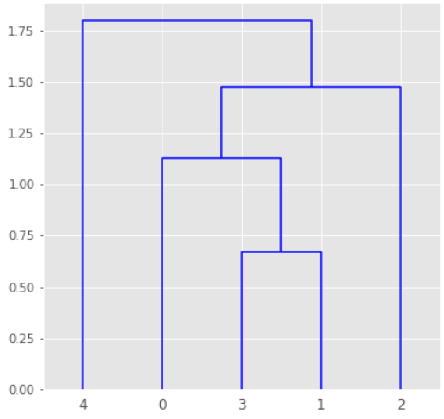
(c)



(d)



(e) Diagram vnořených shluků.



(f) Dendrogram.

Obr. 3.2: Ilustrační příklad: Průběh aglomerativního hierarchického shlukování (centroidní metoda).

3.3 Nehierarchické shlukování

Cílem nehierarchického shlukování je vytvoření optimálního rozkladu množiny objektů. Jednotlivé shluky se nepřekrývají - každý objekt je zařazen právě do jednoho shluku. Požadujeme tedy dodržení následujících podmínek:

$$\begin{aligned} \bigcup_{j=1}^{k} \mathcal{C}_j &= \mathcal{X}, \\ \mathcal{C}_i \cap \mathcal{C}_j &= \emptyset \quad \text{pro } i \neq j, \\ \emptyset \subset \mathcal{C}_j &\subset \mathcal{X}, \text{ tj. } 0 < |\mathcal{C}_j| < n, \quad j = 1, \ldots, k. \end{aligned}$$

Počet shluků k přitom musíme sami určit. Nalezením vhodného k se budeme zabývat později, nyní předpokládejme, že počet shluků známe.

3.3.1 Algoritmus K-Means

Algoritmus K-Means je další z nejstarších a nejpoužívanějších shlukovacích algoritmů. Cílem tohoto algoritmu je minimalizace účelové funkce, která vyjadřuje kvalitu rozkladu. Tuto účelovou funkci v Euklidovském prostoru definujeme následovně

$$\mathrm{SSE}(\mathcal{C}_0) = \sum_{j=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_j} \left[d(\mathbf{x}, \mathbf{c}_j) \right]^2, \tag{3.3}$$

kde d je euklidovská vzdálenost či nepodobnost objektů, $\mathbf{x} \in \mathcal{C}_j$ je objekt patřící do shluku $\mathcal{C}_j$ a $\mathbf{c}_j$ značí centroid shluku $\mathcal{C}_j$. Označení SSE znamená součet čtverců chyb (sum of the squared error). Při využití kosinové nepodobnosti účelovou funkci označujeme jako úplnou kohezi (total cohesion) a nabývá tvaru

$$\mathrm{TC}(\mathcal{C}_0) = \sum_{j=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_j} d(\mathbf{x}, \mathbf{c}_j), \tag{3.4}$$

kde d je kosinová nepodobnost objektů.

V případě Euklidovské i kosinové nepodobnosti se dá ukázat (viz [13]), že centroidy minimalizující účelovou funkci jsou těžiště jednotlivých shluků. Tedy centroid $\mathbf{c}_j$ vypočítáme jako

$$\mathbf{c}_j = \frac{1}{|\mathcal{C}_j|} \sum_{\mathbf{x} \in \mathcal{C}_j} \mathbf{x}.$$

Postup této metody je popsán Algoritmem 2.

Výsledek metody K-Means závisí na počátečním výběru centroidů. Tato metoda tedy konverguje pouze k lokálnímu optimu, proto se často algoritmus použije několikrát s různými počátečními centroidy a vybere se takový rozklad $\mathcal{C}_0$, který nabývá nejmenší hodnoty účelové funkce $J(\mathcal{C}_0)$. Kromě toho, K-Means je také citlivý na outliery. S těmi se lépe vypořádá algoritmus K-Medoids, viz sekce 3.3.6.

Algoritmus 2 K-Means

1: Vyber k objektů jako počáteční centroidy.
2: **repeat**
3: Vytvoř k shluků přiřazením každého objektu k jeho nejbližšímu centroidu.
4: Přepočti centroidy každého shluku.
5: **until** Centroidy se nezmění.

Výběr počátečních centroidů

Nejjednodušším způsobem jak vybrat počáteční centroidy je vybrat k náhodných bodů z prostoru $\mathbb{E}_m$. V tomto případě však může vzniknout problém, kdy dostaneme jeden nebo více prázdných shluků. Druhým způsobem, který tento problém řeší, je vybrat jako centroidy k objektů z $\mathcal{X}$. Další možností je použít algoritmus K-Means++, viz 3.3.3.

3.3.2 Ilustrační příklad

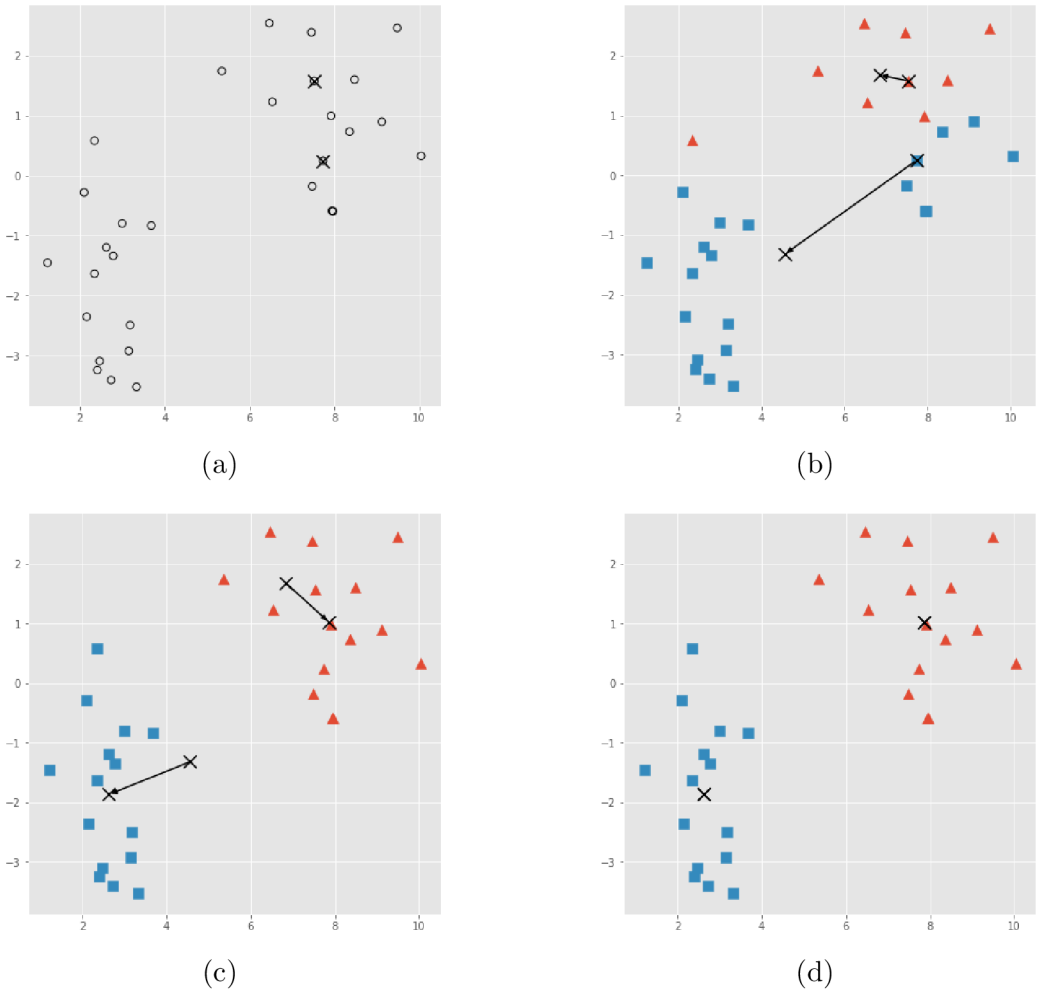
Na ilustračním příkladě ukážeme jednotlivé kroky algoritmu K-Means. Na Obr. 3.3a máme 30 dvourozměrných bodů a dva náhodně vybrané počáteční centroidy označené křížkem. Na Obr. 3.3b vidíme, jak se body přiřadily k nejbližšímu počátečnímu centroidu podle Euklidovské vzdálenosti a vypočítaly se nové centroidy (nová těžiště). Na Obr. 3.3c se některé body přesunuly do jiného shluku a opět se tak změnily centroidy. Na Obr. 3.3d se již žádné body nepřesunuly, centroidy zůstaly nezměněny, a proto byl algoritmus ukončen.

3.3.3 K-Means++

Algoritmus K-Means++ se od obyčejného K-Means liší v počátečním výběru centroidů a obecně dosahuje lepších výsledků (menší hodnoty účelové funkce). Princip spočívá v tom, že se první centroid vybere náhodně, jako druhý centroid se vybere objekt co nejvíce vzdálený (odlišný/nepodobný) od prvního centroidu a třetím centroidem se pak stane objekt co nejvíce vzdálený od obou předchozích. Takhle se postupuje dále dokud není stanoveno k počátečních centroidů. Ve všech těchto krocích má každý objekt pravděpodobnost že bude vybrán jako nový centroid úměrnou čtverci vzdálenosti k jeho nejbližšímu centroidu. Tento postup popisuje Algoritmus 3.

Algoritmus 3 K-Means++

1: Jako první centroid náhodně vyber jeden objekt.
2: **repeat**
3: Pro každý objekt $\mathbf{x}$ vypočti jeho vzdálenost $d(\mathbf{x})$ k nejbližšímu centroidu.
4: Vyber náhodně jeden nový objekt jako nový centroid užitím váženého pravděpodobnostního rozdělení, kde objekt $\mathbf{x}$ je vybrán s pravděpodobností úměrné $d(\mathbf{x})^2$.
5: **until** Je vybráno k centroidů.
6: Nyní pokračuj standardním K-Means algoritmem.



Obr. 3.3: Ilustrační příklad: Průběh algoritmu K-Means.

3.3.4 Bisecting K-Means

Bisecting K-Means je rozšíření algoritmu K-Means. Pro získání k shluků je třeba rozdělit množinu objektů na dva shluky, vybrat jeden shluk který bude opět rozdělen na dva shluky a tak dále dokud není vytvořeno k shluků. Tento postup popisuje Algoritmus 4. Existuje několik různých postupů jak vybrat který shluk rozdělit, např. může být vybrán největší shluk nebo shluk s nejhorší hodnotou účelové funkce $J(\mathcal{C}_0)$. Protože je algoritmus K-Means používán pouze „lokálně“, výsledný rozklad nepředstavuje lokální minimum vzhledem k účelové funkci. Proto se centroidy výsledných shluků používají jako počáteční centroidy pro obyčejný algoritmus K-Means.

3.3.5 Snižování hodnoty účelové funkce postprocessingem

K-Means obvykle konverguje k lokálnímu minimu a protože účelová funkce je součtem jakéhosi měřítka kvality jednotlivých shluků, můžeme její hodnotu snížit úpravou

Algoritmus 4 Bisecting K-Means

```
1: Vytvoř seznam shluků obsahující jeden shluk obsahující všechny objekty.
2: repeat
3:     Odeber shluk ze seznamu shluků.
4:     {Proveď několik pokusů bisekce vybraného shluku.}
5:     for i = 1 to počet pokusů do
6:         Rozděl vybraný shluk na dva využitím algoritmu K-Means.
7:     end for
8:     Vyber dva shluky z bisekce s nejmenší hodnotou účelové funkce.
9:     Přidej tyto dva shluky do seznamu shluků.
10: until Seznam shluků obsahuje k shluků.
```

jednotlivých shluků. Shluky můžeme dělit nebo spojovat s jinými shluky, tím se však mění počet shluků. Pro zachování počtu shluků tak můžeme tyto dva přístupy kombinovat.

Pro snížení hodnoty účelové funkce zvýšením počtu shluků můžou být použity následující metody.

- **Rozdělení shluku:** Obvykle je vybrán shluk s největší hodnotou SSE/TC nebo jiným měřítkem kvality rozkladu.
- **Vytvoření nového centroidu:** Často je vybrán objekt nejvzdálenější od libovolného centroidu. Dalším přístupem je náhodně vybrat objekt ze všech objektů nebo z objektů, které jsou nejvíce nepodobné „svým“ centroidům.

Pro snížení počtu shluků s minimálním přírůstkem hodnoty účelové funkce se používají následující metody.

- **Rozptýlení shluku:** Centroid shluku, který přispívá k hodnotě účelové funkce nejméně je odstraněn a objekty z tohoto shluku jsou připsány k ostatním shlukům.
- **Spojení dvou shluků:** Obvykle jsou vybrány shluky, jejichž centroidy jsou si nejblíže. Tento přístup odpovídá přístupu hierarchického shlukování centroidní metodou.

3.3.6 K-Medoids

Myšlenka metody K-Medoids je podobná K-Means. Reprezentantem shluku však není centroid, ale medoid, který je přímo objektem shlukovaných dat. Medoidem $\mathbf{m}_j$ shluku $\mathcal{C}_j$ se stává objekt, který minimalizuje součet nepodobností tohoto objektu s ostatními objekty shluku, tedy

$$\mathbf{m}_j = \arg\min_{\mathbf{x} \in \mathcal{C}_j} \sum_{\mathbf{y} \in \mathcal{C}_j} d(\mathbf{x}, \mathbf{y}).$$

Jak již bylo dříve zmíněno, K-Medoids je, na rozdíl od K-Means, méně citlivý na outliery. Metodu popisuje Algoritmus 5. Výhodou je, že se v každé iteraci nemusí znova přepočítávat vzdálenosti objektů od centroidů, stačí pouze na začátku algoritmu vypočítat matici nepodobnosti.

Algoritmus 5 K-Medoids

1: Vyber k objektů jako počáteční medoidy.
2: **repeat**
3: Vytvoř k shluků přiřazením každého objektu k jeho nejbližšímu medoidu.
4: Přepočti medoidy každého shluku.
5: **until** Medoidy se nezmění.

3.4 Fuzzy shlukování

Ve většině případů objekty netvoří dobře separované shluky. Proto může být adekvátnější objektům přiřadit určitou váhu s kterou do daného shluku patří. Objekt velmi podobný centroidu shluku tak bude mít vyšší váhu než objekt ležící na hranici shluku. Tento přístup je možný s využitím teorie fuzzy množin, jenž pracuje s funkcí příslušnosti, která nabývá hodnot z intervalu $\langle 0, 1 \rangle$.

3.4.1 Teorie fuzzy množin

Teorie fuzzy množin a fuzzy logiky je přesná matematická disciplína, která umožňuje zahrnout nepřesnost a pracovat tak s významy slov přirozeného jazyka. Za zakladatele této teorie se považuje profesor kalifornské univerzity v Berkeley Lotfi A. Zadeh, jehož první práce na toto téma byla vydána v roce 1965. Dnes má tato teorie již řadu využití, a to především v řízení a regulaci (expertní systémy), dále např. v rozpoznávání obrazů či klasifikaci a shlukování.

Pojem *fuzzy* se do češtiny překládá jako mlhavý, vágní, rozmazaný, neostrý, neurčitý. Neurčitost obecně zahrnuje vágnost a nejistotu, o kterých by se dalo říct, že jsou vzájemně komplementární. Jaký je tedy mezi teorií fuzzy množin, která se zabývá vágností, a teorií pravděpodobnosti a statistiky, která se zabývá nejistotou, rozdíl?

Nejistota charakterizuje jev, který může, avšak nemusí nastat. Základním pojmem v teorii pravděpodobnosti je rozdělení pravděpodobnosti. Ten popisuje, s jak velkou pravděpodobností může nastat jev z nějaké množiny různých jevů, o nichž víme, že jeden z nich určitě nastane. Pravděpodobnost nám pak říká, zda nastání určitého jevu z uvažovaných jevů můžeme očekávat s vyšší jistotou, než nastání jiného jevu.

Teorie fuzzy množin pracuje s fuzzy množinami objektů a stupni jejich příslušnosti do ní. Jde o vymezení jevu, ne o charakterizaci toho, zda nastane či ne. (Např. chceme pracovat s množinou vysokých lidí. Otázkou je, co je to „vysoký“? Každý má totiž svou vlastní představu vysokého člověka. Navíc nemůžeme určit přesný interval výšky od - do, který tuto skupinu lidí charakterizuje.)

Jednoduše tedy můžeme říct, že pravděpodobnost odpovídá na otázku, zda „něco nastalo“, kdežto teorie fuzzy množin odpovídá na otázku „co nastalo“. Více o teorii fuzzy množin a její aplikaci lze najít např. v knihách [9] a [12].

3.4.2 Fuzzy C-Means

Fuzzy C-Means, označovaný taky jako FCM, je fuzzy verze K-Means. Požadavkem při tomto shlukování je vytvoření fuzzy pseudo-rozkladu. Aby toho bylo dosaženo, klademe následující podmínky:

- Součet příslušností objektu k jednotlivým shlukům je roven jedné.

$$\sum_{j=1}^{k} w_{ij} = 1 \tag{3.5}$$

- Každý shluk obsahuje alespoň jeden objekt s nenulovou příslušností a zároveň neobsahuje všechny objekty s příslušností rovné jedné.

$$0 < \sum_{i=1}^{n} w_{ij} < n \tag{3.6}$$

Stejně jako K-Means, FCM můžeme chápat jako minimalizaci účelové funkce:

$$J(\mathcal{C}_0) = \sum_{j=1}^{k} \sum_{i=1}^{n} w_{ij}^{p} \left[d(\mathbf{x}_i, \mathbf{c}_j)\right]^2 .$$

Fuzzy parametr p

Parametr p určuje vliv vážení a může nabývat hodnot z intervalu $(1, \infty)$. Pokud je tento parametr blízký 1, pak se FCM blíží tradičnímu K-Means. Naopak, se zvětšováním p se centroidy všech shluků blíží globálnímu centroidu všech objektů. Jinými slovy, s rostoucím parametrem p se rozklad stává více fuzzy.

Inicializace

Váhy jsou obvykle voleny náhodně, avšak musí splňovat podmínku (3.5). Tato počáteční volba však většinou vede k lokálnímu minimu účelové funkce. Mohou být také použity metody výběru počátečních centroidů popsané v kapitole K-Means.

Výpočet centroidů

Centroid $\mathbf{c}_j$ shluku $\mathcal{C}_j$ je definován rovnicí (3.7).

$$\mathbf{c}_j = \sum_{i=1}^{n} w_{ij}^{p} \mathbf{x}_i / \sum_{i=1}^{n} w_{ij}^{p}. \tag{3.7}$$

Definice fuzzy centroidu je podobná definici centroidu pro K-Means, rozdílem však je zohlednění všech objektů (každý objekt může patřit do libovolného shluku) a příspěvek každého z nich je dán jeho stupněm příslušnosti do daného shluku. V případě obyčejných množin, kde váhy nabývají pouze hodnot 0 nebo 1, se tato definice redukuje na definici tradičního centroidu.

Přepočítání fuzzy pseudo-rozkladu

Fuzzy pseudo-rozklad je definován pomocí vah, proto tento krok zahrnuje přepočítání váhy w_{ij} odpovídající i-tému objektu a j-tému shluku pro všechna i, j pomocí vztahu (3.8).

$$w_{ij} = (1/d(\mathbf{x}_i, \mathbf{c}_j)^2)^{\frac{1}{p-1}} \Big/ \sum_{q=1}^{k} (1/d(\mathbf{x}_i, \mathbf{c}_q)^2)^{\frac{1}{p-1}} \ . \tag{3.8}$$

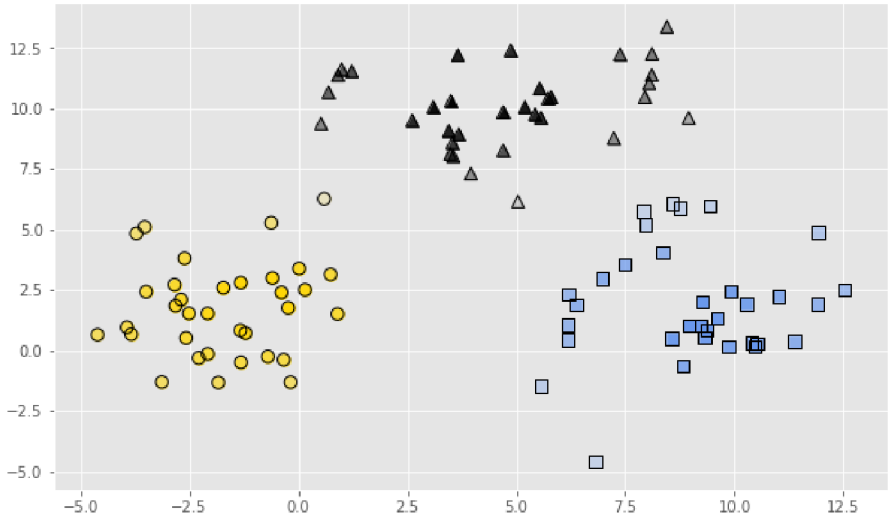
Postup fuzzy C-Means popisuje Algoritmus 6.

Algoritmus 6 Fuzzy C-Means

1: Vyber počáteční fuzzy pseudo-rozklad (tj. přiřaď hodnoty všem w_{ij}).
2: **repeat**
3: Vypočti centroid každého shluku s použitím fuzzy pseudo-rozkladu.
4: Přepočti fuzzy pseudo-rozklad.
5: **until** Centroidy se nezmění.

Ilustrační příklad

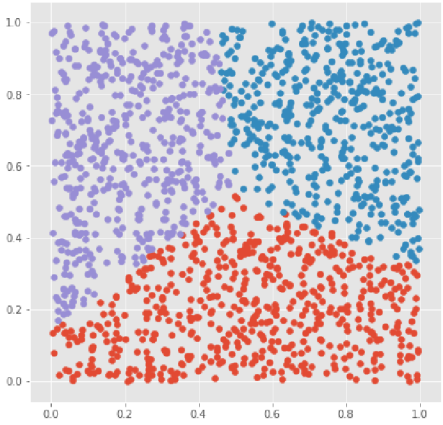
Obr. 3.4 ukazuje 3 shluky nalezené algoritmem FCM. Objekty byly přiřazeny do takového shluku, u kterého měly nejvyšší stupeň příslušnosti. Tento stupeň (váha) odráží průhlednost daného objektu. Lze vidět, že nejprůhlednější (a tedy s nejmenším stupněm příslušnosti) jsou objekty na hranicích shluků, pak objekty na odlehlých okrajích a největší váhu mají objekty v blízkosti centroidu.



Obr. 3.4: Fuzzy C-Means.

3.5 Validace shluků

V této kapitole bude uvedeno několik metod jak vyhodnotit vytvořené shluky. Vyhodnocení těchto shluků je důležitou součástí shlukové analýzy. Každý shlukovací algoritmus může totiž najít v datech shluky i přesto, že struktura dat neobsahuje žádné přirozené shluky. Příkladem můžou být náhodně distribuovaná data - Obr. 3.5 ukazuje 3 shluky nalezené algoritmem K-Means v takovýchto datech.



Obr. 3.5: 3 shluky nalezené algoritmem K-Means v rovnoměrně rozložených datech.

Mezi problémy validace shluků patří:

1. Určení existence nenáhodné struktury v datech.
2. Určení správného počtu shluků.
3. Vyhodnocení výsledků shlukové analýzy bez externích informací.
4. Porovnání výsledků shlukové analýzy s externě známými výsledky (např. předem určené označení objektu podle kategorie do které patří).
5. Porovnání dvou rozkladů shlukové analýzy a vyhodnocení který z nich je lepší.

Vyhodnocovací měřítka využívaná k validaci se dělí na měřítka:

- **bez učitele** (unsupervised): K vyhodnocení není použita žádná externí informace, příkladem může být účelová funkce. Obvykle se rozlišují dvě měřítka: **koheze (kompaktnost) shluku**, které určuje jak moc jsou si objekty v daném shluku blízké (podobné) a **separace (izolace) shluku**, které určuje jak vzdálený (nepodobný) nebo dobře separovaný je daný shluk od ostatních shluků.
- **s učitelem** (supervised): Měří jak moc se výsledky shlukové analýzy shodují s externí informací, příkladem může být entropie. Těmito metodami se práce dále nebude zabývat, známá měřítka jsou popsána např. v [13].

3.5.1 Koheze a separace shluku

Koheze shluku může být definována pomocí (3.9) jako součet nepodobností objektů shluku s jeho prototypem (centroidem nebo medoidem). Podobně, separace mezi dvěma shluky může být měřena nepodobností prototypů těchto shluků ((3.10)).

$$\text{koheze}(\mathcal{C}_i) = \sum_{\mathbf{x}\in\mathcal{C}_i} d(\mathbf{x}, \mathbf{c}_i), \tag{3.9}$$

$$\text{separace}(\mathcal{C}_i, \mathcal{C}_j) = d(\mathbf{c}_i, \mathbf{c}_j). \tag{3.10}$$

3.5.2 Koeficient siluet

Jedna z nejznámějších metod kombinující kohezi a separaci je metoda siluet. Následující kroky popisují jak vypočítat koeficient siluety objektu $\mathbf{x}_i \in \mathcal{A}$ zařazeného do shluku $\mathcal{A}$.

1. Vypočítej průměrnou nepodobnost objektu $\mathbf{x}_i$ se všemi ostatními objekty ze shluku $\mathcal{A}$:
$$a(\mathbf{x}_i) = \frac{1}{|\mathcal{A}| - 1} \sum_{\mathbf{x}_j\in\mathcal{A}} d(\mathbf{x}_i, \mathbf{x}_j).$$

2. Vypočítej průměrnou nepodobnost objektu $\mathbf{x}_i$ se všemi objekty z libovolného shluku neobsahující tento objekt:
$$\bar{d}(\mathbf{x}_i, \mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{\mathbf{x}_j\in\mathcal{C}} d(\mathbf{x}_i, \mathbf{x}_j) \qquad \text{pro } \mathcal{C} \neq \mathcal{A}.$$
Tuto hodnotu vypočítej pro všechny takové shluky a najdi minimum z těchto hodnot:
$$b(\mathbf{x}_i) = \min_{\mathcal{C}\neq\mathcal{A}} \bar{d}(\mathbf{x}_i, \mathcal{C}).$$

3. Vypočítej koeficient siluety objektu $\mathbf{x}_i$:
$$s(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{\max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}}. \tag{3.11}$$

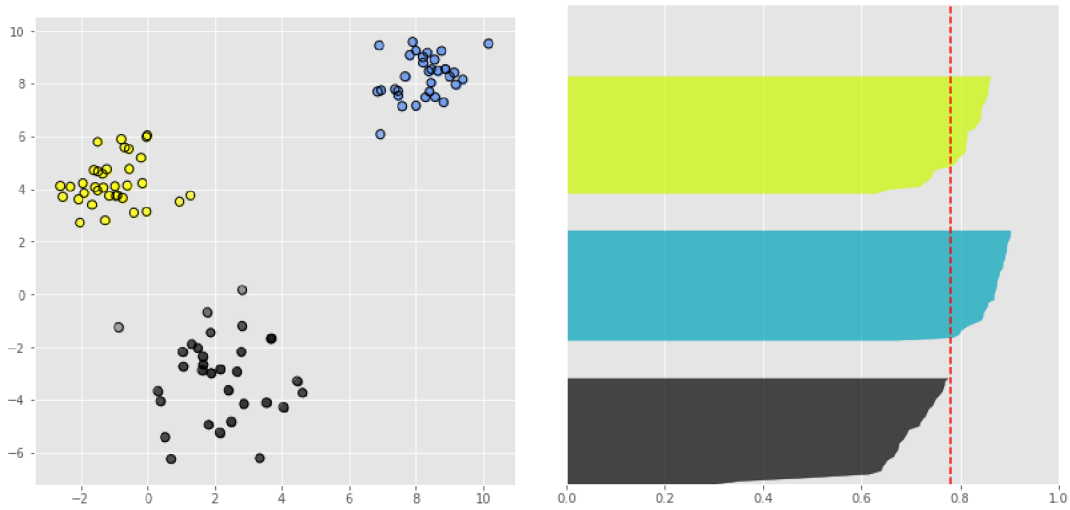
Koeficient siluety (3.11) může být ekvivalentně vyjádřen jako

$$s(\mathbf{x}_i) = \begin{cases} 1 - \frac{a(\mathbf{x}_i)}{b(\mathbf{x}_i)} & \text{pro } a(\mathbf{x}_i) < b(\mathbf{x}_i) \\ 0 & \text{pro } a(\mathbf{x}_i) = b(\mathbf{x}_i) \\ \frac{b(\mathbf{x}_i)}{a(\mathbf{x}_i)} - 1 & \text{pro } a(\mathbf{x}_i) > b(\mathbf{x}_i). \end{cases}$$

Z tohoto zápisu je zřejmé, že $s(\mathbf{x}_i) \in \langle -1, 1\rangle$. Je žádoucí, aby $a(\mathbf{x}_i) \ll b(\mathbf{x}_i)$, tedy aby se koeficient blížil k hodnotě 1. V tomto případě by to znamenalo, že daný objekt je podobný ostatním objektům ve stejném shluku (koheze) a zároveň není moc podobný objektům ze sousedního (nejbližšího/nejpodobnějšího) shluku (separace). Záporná hodnota koeficientu siluet naopak upozorňuje na to, že průměrná vzdálenost (nepodobnost) objektu k ostatním objektům ze stejného shluku je větší než minimální průměrná vzdálenost (nepodobnost) objektu k objektům z jiného shluku - tedy daný objekt patří spíše do sousedního shluku.

Jednotlivé shluky tak mohou být hodnoceny podle průměrného koeficientu siluet objektů patřících do daného shluku, celkový rozklad pak podle průměrného koeficientu siluet všech objektů.

Siluety bývají také znázorňovány dvourozměrným grafem. Siluetu shluku vytvoříme nanesením sestupně či vzestupně seřazených objektů podle jejich hodnoty koeficientu siluety na jednu osu, na druhou osu pak vyneseme příslušnou hodnotu $s(\mathbf{x}_i)$ pro $\forall i$. Ukázku lze vidět na Obr. 3.6. Objekty na Obr. 3.6a jsou barevně odlišeny podle shluků do kterých patří a tato barva také odpovídá barvě siluety patřící danému shluku na Obr. 3.6b. Dále, objekty s nižší hodnotou koeficientu siluet jsou průhlednější. Vidíme, že černý shluk je méně kompaktní než další dva shluky, a proto má také nižší siluetu. Na obrázku siluet je také červenou přerušovanou čarou znázorněn průměrný koeficient siluet rozkladu, v tomto případě nabývá hodnoty $0,78$.



(a) Shluky.

(b) Siluety shluků.

Obr. 3.6: Siluety.

3.5.3 Matice nepodobnosti

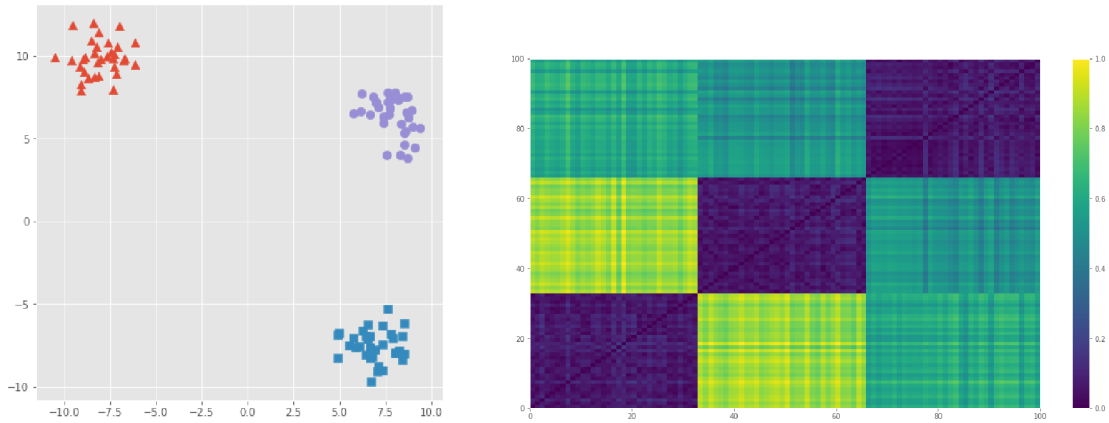
Další metody validace rozkladu shlukové analýzy mohou využívat matici nepodobnosti. Jedním způsobem je výpočet korelačního koeficientu mezi touto maticí a ideální maticí nepodobnosti pro daný rozklad. V ideálním případě objekty jednoho shluku mají hodnotu nepodobnosti se všemi ostatními objekty ze stejného shluku rovnu nule a se všemi objekty ostatních shluků rovnu jedné. Proto, když seřadíme řádky a sloupce matice nepodobnosti tak, aby objekty patřící do jednoho shluku byly u sebe, ideální matice nepodobnosti bude mít blokově diagonální strukturu s nulami uvnitř jednotlivých bloků a jedničkami všude jinde. Ideální matici nepodobnosti vytvoříme jako matici s jedním řádkem a jedním sloupcem pro každý objekt (bude mít tedy velikost $n \times n$) a buňce odpovídající objektům $\mathbf{x}_i, \mathbf{x}_j$ přiřadíme hodnotu 0 pokud jsou tyto objekty v jednom shluku. V opačném případě přiřadíme buňce hodnotu 1. (Pokud nepodobnosti nabývají vyšších hodnot než 1, přepočítáme je pomocí vztahu $(d - d_{min})/(d_{max} - d_{min})$, kde d značí přepočítávanou nepodobnost,

d_{min} je nejmenší hodnota v matici nepodobnosti a d_{max} je největší hodnota v matici nepodobnosti.)

Vysoký korelační koeficient mezi ideální a reálnou maticí nepodobnosti poukazuje na vysokou podobnost objektů s objekty ve stejném shluku, nízký korelační koeficient značí opak.

Vizualizace matice nepodobnosti

Seřadíme matici nepodobnosti tak, aby byly objekty patřící do jednoho shluku vedle sebe, a pak tuto matici vykreslíme. Pokud jsou shluky dobře separované, pak by matice nepodobnosti měla být přibližně blokově-diagonální. Příkladem jsou data na Obr. 3.7a a jejich matice nepodobnosti na Obr. 3.7b (protože osa y směřuje vzhůru, hlavní diagonála vede z levého dolního rohu do pravého horního rohu). Vidíme, že data tvoří tři dobře separované shluky. Shluku vykreslený kolečky odpovídá pravý horní roh matice nepodobnosti a shluku vykreslený čtverečky odpovídá prostřední blok. Tmavě modrá barva odpovídá nulové hodnotě nepodobnosti, žlutá naopak hodnotě 1. Podle světle modré barvy v matici můžeme říct, že „kolečkový“ shluk je do jisté míry podobný (blízký) oběma dalším shlukům. Podle žluté barvy naopak můžeme říct, že „čtverečkový“ a „trojúhelníčkový“ shluk jsou si nejméně podobné (blízké). Korelační koeficient této matice s ideální maticí nepodobnosti nabývá vysoké hodnoty $0,94$.



(a) Shluky. (b) Seřazená matice nepodobnosti.

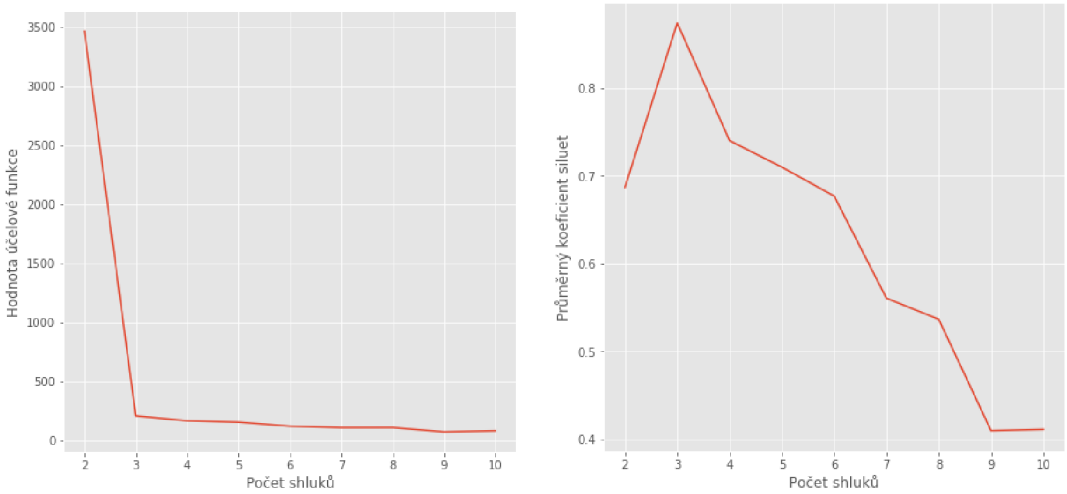
Obr. 3.7: Matice nepodobnosti.

3.5.4 Určení počtu shluků

Určení správného počtu shluků v datech není lehká úloha. Existuje však několik metod, které můžou při rozhodování kolik shluků vytvořit pomoct. Tyto metody spočívají ve vytvoření několika rozkladů pro různý počet shluků a liší se ve výběru měřítka kvality shluků.

Data na Obr. 3.7a mají 3 přirozené shluky. Na Obr. 3.8a je vykreslena hodnota účelové funkce pro $2-10$ shluků vytvořených z těchto dat. Vidíme, že tato hodnota je pro $3-10$ shluků téměř konstantní. Když je počet shluků roven 3, je v tomto grafu vidět výrazné „koleno“ či „loket“, a právě tato hodnota udává vhodný počet shluků. Tato metoda bývá označována jako *elbow method*.

Na Obr. 3.8b je vykreslen průměrný koeficient siluet pro stejný rozsah počtu shluků a stejná data. Nejvyšší hodnoty tento koeficient nabývá pro 3 vytvořené shluky - toto je hledaná hodnota.



(a) Hodnota účelové funkce.

(b) Průměrný koeficient siluet.

Obr. 3.8: Určení počtu shluků.

Stejně může být použita hodnota korelace z předchozí části kapitoly, koheze shluků či jiné měřítko kvality rozkladu. Tyto metody jsou však pro data vysoké dimenze nebo data s překrývajícími se shluky (či s jinak komplikovanou strukturou) nespolehlivé. Můžou však přinést jistý pohled na věc pokud o shlukovaných datech nic nevíme.

4 APLIKACE NA REÁLNÁ DATA

Tato kapitola ukazuje praktické využití shlukové analýzy v oblasti text miningu. Analýza je provedena na hovorech z kontaktního centra České pojišťovny a.s.

4.1 Česká pojišťovna a.s.

> „Česká pojišťovna a.s. je univerzální pojišťovnou poskytující celé spektrum služeb, pokrývající jak individuální životní a neživotní pojištění, tak pojištění pro malé, střední a velké klienty v oblasti průmyslových a podnikatelských rizik a také v zemědělství. Česká pojišťovna a její dceřiné společnosti jsou od roku 2008 součástí skupiny Generali Group.“[3]

Kontaktní centrum České pojišťovny (dále jen ČP) využívá k analýze dat software EMMA (Expert Multichannel Multilingual Analyst) od společnosti DATERA, jejímž základem je kognitivní platforma IBM® Watson™. Tento software dokáže zpracovat a analyzovat hovory, e-maily, chaty a další komunikační kanály. Při přepisu řeči do textu má asi 70% úspěšnost.

4.2 Popis problému

Současným problémem, který ČP řeší je tzv. kategorizace hovorů - cílem je vytvořit vhodné kategorie a jednotlivé dokumenty do nich zařadit tak, aby bylo možné provádět kvalitnější analýzy. Nynější kategorie jsou tvořeny na základě tzv. pravidel. Pravidlem se rozumí query neboli dotaz a je vytvářeno člověkem podobně jako dotaz ve fulltextovém webovém vyhledávači. Dotaz může být tvořen:

- klíčovými slovy,
- booleovskými operátory AND, OR, NOT/-,
- přesnými frázemi (uvádějí se do uvozovek),
- dvěma a více slovy se specifikovanou vzdáleností (`(sjednání smlouvy)WITHIN 5` vyhledá dokumenty, ve kterých se vyskytují tvary slov „sjednání“ a „smlouva“ ve vzájemné vzdálenosti rovné maximálně 5),
- dvěma a více slovy v určeném pořadí (`(sjednání smlouvy)WITHIN 5 INORDER` vyhledá dokumenty jako v předchozím příkladě, avšak slovo „sjednání“ musí být před slovem „smlouva“),
- slovy či frázemi vyhledávaných v předem definovaných polích (datum vzniku dokumentu, linka operátora, ...).

Tento systém vyžaduje kromě vytvoření témat také přesné definování slovních spojení, která danou kategorii vystihují. Daný postup je zdlouhavý a ne zcela objektivní a pokrývá pouze malé procento dokumentů z celé kolekce odpovídajících danému tématu. Vhodné řešení tohoto problému by mohla přinést shluková analýza.

Na Obr. 4.1 je zobrazen zjednodušený dotaz pro vyhledávání hovorů, kde klient reaguje na reklamní kampaň. Pomocí definovaných polí je také specifikováno datum a kdo je volající.

```
(keyword::/"Telefonát"/"Volající"/"CLIENT") AND (date >="2016-10-01
00:00:00") AND (("K:" reklama televize)WITHIN SENTENCE OR ("K:" vidět
reklama)WITHIN SENTENCE OR ("K:" reklama záškodník)WITHIN SENTENCE OR
(v televize)WITHIN 2 INORDER OR ("K:" vidět televize)WITHIN SENTENCE
OR ("K:" v reklama)WITHIN SENTENCE INORDER) -(pojistná událost)WITHIN
3 -(škodní událost)WITHIN 3 -(nahlásit událost)WITHIN 5 -(nahlásit
škodu)WITHIN SENTENCE
```

Obr. 4.1: Ukázka dotazu pro vyhledání reklamní odezvy.

4.3 Popis dat

Nahrané hovory z kontaktního centra ČP přepsané využitím umělé inteligence do textu byly anonymizovány (byly vynechány části s citlivými údaji klienta jako je jméno, kontaktní údaje, číslo smlouvy, číslo pojistné události). Data byla dodána v původním tvaru a v lemmatizovaném tvaru. Každý takovýto dokument se skládá ze dvou zvláštních souborů, kde jeden obsahuje řeč klienta, druhý řeč operátora:

- `576397.traleft`,
- `576397.traright`.

Soubor s koncovkou `.traleft` patří volajícímu (obvykle klient), soubor s koncovkou `.traright` volanému (obvykle operátor). Kolekce obsahuje také hovory, kdy zaměstnanec ČP volá klientovi, případně se může stát, že jsou tzv. *kanály* přehozené. Proto nebudou soubory rozlišovány podle osoby ke které patří. Soubor s koncovkou `.traleft` bude označován jako *levý kanál* (L:), s koncovkou `.traright` jako *pravý kanál* (R:). Ignorováním dokumentů jejichž lemmatizovaný tvar chybí (jeden kanál nebo oba) zůstává celkem 774 hovorů obsažených v 3096 souborech (2 kanály původního přepisu a 2 kanály lemmatizovaného tvaru pro jeden hovor).

Každý soubor obsahuje časové značky, které označují úsek audionahrávky, kdy byl daný úryvek hovoru pronesen. Jsou důležité pro rekonstrukci hovoru (sestavení původního hovoru spojením levého a pravého kanálu). Dále se v přepisu vyskytují značky jako `C_SILENCE`, `C_SURNAME`, `C_NOISE` apod., které v této práci nejsou důležité. Ukázka původního souboru `576397.traleft` je na Obr. 4.2.

Odstraněním všech značek, rozdělením jednotlivých slov od sebe (nahrazením podtržítka _ mezerou) a využitím časových značek je možno získat přepis audionahrávky ve tvaru vhodném ke čtení. Ukázka takto transformovaného předchozího hovoru je na Obr. 4.3.

4.4 Postup a vyhodnocení

Zpracování dat a následná shluková analýza byla provedena v programovacím jazyce Python s využitím knihoven scikit-learn[1], NumPy[2] a SciPy[3]. Pro analýzu byly

[1] http://scikit-learn.org/stable/
[2] http://www.numpy.org/
[3] https://www.scipy.org/

```
0.0000000 2.5441221 C_SILENCE
2.5441221 3.5104763 dobrý_den,_tady_je
3.5104763 3.8768665 C_SURNAME
3.8768665 4.8569603 C_TOWN
4.8569603 5.4179953 C_NOISE
5.4179953 7.4285617 prosím_vás,_potřeboval_bych_informaci
7.4285617 7.7835022 C_SILENCE
7.7835022 9.4322582 já_jsem_zdědil
9.4322582 9.9635240 C_NOISE
9.9635240 11.1199431 zdědil_vozidlo
11.1199431 11.3947358 C_SILENCE
11.3947358 12.7114506 po_zemřelém_otci
12.7114506 14.6670584 C_SILENCE
(...)
```

Obr. 4.2: Ukázka původního souboru `576397.traleft`.

```
R: u telefonu (...), dobrý den
L: dobrý den, tady je
L: prosím vás, potřeboval bych informaci
L: já jsem zdědil
R: ano
L: zdědil vozidlo
L: po zemřelém otci
R: ano
(...)
```

Obr. 4.3: Ukázka rekonstrukce hovoru `576397`.

použity lemmatizované texty a výsledky (nejčastější slova jednotlivých shluků, jejich reprezentující dokument a označení zařazení jednotlivých dokumentů do shluků) včetně použitých parametrů při zpracování byly uloženy do databáze.

Data byla nejprve předzpracována - byly spojeny oba kanály hovoru, dále všechny značky začínající `C_` byly spolu s časovými značkami a interpunkčními znaménky odstraněny a podtržítka byla zaměněna za mezery. Ukázka předzpracovaného textu je na Obr. 4.4.

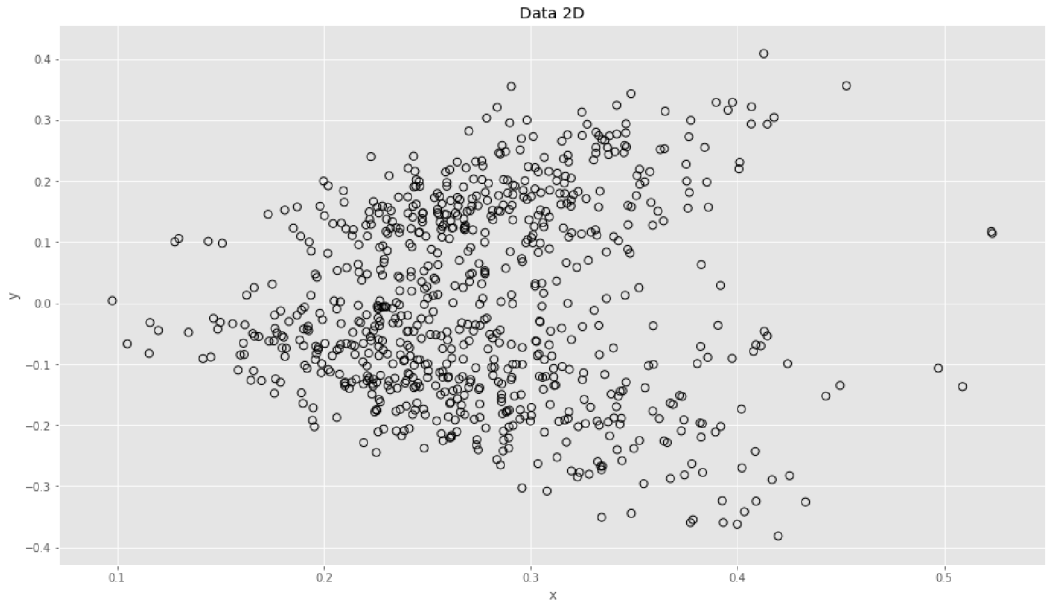
```
u telefon (...) dobrý den dobrý den tady být prosit vás potřebovat
být informace já být zdědit ano zdědit vozidlo po zemřelý otec (...)
```

Obr. 4.4: Ukázka předzpracování hovoru `576397`.

Texty byly dále převedeny do dokument-term matice, přičemž byla provedena redukce dimenze pomocí seznamu stopslov (viz příloha C) a prahování. Seznam stopslov byl převzat z [4] a dále byl postupně v průběhu analýzy doplňován. Původní

velikost dimenze (slovníku) originálních přepisů hovorů byla asi 23 800, lemmatizací byla snížena na hodnotu 13 300. Využitím seznamu stopslov a prahování, kdy hodnota horního prahu byla nastavena na $0,95$ (vynechání slov, která se vyskytují ve více než 95 % dokumentů) a dolního prahu na 3 (vynechání slov, která se vyskytují nejvýše ve dvou dokumentech), byla velikost výsledného slovníku 4 200. Celkově tak byla dimenze snížena o asi 19 600 různých tvarů slov.

Využitím singulárního rozkladu matice mohou být data zobrazena ve 2D. Data při použití TF-IDF vážení a prahování (horní práh = $0,95$, dolní práh = 3) jsou na Obr. 4.5.

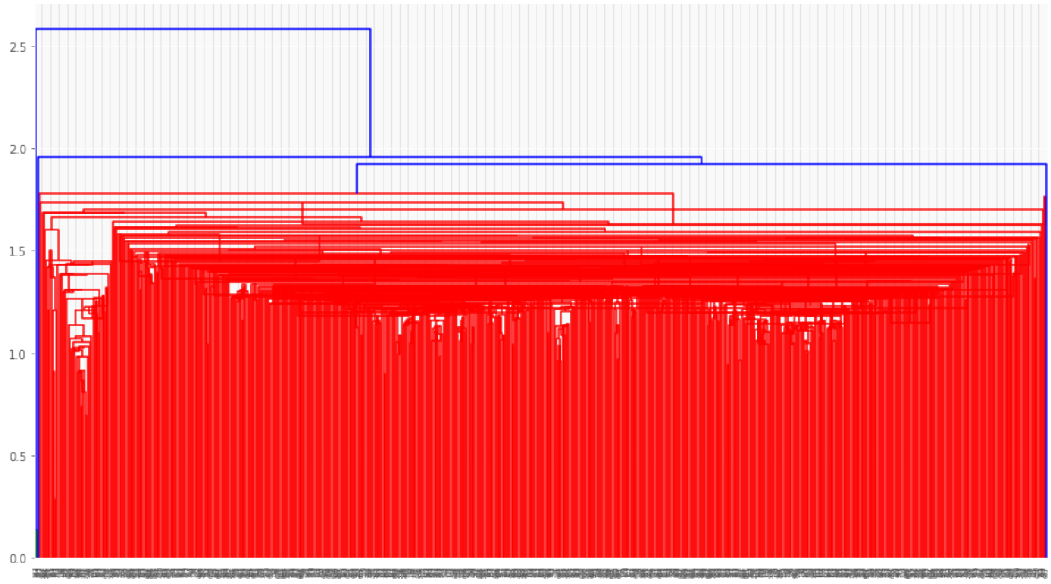


Obr. 4.5: Data převedena do 2D pomocí singulárního rozkladu dokument-term matice.

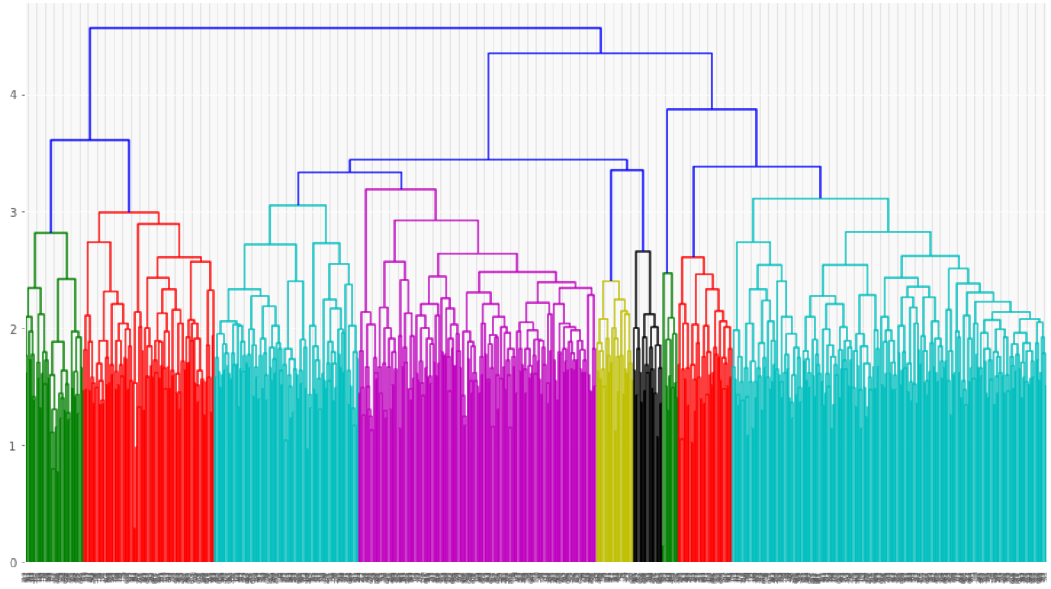
Dále budou popsány použité shlukovací metody. Přitom byla použita kosinová nepodobnost.

Ze shlukovacích metod byly nejprve vyzkoušeny hierarchické aglomerativní. Pro určení vhodného počtu shluků byly pro jednotlivé metody vykresleny dendrogramy. Na Obr. 4.6 je zobrazen dendrogram centroidní metody, dendrogram metody nejbližšího souseda lze nalézt v příloze (Obr. A.1). Z dendrogramů lze usoudit, že tyto metody nejsou vhodné. Dendrogram metody nejvzdálenějšího souseda je na Obr. 4.7. Při vytvoření 9 shluků však podle siluet i matice nepodobnosti (Obr. A.2 a A.3 v příloze) není nejvhodnější ani metoda nejvzdálenějšího souseda. Průměrný koeficient siluet je dokonce záporný a korelační koeficient matice nepodobnosti je v tomto případě roven $0,2846$.

Při shlukování pomocí K-Means++ bylo využito latentní sémantické analýzy. Byl proveden singulární rozklad matice a dimenze dat byla zmenšena na počet vytvářených shluků (pro vytvoření 2 shluků byly vektory dokumentů redukovány na dim = 2 apod.). Kromě urychlení procesu shlukování tak byly dosaženy výsledky



Obr. 4.6: Dendrogram centroidní metody.

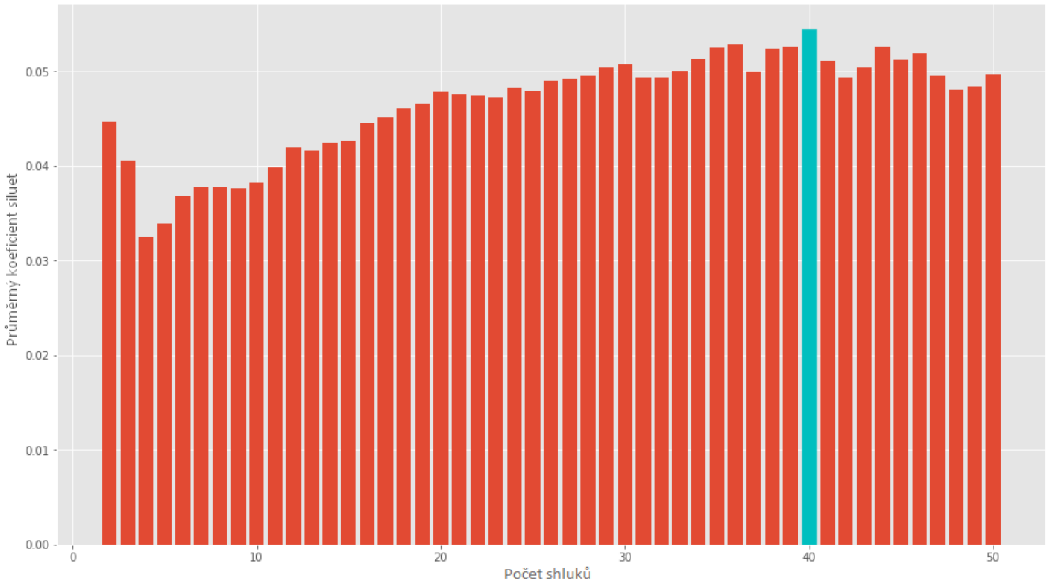


Obr. 4.7: Dendrogram metody nejvzdálenějšího souseda.

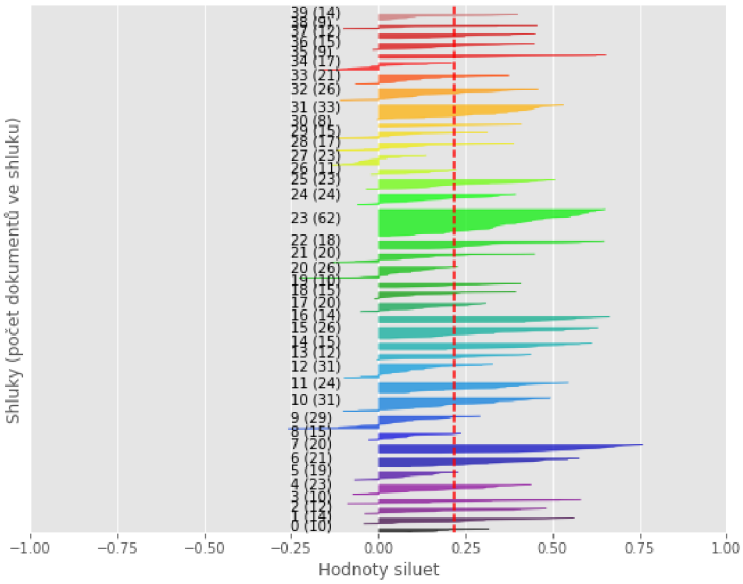
s lepším průměrným koeficientem siluet i korelačním koeficientem mezi maticí nepodobnosti a ideální maticí.

Pro určení vhodného počtu shluků byl algoritmus spuštěn $30\times$ pro $2-50$ shluků a z těchto opakování byl vybrán rozklad s nejlepším průměrným koeficientem siluet. Pro jednotné měřítko byla následná analýza průměrného koeficientu siluet provedena na redukovaných datech s dim $= 500$ (takto redukovaná matice popisuje původní data z 81 %). Na Obr. 4.8 lze vidět, že nejvyššího koeficientu bylo dosaženo pro 40

shluků. Analýza hodnoty účelové funkce pomocí elbow method je v tomto případě nepoužitelná, viz Obr. A.4 v příloze. Průměrný koeficient siluet nabývá hodnoty $0,2196$ a koeficient korelace matice nepodobnosti s ideální maticí $0,4291$. Jednotlivé siluety jsou zobrazeny na Obr. 4.9, matici nepodobnosti lze nalézt v příloze (Obr. A.5).



Obr. 4.8: Analýza průměrného koeficientu siluet při použití K-Means++.



Obr. 4.9: Siluety 40 shluků vytvořených K-Means++.

Pomocí top slov (nejvýznamnější slova popisující shluk) a prohlédnutí několika textů zařazených do příslušného shluku bylo odhadnuto téma a vytvořen název každého vzniklého shluku. Výsledky s číselným označením shluku jsou uvedeny v Tab. 4.1.

U některých shluků bylo těžké zjistit, v čem se liší od jiných (např. shluky č. 8 a 26 nebo 10 a 25). Proto bylo na shluky aplikováno hierarchické shlukování, kde jako počáteční objekty (shluky) byly zvoleny centroidy původních shluků. Konkrétně byla využita metoda nejbližšího souseda a vážená metoda průměrné vazby. Dendrogram metody nejbližšího souseda je zobrazen v příloze na Obr. A.6. Podle této shlukovací metody bylo vytvořeno 24 shluků, Tab. 4.2 ukazuje, jak byly původní shluky sloučeny. Výsledky vážené metody průměrné vazby jsou v Tab. 4.3, dendrogram lze nalézt v příloze (Obr. A.7). Touto metodou bylo získáno 18 témat.

Kvůli možnosti rozlišit více témat jednoho dokumentu bylo provedeno fuzzy shlukování pomocí algoritmu fuzzy C-Means. V tomto kroku je důležité zvolit vhodnou hodnotu parametru p určující velikost vlivu fuzzifikace. Pro příliš velké p se všechny příslušnosti dokumentu do jednotlivých shluků blíží hodnotě $1/k$, kde k je počet shluků. Stejný problém nastává pro příliš velkou dimenzi úlohy. Proto byla opět využita latentní sémantická analýza a pro vytvoření k shluků byly vektory dokumentů redukovány na $\dim = k$.

Pro vytvoření tématicky podobných shluků získaných v předchozím kroku (40 shluků metodou K-Means++) byly pro algoritmus FCM jako počáteční centroidy zvoleny centroidy předchozích shluků. Fuzzy parametr byl nastaven na hodnotu $1,2$, pro větší fuzzifikaci je možné tento parametr zvýšit.

Výsledné kategorie zůstaly nezměněny, rozdílné je pouze zařazení některých dokumentů (podle nejvyššího stupně příslušnosti). Kromě toho byl každý dokument zařazen do každé takové kategorie, u které byl jeho stupeň příslušnosti větší než $1/k = 1/40$. Matice nepodobnosti pro tento rozklad (s přiřazením dokumentů do kategorie s největším stupněm příslušnosti) se nachází v příloze na Obr. A.8.

Ukázku zařazení dokumentů do více shluků lze vidět v Tab. 4.4. Ve sloupci *id* se nachází identifikátor souboru, sloupec *cl_1* obsahuje číslo shluku, do kterého byl dokument zařazen na základě nejvyššího stupně příslušnosti, *cl_2* podobně obsahuje číslo shluku, do kterého byl dokument zařazen, avšak s nižším stupněm příslušnosti atd. Buňky obsahující „-“ naznačují stupeň příslušnosti nižší než 1/40.

Lze vidět, že dokument `281100` byl podle FCM zařazen do shluků č. 23, 6 a 12, tedy *Úrazové pojištění (info, nabídka)*, *PU úraz* a *Kontakt na likvidátora*. Podle obyčejného K-Means++ byl zařazen do shuku č. 23. Ačkoli toto téma danému souboru odpovídá (operátor se klienta na konci hovoru ptá, zda má sjednáno úrazové pojištění), není to jediné téma (klient se nemůže dovolat likvidátorovi, chce vyřešit trvalé následky úrazu). Přepis tohoto hovoru lze nalézt v příloze.

Výsledky získané metodou K-Means++ je možné vylepšit zvýšením průměrného koeficientu siluet, a to převedením objektů se zápornou hodnotou siluety do nejbližšího sousedního shluku. Tab. 4.5 ukazuje, že použitím algoritmu fuzzy C-Means se většinou tyto „chyby“ samy opraví. Sloupec *cluster* obsahuje číslo shluku, do kterého byl dokument zařazen metodou K-Means++, sloupec *new-label* číslo nejbližšího sousedního shluku do kterého by dokument měl být přeřazen a zbytek označení je stejné jako v předchozí tabulce.

Tab. 4.1: Vytvořené shluky metodou K-Means++.

č. shluku	**Název**
0	Zrušení pojištění (odpovědnost, domácnost)
1	Kancelář pojistitelů
2	Zasílání dokumentů přes web
3	Krycí dopis
4	Platby
5	Smlouvy a PU vedené na firmu, odpovědnost z výkonu povolání
6	PU úraz
7	Indexace
8	Platby (variabilní symbol)
9	PU za firmu/klienta, odpovědnost
10	Hlášení PU (HAV/POV)
11	Nabídka pojištění domácnosti
12	Kontakt na likvidátora
13	Doklady k PU (HAV/POV)
14	PU čelní sklo
15	Platby (složenka)
16	PU (domácnost, stavba)
17	PU jiná pojišťovna/zahraničí
18	Pobočka / web / cestovní pojištění
19	HAV/POV bonus
20	Zánik/zrušení pojištění
21	PU faktury
22	Zelená karta
23	Úrazové pojištění (info, nabídka)
24	Odpovědnost z výkonu povolání (PU, platby, nabídka)
25	Hlášení PU (HAV/POV)
26	Platby (variabilní symbol)
27	Zrušení pojištění, platby
28	Platby, nabídky (domácnost, stavba, odpovědnost)
29	Zrušení pojištění (HAV/POV), platby
30	PU HAV/POV - firma / makléřská společnost
31	Zrušení životní pojistky
32	PU - firma
33	HAV/POV smlouvy
34	Domácnost, stavba (info, nabídka)
35	Změna adresy
36	Volání za někoho jiného
37	Zrušení pojištění (prodej, úmrtí)
38	PU dopravní nehoda
39	Platby

Tab. 4.2: Sloučení shluků metodou nejbližšího souseda.

č. shluku	**Název**
1	Změna adresy
2	Indexace
3	Zrušení pojištění (odpovědnost, domácnost)
4	PU úraz
5	Volání za někoho jiného
6	Domácnost, stavba (info, nabídka) Nabídka pojištění domácnosti
7	Platby, nabídky (domácnost, stavba, odpovědnost) Platby Platby Platby (variabilní symbol) Platby (variabilní symbol)
8	Platby (složenka)
9	Zánik/zrušení pojištění Zrušení pojištění, platby Zrušení životní pojistky
10	Úrazové pojištění (info, nabídka)
11	Zrušení pojištění (HAV/POV), platby
12	Zrušení pojištění (prodej, úmrtí)
13	Pobočka / web / cestovní pojištění
14	HAV/POV smlouvy
15	HAV/POV bonus
16	Zelená karta
17	Krycí dopis
18	PU čelní sklo
19	Hlášení PU (HAV/POV) Hlášení PU (HAV/POV) PU dopravní nehoda Doklady k PU (HAV/POV)
20	Kancelář pojistitelů
21	PU HAV/POV - firma / makléřská společnost Odpovědnost z výkonu povolání (PU, platby, nabídka) Smlouvy a PU vedené na firmu, odpovědnost z výkonu povolání PU za firmu/klienta, odpovědnost Kontakt na likvidátora PU - firma PU jiná pojišťovna/zahraničí
22	PU faktury
23	Zasílání dokumentů přes web
24	PU (domácnost, stavba)

Tab. 4.3: Sloučení shluků váženou metodou průměrné vazby.

č. shluku	**Název**
1	Krycí dopis
2	PU čelní sklo
3	Hlášení PU (HAV/POV)
	Hlášení PU (HAV/POV)
	PU dopravní nehoda
	Doklady k PU (HAV/POV)
	Kancelář pojistitelů
4	Zasílání dokumentů přes web
5	PU HAV/POV - firma / makléřská společnost
	PU - firma
	PU jiná pojišťovna/zahraničí
	PU faktury
	Kontakt na likvidátora
	PU za firmu/klienta, odpovědnost
	Smlouvy a PU vedené na firmu, odpovědnost z výkonu povolání
	Odpovědnost z výkonu povolání (PU, platby, nabídka)
6	PU (domácnost, stavba)
7	Zrušení pojištění (odpovědnost, domácnost)
8	Zrušení pojištění (prodej, úmrtí)
	Zrušení pojištění (HAV/POV), platby
9	Pobočka / web / cestovní pojištění
10	Úrazové pojištění (info, nabídka)
	Zrušení životní pojistky
	Zrušení pojištění, platby
	Zánik/zrušení pojištění
11	Platby (variabilní symbol)
	Platby (variabilní symbol)
	Platby (složenka)
	Platby
	Platby
	Platby, nabídky (domácnost, stavba, odpovědnost)
12	Domácnost, stavba (info, nabídka)
	Nabídka pojištění domácnosti
	Volání za někoho jiného
13	HAV/POV smlouvy
14	HAV/POV bonus
15	Zelená karta
16	PU úraz
17	Změna adresy
18	Indexace

Tab. 4.4: Ukázka zařazení dokumentů do více než jednoho shluku podle FCM.

id	cl_1	cl_2	cl_3	cl_4	cl_5	cl_6
576397	33	37	29	22	19	27
270516	12	30	32	-	-	-
362044	36	-	-	-	-	-
908964	6	17	9	-	-	-
281100	23	6	12	-	-	-

Tab. 4.5: Porovnání zařazení pomocí K-Means++, vylepšení podle siluet a FCM.

id	cluster	new-label	cl_1	cl_2	cl_3	cl_4
752562	20	29	29	27	20	0
899114	26	8	8	26	-	-
493223	28	15	15	28	4	39
781296	36	19	19	36	22	-
186646	29	27	34	33	27	29

4.5 Srovnání výsledků s kategorizací ČP

Z důvodu tvaru pravidel tvořících kategorizaci v ČP (která využívají značky `K:` a `O:` pro rozlišení klienta a operátora) bylo nutné najít hovory, ve kterých levý kanál patří operátorovi (nebo jinému zaměstnanci ČP), aby bylo možné hovory nahrát do analytického prostředí ČP a porovnat výsledky shlukové analýzy se současnou kategorizací hovorů ČP. K tomuto byla opět využita shluková analýza, konkrétně byla provedena pouze na přepisu levého kanálu pomocí K-Means++. Kromě kategorií podobných těm výsledným byly nalezeny shluky, u kterých byla jako top slova vypsána slova často používaná právě operátory. Celkově tak bylo nalezeno 134 hovorů, u kterých levý kanál „patřil“ zaměstnanci ČP (jednalo se o hovory likvidátorů, domluvené callbacky či přehozené kanály).

4.5.1 Zastoupení jednotlivých témat

Současné kategorie vytvořené pomocí pravidel (dotazů/query) mají jednoduchou strukturu (nejsou zde žádné podkategorie). Kromě samotných kategorií byla vytvořena také pravidla pro zaznamenání produktů ČP. Obojí je možné vidět v Tab. 4.6.

Hierarchickým shlukováním původních 40 shluků došlo ke sloučení tématicky podobných shluků, avšak některé shluky by mohly zůstat odděleně či by mohly být přiřazeny do jiné skupiny. Vytvořené shluky byly nakonec rozřazeny do 9 hlavních kategorií a celkově byly vytvořeny 3 úrovně, viz Tab. 4.7.

V současném systému může být každý dokument zařazen do více než jedné kategorie. Sečtením podílů jednotlivých kategorií (počet dokumentů v kategorii / počet dokumentů v kolekci) je tak možno získat více než 100 %, v případě analyzovaných dokumentů je to dokonce 351 %. Analýza zastoupení témat hovorů podle hlavního

Tab. 4.6: Současná kategorizace (na základě pravidel).

Kategorie	**Produkty**
Neobdržel jsem / nemám / nevím	Pojištění staveb
Chci ověřit / nerozumím	Pojištění domácnosti
Chci nahlásit	Povinné ručení
Chci sjednat/aktualizovat PS	Odpovědnost
Chci zaslat	Úrazové pojištění
Přepojení	Havarijní pojištění
Opakované hovory	Život - běžně placený
Chci mluvit	Cestovní pojištění
Chci zrušit / stornovat PS	ČP Penzijní společnost
WWW	Život - jednorázově placený
Chci změnit	Pojištění domácích zvířat (PET)
Udělali jste chybu	Podnikatelské pojištění
Měl jsem zavolat/mám zmeškaný hovor	ČP Zdraví
Chci vrátit	
Omyl	
Nejsem spokojený/nesouhlasím	
Klientská zóna	
Chci podat stížnost	
Půjčovné	
Penzijní fond/ČP Zdraví/ČP Invest	

tématu rozhovoru tak není možná. Rozložení jednotlivých témat je v Tab. 4.8, největší podíl má kategorie *Neobdržel jsem / nemám / nevím* (49 %), *Chci ověřit / nerozumím* (47 %) a *Chci nahlásit* (42 %).

Zařazením dokumentů do jednotlivých kategorií získaných shlukovou analýzou podle nejvyššího stupně příslušnosti je možné určit procentuální zastoupení témat hovorů podle hlavního tématu. Největší podíl v kolekci mají hovory s tématem *PU* (pojistné události, 42 %), poté *Nabídka/sjednání/info* (19 %), *Platby* (16 %) a *Zrušení pojištění* (12 %), viz Tab. 4.9, sloupec *Podíl (1 kat.)*. Zařazením hovorů do více než jedné kategorie se podíl příliš nezměnil, viz sloupec *Podíl (více kat.)* v téže tabulce.

Tab. 4.7: Vytvořené kategorie podle shlukové analýzy.

1. úroveň	**2. úroveň**	**3. úroveň**
PU	PU HAV/POV	Hlášení PU (HAV/POV) Hlášení PU (HAV/POV) PU dopravní nehoda Doklady k PU (HAV/POV) PU čelní sklo Kancelář pojistitelů
	PU domácnost, stavba	-
	PU úraz	-
	PU ostatní	PU HAV/POV - firma/makléřská společnost PU - firma PU jiná pojišťovna/zahraničí PU faktury PU za firmu/klienta, odpovědnost Smlouvy a PU vedené na firmu, odpovědnost z výkonu povolání Odpovědnost z výkonu povolání (PU, platby, nabídka)
	Kontakt na likvidátora	-
	Krycí dopis	-
Nabídka / sjednání / info	HAV/POV smlouvy	Smlouvy HAV/POV HAV/POV bonus Zelená karta
	Domácnost, stavba (info, nabídka)	Domácnost, stavba (nabídka, info) Nabídka pojištění domácnosti
	Úrazové pojištění (info, nabídka)	-
Zasílání dokumentů přes web	-	-
Pobočka / web / cestovní pojištění	-	-
Zrušení pojištění	Zrušení životní pojistky Zrušení pojištění, platby Zrušení pojištění (odpovědnost, domácnost) Zrušení pojištění (HAV/POV), platby Zrušení pojištění (prodej, úmrtí)	- - - - -
Zánik/zrušení pojištění	-	-
Platby	Platby Platby Platby (variabilní symbol) Platby (variabilní symbol) Platby (složenka) Platby, nabídky (domácnost, stavba, odpovědnost) Indexace	- - - - - - -
Volání za někoho jiného	-	-
Změna adresy	-	-

4.5.2 Korelace mezi kategoriemi ČP a shluky

Kategorie vytvořené shlukovou analýzou vytváří jiný pohled než současná kategorizace, a proto jsou těžko porovnatelné. Nějaké překryvy témat zde však jsou, mohou být porovnány podle korelace zobrazované v analytickém softwaru. (Tato korelace neodpovídá klasickému matematickému pojetí, např. nenabývá hodnot z intervalu

Tab. 4.8: Zastoupení současných kategorií.

Kategorie	**Podíl**
Neobdržel jsem / nemám / nevím	49 %
Chci ověřit / nerozumím	47 %
Chci nahlásit	42 %
Chci sjednat/aktualizovat PS	34 %
Chci zaslat	29 %
Přepojení	26 %
Opakované hovory	25 %
Chci mluvit	21 %
Chci zrušit / stornovat PS	15 %
WWW	14 %
Chci změnit	9 %
Udělali jste chybu	9 %
Měl jsem zavolat/mám zmeškaný hovor	8 %
Chci vrátit	6 %
Omyl	6 %
Nejsem spokojený/nesouhlasím	5 %
Klientská zóna	2 %
Chci podat stížnost	2 %
Půjčovné	1 %
Penzijní fond/ČP Zdraví/ČP Invest	1 %
Celkem	**351 %**

Tab. 4.9: Zastoupení nových kategorií.

Kategorie	**Podíl (1 kat.)**	**Podíl (více kat.)**
PU	42 %	46 %
Nabídka / sjednání / info	19 %	25 %
Platby	16 %	20 %
Zrušení pojištění	12 %	17 %
Zánik/zrušení pojištění	3 %	7 %
Volání za někoho jiného	2 %	3 %
Pobočka / web / cestovní pojištění	2 %	3 %
Zasílání dokumentů přes web	2 %	3 %
Změna adresy	1 %	2 %
Celkem	**100 %**	**126 %**

$\langle -1, 1 \rangle$. Vyjadřuje však jistou závislost (podobnost) daných kategorií.) V následujících tabulkách na obrázcích se v prvním sloupci vyskytují současné kategorie (Kategorizace a Produkty ČP), ve druhém pak jsou kategorie vytvořené na základě shlukové analýzy. Řádky jsou seřazeny sestupně podle Korelace.

V tabulce na Obr. 4.10 je možné vidět podobnost kategorií *Chci zrušit / stornovat PS* a *Zrušení pojištění*, dále *Chci sjednat/aktualizovat PS* a *Nabídka / sjednání / info* atd.

Řádky: Kategorizace	Sloupce: Shluky	Frekvence	Korelace
Chci zrušit / stornovat PS	Zrušení pojištění	82	3,2
Chci sjednat/aktualizovat PS	Nabídka / sjednání / info	109	1,4
Chci zaslat	Zasílání dokumentů přes web	15	1,3
Chci změnit	Platby	29	1,3
WWW	Zasílání dokumentů přes web	9	1,3
Chci vrátit	Zrušení pojištění	18	1,2
Chci nahlásit	PU	193	1,1
Nejsem spokojený/nesouhlasím	Platby	16	1,1
Chci změnit	Nabídka / sjednání / info	30	1,1
Chci mluvit	Nabídka / sjednání / info	57	1,0
Chci mluvit	Pobočka / web / cestovní pojištění	12	1,0
Neobdržel jsem / nemám / nevím	Zrušení pojištění	84	1,0

Obr. 4.10: Korelace mezi současnou Kategorizací a Shluky.

Tabulka na Obr. 4.11 ukazuje korelaci mezi Produkty ČP a vytvořenými shluky v kategorii *Nabídka / sjednání / info*. Lze vidět vysokou korelaci mezi *Úrazové pojištění* a *Úrazové pojištění (info, nabídka)*, *Pojištění domácnosti* a *Domácnost, stavba (info, nabídka)* atd. Další ukázky korelace lze nalézt v příloze na obrázcích A.9, A.10, A.11.

Řádky: Produkty ČP	Sloupce: Nabídka / sjednání / info	Frekvence	Korelace
Úrazové pojištění	Úrazové pojištění (info, nabídka)	60	3,9
Pojištění domácnosti	Domácnost, stavba (info, nabídka)	45	2,8
Povinné ručení	HAV/POV smlouvy	48	2,5
Havarijní pojištění	HAV/POV smlouvy	36	2,4
Odpovědnost	Domácnost, stavba (info, nabídka)	31	2,0
Pojištění staveb	Úrazové pojištění (info, nabídka)	32	1,5
Pojištění staveb	Domácnost, stavba (info, nabídka)	26	1,4
Cestovní pojištění	Úrazové pojištění (info, nabídka)	5	1,1
Život - běžně placený	Úrazové pojištění (info, nabídka)	5	1,1

Obr. 4.11: Korelace mezi Produkty ČP a kategorií *Nabídka / sjednání / info*.

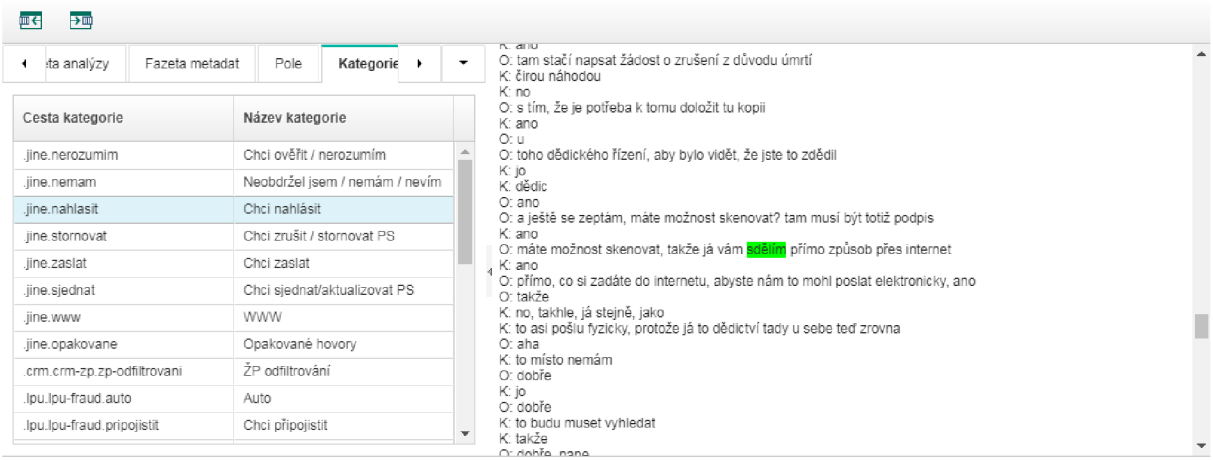
4.5.3 Porovnání zařazení určitých dokumentů

V Tab. 4.10 lze vidět do jakých kategorií byl zařazen hovor `576397`. Kategorie podle shlukové analýzy (SHA) odpovídají tématu dokumentu. Dále bude ukázáno špatné zařazení hovoru do některých současných kategorií.

Tab. 4.10: Kategorizace dokumentu `576397`.

Současná kategorizace a produkty	**Kategorie podle SHA**
WWW	Smlouvy HAV/POV
Chci zrušit / stornovat PS	Zrušení pojištění (prodej, úmrtí)
Opakované hovory	Zrušení pojištění (HAV/POV), platby
Chci zaslat	Zelená karta
Chci sjednat/aktualizovat PS	HAV/POV bonus
Chci nahlásit	Zrušení pojištění, platby
Chci ověřit / nerozumím	
Neobdržel jsem / nemám / nevím	
Úrazové pojištění	
Povinné ručení	

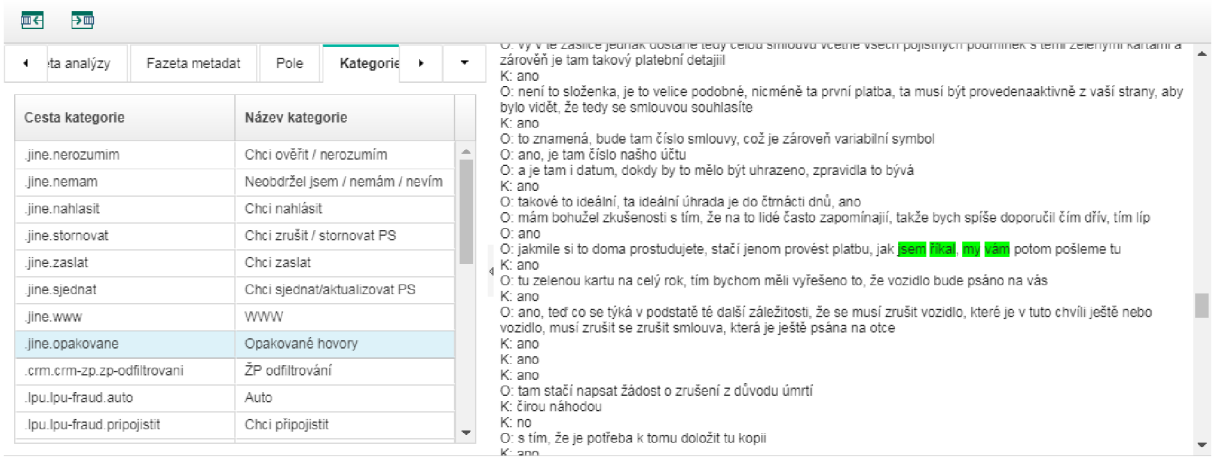
Na následujících obrázcích lze v levé části vidět kategorie do kterých byl hovor zařazen (Cesta kategorie pro sledované kategorie ČP začíná `.jine.`), vpravo je přepis hovoru. Zvýrazněná slova v přepisu odpovídají pravidlu zvýrazněné kategorie vlevo. Na Obr. 4.12 si lze všimnout, že hovor `576397` byl zařazen do kategorie *Chci nahlásit* na základě slova „sdělit“ vyřčeným operátorem. Hovor byl tedy do této kategorie zařazen chybně. Podobně, podle Obr. 4.13, byl špatně zařazen do kategorie *Opakované hovory*.



Obr. 4.12: Chybné zařazení hovoru do kategorie *Chci nahlásit*.

Další ukázkou je hovor `270516`. Tab. 4.11 ukazuje do kterých kategorií byl hovor zařazen. Obrázky A.12, A.13, A.14 v příloze ukazují chybné zařazení do kategorie *Měl jsem zavolat/mám zmeškaný hovor*, *Chci nahlásit*, *Chci vrátit*. Tento dokument byl tedy zařazen pouze do špatných kategorií.

Tab. 4.12 ukazuje poslední příklad zařazení hovoru do kategorií, konkrétně jde o hovor `438286`. Obrázky A.15 a A.16 v příloze ukazují chybné zařazení do kategorie *Chci zrušit/stornovat PS* a do kategorie *Chci sjednat/aktualizovat PS*.



Obr. 4.13: Chybné zařazení hovoru do kategorie *Opakované hovory*.

Tab. 4.11: Kategorizace dokumentu `270516`.

Současná kategorizace a produkty	**Kategorie podle SHA**
Chci vrátit	Kontakt na likvidátora
Měl jsem zavolat/mám zmeškaný hovor	PU HAV/POV - firma / makléřská společnost
Chci nahlásit	PU - firma

Tab. 4.12: Kategorizace dokumentu `438286`.

Současná kategorizace a produkty	**Kategorie podle SHA**
Chci změnit	Úrazové pojištění (info, nabídka)
Chci zrušit/stornovat PS	Indexace
Chci sjednat/aktualizovat PS	
Úrazové pojištění	

Nezařazení některých dokumentů do žádné kategorie je dalším problémem současné kategorizace. Příkladem nezařazených dokumentů může být dokument `367250`, který byl pomocí SHA zařazen do kategorie *PU jiná pojišťovna/zahraničí* nebo `340812`, jenž byl shlukovou analýzou zařazen do *PU – firma*. Tyto dokumenty však kromě toho také patří do kategorie *PU HAV/POV*, ale podle SHA do nich zařazeny nebyly. Problémem nově vytvořené kategorizace tak může být to, že dokumenty nejsou zařazeny do všech vhodných kategorií. Vždy jsou ale zařazeny alespoň do jedné, a ta je podle provedeného pozorování většinou správná (celkovou přesnost zařazení dokumentů do kategorií není možné určit).

4.6 Shrnutí

U některých shluků bylo těžké odhadnout v čem se liší (např. shluky týkající se plateb) a z hlediska produktů by některé z nich mohly být konkrétnější (např. hovory v *Platby, nabídky (domácnost, stavba, odpovědnost)* by mohly být rozděleny do tří zvláštních shluků). Stejně tak by například bylo vhodné rozlišit hovory, ve kterých se vyskytuje nabídka pojištění a ve kterých dochází ke sjednání smlouvy. Dalším požadavkem by mohlo být vytvoření kategorií které zohledňují způsob další komunikace, např. přes pobočku, internet či obchodníka. Takových pohledů by mohlo být více, a proto by se mohlo zdát, že shluková analýza neposkytuje požadované řešení problému. Avšak je důležité uvědomit si, že analýza byla provedena pouze na 774 hovorech. Pro srovnání, za pracovní dny v dubnu 2018 bylo v průměru na KC ČP zaznamenáno asi 3 300 hovorů denně. Kdyby tedy byla analýza provedena na větším množství dat, je možné, že by bylo dosaženo přesnějších a kvalitnějších výsledků.

Závěrem uvedeme některé výhody a nevýhody současného systému kategorizace a kategorizace s využitím shlukové analýzy.

Výhodou kategorizace podle pravidel je možnost vybrat si určitý pohled pro jednotlivé kategorie (např. kategorizace podle typu produktů, typu komunikace, konkurence apod.).

Nevýhodou je složité vytváření pravidel (nutnost vymyslet všechna možná slovní spojení charakterizující dané téma, subjektivní výběr omezení typu `WITHIN 5`, `WITHIN SENTENCE` apod., postupná filtrace – vyřazení hovorů obsahující daná slovní spojení, ale netýkající se daného tématu, časová náročnost), což vede k více chybám (viz předchozí příklady) např. z nepozornosti či nevědomosti. Další nevýhodou je možnost nezařazení dokumentu do žádné kategorie.

Výhodou vytvoření kategorií pomocí shlukové analýzy je větší objektivnost a přesnost (menší chybovost) než je v případě kategorizace podle pravidel, dále zařazení každého dokumentu do alespoň jedné kategorie a relativní jednoduchost vytvoření těchto kategorií.

Nevýhodou je nemožnost určit si předem vytvářenou kategorii a nevědomost kritéria podle kterého byl dokument zařazen do dané kategorie. Někdy také může být trochu obtížné určit téma daného shluku.

5 ZÁVĚR

Tato diplomová práce se zabývala shlukovou analýzou v oblasti text miningu s aplikací na nahraných hovorech call centra České pojišťovny a.s. přepsané do textu. V první části práce jsme poukázali na důležitost tzv. lemmatizace textu, tj. převedení slov na jejich základní tvar. Ukázali jsme, jak textové dokumenty reprezentovat vektory pomocí několika různých druhů vah, mezi které patří např. booleovské, TF, IDF a nejčastěji používané TF-IDF. Pro využití metod shlukové analýzy je velice důležitá redukce dimenze vektorového prostoru, které dosáhneme tzv. prahováním DF a seznamu stopslov, případně také singulárním rozkladem matice.

Další část práce byla zaměřena na samotnou shlukovou analýzu. Po zavedení základních pojmů jako je shluk a nepodobnost jsme se věnovali hierarchickému aglomerativnímu shlukování, které může být reprezentováno pomocí dendrogramu. Jednotlivé metody tohoto přístupu určuje použitý koeficient nepodobnosti shluků, mezi nejznámější patří metoda nejbližšího a nejvzdálenějšího souseda a centroidní metoda. V některých případech lze využít Lance-Williamsův vzorec, který je užitečný zejména při implementaci těchto metod.

Z nehierarchických metod jsme si uvedli základní algoritmus K-Means a K-Means++, které se od sebe liší výběrem počátečních centroidů. Bisecting K-Means je další možný způsob, jak vybrat počáteční centroidy. Vzhledem k lokální konvergenci těchto algoritmů je občas vhodné výsledky upravit postprocessingem. Méně citlivá na outliery je pak např. metoda K-Medoids, která místo centroidů využívá medoidy, jenž jsou přímo objekty shlukovaných dat.

V části věnované fuzzy shlukování jsme si stručně představili pojem *fuzzy* a uvedli jsme algoritmus Fuzzy C-Means, který je fuzzy verzí K-Means. Každé dvojici objektu a shluku přiřadí stupeň příslušnosti, se kterým objekt do daného shluku náleží. Jak moc fuzzy výsledný rozklad bude pak určuje fuzzy parametr p.

Důležitou součástí shlukové analýzy je validace shluků neboli vyhodnocení vytvořeného rozkladu. Mezi základní měřítka shluků patří koheze a separace shluku. Jejich kombinaci pak využívá koeficient siluet, pomocí kterého můžeme také určit kolik shluků vytvořit. Dalším měřítkem může být koeficient korelace mezi ideální a reálnou maticí nepodobnosti, případně může pomoct vizualizace této matice.

Poslední část práce byla věnována aplikaci popsaných metod na reálná data, konkrétně na hovory z kontaktního centra ČP. Cílem bylo vytvořit vhodné kategorie a jednotlivé dokumenty do nich zařadit tak, aby bylo možné provádět kvalitnější analýzy. Data byla nejprve předzpracována, byla snížena dimenze prostoru a na základě analýzy kvality rozkladu bylo vytvořeno 40 shluků metodou K-Means++. Centroidy těchto shluků pak byly použity jako počáteční centroidy pro Fuzzy C-Means a 40 výsledných shluků bylo následně rozděleno do devíti hlavních kategorií. Pro porovnání zařazení dokumentů do jednotlivých kategorií vytvořených na základě shlukové analýzy a současné kategorizace ČP bylo nutné rozlišit, který kanál patří operátorovi a který klientovi. K tomu pomohla opět shluková analýza provedena na levém „kanálu“. V práci byly uvedeny ukázky zařazení hovorů oběma přístupy a na závěr byly uvedeny jejich výhody a nevýhody.

Za přínos této práce by se dalo považovat vytvoření kategorií přinášející jiný pohled na data, než jaký dává současná kategorizace, a také využití jiného přístupu,

který má jisté výhody oproti původnímu systému.

Výsledky této práce by mohly být využity pro sestavení klasifikačního modelu, který patří mezi metody strojového učení s učitelem, avšak vzhledem k malému množství dat by nejspíš nebyl příliš vhodný. Představuje však další možnost kategorizace hovorů, která může vyřešit nevýhody shlukové analýzy i současného systému kategorizace v ČP. Nutností je ale dodat velké množství dat předem opatřených popisem, do které kategorie daný hovor patří.

LITERATURA

[1] BONNER, R. E. On Some Clustering Techniques. *IBM Journal of Research and Development*. 1964, 8(1), 22-32 [cit. 2018-03-09]. DOI: 10.1147/rd.81.0022. ISSN 0018-8646. Dostupné z: http://ieeexplore.ieee.org/document/5392274/

[2] *CSE 634 – Data Mining: Text Mining* [online]. NY, 2016 [cit. 2018-01-14]. Dostupné z: http://www3.cs.stonybrook.edu/~cse634/presentations/TextMining.pdf. Stony Brook University.

[3] ČESKÁ POJIŠŤOVNA A.S. *Výroční zpráva 2016* [online]. Praha: Česká pojišťovna, 2017 [cit. 2018-04-15]. Dostupné z: https://www.ceskapojistovna.cz/documents/10262/9381955/VZ_CP_2016_CZ.pdf

[4] Český stoplist. *Centrum zpracování přirozeného jazyka* [online]. Brno: Fakulta informatiky Masarykovy univerzity, 2018 [cit. 2018-05-16]. Dostupné z: https://nlp.fi.muni.cz/cs/StopList

[5] *CzechEncy – Nový encyklopedický slovník češtiny* [online]. Brno: Centrum zpracování přirozeného jazyka, ©2012-2017 [cit. 2018-01-14]. Dostupné z: https://www.czechency.org/

[6] FELDMAN, Ronen a James SANGER. *The text mining handbook: advanced approaches in analyzing unstructured data*. New York: Cambridge University Press, 2007. ISBN 978-0-521-83657-9.

[7] JOVANOVIC, Jelena. *Introduction to Text Mining* [online]. Srbija, 2015 [cit. 2018-01-20]. Dostupné z: http://ai.fon.bg.ac.rs/wp-content/uploads/2015/04/Intro-to-TM.pdf. Univerzitet u Beogradu.

[8] KJERSTI, Aas a Eikvil LINE. *Text Categorisation: A Survey* [online]. Oslo, 1999 [cit. 2018-01-20]. Dostupné z: https://www.nr.no/~eikvil/tm_survey.pdf. Norwegian Computing Center.

[9] KOLESÁROVÁ, Anna a Monika KOVÁČOVÁ. *Fuzzy množiny a ich aplikácie*. Bratislava: Slovenská technická univerzita v Bratislave, 2004. ISBN 80-227-2036-4.

[10] KRÁTKÝ, Michal. *Využití SVD pro indexování latentní sémantiky* [online]. Ostrava, 2002 [cit. 2018-01-22]. Dostupné z: http://www.cs.vsb.cz/arg/techreports/lsi-svd_ma.pdf. Technická zpráva. Department of Computer Science, VŠB-Technical University of Ostrava.

[11] LUKASOVÁ, Alena a Jana ŠARMANOVÁ. *Metody shlukové analýzy*. Praha: Státní nakladatelství technické literatury, 1985.

[12] NOVÁK, Vilém. *Základy fuzzy modelování*. Praha: BEN - technická literatura, 2000. ISBN 80-7300-009-1.

[13] TAN, Pang-Ning, Michael STEINBACH a Vipin KUMAR. *Introduction to data mining.* Boston: Pearson Addison Wesley, ©2006. ISBN 0-321-32136-7.

[14] SEDLÁČEK, Petr. Text mining a jeho možnosti (aplikace). In: *PV109 Historie a vývojové trendy ve výpočetní technice: Kolokviální práce zpracované studenty* [online]. Brno: Jan Kučera, 2003 [cit. 2018-01-14]. Dostupné z: https://www.fi.muni.cz/usr/jkucera/pv109/2003p/xsedlac5.htm

[15] Sémantická analýza textů (4). In: *Blog Seznam.cz* [online]. Praha: Seznam.cz, 2018 [cit. 2018-01-22]. Dostupné z: https://blog.seznam.cz/2011/10/semanticka-analyza-textu-4/

[16] SCHULTZ, Jeff. How Much Data Is Created on the Internet Each Day? In: *Micro Focus Blog* [online]. Micro Focus, 2017 [cit. 2018-01-14]. Dostupné z: https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/

[17] *Terra Incognita* [online]. Montreal, QC, Canada: Christian S. Perone, ©2018 [cit. 2018-01-21]. Dostupné z: http://blog.christianperone.com/

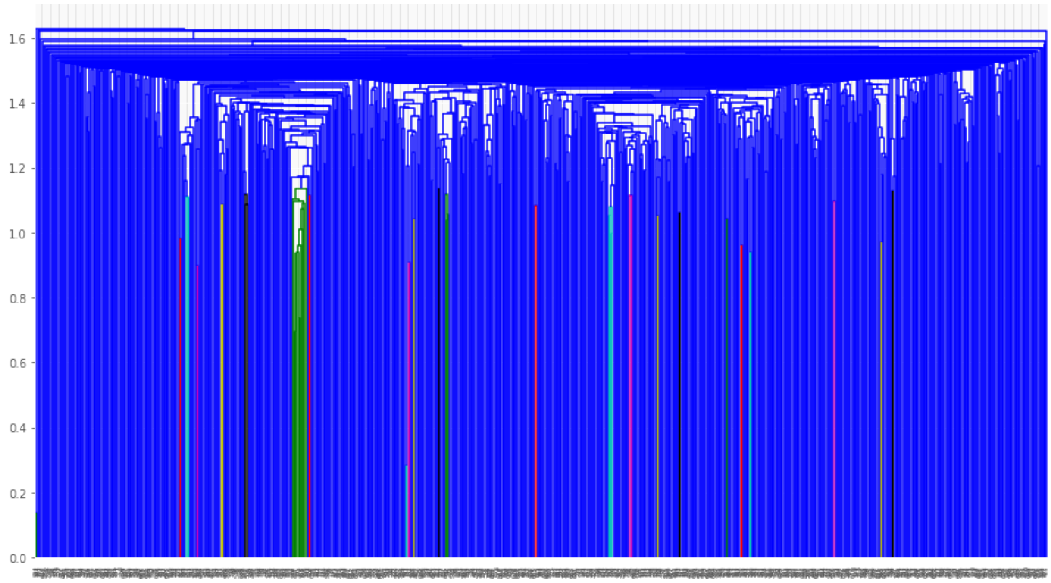
[18] TOLKIEN, J. R. R. *Hobit, aneb, Cesta tam a zase zpátky.* Ilustrovala Jemima CATLIN, přeložil František VRBA. Praha: Argo, 2013. ISBN 978-80-257-0884-2.

[19] TRYON, R. C. *Cluster Analysis.* Ann. Arbor, Edwards Bros. 1939.

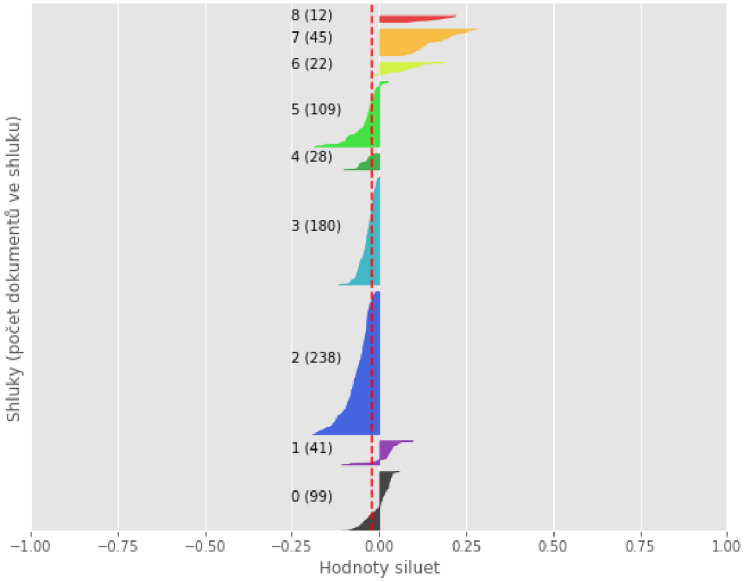
[20] WITTE, René. *Introduction to Text Mining: Tutorial at EDBT'06* [online]. Germany: Institute for Program Structures and Data Organization (IPD), 2006 [cit. 2018-01-14]. Dostupné z: http://www.edbt2006.de/edbt-share/IntroductionToTextMining.pdf

SEZNAM PŘÍLOH

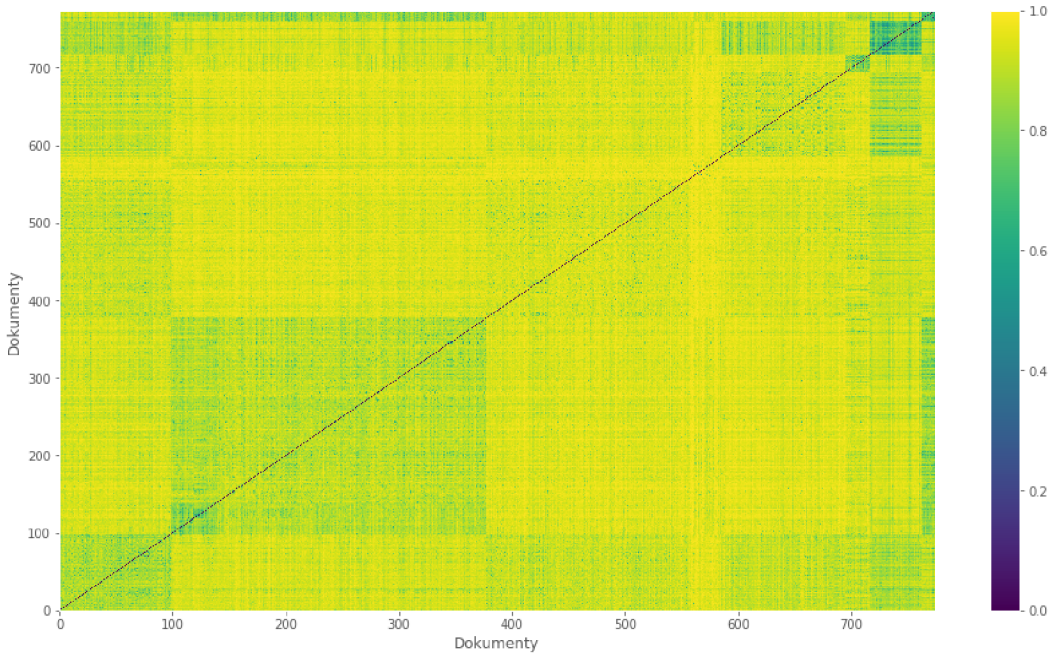
A DOPLŇUJÍCÍ OBRÁZKY



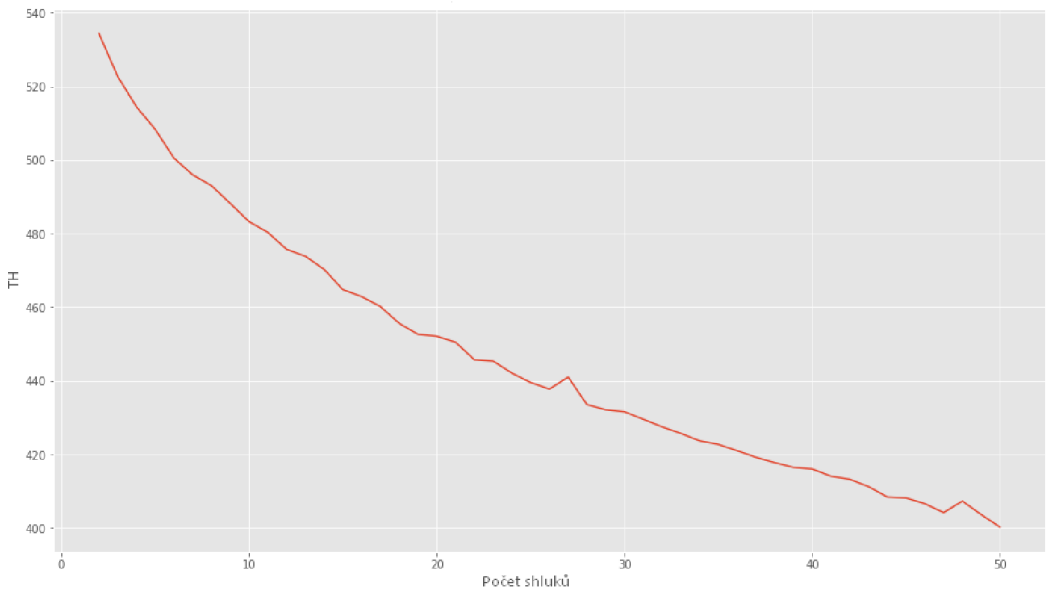
Obr. A.1: Dendrogram metody nejbližšího souseda.



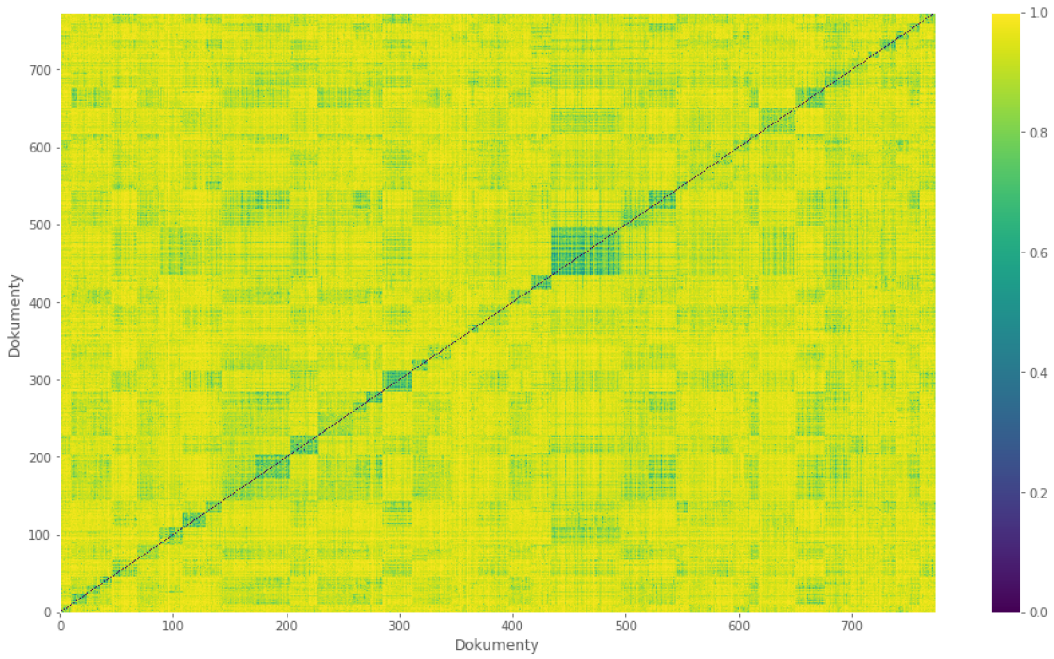
Obr. A.2: Siluety metody nejbližšího souseda.



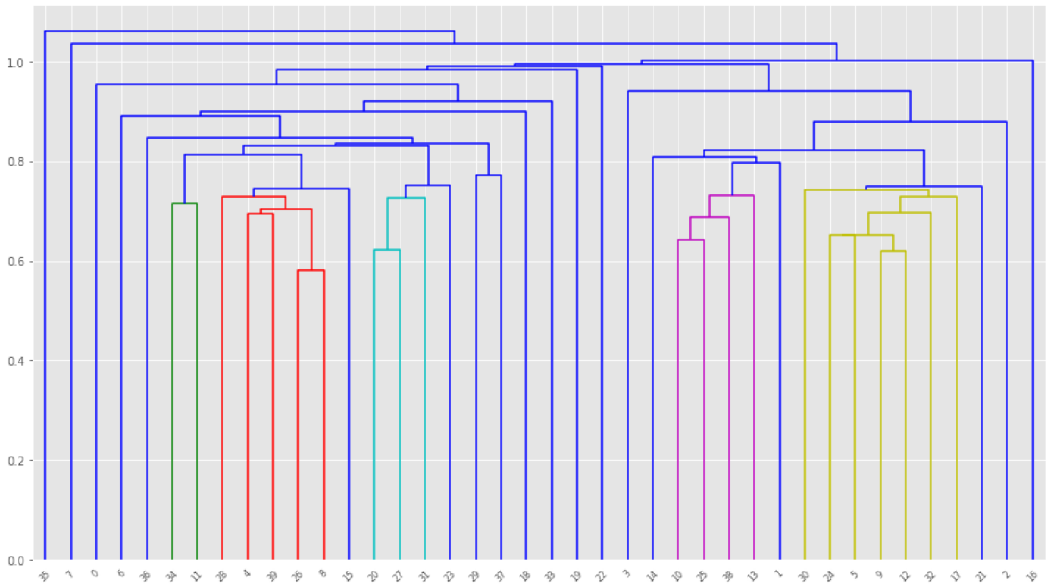
Obr. A.3: Matice nepodobnosti metody nejbližšího souseda.



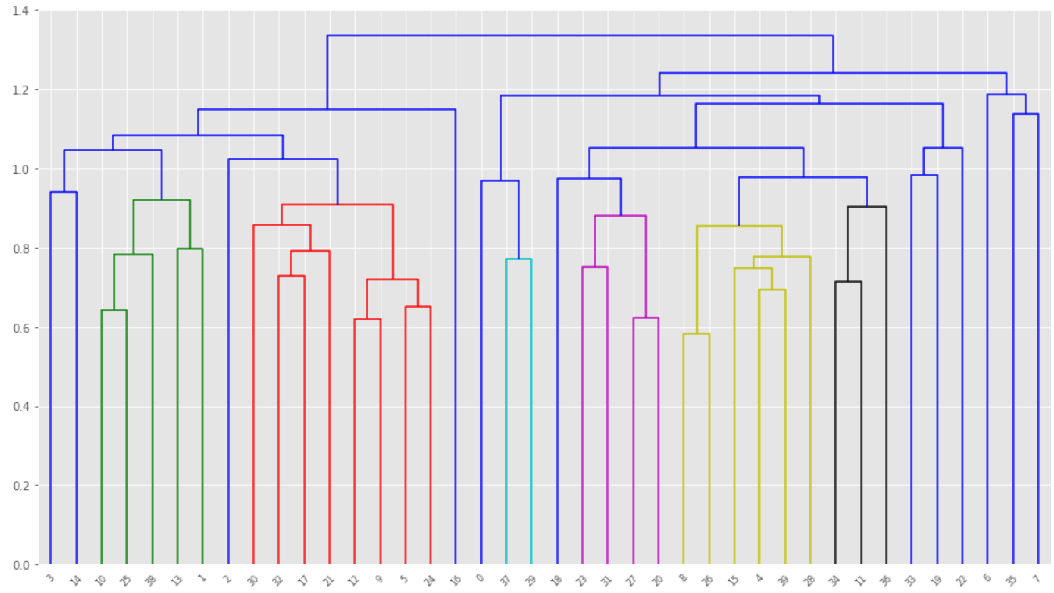
Obr. A.4: Analýza hodnoty účelové funkce při použití K-Means++.



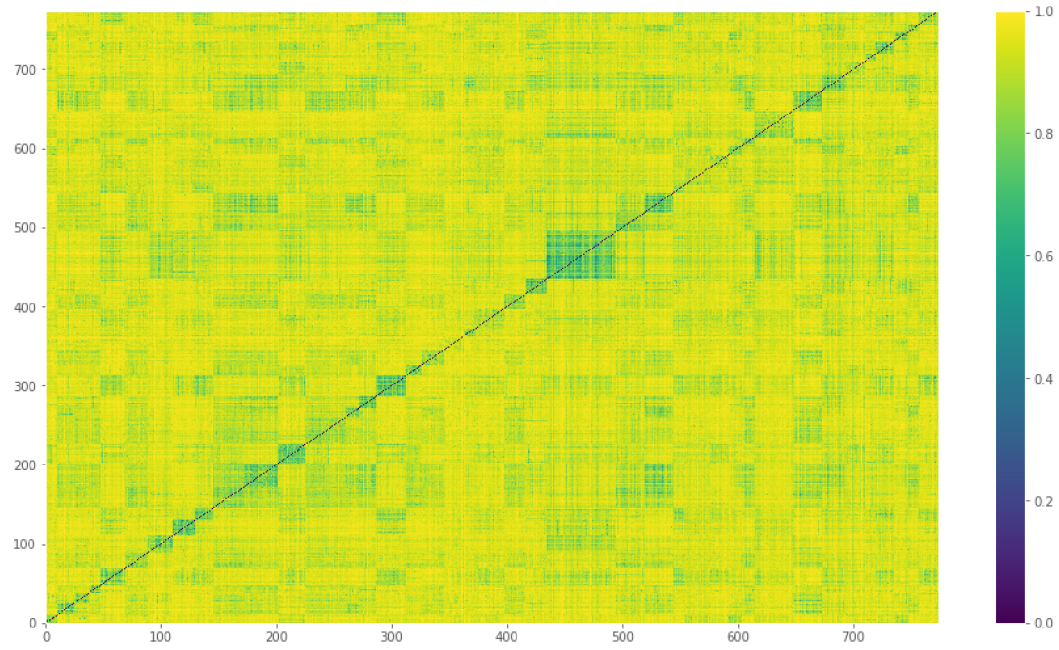
Obr. A.5: Matice nepodobnosti rozkladu na 40 shluků pomocí K-Means++.



Obr. A.6: Dendrogram metody nejbližšího souseda aplikované na výsledky z K-Means++.



Obr. A.7: Dendrogram vážené metody průměrné vazby aplikované na výsledky z K-Means++.



Obr. A.8: Matice nepodobnosti rozkladu na 40 shluků pomocí FCM.

Řádky: Kategorizace	Sloupce: Platby	Frekvence	Korelace
Chci změnit	Indexace	21	6,4
Nejsem spokojený/nesouhlasím	Indexace	11	4,6
Neobdržel jsem / nemám / nevím	Platby (složenka)	25	1,0

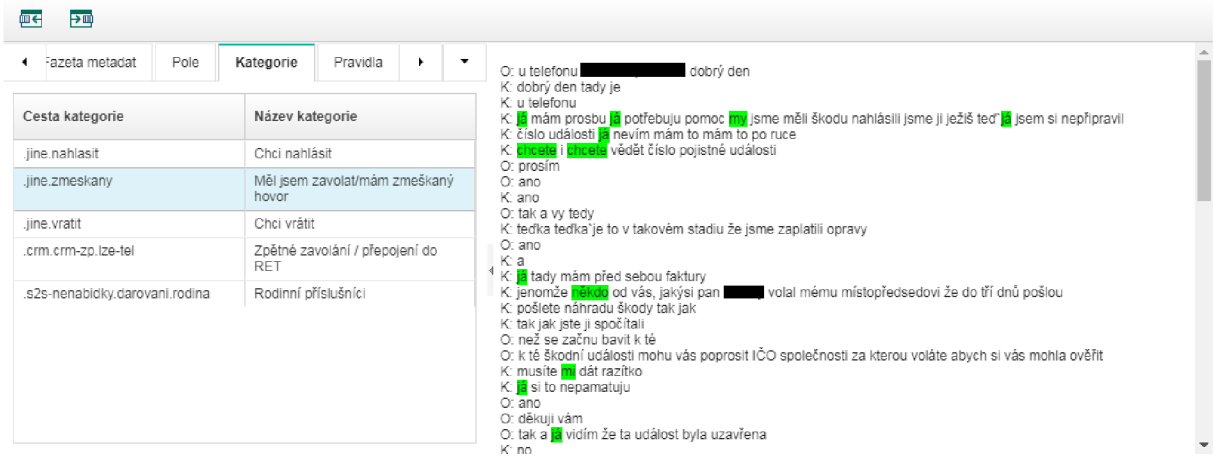
Obr. A.9: Korelace mezi Kategorizací a kategorií *Platby*.

Řádky: Produkty ČP	Sloupce: PU	Frekvence	Korelace
Havarijní pojištění	PU HAV/POV	41	1,6
Povinné ručení	PU HAV/POV	49	1,5
Pojištění domácnosti	PU domácnost, stavba	11	1,4
Úrazové pojištění	PU úraz	11	1,2

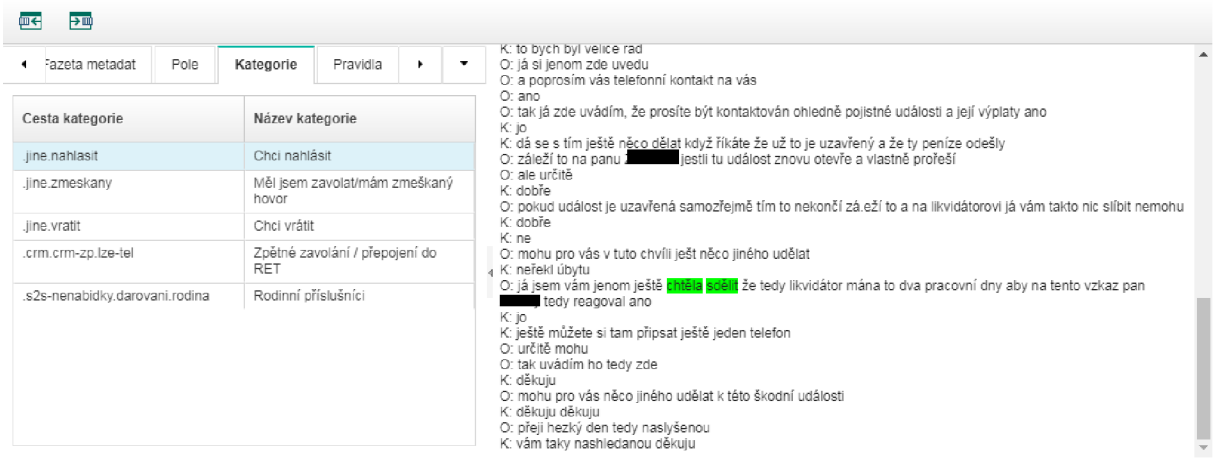
Obr. A.10: Korelace mezi Produkty ČP a kategorií *PU*.

Řádky: Produkty ČP	Sloupce: Zrušení pojištění	Frekvence	Korelace
Pojištění domácnosti	Zrušení pojištění (odpovědnost, domácnost)	11	1,6
Havarijní pojištění	Zrušení pojištění (HAV/POV), platby	12	1,2
Úrazové pojištění	Zrušení životní pojistky	19	1,2
Život - jednorázově placený	Zrušení pojištění, platby	3	1,1

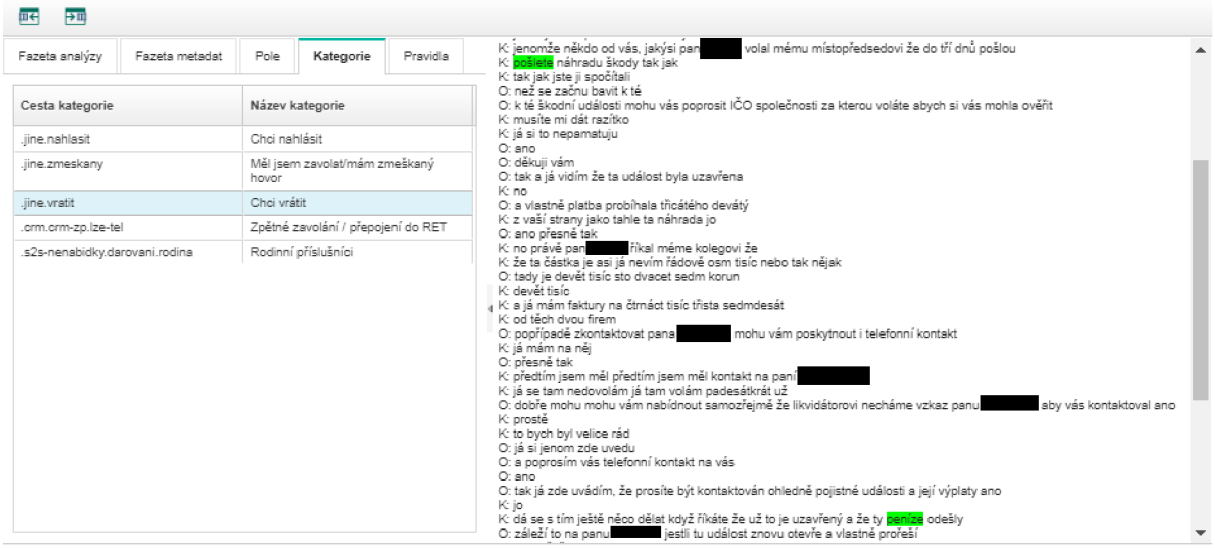
Obr. A.11: Korelace mezi Produkty ČP a kategorií *Zrušení pojištění*.



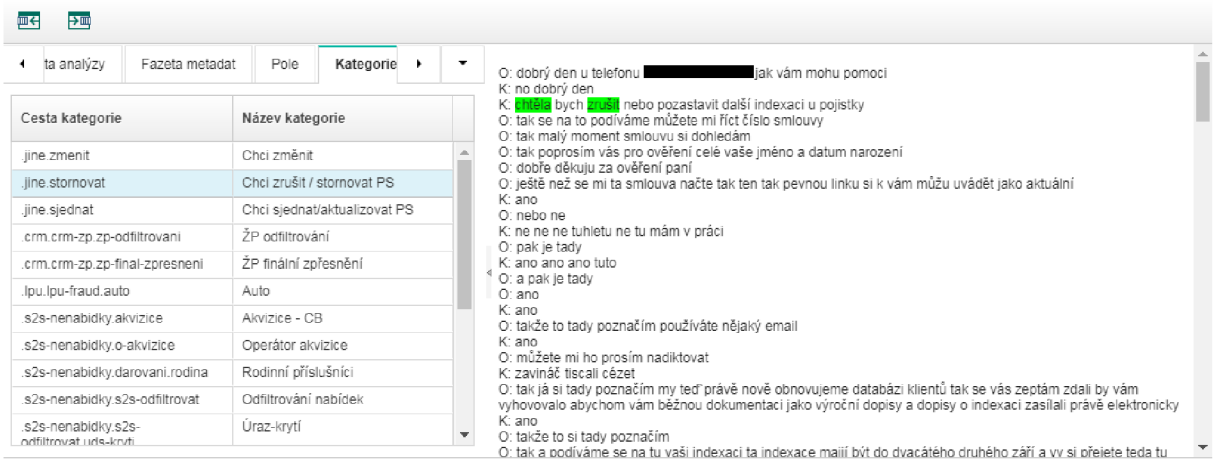
Obr. A.12: Chybné zařazení hovoru do kategorie *Měl jsem zavolat/mám zmeškaný hovor*.



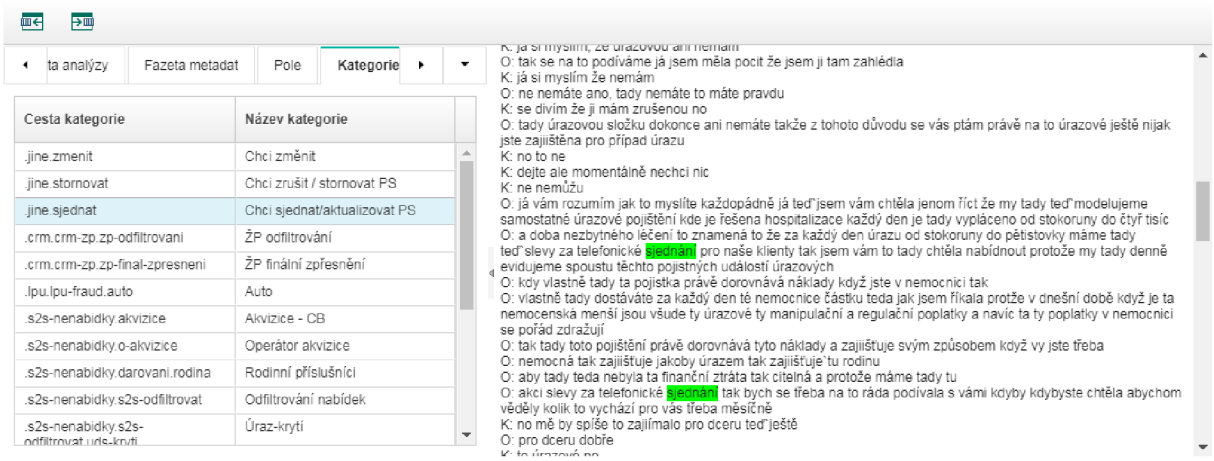
Obr. A.13: Chybné zařazení hovoru do kategorie *Chci nahlásit*.



Obr. A.14: Chybné zařazení hovoru do kategorie *Chci vrátit*.



Obr. A.15: Chybné zařazení hovoru do kategorie *Chci zrušit/stornovat PS*.



Obr. A.16: Chybné zařazení hovoru do kategorie *Chci sjednat/aktualizovat PS*.

B Přepis hovoru 281100

R: u telefonu (...), dobrý den
L: dobrý den tady
L: já by sem potřebovala mluvit s paní (...), jestli tam je někde
R: a paní (...) je operátor nebo likvidátor události?
L: likvidátor
R: tak
R: paní (...) zřejmě nemohla zvednout telefon a bylo to přesměrováno na komunikační centrum České pojišťovny
R: vy jste se teďka dovolala vlastně na komunikační centrum
R: a vy
R: jedině jestli můžete to zkusit třeba někdy později
R: jí zavolat znovu
L: no, zkusím no, už to zkouším od včerejška, tak snad se dovolám, no
R: od včerejška, nebo teda ještě můžeme
R: se podívat na tu událost a dát jí požadavek, aby se ona spojila s vámi
R: že s ejí nemůžete dovolat, jo?
L: no, to by bylo
L: to by bylo lepší, no
R: tak já vás poprosím o číslo té škody
R: ano
R: tak
R: já jenom musím poprosit
R: vy voláte za koho?
L: za
L: za syna
R: jeho datum narození, nebo adresu
L: adresa
R: děkuju a ještě teda vaše celé jméno a datum narození
R: děkuju
R: dívám se je tady na vás uvedený telefon, takže to se neměnilo, všechny kontakty jsou stále aktuální
R: a vy s ní potřebujete probrat co, jenom se zeptám
L: no, on měl syn úraz před rokem
L: bolestný jsme normálně dostali a pak jsme dávali ještě požadavek na trvalý následky a už je to přes dva měsíce a nikdo se neozval
R: nikdo se neozval, já se ještě podívám do toho spisu
R: ano, ono to tady doložené je, máte pravdu
R: já jí teda předám, že byste se s ní potřebovala ještě domluvit, tak aby, aby se, aby vás prostě co nejdříve kontaktovala, že se jí nemůžete dovolat, tak ať se ona spojí s vámi, ano
L: dobře
L: tak děkuju moc
R: a můžu teda ještě pro vás něco jiného udělat v tuto chvíli?

L: ne ne ne, to je všechno, to je všechno, děkuju
R: dobře
R: a mohla bych jenom dotaz, já jse se chtěla zeptat, jestli i vy smaa máte třeba nějaké úrazové pojištění
R: kdyby se vám třeba stalo něco při výkonu povolání, nebo na ulici vás třeba srazilo auto, nebo ve vozidle
R: nějaká dopravní nehoda
R: máte nějaké úrazové pojištění, z kterého byste pak mohla třeba uplatňovat ušlý zisk, nebo
R: nějaké trvalé následky, kdyby vám to zanechal ten úraz
L: no, já jsem si ted'kom nechávala udělat pojištění i jako se synem dohromady
L: ale jsem v neschopnosti, jsem po úrazu, takže čekám až
L: budu schopná práce a pak mi to
L: odklepnou, že je to platný
R: aha,tak a vy a vy jste si to nechala udělat u nějaké jiné pojišťovny než u České, nebo u nás?
L: já myslím, že je to zrovna Česká, no
R: u nás
L: je, no
R: tak to je dobře, že jste na to teda takhle, vám se tedy asi stal nějaký úraz paní (...), pak pak jste si teprve sjednala to pojištění?
L: no, no
R: tak to je zrovna škoda,no, že jste ho neměla už předtím
L: to je velká škoda, no
R: ale aspoň teda že jste se rozhodla, že si to takto zařídíte, protože
R: ono tedy nikdy nevíte a právě ten úraz se vám může stát kdekoliv, kdykoliv, tam je to takové vlastně, můžete se opravdu jít třeba jenom po ulici a špatně došlápnout, zlomit si nohu a už
L: no
R: vlastně potom nemůžete chodit do práce, a to je potom problém
R: takže
R: to je velmi, velmi dobře, to můžu jedině pochválit, že jste si to takhle sjednala, takže já vám zatím moc poděkuju za váš čas, přeji vám, pěkný pěkný den, mějte se hezky
L: vám taky, děkuju moc
R: děkuju za zavolání, nashledanou
L: nashledanou

C SEZNAM STOPSLOV

a
aby
abych
abysme
abyste
aha
akorát
ale
alice
ani
aniž
ano
asi
aspoň
ať
až
bez
březen
buď
bude
by
byl
byla
bylo
být
celý
co
cokoliv
což
čeho
červen
červenec
český
či
čím
člověk
čtyři
čtyřicet
další
dana
dát
datum
děkovat
den
deset
devátý
devět
do
dobře
druhý
duben
dva
dvacátý
dvacet
dvěstě
gabriela
hned
ho
hodně
chtít
i
ivana
já
jak
jaká
jaké
jakého
jakémkoliv
jakmile
jako
jakoby
jakože
jaksi
jaký
jakým
jan
jasně
jasný
je
jeden
jedenáct
jeho
její
jejich
jen
jenom
jenž
jestli
jestliže
ještě
ji
jí
jim
jinak
jiný
jít
již
jméno
jo
jo
jsem
jsme
jsou
k
kam
karla
každopádně
každý
kde
kdo
kdy
kdybych
kdybyste
kdykoliv
když
kdyžtak
ke
klára
klidně
koho
kolem
kolik
kristýna
která
které
kterého
kterou
který
kteří
květen
ladislava
leden
lenka
let
listopad
lucie
lukáš
lukáš
má
malý
marcela
markéta
martin
mé
mě
mezi
mi
michaela
miluše
mirka
mít
mně
mnou
moc
moci
moje
moment
monika
možná
možný
mu
můj
muset
může
my
na
nad
nám
narození
nás
nashledanou
naschledanou
náš
naše
naší
ne
nebo
nebýt
něco
něho
nej
nějak
nějaká
nějaké
nějakou
nějaký
nějakých
nějakým
někam
někde
někdo
někoho
některá
není
než
ní
nic
nicméně
nich
ním
nimi
no
o
od
ohledně
on
ona
oni
ono
opravdu
osm
padesát
pak
pan
pán
pane
panem
paní
pěkně
pět
petr
petra
pětset
po
pod
podívat
podle
podobně
podstata
pojistka
pojištění
pojišťovna
pokud
poprosím
poprosit
pořádek
potom
potřeba
pouze
právě
pro
proč
prosinec
prosit
prostě
proti
proto
protože
první
přát
přece
před
přes
přesně
při
přímo
případně
půl
rád
radmila
renata
rodný
rok
romana
řekl
říci
říjen
říkat
s
sám
samozřejmě
se
sebe
sedm
sedmý
sem
set
si
simona
smlouva
spíše
spolu
srpen
sta
stě
sto
své
svého
svou
svůj
šárka
šedesát
šest
šestý
ta
tady
tak
také
takhle
takový
takto
taky
takže
tam
tato
té
teď
teda
teďka
tedy
těch
těchto
telefon
těm
ten
tento
tereza
této
ti
tím
tímto
tisíc
tisíce
to
todle
todleto
tohle
tohleto
toho
tohoto
tom
tomto
tomu
totiž
toto
tou
třeba
tři
třicet
třicet
tu
tuhle
tuto
ty
tý
tyto
u
únor
určitě
uvyst
už
v
valerie
vám
vámi
vás
váš
vaše
vašeho
vašem
vaši
vaší
ve
věc
věra
veronika
více
vlastně
volat
však
vše
všechen
všechno
vůbec
vy
vždycky
z
za
září
zas
zase
zatím
zde
zdeněk
ze
zeptat
zuzana
žádný
že