UNIVERZITA PARDUBICE

FAKULTA EKONOMICKO-SPRÁVNÍ

BAKALÁŘSKÁ PRÁCE

2008 **Karel ŠŤOVÍČEK**

UNIVERZITA PARDUBICE
FAKULTA EKONOMICKO - SPRÁVNÍ
ÚSTAV SYSTÉMOVÉHO INŽENÝRSTVÍ A INFORMATIKY

Komparace nástrojů pro zpracování dokumentů

Karel Šťovíček

BAKALÁŘSKÁ PRÁCE

2008

Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav systémového inženýrství a informatiky
Akademický rok: **2007/2008**

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení: **Karel ŠŤOVÍČEK**

Studijní program: **B6209 Systémové inženýrství a informatika**

Studijní obor: **Informační a bezpečnostní systémy**

Název tématu: **Komparace nástrojů pro zpracování dokumentů**

Zásady pro vypracování:

Cílem práce je popis a komparace dvou nástrojů pro zpracování dokumentů za použití standardních datových sad.

- Vymezení základních pojmů
- Nástroje pro zpracování dokumentů
- Testování funkčnosti dvou nástrojů pro zpracování dokumentů
- Komparace dostupných metod na standardních datových sadách
- Hodnocení použitých nástrojů

Rozsah grafických prací:

Rozsah pracovní zprávy:

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

WEISS, Sholom, et al. Text Mining: Predictive Methods for Analyzing Unstructured Information. 1st edition. : Springer, 2004. 236 s. ISBN 0387954333.

POKORNÝ, Jaroslav, et al. Dokumentografické informační systémy. 2. vyd. : Karolinum, 2006. 185 s. ISBN 80-246-1148-1.

VAN RIJSBERGEN , Cornelis Joost , CRESTANI, Fabio, LALMAS, Mounia. Information Retrieval: Uncertainty and Logics. : Springer, 1998. 352 s. ISBN 0792383028.

PÉLADEAU, Normand. SIMSTAT for Windows : User's Guide. : [s.n.], 1996. 255 s.

Vedoucí bakalářské práce: **Ing. Hana Kopáčková, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání bakalářské práce: **30. října 2007**

Termín odevzdání bakalářské práce: **19. května 2008**

prof. Ing. Jan Čapek, CSc.
děkan

L.S.

doc. Ing. Pavel Petr, Ph.D.
vedoucí ústavu

V Pardubicích dne 30. října 2007

Souhrn

Bakalářská práce se zabývá souhrnem kategorií programů používaných v text miningu, komparací dvou nástrojů pro zpracování dokumentů za použití standardní datových sady a demonstrační datové sady a testováním funkčnosti těchto nástrojů k využití pro text mining. Druhá část je zaměřena na testování nástrojů pomocí vlastní datové sady a hodnocení kvality klasifikace pro další možné využití.

Klíčová slova

data mining, text mining, kategorizace, shlukovaní, datová sada

Abstrakt

Bachelor work describes categories of programs, which are used for text mining. It compares two programs for analyzing text documents with using standard data set and demo data set and tests funkcionality of these programs for text mining. Second part describes special data set using text mining programs and posibility of implementing of categorization in the future.

Keywords

data mining, text mining, categorization, clustering, data set

Obsah

Seznam obrázků

Seznam tabulek

1. Úvod

Zpracování dat a získávání netriviálních informací z ních nabylo na svém významu společně s rozvojem oblasti informatiky, ukládáním dat a potřebou s těmito daty dále nakládat. S pokrokem v oblasti informačních technologiích se začaly vyvíjet i nástroje specializované pro oblast data miningu a text miningu, ty nabídly využití metod, analýz a práce s velkým objemem dat efektivně a uživatelům přístupnou formou.

Nejen díky internetů je současná doba ve znamení prudkého nárůstu dat, velmi se zjednodušil přístup k nim pro kohokoliv a zároveň se data v elektronické podobě stala nepsaným standardem. Existence dat však jejich potenciálnímu vlastníkovi zdaleka nemusí přinášet užitek, stejně důležitá je schopnost data zpracovat, umět analyzovat a získávat z nich tak jejich skrytou hodnotu. Data mining a text mining se tak pro tyto činnosti stávají stále častěji nepostradatelným.

Cílem této práce je

- vymezení základních pojmů,
- nástroje pro zpracování dokumentů,
- testování funkčnosti dvou nástrojů pro zpracování dokumentů,
- komparace dostupných metod na standardních datových sadách,
- hodnocení použitých nástrojů.

Ve třetí kapitole jsou uvedeny skupiny nástrojů v současné době používané a zároveň vhodné pro zpracování textových dat uživatelem v širším měřítku. Z nástrojů specializovaných pro text mining byly dále vybrány dva přibližně si odpovídající úrovně a ve čtvrté a páté kapitole byla provedena jejich komparace jak po funkční, tak po uživatelské stránce. Jeden z těchto nástrojů byl hodnocen též pro možné použití a další výzkumnou činnost fakultou ekonomicko-správní Univerzity Pardubice. Vzhledem k předběžným výsledkům testování však bylo od pořízení ustoupeno a program byl dále analyzován v demoverzi. V tomto případě nemohlo být přistoupeno k testování standardní datovou sadou z důvodů licenčních ujednání, program tak byl testován demonstrační sadou výrobce.

Naopak byla využita možnost vytvoření vlastní datové sady na základě získaných dat z výrobně-obchodní firmy. Specifika této datové sady jsou především ne příliš dlouhé texty obchodní korespondence a dále dvojjazyčnost dat, pocházející ze současných potřeb managementu. Tato sada umožnila v páté kapitole otestovat kategorizaci reálných dat a zhodnotit možný přínos pro širší uplatnění text miningu.

2. Vymezení základních pojmů

Data mining

Počátek rozvoje data miningu lze datovat do osmdesátých až devadesátých let minulého století. Je přímo spjat jednak s vývojem výpočetní techniky, bez které je velká většina metod v něm užívaných velmi těžko realizovatelná, stejně tak jako s potřebou hlubší analýzy dat a získávání informací z nich. Proces data miningu lze definovat jako netriviální získávání implicitních, dříve neznámých a potenciálně užitečných informací z dat [14]. Díky prudkému nárůstu dat v poslední době lze již nalézt uplatnění data miningu ve většině oborů závislých na zpracovávání velkých objemů dat.

Data mining využívá širokou škálu matematických i statistických technik [3], například shlukovou analýzu [2], klasifikaci, rozhodovací stromy, neuronové sítě, genetické algoritmy a jiné. Přesto však okruh potencionálních uživatelů data miningu je poměrně široký a nároky na ně se mohou i velmi lišit, uživateli mohou být jak specialisté, tak manažeři.

Úlohy data miningu se člení dle úloh a metod k jejich řešení uvedených v tabulce (Tab. 1), či je možné data mining členit dle Weisse v tabulce (Tab. 2) na dvě základní skupiny: Predikci a Objevování znalostí. Avšak ani tato uvedená členění nejsou jediná, či jediná správná. V mnoha zdrojích je možné najít i další.

Tab. 1 - Úlohy a metody data miningu [5]

Úloha	**Metoda**
Klasifikace	Diskriminační analýza
	Logistická regresní analýza
	Klasifikační (rozhodovací) stromy
	Neuronové sítě (algoritmus "back propagation")
Odhady hodnot vysvětlované proměnné	Lineární regresní analýza
	Nelineární regresní analýza
	Neuronové sítě (RBF -- "radial basis function")
Segmentace (shlukování)	Shluková analýza
	Genetické algoritmy
	Neuronové shlukování (Kohonenovy mapy)
Analýza vztahů	Asociační algoritmus pro odvozování pravidel typu If X, then Y
Predikce v časových řadách	Boxova-Jenkinsova metodologie
	Neuronové sítě ("recurrent back propagation")
Detekce odchylek	Vizualizace
	Statistické postupy

Tab. 2 - Rozdělení data miningových problémů dle Weisse [5]

Predikce	Objevování znalostí (deskriptivní analýzy)
Klasifikace	Zjišťování odchylek
Regresní analýza	Segmentace databáze
Analýza časových řad	Shlukování
	Asociační pravidla
	Sumarizace
	Vizualizace
	Dolování v textu

Text mining

Definice text miningu není jednoznačně dána, Nahm a Mooney například popisují text mining jako hledání zákonitostí v nestrukturovaném textu [12] či dle Tana je text mining proces extrahování netriviálních zákonitostí či znalostí z textových dokumentů [18]. V současnosti text mining představuje specifickou oblast zahrnující různé nástroje pro klasifikaci, filtrování textů, shlukování, extrakci informací, sumarizaci a další. Text mining se zabývá zpracováním nestrukturovaných dat. To jsou data, která nemají předem danou strukturu, například délky slov, přesto však mohou být použita ve strukturách jako věty, odstavce, slovní spojení. Na rozdíl od zpracování číselných dat je však text mining plně závislý na národních zvyklostech, tedy jazyku, ve kterém je text k dispozici, struktuře textu, použité gramatice, která se například i u jednoho jazyka vzhledem k zeměpisné oblasti může lišit. To velmi znesnadňuje rychlé šíření efektivních nástrojů pro jazyky málo používané v globálním měřítku.

Úloha text miningu spočívá v získávání a rozkrývání užitečných informací obsažených v textu, které ani nemusí být na první pohled zjevné. Získávání informací probíhá velmi často ne z jednoho textu, ale ze souborů textů. Text mining sestává ze dvou částí, předzpracování a získávání znalostí [19].

Předzpracování může být složeno z mnoha různých činností především v závislosti na dostupných činnostech pro daný jazyk textu (například lemmatizace). Prvním krokem je extrahování samotného textu z dokumentu a tím očištění od dalších dat jako obrázky, značky či jiné netextové informace. Odstraněny jsou informace o fontu, velikosti, barvě a dalších atributech písma, naopak struktura textu může být zachována, je-li to užitečné a dle vybraného nástroje, pro další analýzu. Nyní je možné text rozdělit na termy a slova, která je možné vyjmout z dalšího zpracování vzhledem k jejich četnosti či nedůležitosti (především předložky, spojky, velmi často se opakující slova). Tím lze získat tvar textu určený pro samotné získávání znalostí. Získávání znalostí pak probíhá již pouze na tomto výběru dle požadovaného typu zpracování.

Specifika zpracování textů spočívají především v sémantice jazyků a také naprostou nevhodností jazyka pro počítačové zpracování, což je základní rozdíl vůči číselným datům [16]. Textová data obsahují velmi mnoho šumu, který je nutný odstranit. To je možné zajistit jak pomocí filtrování slov s příliš vysokou četností v textu, tak i pomocí slovníků slov pro vyřazení. Dále pak většina slov existuje ve více gramatických tvarech, k jejich převedení do základního tvaru slouží lemmatizace. Dalším problémem, který však nelze odstranit prostým zpracováním jednotlivých slov, jsou různé výrazy jednoho slova.

Term

Pojem term nemá přesnou definici, v text miningu vyjadřuje slovo či spojení, které je použito pro analýzu a zpracování textů. Při analýze textu získáme termy po fázi předzpracování dokumentu, jeho očištění, případně po použití lemmatizace či dalších nástrojů. Seznam termů, který je výchozím bodem pro fázi získávání znalostí, obsahuje jedinečné výskyty slov.

Lemmatizace

Tímto slovem je nazývána činnost, kterou převádíme slova na základní gramatický tvar (nazývaný též lemma). Způsob, kterým je tohoto stavu dosahováno je závislý především na použitém jazyku a jeho gramatice. Čeština patří mezi jazyky gramaticky velmi ohebné, což ztěžuje práci lemmatizátoru, nástroje pro lemmatizaci. K lemmatizaci je možné přistupovat dvěma základními metodami, pro dosažení co nejlepších výsledků zpravidla jejich kombinací, a to lemmatizací na základě výčtu všech tvarů daného slova (slovníku všech tvarů slov) nebo lemmatizaci na základě gramatiky jazyka (způsobu tvorby jednotlivých tvarů slov).

Takto popsané zpracování je však pouze na úrovni slov, neřeší otázku mnohoznačnosti slov, či různých tvarů stejně psaných, avšak s rozdílným významem. Při používání jazyka člověk automaticky rozpozná správný význam z kontextu, pro strojové zpracování je však takovéto rozpoznání velmi složité jak co do metod realizace tak i výpočetně vzhledem k nárůstu možných kombinací. Přesný výběr adekvátního významu slova není dosud při automatickém zpracování uspokojivě vyřešen.

Stemming

Stemming se zabývá zkracováním slov, extrakcí jejich kořenů. V jazycích jako je např. angličtina, kde je tvar slova dán neměnným kořenem a koncovkou určující bližší tvar slova (pád, číslo, aj.) je stemming pro použití vhodnější a jednodušší než například v češtině, kde se pro jedno slovo mění i tvar kořene slova. S automatickým stemmingem použitým pro česká slova se můžeme setkat například v lokalizovaném vyhledávači Google, kde byl zaveden v roce 2007 [13].

Syntaktická analýza

Tato analýza se zabývá rozpoznáním zákonitostí a vztahů mezi slovy například na základě gramatických pravidel, strukturou věty. Předpokladem je rozeznání textu v daném jazyce. Syntaktická analýza je základem pro další práci s takto získanými informacemi například pro strojové překlady, gramatické korektory a další. Čeština patří mezi jazyky velmi obtížně analyzovatelné především díky volnému slovosledu a tvarům slov.

Sémantická analýza

Cílem je zkoumání slov, spojení, vět či textů. Pro toto zkoumání jsou využívány výsledky různých dalších analýz jako například syntaktické, lemmatizace a jiných. Jedná se o nejkomplexnější úroveň zpracování spolu s pragmatickou analýzou. Typickým využitím sémantické analýzy jsou automatické překlady textů. Vzhledem k jazyku, jeho složitosti a nevhodnosti pro klasické počítačové zpracování však zatím tato analýza není uspokojivě zvládnuta.

Tokenizace

Tokenizací korpusu (velkého textového souboru) provádíme jeho rozložení na základní prvky, se kterými budeme dále pracovat. Prvky mohou být slova, čísla, interpunkce, spojení slov. Tokenizace korpusu ovlivňuje další práci s daty a jsou na ní závislé výsledky následných analýz.

Indexace

Proces vyjádření obsahu dokumentu pomocí prvků selekčního jazyka, obvykle s cílem umožnit zpětné vyhledávání. Podle použitých metod se rozlišuje pojmová a slovní indexace, podle použitých postupů se rozlišuje intelektuální, automatická a poloautomatická indexace. Z hlediska použitých selekčních jazyků se rozlišuje prekoordinovaná indexace a postkoordinovaná indexace [1].

Mezi nejčastější metody vážení patří

- Booleova,
- četnost TF,
- relativní četnost,
- četnost TFIDF,
- četnost TFC,
- četnost LTC.

V Booleovském vážení probíhá přiřazování binárních hodnot 0 a 1 termům dle toho, zda se v dokumentech vyskytují či nikoliv. Tato metoda nijak nezkoumá kvalitativní hodnotu výskytu termu v dokumentu, pouze jeho existenci.

Četnost TF (term frequency) používá pro zvětšení vypovídací hodnoty charakteristického znaku vážení. Předpokladem této metody je hypotéza, že term w_k s vyšší četností výskytu je pro dokument $\vec{d}_i$ charakteristický

$$TF = (w_k, \vec{d}_i), \tag{1}$$

kde vyjadřuje $\vec{d}_i$ - dokument (s pořadovým číslem i) a w_k – Slovo (term) v dokumentu.

Relativní četnost rozšiřuje předchozí metodu o zohlednění délky dokumentu

$$d_{ik} = \frac{TF(w_k, \vec{d}_i)}{\sum_{\vec{d}_i} w_k}, \tag{2}$$

ve které je term obsažen.

Četnost TFIDF (term frequency, inverse document frequency) ve výpočtu zahrnuje četnost v rámci dokumentu a počtu dokumentů, ve kterých se term vyskytuje

$$d_{ik} = TF(w_k, \vec{d}_i) \cdot IDF(w_k), \tag{3}$$

$$IDF(w_k) = \log\left(\frac{|\mathrm{D}|}{DF(w_k)}\right). \tag{4}$$

Četnost TFC ve svém výpočtu zohledňuje délku dokumentů a tu vyrovnává délkovou normalizací

$$d_{ik} = \frac{TF(w_k, \vec{d}_i) \cdot IDF(w_k)}{\sqrt{\sum_{j=1}^{N_T} \left(TF(w_j, \vec{d}_i) \cdot IDF(w_j)\right)^2}}. \tag{5}$$

V četnosti LTC je použit logaritmus pro četnost znaku k redukci výrazných rozdílů v četnostech

$$d_{ik} = \frac{\log\left(TF(w_k, \vec{d}_i)+1\right) \cdot IDF(w_k)}{\sqrt{\sum_{j=1}^{N_T} \left(\log\left(TF(w_j, \vec{d}_i)+1\right) \cdot IDF(w_j)\right)^2}} \tag{6}$$

Snížení dimenze

Snížení dimenze provádíme především z důvodu příliš velkého množství termů extrahovaných z dokumentů při předzpracování textu. Redukce je vhodná jednak z důvodů výpočetních, dále pak očištění vektorového prostoru od termů, které nejsou pro další zpracování významné z důvodu příliš velké či příliš malé četnosti. Dimenzi je možné provádět buď selekcí, nebo extrakcí. Metody selekce jsou různé [1], například

- četnost TF,
- dokumentová četnost DF,
- χ^2 statistika,
- mutual information MI,
- information gain IG.

Pomocí četnosti TF lze redukovat termy, které mají minimální výskyt v dokumentu a tím jsou pro následnou kategorizaci nevýznamné. Počet výskytu termů, které mají být vyjmuty, či naopak ponechány, není přesně dán a závisí především na typu a rozsahu zpracovávaného textu. Dokumentová četnost se naopak zabývá termy s příliš vysokou hodnotou výskytu ve všech dokumentech. Pro další zpracování nemají velký význam obdobně jako termy s minimálním výskytem.

χ^2 statistika používá pro hodnocení termů výpočet

$$\chi^2(w_k, c_i) = \frac{|D| \cdot (P(w_k, c_i) \cdot P(\overline{w}_k, \overline{c}_i) - P(w_k, \overline{c}_i) \cdot P(\overline{w}_k, c_i))^2}{P(w_k) \cdot P(\overline{w}_k) \cdot P(c_i) \cdot P(\overline{c}_i)}, \tag{7}$$

kde:

- D je množina všech dokumentů,
- w_k označení jednotlivého termu,
- c_i označení jednotlivé kategorie,
- $P(c_i)$ pravděpodobnost, že dokument padne do kategorie c_i,
- $P(w_k)$ pravděpodobnost, že se v dokumentu vyskytuje slovo w_k,
- $P(w_k, c_i)$ sdružená pravděpodobnost, že dokument padne do kategorie c_i a zároveň obsahuje slovo w_k,
- $P(\overline{w}_k, \overline{c}_i)$ sdružená pravděpodobnost, že dokument nepadne do kategorie c_i a zároveň neobsahuje slovo w_k,
- $P(w_k, \overline{c}_i)$ sdružená pravděpodobnost, že dokument nepadne do kategorie c_i, ale obsahuje slovo w_k,
- $P(\overline{w}_k, c_i)$ sdružená pravděpodobnost, že dokument padne do kategorie c_i, ale neobsahuje slovo w_k.

MI – Mutual information (míra vzájemné informace) je definována vztahem

$$MI(w_k, c_i) = \log \frac{P(w_k, c_i)}{P(w_k) \cdot P(c_i)}, \tag{8}$$

IG – Information gain (informační zisk) je založen na minimalizaci entropie a vypočítá se dle vztahu

$$IG(w_k, c_i) = \sum_{c_i \in \{c_i, \bar{c}_k\}} \sum_{w_k \in \{w_k, \bar{w}_k\}} \log \frac{P(w_k, c_i)}{P(w_k) \cdot P(c_i)}. \tag{9}$$

Shlukování

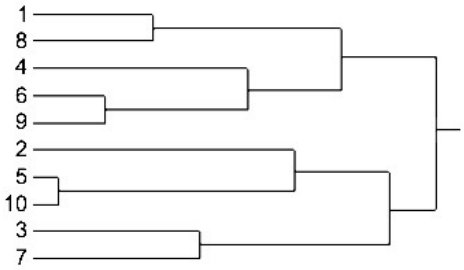
Proces rozkladu celku na několik podmnožin (shluků), jejichž prvky si jsou v rámci podmnožiny co nejvíce podobny a mezi podmnožinami podobny co nejméně, nazýváme shlukování.

Shlukování obsahuje velkou škálu metod a výsledky se mezi nimi mohou i významně lišit především i v požadavcích na výsledné shluky [4]. Jednou ze základních vlastností shluků je, zda jsou disjunktní či nedisjunktní neboli zda mohou existovat takové prvky, které jsou obsaženy ve více shlucích. Shluky lze zobrazit dle typu zobrazovaného prostoru jako kruhy či koule. Jejich charakteristickými znaky jsou střed shluku a jeho poloměr. Metody se pak mohou zaměřovat na rozdělení prostoru shluky o stejných velikostech, ve kterých však může velmi kolísat počet prvků, popřípadě mohou vznikat prázdné shluky. Další možností jsou shluky se stejným počtem prvků. Odpadá zde problém s nerovnoměrným rozložením souborů ve shlucích, avšak vytváření shluků je složité.

Shlukování lze rozdělit na

- hierarchické,
- nehierarchické.

Hierarchické shlukování lze dále dělit na aglomerativní či divizivní. Aglomerativní shlukování vychází z jednotlivých prvků reprezentujících shluky a mechanismus vytváření shluků hledá vždy dva nejbližší prvky, ze kterých pak vytvoří nový shluk [6]. Zjišťování vzdálenosti mezi aktuálně vytvořenými shluky a shlukování nejbližších prvků či shluků probíhá tak dlouho, existují-li alespoň dva shluky. Divizivní shlukování postupuje opačně, kdy počátek je v jednom jediném shluku, který se dále dělí až na výsledné jednotlivé prvky. Měření vzdálenosti mezi shluky je možné provádět více metodami, například metodou nejbližšího souseda, metodou nejvzdálenějšího souseda či metodou centroidní. Výsledkem hierarchického shlukování je též graf nazývaný dendogram na obrázku (Obr. 1). Dendogram zároveň zachycuje jednotlivé kroky při hierarchickém shlukování.



Obr. 1 - Dendogram

Nehierarchické shlukování se vyznačuje postupnou optimalizací prvků ve shlucích. Na počátku je stanoven počet shluků, určeny polohy jejich centroidů a výchozí přiřazení prvků k těmto shlukům. Dále následují iterace, kdy jsou prováděny výpočty nových hodnot centroidů a následnému přeskupování prvků dle vzdálenosti. Iterace probíhají tak dlouho, dokud existují rozdíly mezi rozdělení prvků mezi shluky. U nehierarchického shlukování může docházet i cyklickému přesouvání prvku mezi shluky, pak nelze jednoznačně určit optimum, tak může tato metoda na stejných datech dosáhnout jiných optimálních řešení způsobeným různým výchozím přiřazením prvků a volbě výchozí volbě shluků. Mezi nehierarchické metody shlukování patří například K-Means [11] nebo Kohonenovy mapy [9].

Kategorizace

Kategorizací dokumentů nazýváme jejich automatickou klasifikaci do definovaného počtu tříd. Zařazování dokumentů do tříd provádí klasifikátor, pro něj jen nutné nejprve vytvořit množinu oklasifikovaných dat pro učení. Pomocí charakteristických slov v těchto dokumentech a roztřídění těchto dokumentů do tříd je možné provést fázi učení klasifikátoru. Kategorizaci je možné provádět různými metodami [21], blíže budou popsány tři z nich

- metoda nejbližšího souseda,
- Bayesův naivní klasifikátor,
- kosínová podobnost.

Metoda nejbližšího souseda spočívá v hledání nejpodobnějších dokumentů vůči dokumentu testovanému. Měření probíhá pomocí výpočtu Euklidovské vzdálenosti

$$d_e(x_1, x_2) = \sqrt{\sum_{j=1}^{m} (x_{1j} - x_{2j})^2} \,. \tag{10}$$

Při této metodě je nutné jednak stanovit počet nejpodobnějších dokumentů, kdy tato hodnota významně ovlivňuje kvalitu zpracování dokumentů, dále pak určit způsob ohodnocení blízkosti dokumentů.

Bayesův naivní klasifikátor [3] je založen na pravděpodobnostním modelu a i přes své nedostatky dosahuje poměrně dobrých výsledků. Mezi jeho přednosti patří schopnost zařazení i

neúplně popsaných případů, rychlost zpracování díky relativní nenáročnosti výpočtu. Naopak jeho použití nedosahuje optima při málo početných třídách, reprezentace znalostí pomocí pravděpodobností je méně srozumitelná.

Kosínová podobnost (cosine similarity) provádí výpočet na základě slov či lemmat ve vektorovém prostoru. Pro výpočet slouží vzorec [7]

$$\cos(X,Y) = \frac{\sum x_i * y_i}{\sqrt{\sum (x_i)^2} * \sqrt{\sum (y_i)^2}}. \tag{11}$$

3. Nástroje pro zpracování dokumentů

Již od počátku práce s daty musel být řešen problém s jejich uchováváním. V oblasti výpočetní techniky pak s uložením ve vhodném typu dat tak, aby mohla být opětovně zpracována. Problém obsahuje dvě části: jak data uchovávat v čase a v jakém typu data ukládat. Typ dat pak předurčuje, jaké informace jsou uchovávány a nepřímo říká, jakým nástrojem mohou být data dále zpracována. Jednotlivé typy dat mohou být například

- data v textových souborech,
- data v grafických souborech,
- multimediální záznamy (zvukový záznam, videozáznam),
- databáze dat.

Společně s ukládáním dat se vyvíjely i nástroje pro práci s nimi. Je možné je rozdělit na několik skupin, a to nástroje umožňující nám

- zobrazení dat,
- převod mezi typy dat či mezi jejich formáty,
- zpracování, ukládání a analýza dat.

Nástroje mohou obsahovat i více těchto skupin současně (a zpravidla tomu tak i je), toto se pak zpravidla projevuje i v pořizovacích nákladech na ně.

3.1. Nástroje pro zobrazení dat

Většina z těchto nástrojů patří do oblasti freeware, mají velmi omezené funkce pro další práci s dokumenty, zpravidla pouze umožňují náhled a tisk dat. V oblasti data miningu a text miningu slouží především pro rychlý náhled jednotlivých souborů. Mezi nejčastěji používané patří

- grafické prohlížeče,
- multimediální přehrávače,
- nástroje pro zobrazení dat ve standardních formátech.

3.2.Nástroje pro převod mezi typy dat

Při práci s daty, jejich převodem do elektronické podoby či jejich záznamem narazíme na nemožnost uložit vstupní data do různých typů. Převod mezi typy dat je velmi složitý a výpočetně náročný. V současnosti jde především o převod grafických dat obsahujících obraz textu (OCR - Optical Character Recognition) a zpracování lidské řeči do textové podoby.

Programy pro OCR jsou již delší dobu rozšířené a použitelné, od jednoduchých freewarových nástrojů až po automatizované komerční programy sdružující i další funkce pro práci s textem. Mezi nejčastěji používané patří

- ABBYY FineReader (ABBYY Software House),
- Readiris (I.R.I.S. s.a.),
- Omnipage (Recognita),
- FreeOCR.net.

Oblast rozpoznávání a převodu záznamu hlasu se díky své náročnosti zpracování začala vyvíjet až v posledních letech. Projevuje se zde však jazyková bariéra, kde lokalizace nelze jednoduše provést a pokrok v této oblasti je tedy velmi závislý na používaném jazyku. Nejlepší systémy nyní dosahují pro anglický jazyk úspěšnost zpracování asi 98% [8].

3.3.Nástroje pro převod formátu dat

Velmi často nazývané spíše konvertory patří naopak mezi jednodušší programy, specializované na převod mezi formáty, v nichž jsou data uložena. Ty jsou velmi často používané především ve spojení s programy, do nichž nelze importovat data různých formátů či umožňujících načtení pouze jediného typu. Pro text mining jsou zajímavé především textové konvertory, které umožňují převod mezi všemi nejpoužívanějšími formáty včetně html. Nejčastěji používaným výstupem dat pro další zpracování v oblasti text miningu je formát txt. Mezi ně například patří

- Text Mining Tool,
- All Office Converter Pro,
- Leadtools ePrint Professional.

Specifikem pak je převod souborů, které sice obsahují vnitřní strukturu textového souboru, avšak z různých důvodů mají buď jinou příponu, či nemají příponu vůbec. Zde se pak jedná pouze o hromadné přejmenovávání souborů, pro které lze využívat průzkumníky či souborové managery.

3.4.Nástroje pro zpracování a analýzu textových dat

Tyto nástroje lze rozdělit na nástroje pro data mining s možností zpracovávat texty a na specializované nástroje pro text mining. Poptávka po data miningu se začala utvářet v 90. letech minulého století jako reakce na potřebu zpracování velkého množství dat již existujících dat. Přestože nelze pro data mining vymezit jednoznačný návod k postupu, vznikly dvě metodologie dosud užívané a uznávané, a to

- CRISP-DM,
- SEMMA.

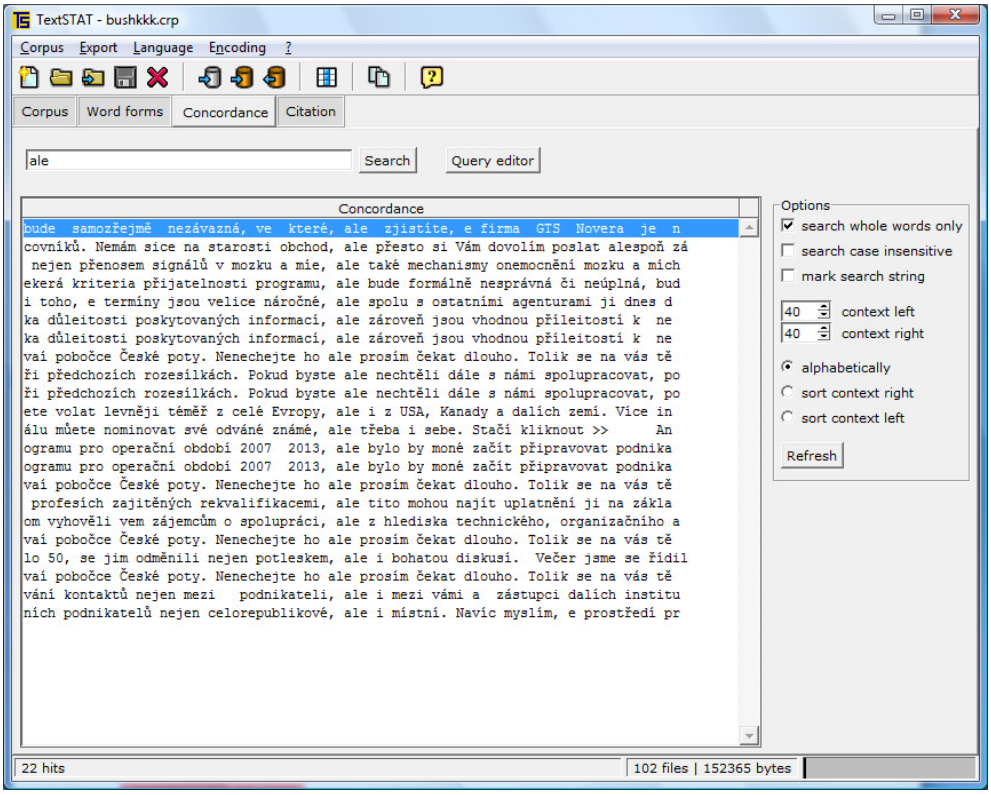
Tyto metodologie využívají i v současnosti jak komerční nástroje, tak nástroje freewarové. Mezi nejznámější komerční nástroje patři

- SAS Enterprise Miner,
- SPSS Clementine,
- STATISTICA Data Miner.

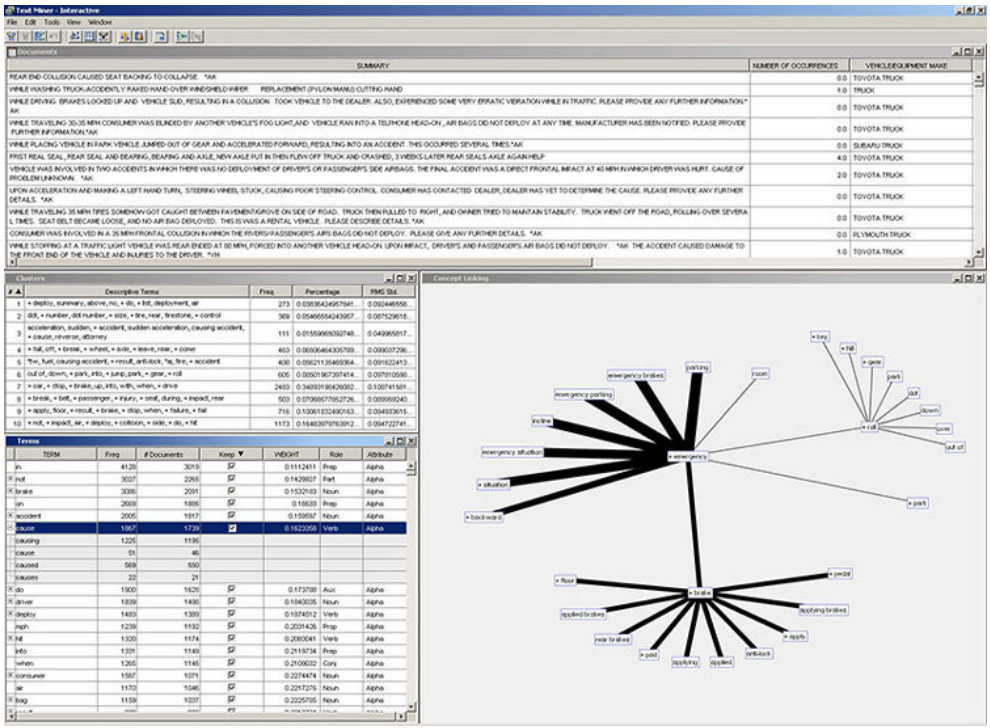
Freewarové nástroje určené pro data mining pak jsou například

- Weka,
- RapidMiner,
- LISp-Miner [10].

Nástroje specializované pro text mining jsou ve velké míře na trhu především pro světové jazyky jako angličtina, francouzština, španělština. Dle místa původu vzniku nástroje pak lze kromě těchto jazyků nalézt též italštinu, němčinu a další. Dle zpracování dat je lze rozdělit na statistické a sémantické. Použití českých textů není buď podporováno vůbec, nebo s texty lze pracovat, avšak bez využití funkcí jako lemmatizace či automatické slovníky. Nástroje pro text mining lze nalézt jak freewarové, zpravidla pouze základními funkcemi, až po profesionální nástroje. Freeware nástroj pro text mining je například TextSTAT od Matthiase Hüninga [20] na obrázku (Obr. 2) se základními funkcemi, mezi profesionální nástroje například SAS® Text Miner od společnosti SAS Institute Inc. na obrázku (Obr. 3) s cenou 32.000 $ [15].



Obr. 2 - Freeware nástroj TextSTAT



Obr. 3 - SAS® Text Miner

Nejvíce nástrojů je však ve skupině shareware, či komerčních nástrojů s pořizovací cenou v řádu stovek až tisíců dolarů. Ty jsou již vybaveny i pokročilými funkcemi pro zpracování dat a umožňují širokou škálu výstupů z analýz pro další použití. Mezi ně patří i oba programy

vybrané pro následnou komparaci, a to program Wordstat od společnosti Provalis Research a program T-Lab od stejnojmenné společnosti. Oba programy jsou komerční a ve stejné cenové kategorii. Akademickou licenci programu Wordstat vlastní Univerzita Pardubice, program T-Lab byl testovám v demoverzi, neboť licence nebyla zakoupena.

4. Testování funkčnosti dvou nástrojů pro zpracování dokumentů

4.1. Wordstat a QDA Miner

Program Wordstat [22] od společnosti Provalis Research je softwarový modul pro rozšíření jednoho z programů, a to buď Simstat nebo QDA Miner. Tento software je plně komerční i pro domácí či studijní použití. Dále popisována bude kombinace pragramů Wordstat 5.1 a QDA Miner 2.0.1.

Data pro analýzu se importují nejprve do programu QDA Miner, kde s nimi lze provádět další úpravy. Program Wordstat neumožňuje přímý import dat.

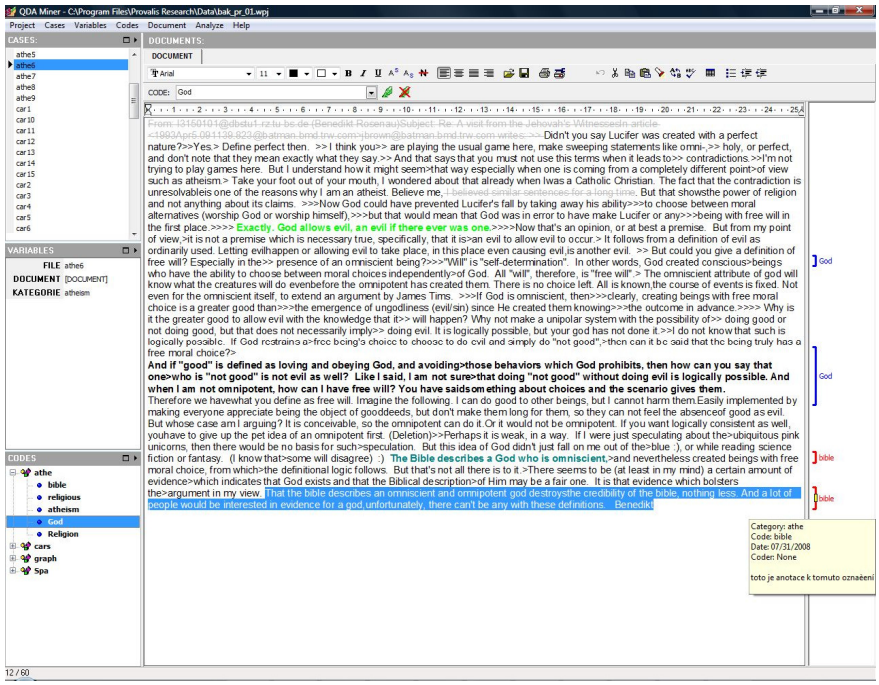
4.1.1. QDA Miner

Software umožňuje načítat formáty txt, doc, pdf, rtf, vpd a htm(l). Dále lze načítat data ze souborů obsahující databáze a je možné se připojit též na spuštěnou databázi pro import dat.

Dokumenty načtené do projektu v prostředí QDA Mineru, jejich seznam je na obrázku (Obr. 4) vlevo nahoře, je možné libovolně editovat. Dále jsou k dispozici nástroje hledání v textu, nahrazení textu, jednoduchá tabulka, odrážky či formátování odstavce.

Vlastnosti každého dokumentu jsou uvedeny v tabulce atributů (variables) na obrázku (Obr. 4) vlevo uprostřed. Tyto vlastnosti jsou plně editovatelné, je možné přidávat vlastní skupiny pro další možnosti třídění především v modulu Wordstat. Podporované jsou tyto typy vlastností

- ordinální,
- numerické,
- datum,
- boolean,
- krátký řetězec (definovatelná délka),
- dokument.



Obr. 4 - Uživatelské rozhraní programu QDA Miner

Hlavním nástrojem programu QDA Miner je především možnost tvorby kódovací knihy a anotaci částí textu či slov těmito kódy. Kódovací kniha je zobrazena na obrázku (Obr. 4) v levém dolním rohu a anotace se zobrazují v pravém sloupci vedle textu. Kódy je možné barevně rozlišovat. K jednotlivým přiřazením je možné psát jednoduché komentáře. Přesnou část textu, ke které je kód přiřazen je možné vyvolat pomocí kontextového menu.

Nástroje pro analýzu dokumentů umožňují klasické vyhledávání slov, dále vyhledávání částí textu definovanými počátečními slovy a různými možnostmi délky textu či inteligentního rozpoznávání odstavců. Další možností je vyhledávat jedno či více slov s pomocí booleanovských funkcí AND, OR a NOT z definované skupiny slov. Třídění výsledků je dle souborů, náhledů textu či výsledků v jednotlivých skupinách atributů. Přímo z výsledků hledání se lze přepnout přímo do jednotlivých náhledů souborů.

Další nástroje pro analýzu již využívají kódovací knihu. Mezi nástroje patří zobrazení procentuální a číselné statistiky použití kódů v dokumentech, včetně zobrazení v grafu, dále vyhledávání jednotlivých kódů v dokumentech a možnost exportování výsledků do souborů.

Současný výskyt kódů je možné zobrazovat z vybrané skupiny kódů a dokumentů. K zobrazení výsledků je k dispozici dendogram, 2D a 3D mapa shluků, graf a tabulka blízkosti a statistika výsledků.

Hledání sekvence kódů je prováděno jak pro všechny, tak s definovaným prvním, případně i druhým kódem. Lze volit též blízkost a maximální vzdálenost dvou kódů. Výsledky jsou zobrazovány v seznamu, křížovou tabulkou a náhledy jednotlivých výsledků.

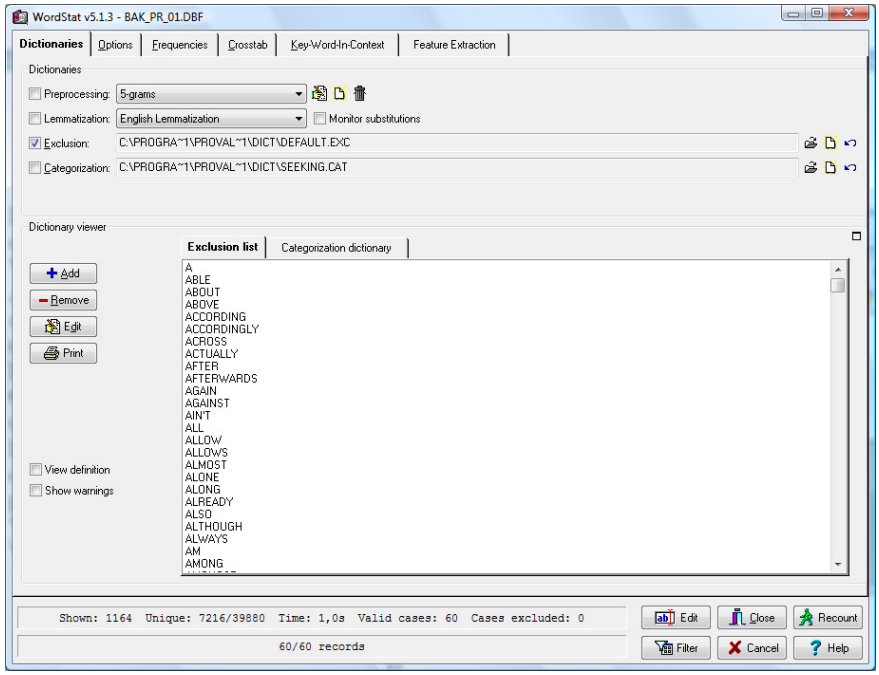
Analýza četnosti kódů umožňuje tabulkové i grafické zobrazení kódů v závislosti na souborech a na atributech. Poslední volbou je zobrazení a statistiky kódů v závislosti na uživatelích v případě, že projekt obsahuje více uživatelů než jednoho.

4.1.2. Wordstat

Další zpracování dat již probíhá v programu Wordstat. Ten je spouštěn z prostředí QDA Mineru a analýzy jsou přímo závislé na nastavení a úpravách aktuálního projektu QDA Mineru.

Při spuštění Wordstatu je možné zvolit rozsah dat z aktuálního projektu, která budou použita a dále dle typu analýzu popisnou, analýzu v závislosti na vlastnostech dokumentů či přiřazených kódech a kategoriích. Tato volba ovlivňuje dostupnost některých funkcí Wordstatu.

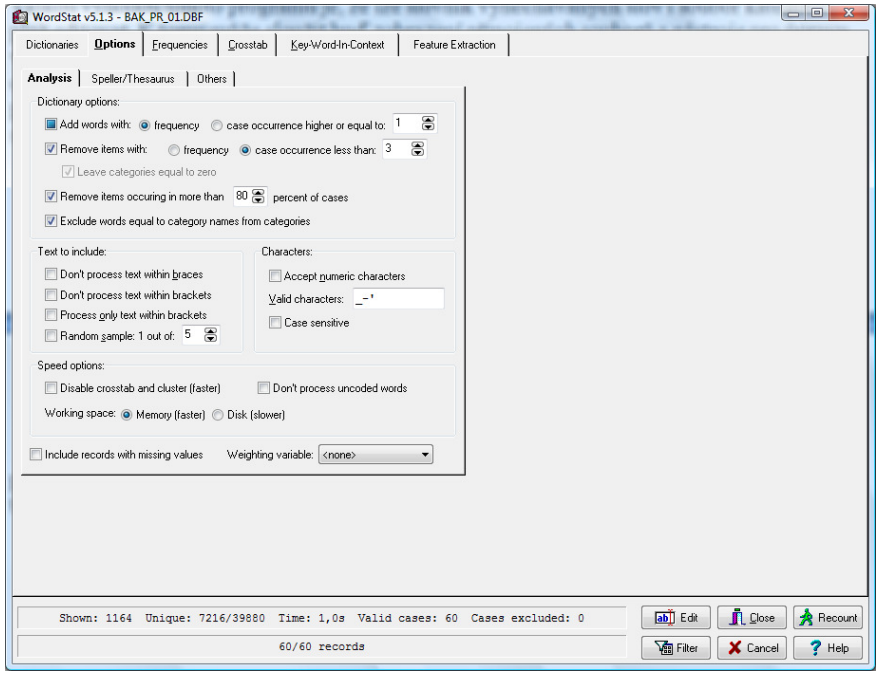
V záložce slovníků na obrázku (Obr. 5) je možné zvolit preprocessing, lemmatizaci (výběr je z anglické, francouzské, italské a španělské), soubor se slovy vynechávanými při zpracovávání a soubor s kategoriemi včetně příslušných charakteristických slov.



Obr. 5 - Nastavení parametrů pro práci s textem ve Wordstatu

Obsah vybraného slovníku vynechávaných slov je zobrazen v dolní polovině okna na obrázku (Obr. 5) společně s nástroji pro jeho úpravu po levé straně.

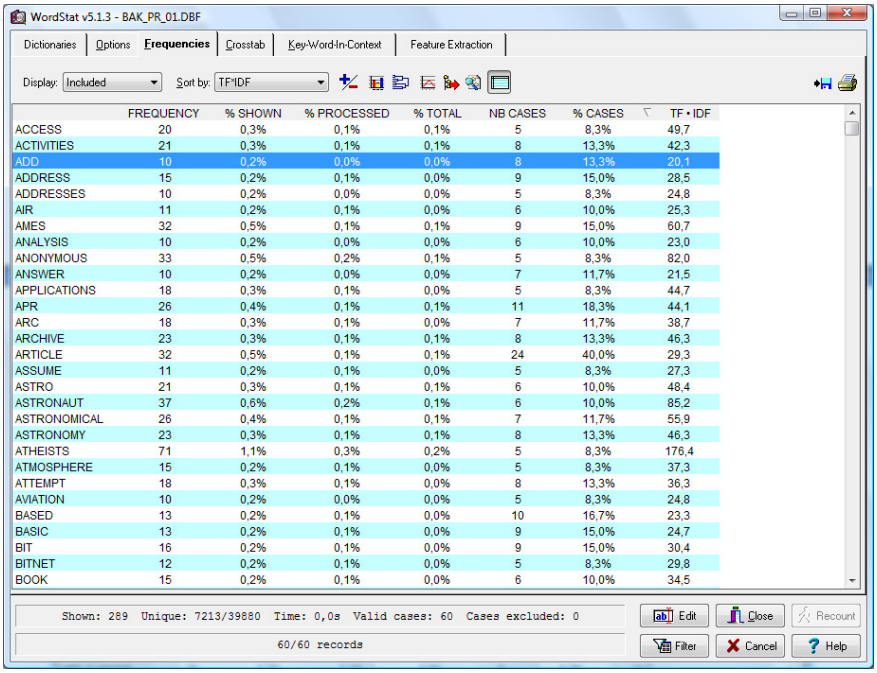
Záložka Options (vlastnosti) na obrázku (Obr. 6) umožňuje přidání či odstranění slov větších či menších než definovaná délka nebo počet výskytů, dále vynechání slov s volitelnou frekvencí výskytu.



Obr. 6 - Záložka Options programu Wordstat

Další možnosti jsou volby přidání či ignorování určitých typů textových a numerických znaků, možnost rozlišování velkých a malých písmen, zapnutí pravopisu a slovníku (opět pouze pro výše definované jazyky), nastavení způsobu zobrazení formátovacích značek a volba zobrazení prázdných kategorií.

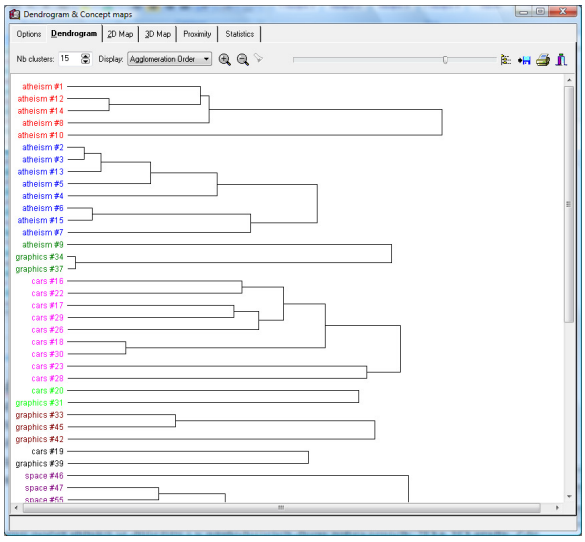
Záložka Frequencies (četnosti) na obrázku (Obr. 7) je již výpočtová. Její aktualizace proběhne při jejím prvním otevření a pak při každé změně předchozích záložek s nastaveními. Ukazuje přehled termů dle jedné z možností zobrazení (odpovídající výběru, všechny, mimo výběr) a statistiky dle jednotlivých kategorií. Pomocí pravého tlačítka myši je možnost dále pracovat s termy zařazováním či odstraňováním do/z kategorií, slovníku vynechávaných slov a možnostmi vyhledání termu v textech a jejich zobrazení. Pro označené termy je možné vytvořit sloupcový či výsečový graf a dále dendogram.



	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NB CASES	% CASES	TF • IDF
ACCESS	20	0,3%	0,1%	0,1%	5	8,3%	49,7
ACTIVITIES	21	0,3%	0,1%	0,1%	8	13,3%	42,3
ADD	10	0,2%	0,0%	0,0%	8	13,3%	20,1
ADDRESS	15	0,2%	0,1%	0,0%	9	15,0%	28,5
ADDRESSES	10	0,2%	0,0%	0,0%	5	8,3%	24,8
AIR	11	0,2%	0,1%	0,0%	6	10,0%	25,3
AMES	32	0,5%	0,1%	0,1%	9	15,0%	60,7
ANALYSIS	10	0,2%	0,0%	0,0%	6	10,0%	23,0
ANONYMOUS	33	0,5%	0,2%	0,1%	5	8,3%	82,0
ANSWER	10	0,2%	0,0%	0,0%	7	11,7%	21,5
APPLICATIONS	18	0,3%	0,1%	0,0%	5	8,3%	44,7
APR	26	0,4%	0,1%	0,1%	11	18,3%	44,1
ARC	18	0,3%	0,1%	0,0%	7	11,7%	38,7
ARCHIVE	23	0,3%	0,1%	0,1%	8	13,3%	46,3
ARTICLE	32	0,5%	0,1%	0,1%	24	40,0%	29,3
ASSUME	11	0,2%	0,1%	0,0%	5	8,3%	27,3
ASTRO	21	0,3%	0,1%	0,1%	6	10,0%	48,4
ASTRONAUT	37	0,6%	0,2%	0,1%	6	10,0%	85,2
ASTRONOMICAL	26	0,4%	0,1%	0,1%	7	11,7%	55,9
ASTRONOMY	23	0,3%	0,1%	0,1%	8	13,3%	46,3
ATHEISTS	71	1,1%	0,3%	0,2%	5	8,3%	176,4
ATMOSPHERE	15	0,2%	0,1%	0,0%	5	8,3%	37,3
ATTEMPT	18	0,3%	0,1%	0,0%	8	13,3%	36,3
AVIATION	10	0,2%	0,0%	0,0%	5	8,3%	24,8
BASED	13	0,2%	0,1%	0,0%	10	16,7%	23,3
BASIC	13	0,2%	0,1%	0,0%	9	15,0%	24,7
BIT	16	0,2%	0,1%	0,0%	9	15,0%	30,4
BITNET	12	0,2%	0,1%	0,0%	5	8,3%	29,8
BOOK	15	0,2%	0,1%	0,0%	6	10,0%	34,5

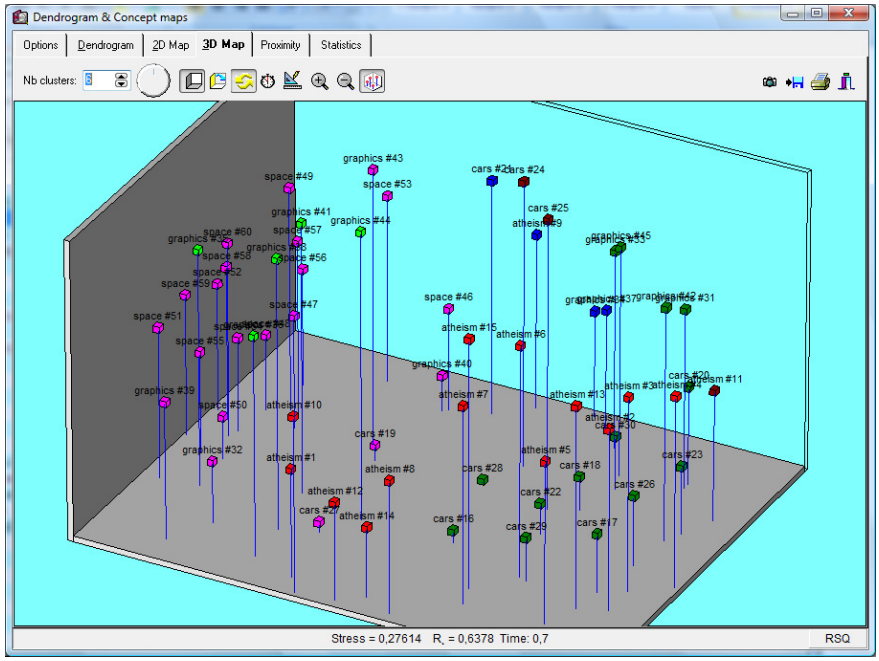
Obr. 7 - Záložka Frequencies

Volba dendogram otevře nové okno s upřesňujícím nastavením pro seskupování dat a následně je dle této volby zobrazen odpovídající dendogram na obrázku (Obr. 8). Posuvníkem nad dendogramem je možné měnit počet zobrazených shluků, ty jsou od sebe odděleny barevně. Dendogram je možné pomocí tlačítka exportovat do rastrového souboru bmp, png či jpg.



Obr. 8 - Dendogram v programu Wordstat

Pro zobrazení 2D a 3D grafu na obrázku (Obr. 9) lze měnit počet shluků obdobným způsobem. Pro snazší orientaci jsou data barevně oddělena. Zobrazení je možné natáčet a exportovat.

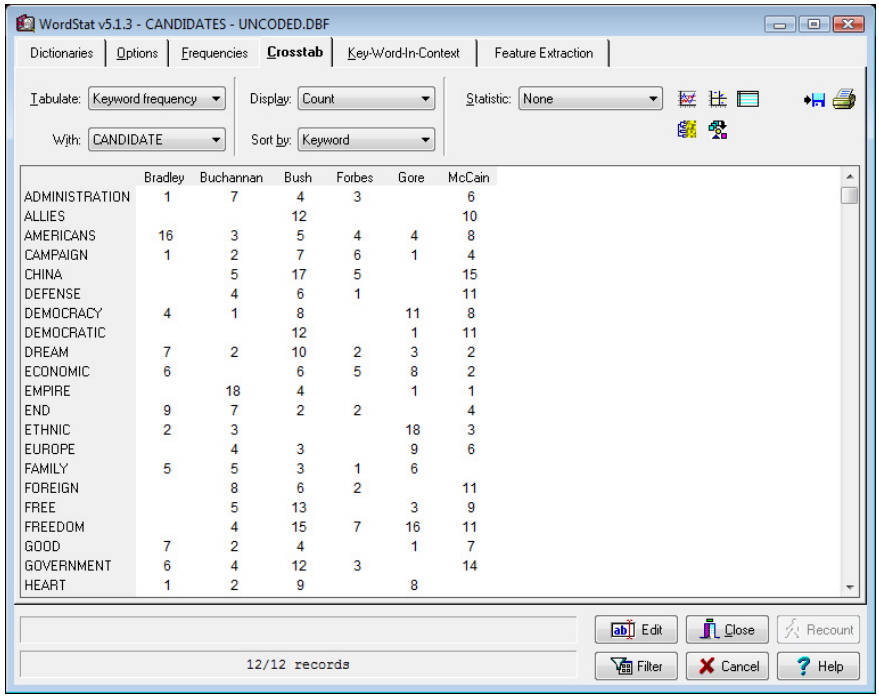


Obr. 9 - 3D graf zobrazení shluků

Zobrazení na záložce Proximity (blízkost) umožňuje zobrazit vztah blízkosti jednotlivých souborů či termů.

Záložka Crosstab (kontingenční tabulka) na obrázku (Obr. 10) nabízí zobrazení v relacích

- term – term,
- term – zvolený atribut,
- term – soubor.



	Bradley	Buchannan	Bush	Forbes	Gore	McCain
ADMINISTRATION	1	7	4	3		6
ALLIES			12			10
AMERICANS	16	3	5	4	4	8
CAMPAIGN	1	2	7	6	1	4
CHINA		5	17	5		15
DEFENSE		4	6	1		11
DEMOCRACY	4	1	8		11	8
DEMOCRATIC			12		1	11
DREAM	7	2	10	2	3	2
ECONOMIC	6		6	5	8	2
EMPIRE		18	4		1	1
END	9	7	2	2		4
ETHNIC	2	3			18	3
EUROPE		4	3		9	6
FAMILY	5	5	3	1	6	
FOREIGN		8	6	2		11
FREE		5	13		3	9
FREEDOM		4	15	7	16	11
GOOD	7	2	4		1	7
GOVERNMENT	6	4	12	3		14
HEART	1	2	9		8	

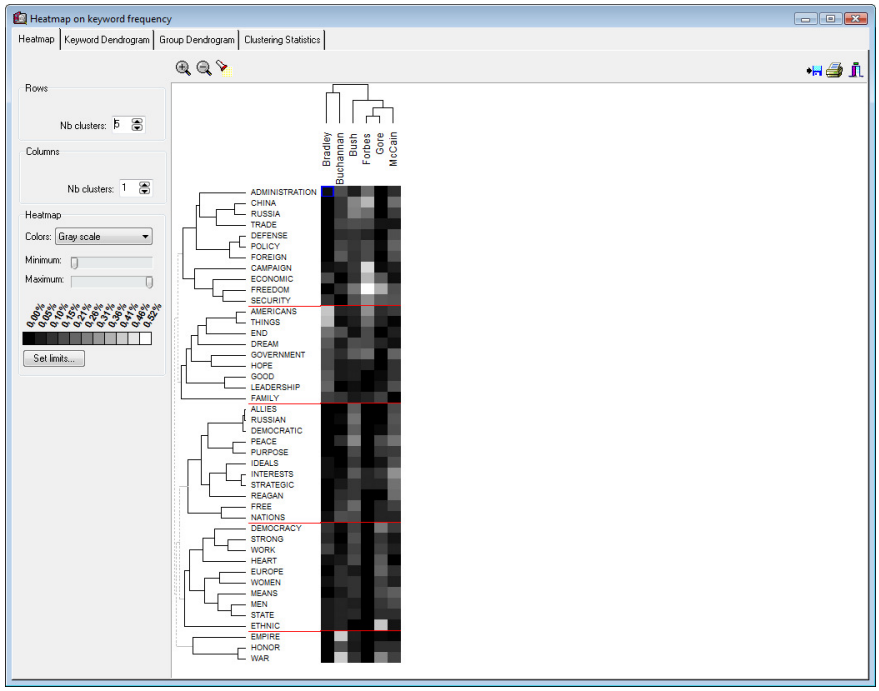
Obr. 10 - Kontingenční tabulka

Výpočet kontingenční tabulky je též spouštěn po kliknutí na tuto záložku. Uživatelsky vybrané řádky z tabulky lze zobrazit jejich graf, 2D mapa, 3D mapa a statistická tabulka.

Ikona Clustering and Heatmap na záložce Crosstab na obrázku (Obr. 10) otevře okno s rozsáhlými možnostmi zobrazení dat a to

- dendogram termů,
- dendogram atributů,
- statistiky v podobě kontingenčních tabulek pro všechny výše jmenované relace,
- heatmap.

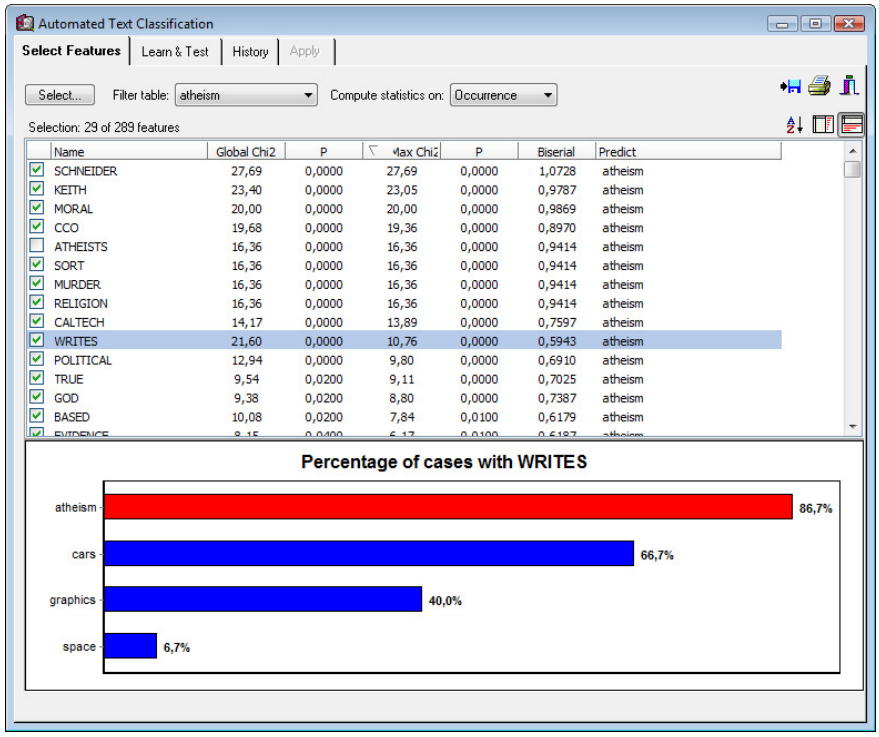
Heatmap na obrázku (Obr. 11) graficky prezentuje závislost termů a zvolených relací. Termy je možné graficky vodorovnou čarou rozdělit do definovaných počtů shluků. Světlá barva v dendogramu zobrazuje vyšší četnosti výskytu termů. Pro zobrazení term – term není tato volba dostupná. Pro zobrazení dendogramu je možné definovat barevnou škálu limitní hodnoty rozsahu.



Obr. 11 - Heatmap pro zobrazení závislostí

Automated text classification (automatická textová klasifikace) slouží k nastavení parametrů, použití projektu pro učení a testování, uživatelsky řízeným experimentům při kategorizaci, zobrazení efektivnosti a v neposlední řadě klasifikace externích dat pomocí nastavených parametrů.

První ze záložek na obrázku (Obr. 12) umožňuje výběr, třídění a filtrování termů dle kritérií.

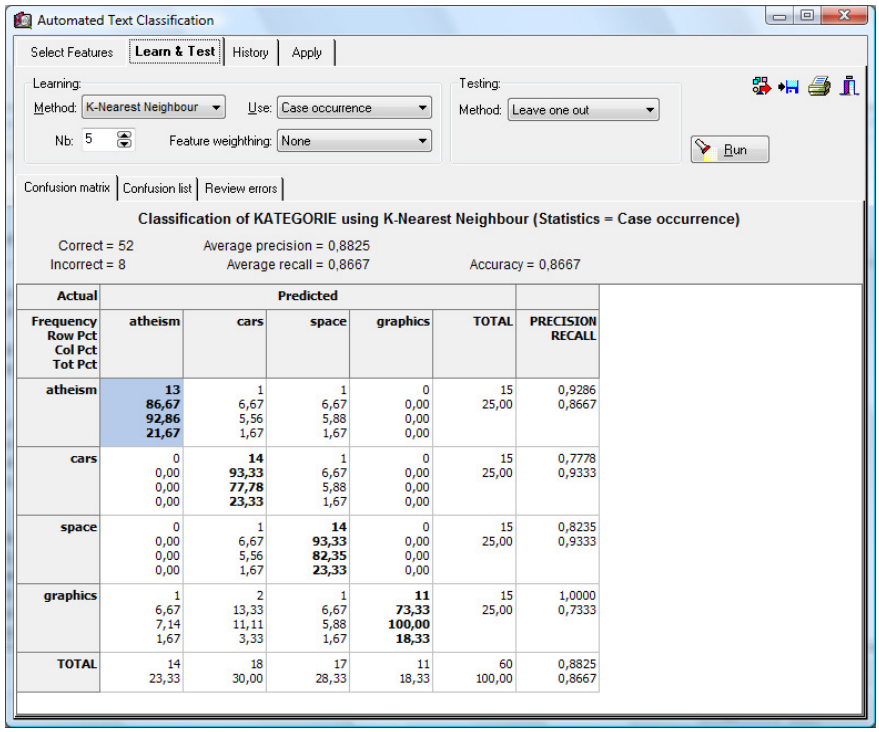


Obr. 12 - Automatická textová klasifikace – Selected Features

Dostupná kritéria jsou:

- globální χ^2,
- maximalizovaný χ^2,
- biserial.

Záložka Learn & test na obrázku (Obr. 13) umožňuje klasifikaci dokumentů a to buď metodou naivního Bayesova algoritmu nebo metodou nejbližšího souseda. Další volby umožňují nastavit podrobněji metodu učení a metodu testování. Výpočet je spouštěn manuálně, výsledky jsou zobrazovány v matici včetně údajů o skutečných hodnotách, předpovězených hodnotách a zobrazení chyb v jednotlivých kategoriích.



Actual	Predicted					
Frequency Row Pct Col Pct Tot Pct	atheism	cars	space	graphics	TOTAL	PRECISION RECALL
atheism	**13** **86,67** **92,86** **21,67**	1 6,67 5,56 1,67	1 6,67 5,88 1,67	0 0,00 0,00 0,00	15 25,00	0,9286 0,8667
cars	0 0,00 0,00 0,00	**14** **93,33** **77,78** **23,33**	1 6,67 5,88 1,67	0 0,00 0,00 0,00	15 25,00	0,7778 0,9333
space	0 0,00 0,00 0,00	1 6,67 5,56 1,67	**14** **93,33** **82,35** **23,33**	0 0,00 0,00 0,00	15 25,00	0,8235 0,9333
graphics	1 6,67 7,14 1,67	2 13,33 11,11 3,33	1 6,67 5,88 1,67	**11** **73,33** **100,00** **18,33**	15 25,00	1,0000 0,7333
TOTAL	14 23,33	18 30,00	17 28,33	11 18,33	60 100,00	0,8825 0,8667

Obr. 13 - Automatická textová klasifikace - Learn & Test

Záložka History slouží pro vytváření experimentů s možností hromadného zpracování dat a volby různých parametrů pro porovnávání úspěšnosti klasifikace při jejím různém nastavení. Výsledky jsou zobrazovány a ukládány do tabulky dat a grafu.

Záložka Apply umožňuje klasifikovat data pomocí nastavení v předchozích záložkách. Pro načtení dat ke klasifikaci lze zvolit jeden ze způsobů

- jeden soubor,
- výběr souborů,
- soubory ve stávajícím projektu,
- jiný projekt se soubory.

Záložka hlavního okna aplikace Key word in kontext umožňuje vyhledávání definovaných či vlastních slov v dokumentech projektu. Výsledky jsou zobrazeny včetně náhledů jednotlivých dokumentů, ty zde lze editovat a ukládat.

Záložka Feature Extraction slouží pro vyhledávání frází, spojení slov a jejich četnosti. Výsledky je možno dále třídit, porovnávat, editovat v dokumentech a přiřazovat do definovaných kategorií či do listu vynechaných slov.

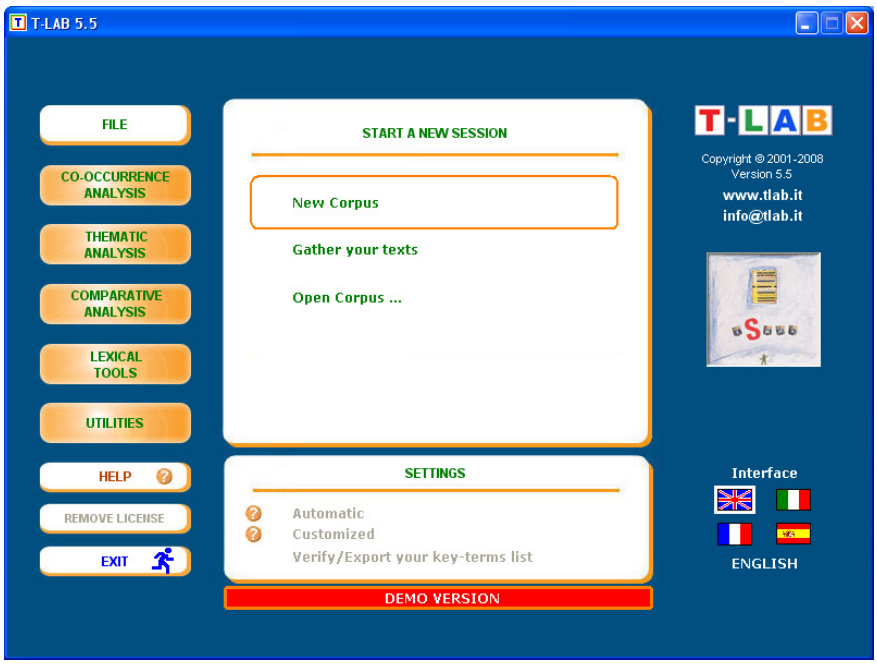
4.2.T-Lab

T-Lab [17] od stejnojmenné společnosti je komerční program distribuovaný ve čtyřech jazykových mutacích: anglické, italské, španělské a francouzské. Program je zaměřen na tyto tři oblasti textové analýzy

- závislostní analýza, kde jsou slovní závislosti, porovnávání párů slov, analýza souvisejících slov, analýza předchůdců a následovníků, výskyt v textu,
- obsahová analýza elementárních spojení, vět a odstavců, sekvenční analýza témat, obsahová analýza dokumentů, klíčové souvislosti slov,
- srovnávací analýza podmnožin projektu, correspondence analysis, jedno- a vícenásobná analýza shodnosti, shluková analýza.

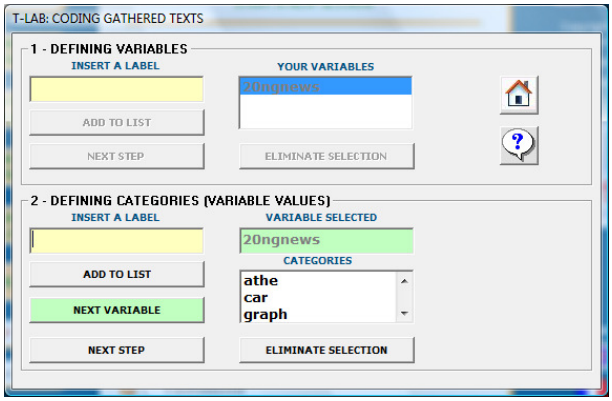
4.2.1. Tvorba korpusu

Pomocí tvorby korpusu se provádí předzpracování textu. Pro začátek práce je možné zvolit na obrázku (Obr. 14) vytvoření nového korpusu (New corpus), Vložení vlastních textů (Gather your texts) a načtení již vytvořeného korpusu (Open corpus).



Obr. 14 - Výchozí okno aplikace T-LAB

Pro práci s vlastními texty, pak je určena volba Gather your text a vytvoření txt souboru se všemi požadovanými texty. Vložit lze pouze txt soubory obsahující jednotlivé texty, vybrány jsou všechny soubory obsažené v dané složce. Dále je možná automatická tvorba korpusu či vlastní nastavení Variables na obrázku (Obr. 15) a přiřazení textů do nich.

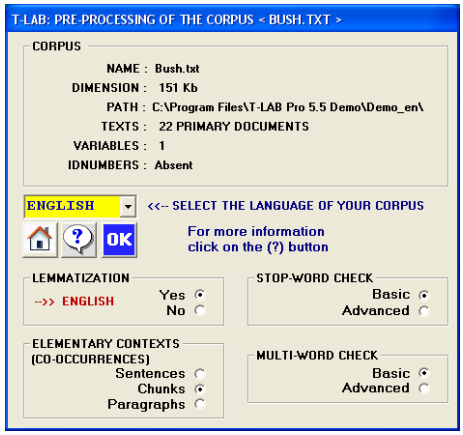


Obr. 15 - Nastavení Variables

Pro takto zpracovaná data je vygenerován soubor MyCorpus.txt obsahující všechny texty doplněné o index souboru, ze kterého text pochází a jeho přiřazení do kategorie. Dalším souborem, který je generován je tabulka s indexy souborů a jejich odpovídajícím původním názvem.

Při volbě New corpus je vybrán textový soubor s připravenými daty, posléze je k dispozici na obrázku (Obr. 16) výčet statistických údajů o souboru a nastavení těchto parametrů

- jazyk korpusu (anglický, italský, francouzský, španělský, jiný),
- lemmatizace,
- společné výskyty,
- seznam slov vyjmutých z analýzy,
- seznam slovních spojení.

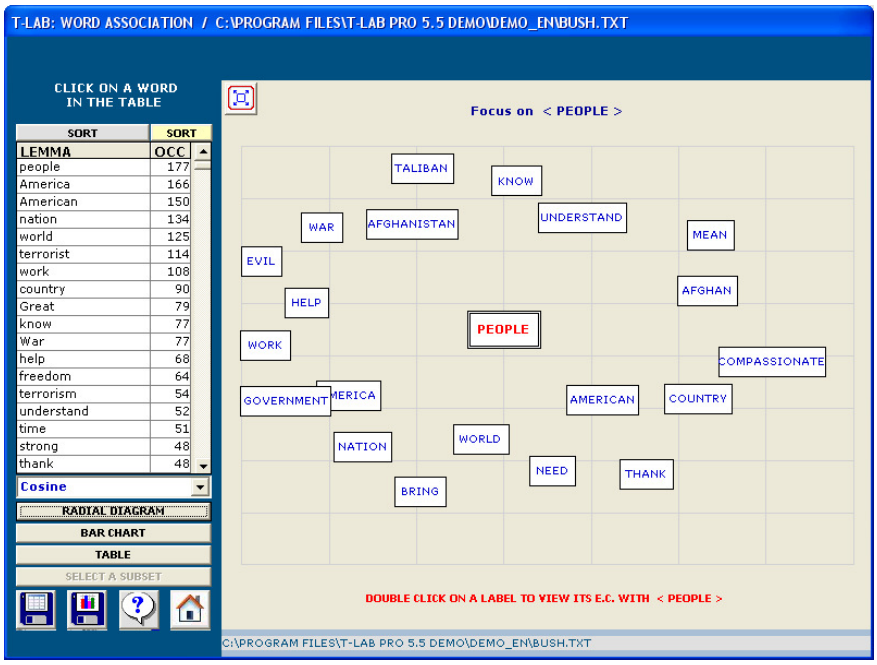


Obr. 16 - Nastavení parametrů korpusu

4.2.2. Závislostní analýza

Závislost mezi slovy je možné zvolit buď v rámci celého korpusu, nebo výběru jeho části. V levém dolním rohu je volba metody výpočtu asociačního koeficientu (Cosine, Dice a Jaccard). Zobrazení diagramu na obrázku (Obr. 17) ukazuje závislost mezi vybraným slovem

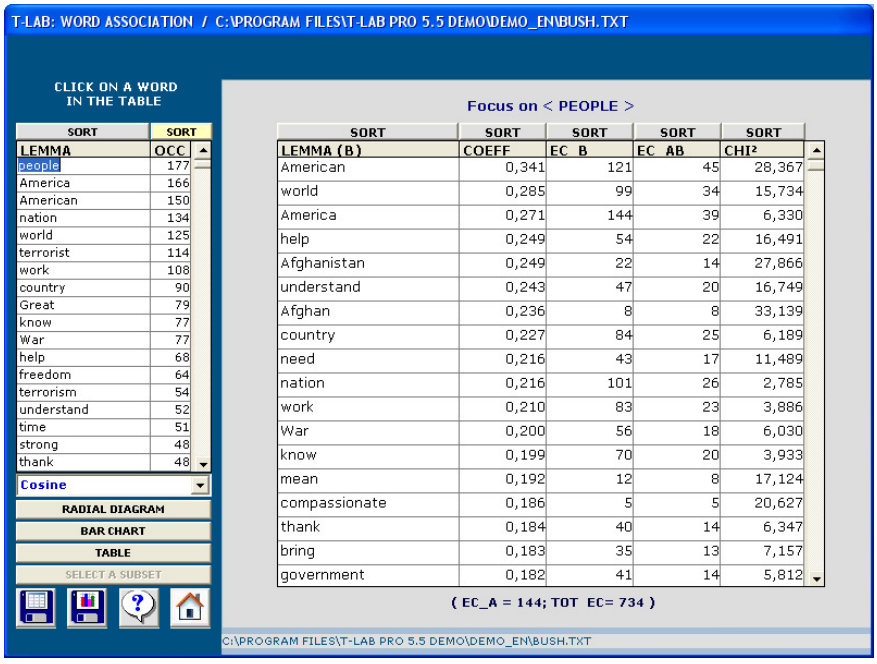
z levého sloupce a ostatními slovy. Poklepáním na kterékoliv ze slov zobrazených v diagramu je zobrazen text v html tvaru se zvýrazněním těchto slov.



Obr. 17 - Diagram závislosti slov

Sloupcový graf graficky znázorňuje hodnoty vypočtených koeficientů slov.

Tabulka na obrázku (Obr. 18) zobrazuje vybrané slovo, související slova, jejich koeficient, celkový výskyt souvisejících slov v korpusu či jeho části, celkový výskyt vybraného slova, celkový počet společných výskytů vybraného slova a slova souvisejícího a hodnotu χ^2 testu.



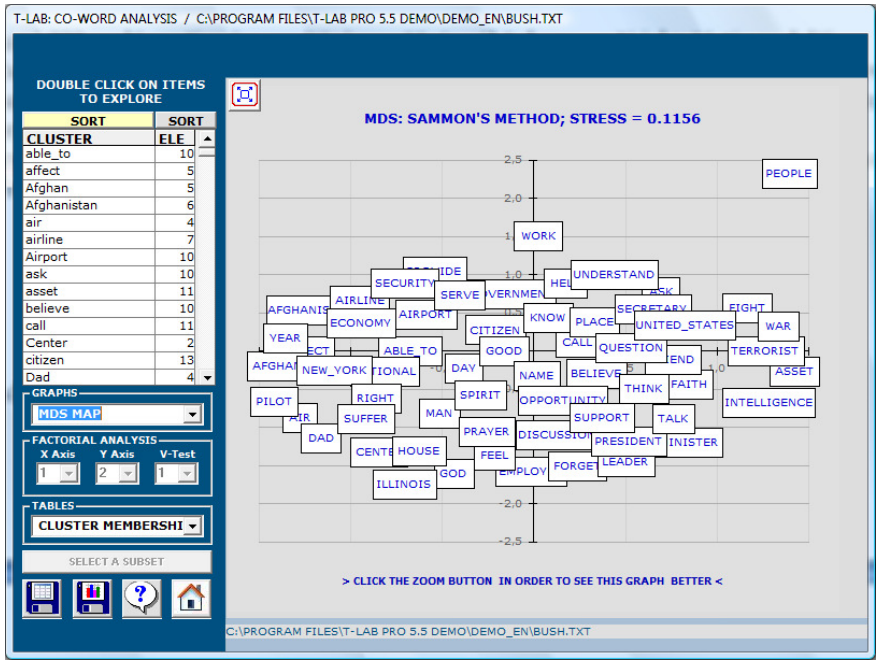
Obr. 18 - Tabulka hodnot pro dvojice slov

Porovnání páru slov v tabulce Intersection umožňuje vybrat vždy pár slov A, B a zobrazit pro ostatní slova hodnoty výskytu společně se slovy z páru, dále je pak zobrazena hodnota χ^2 testu a možnost výběru vztahů pro zobrazení v grafu výskytu zobrazení.
Tabulka Difference uvádí hodnoty výskytu rozdílů slov z párů, nebo-li při výskytu slova A z páru a zároveň bez výskytu slova B a naopak.

Shluková analýza při svém spuštění umožňuje výběr korpusu či jeho části pro další výpočet dle počtu užitých slov. V případě deseti až sto slov program dále pracuje s těmito slovy, v případě více než sto slov program vytvoří automaticky z těchto slov malé shluky, s kterými dále pracuje. Dále je výběr metody výpočtu asociačního indexu (Cosine, Dice, Jaccard) a počet shluků v rozmezí padesát až sto. Po výběru parametrů program provádí tyto kroky

- tabulka současných výskytů,
- výpočet asociačních indexů (Cosine, Dice, Jaccard),
- hierarchické shlukování (pouze v případě více než sto slov),
- grafická prezentace dat.

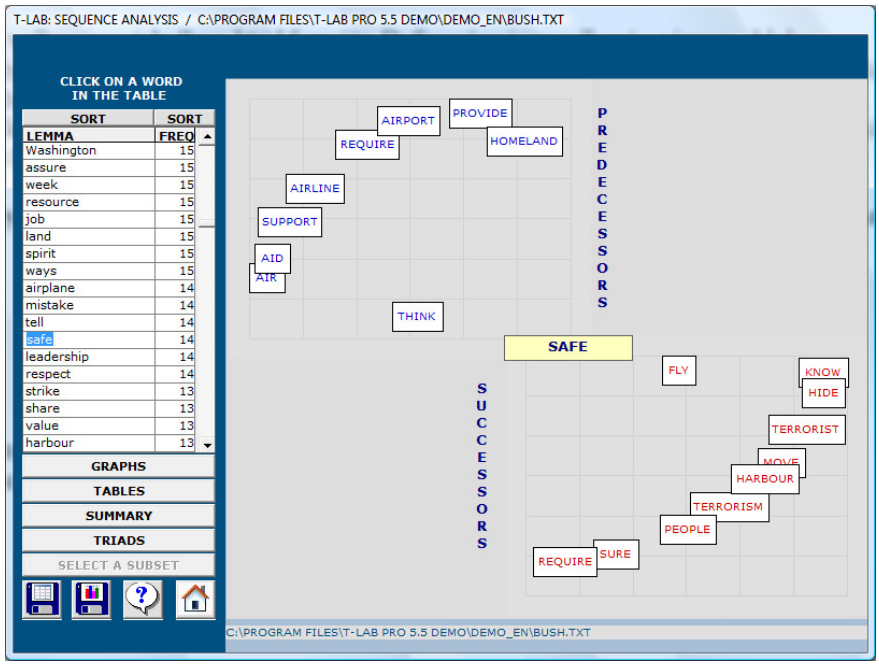
Část výsledků je automaticky exportována do tabulek ve formátu xls a uložena do pracovní složky. V horní části levého sloupce na obrázku (Obr. 19) jsou zobrazeny použité shluky (či slova). Níže je volba grafického výstupu dat, a to MDS mapa, faktorová analýza, asociační diagramy pro jednotlivé shluky a graf polohy ve středu x hustoty shluku.



Obr. 19 - Zobrazení shluků

Třetí volbou jsou tabulky v html formátu. Definovány jsou celková s názvem shluku, množstvím výskytů všech spojení obsažených ve shluku, polohy ve středu, hustoty a slova obsažená ve shluku. Další je tabulka členů v jednotlivých shlucích a poslední tabulkou je tabulka indexů asociace. Zde je před jejím zobrazením možnost výběru mezi shluky a uvnitř shluků.

Analýzu předchůdců a následovníků je opět možné zvolit z celého korpusu či z jeho části. Na obrázku (Obr. 20) jsou graficky zobrazeni předchůdci a následovníci pro vybrané slovo.



Obr. 20 - Předchůdci a následovníci

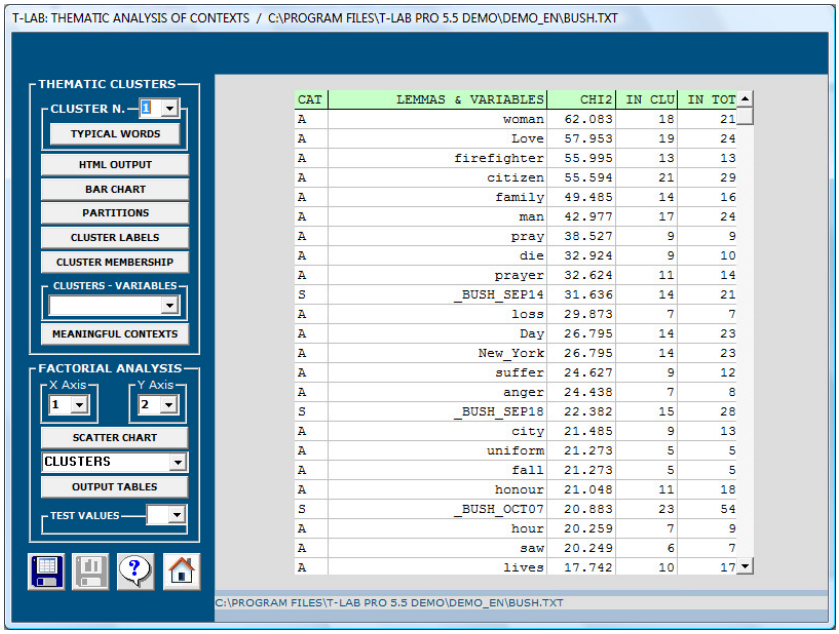
Tabulka pak osahuje informace o předchůdcích a následovnících společně s jejich hodnotou pravděpodobnosti výskytu.

Triads (trojice) umožňují výběr slova a jeho pozice ve spojení tří slov. Pro vybrané slovo jsou pak vypsána nalezená spojení.

Poslední volba umožňuje zobrazovat slova v textu. Všechny výskyty jsou zobrazeny v tabulce, související text je možné zobrazit v okně náhledu pod tabulkou. Možnost exportu všech výskytů daného slova vytváří soubor html.

4.2.3. Obsahová analýza

Obsahová analýza textu na obrázku (Obr. 21) zobrazuje lemma a proměnné ve vybraném shluku, odpovídající hodnotu χ^2 testu, počet výskytů ve vybraném shluku a počet výskytů v korpusu či jeho části, dle uživatelské volby. Dále je pak zobrazeno lemma a proměnné na definované v uživatelském nastavení a na doplňkové, vytvořené programem.

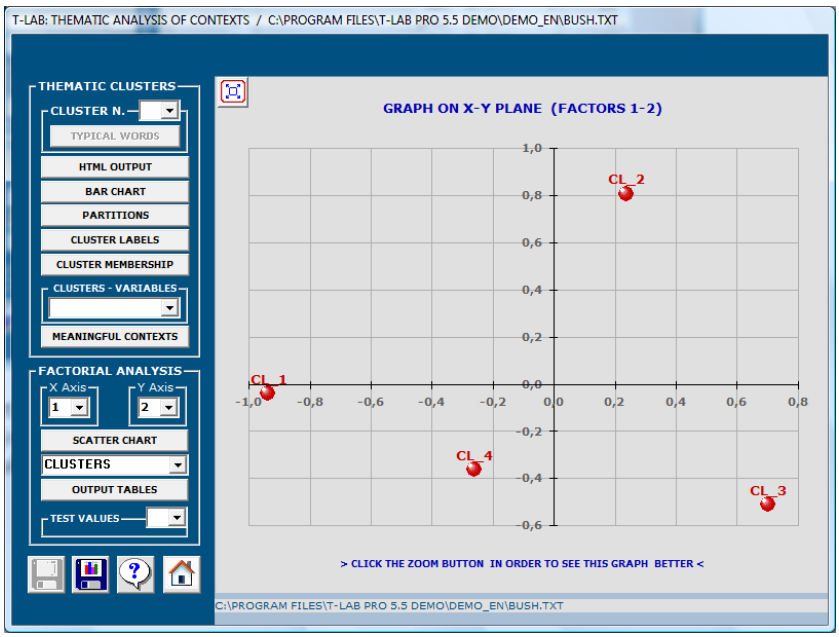


CAT	LEMMAS & VARIABLES	CHI2	IN CLU	IN TOT
A	woman	62.083	18	21
A	Love	57.953	19	24
A	firefighter	55.995	13	13
A	citizen	55.594	21	29
A	family	49.485	14	16
A	man	42.977	17	24
A	pray	38.527	9	9
A	die	32.924	9	10
A	prayer	32.624	11	14
S	_BUSH_SEP14	31.636	14	21
A	loss	29.873	7	7
A	Day	26.795	14	23
A	New_York	26.795	14	23
A	suffer	24.627	9	12
A	anger	24.438	7	8
S	_BUSH_SEP18	22.382	15	28
A	city	21.485	9	13
A	uniform	21.273	5	5
A	fall	21.273	5	5
A	honour	21.048	11	18
S	_BUSH_OCT07	20.883	23	54
A	hour	20.259	7	9
A	saw	20.249	6	7
A	lives	17.742	10	17

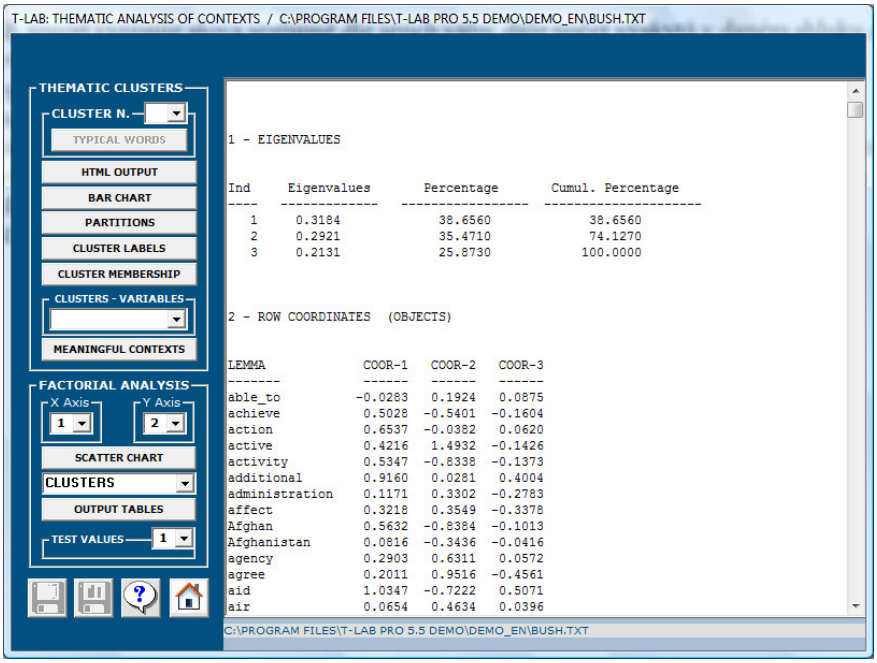
Obr. 21 - Obsahová analýza textu

HTML report vypisuje slova sestupně dle jejich váhy, dále počet výskytů v daném shluku a celkových výskytů.

Histogram zobrazuje procentuální podíl výskytů ve vytvořených shlucích. Zobrazení vztahů mezi shluky ve dvourozměrném prostoru je na obrázku (Obr. 22), zobrazení lze vybrat pro shluky, lemma, shluky a proměnné, shluky a lemma. Názvy shluků je možné editovat. U faktorové analýzy lze měnit nastavení hodnot os X a Y. Výsledek korespondenční analýzy na obrázku (Obr. 23) je zobrazen v tabulce s výpisem charakteristických hodnot a hodnot vypočtených pro každé lemma.

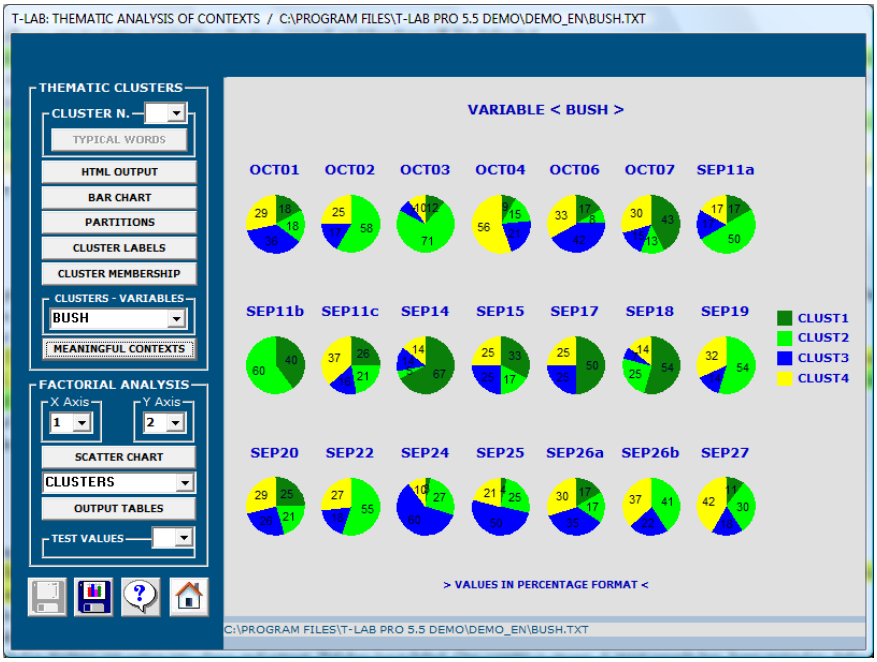


Obr. 22 - Graficky zobrazený vztah mezi shluky



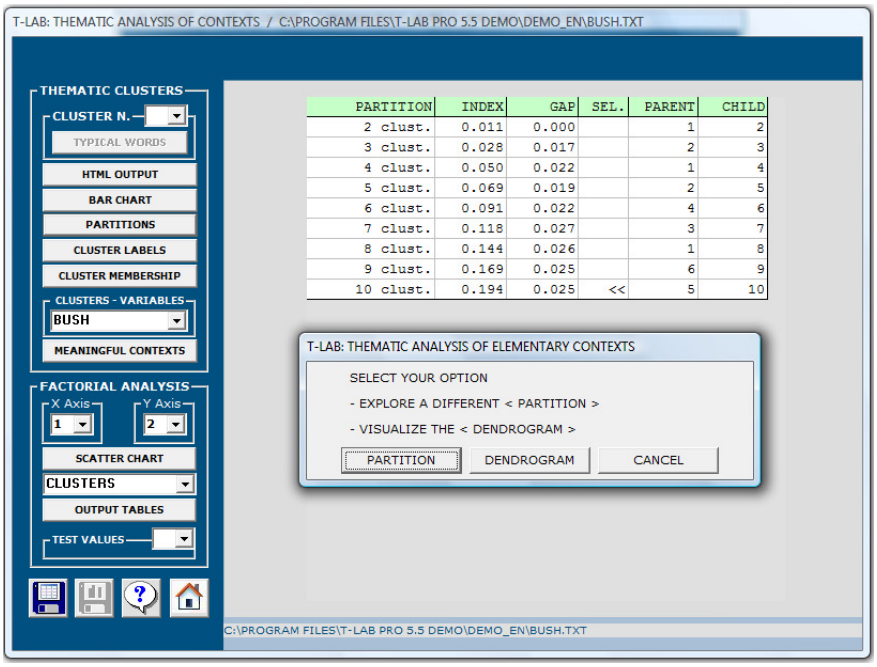
Obr. 23 - Výsledek korespondenční analýzy

Pro jednotlivé části korpusu je na obrázku (Obr. 24) zobrazen graf s poměrným výskytem jednotlivých clusters v částech korpusu. Dále lze celý text exportovat do html formátu s barevným rozlišením.

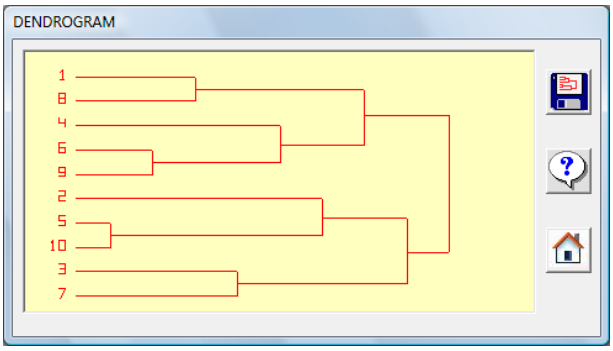


Obr. 24 - Poměrný výskyt clusters v částech korpusu

U shlukování jsou v tabulce zobrazeny hodnoty indexu, GAP (rozdíl mezi dvěma po sobě jdoucími indexy), rodič a dítě na obrázku (Obr. 25) a dendogram na obrázku (Obr. 26).



Obr. 25 - Poměrný výskyt Variable v částech korpusu



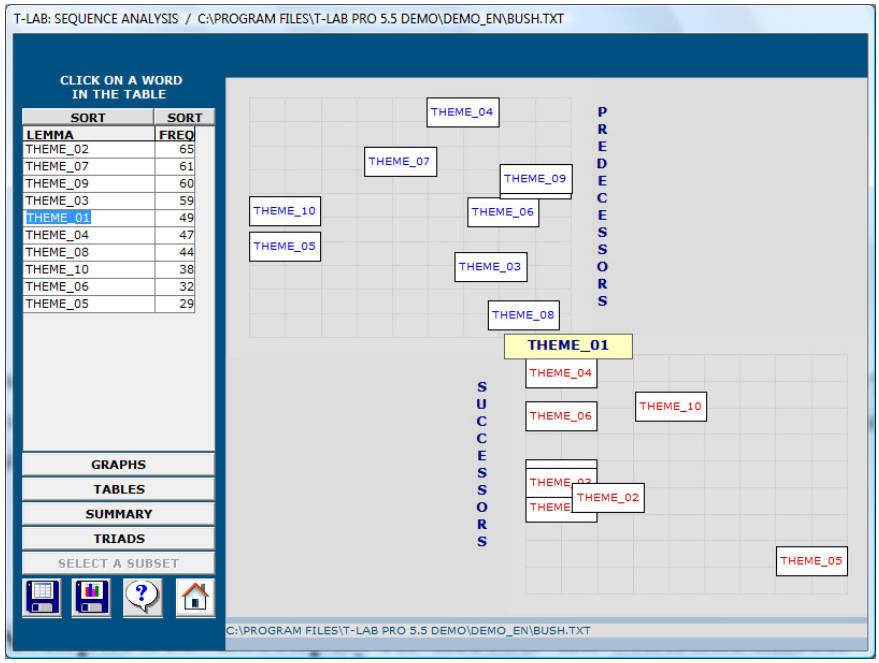
Obr. 26 - Dendogram hierarchického shlukování

Program nabízí analýzu předchůdců a následovníků nejen pro jednotlivá slova, jak bylo uvedeno výše, ale též pro shluky. Podmínkou je minimálně pět shluků vytvořených obsahovou analýzou. Výsledky je možné zobrazit grafem na obrázku (Obr. 27), tabulkou předchůdců a následovníků s hodnotami pravděpodobnosti výskytu, celkovým přehledem a možností zobrazení trojic pří zachování výběru pozice daného shluku.

Klasifikace dle tématu dokumentů obsažených v korpusu lze použít pouze tehdy, jsou-li splněny tyto dvě podmínky

- minimálně 20 textů obsažených v korpusu,
- délka textu musí být dostatečně velká, aby nebyly programem považovány za jednoduchá spojení.

Při splnění těchto podmínek pak jsou k dispozici stejné nástroje jako pro obsahovou analýzu textu.



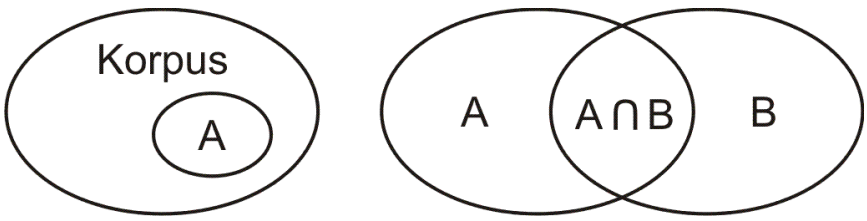
Obr. 27 - Analýza předchůdců a následovníků

Souvislosti charakteristických slov je možné analyzovat v celém korpusu, či v jeho části. Na rozdíl od ostatních metod zde můžeme pro každé charakteristické slovo upravovat list souvisejících slov a tím přesněji definovat jeho význam. Každé charakteristické slovo je možné definovat až padesáti souvisejícími slovy.

4.2.4. Srovnávací analýza

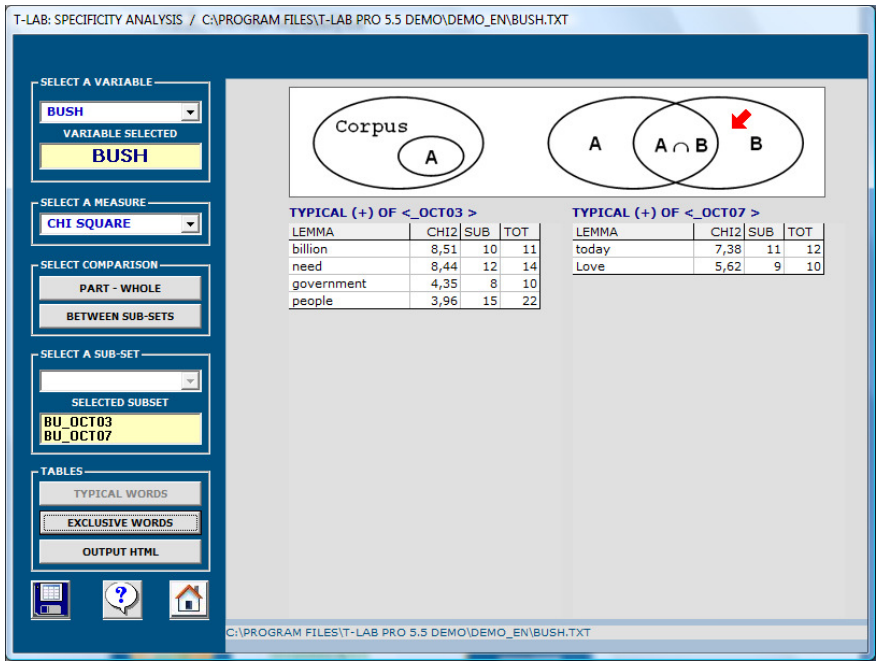
Analýza charakteristických slov umožňuje zjistit, která slova jsou typická či výlučná pro danou část textu. Máme možnost výběru mezi dvěma typy srovnávání v závislosti na výskytu na obrázku (Obr. 28)

- mezi množinou „A“ a zbylými daty korpusu,
- mezi dvěma množinami „A“ a „B“.



Obr. 28 - Typy srovnávání v závislosti na výskytu

Výstup těchto analýz je na obrázku (Obr. 29). Tato analýza je dostupná, pokud jsou v korpusu alespoň dva texty.



Obr. 29 - Lemma charakteristická pro daný dokument

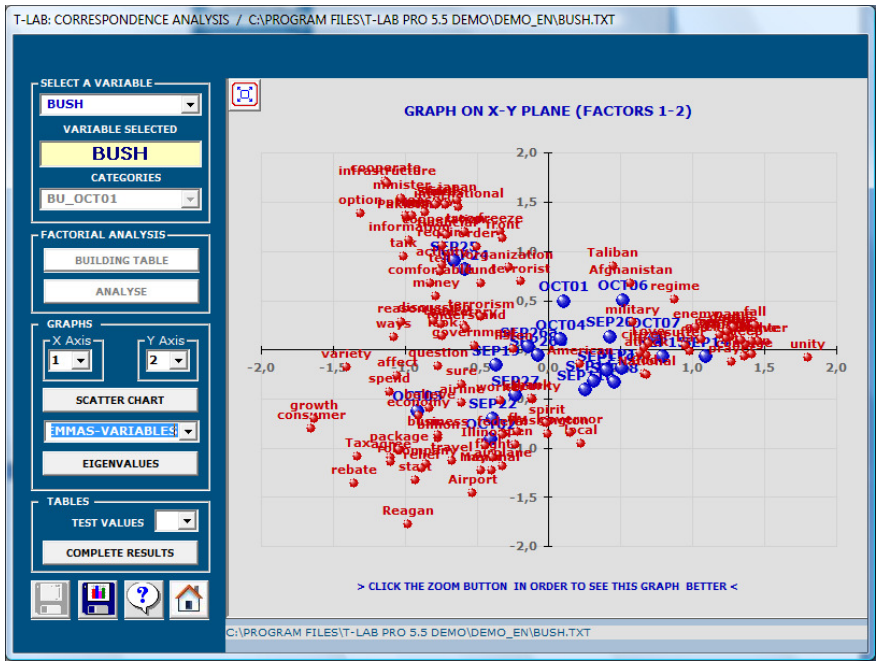
Pro zjištění shodností či rozdílností v textu je určena correspondence analysis. T-Lab umožňuje zjišťovat tyto vztahy pomocí dvou typů tabulek

- s hodnotami výskytů (lemma na proměnné),
- s hodnotami spoluvýskytů (elementární shluky na lemma).

Výstupem pak je graf se zobrazením vztahů mezi lemma, proměnnými či obojím navzájem na obrázku (Obr. 30).

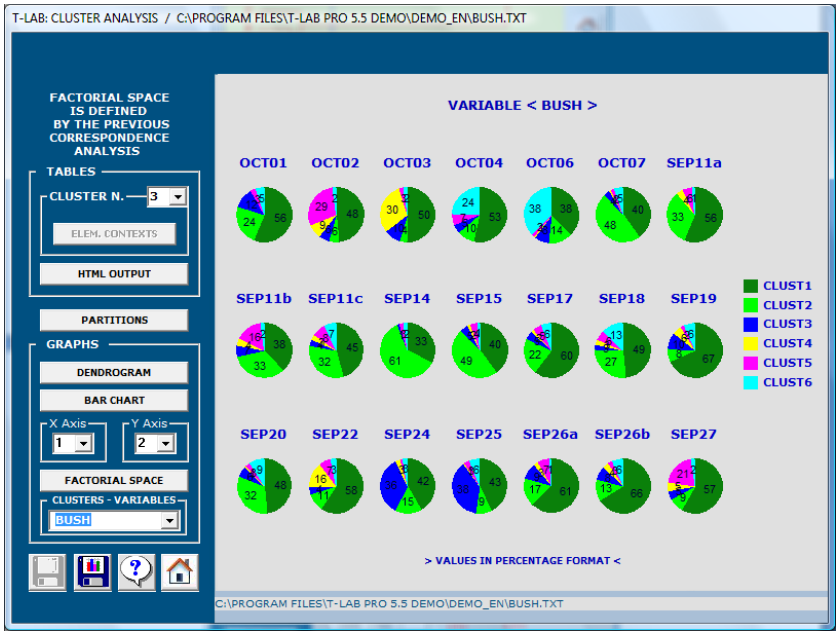
Shluková analýza umožňuje vybrat jednu ze tří technik

- hierarchické shlukování,
- K-means,
- Kohonenovy mapy.

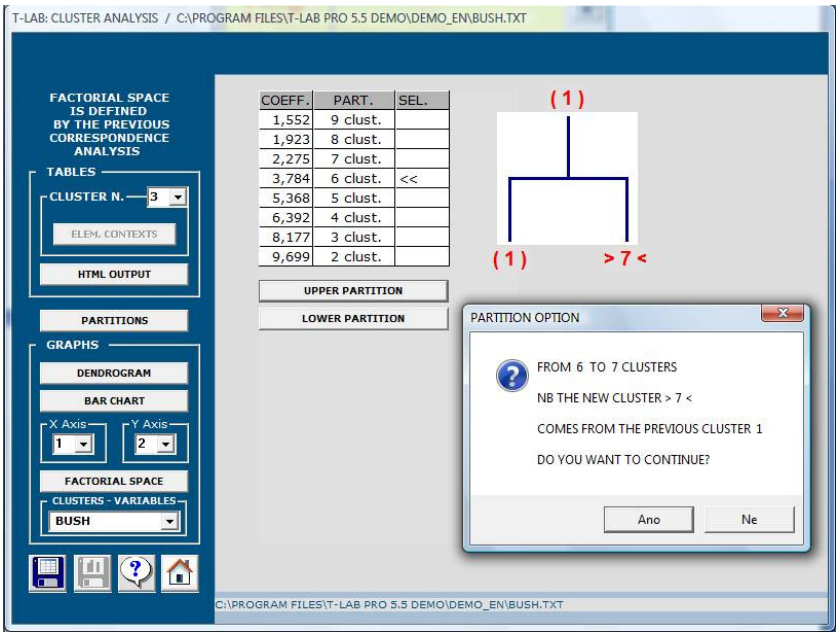


Obr. 30 - Graf vztahu lemma a proměnných

První dvě metody poskytují jako výstup grafy a tabulky výsledků pro dva až devět shluků a údaje v závislosti na typu použité metody. Výsledky je možné zobrazovat více způsoby: sloupcovým grafem, dendogramem, zobrazením jednotlivých shluků výčtem obsažených výrazů a vypočtených hodnot, koláčovými grafy zobrazujícími poměr jednotlivých shluků v částech korpusu na obrázku (Obr. 31), tabulkami s celkovým přehledem a dále pak možností editace počtu shluků na obrázku (Obr. 32).



Obr. 31 - Grafy zastoupení clusters v částech korpusu



Obr. 32 - Zobrazení a editace počtu shluků

V případě volby Kohonenovy mapy je pak výstup pouze jediný, ve formátu html, ve tvaru tabulky (Tab. 3) neuronové sítě a odpovídajících roztříděných výrazů.

Tab. 3 - Výstup Kohonenovy mapy

son (34)	health (610)	worry (28)
help (32)	wife (76)	go_out (13)
contentment (31)	grandchild (30)	sister (8)
world (26)	content (25)	pay (7)
welfare (22)	look (20)	day (7)
well_being (17)	dog (20)	
church (16)	try (16)	
country (15)	stay (15)	
family (708)	work (130)	good (303)
life (162)	healthy (45)	think (49)
live (145)	time (27)	food (24)
child (137)	enjoy (24)	daughter (24)
husband (95)	suppose (23)	music (19)
home (90)		know (18)
peace (77)		hobby (15)
people (63)		get_on (15)
mind (47)		
security (40)		
relationship (27)		
holiday (23)		
job (151)	happiness (228)	able_to (52)
car (22)	money (171)	comfortable (20)
future (17)	happy (136)	state (8)
social (15)	friend (117)	
	freedom (38)	
	standard (37)	
	love (36)	
	house (36)	
	nice (30)	
	important (26)	
	education (25)	
	satisfaction (21)	

5. Komparace dostupných metod na standardních datových sadách

Pro komparaci byly vybrány dva programy, které si dle údajů výrobců byly podobné jak funkčností, tak jejich cenovou hladinou. Byl tudíž předpoklad, že poměr cena/výkon bude u těchto programů srovnatelný. Program Wordstat byl užíván na základě licence zakoupené Univerzitou Pardubice a testován na výběru z datové sady 20 Newsgrous , Program T-Lab byl testován v demoverzi, která však neumožňuje načtení vlastních dat. Testován byl tedy na ukázkové datové sadě textů Bush. Nad rámec zadání byla vytvořena datová sada elektronické korespondence pocházející z firemních dat výrobně obchodní firmy. Tato sada měla za úkol prověření funkčnosti zpracování těchto typů dokumentů v testovaných produktech. Vzhledem k omezení demoverze programu T-Lab pak byla sada testována pouze v programu QDA Miner a Wordstat.

5.1. Testování dat v programu Wordstat a QDA Miner

Výběr z datové sady 20 Newsgroups

Originální datová sada 20 Newsgroups data set obsahuje přibližně dvacet tisíc anglických příspěvků zařazených do dvaceti skupin. Z této sady je vytvořen výběr dokumentů zařazených do čtyř skupin popsaných v tabulce (Tab. 4), který bude dále používán jako standardní sada. Texty jsou uloženy v jednotlivých txt souborech.

Tab. 4 - Výběr z 20 Newsgroups

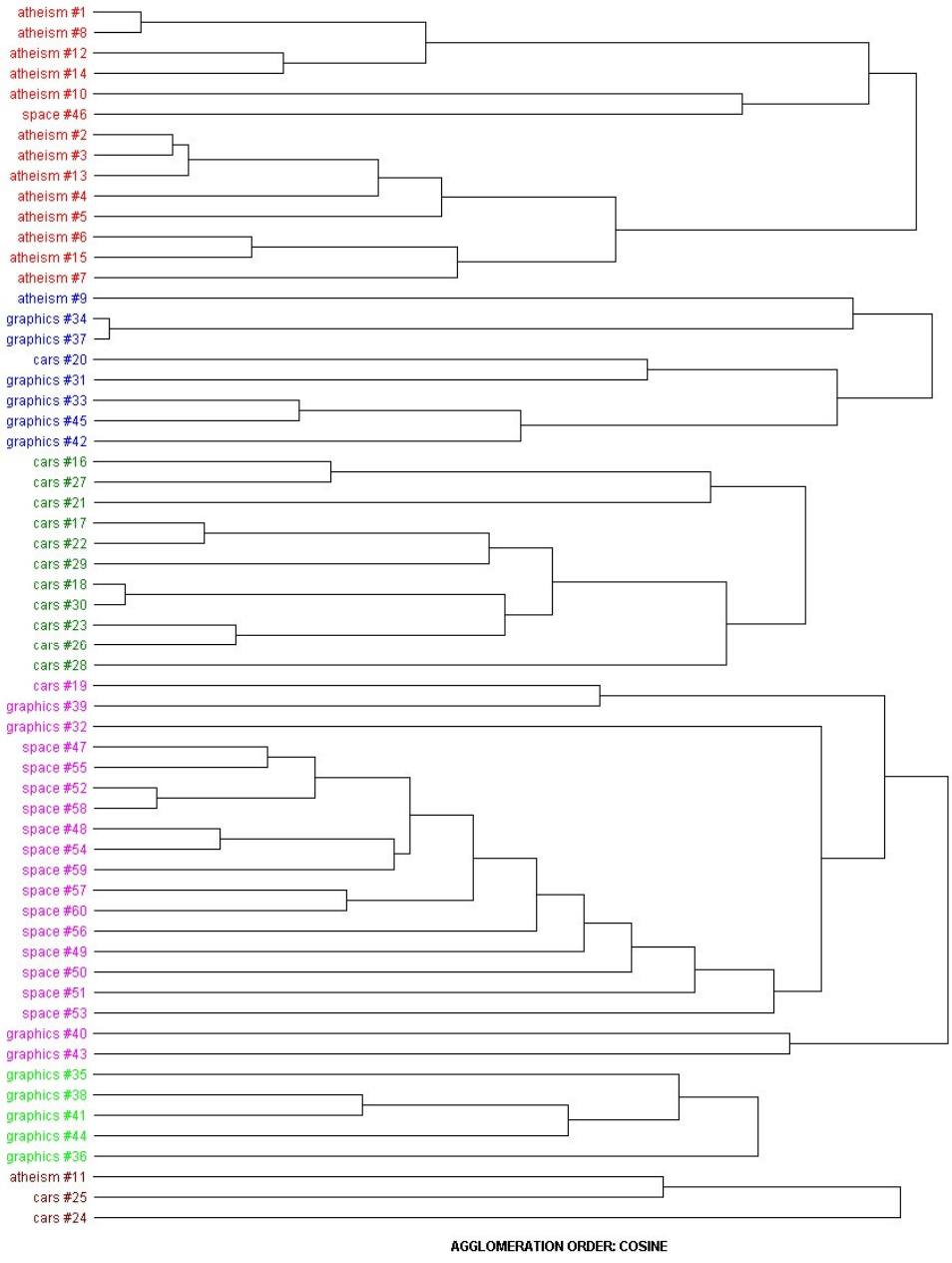
Výběr z 20 Newsgroups				
	Atheism	Space	Cars	Graphics
Počet souborů	15	15	15	15

Práce s programem

Načtení souborů s texty sady Výběru z 20 Newsgroups bylo provedeno v programu QDA Miner, tyto soubory byly zařazeny do kategorií dle tabulky (Tab. 4). Textová data nebyla dále v programu upravována, po spuštění programu Wordstat do něj byla načtena.

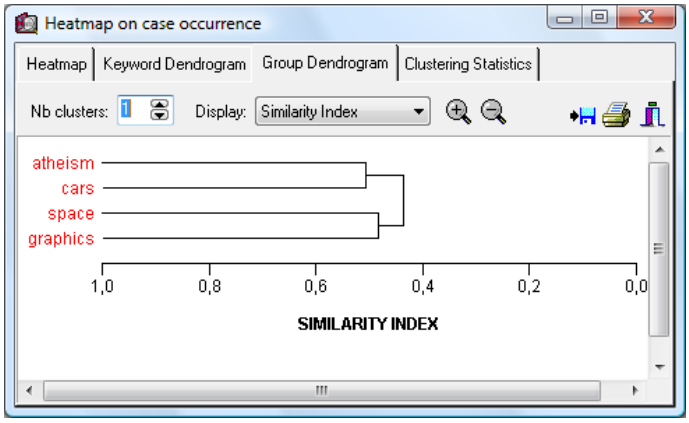
Další testování proběhlo s využitím anglické lemmatizace a slovníku vyjmutých slov obsažených v programu.

Experimentálně byl zkoušen optimální počet shluků v hierarchickém shlukování termů. Při shlukování bylo optimum nalezeno při počtu šesti shluků na obrázku (Obr. 33) oproti původním čtyřem kategoriím, při nižším počtu byly již výsledky shlukování neuspokojivé vzhledem k charakteru a původu datové sady, kdy texty jsou převážně krátké a většina z nich má stejnou či velmi podobnou hlavičku s informacemi.



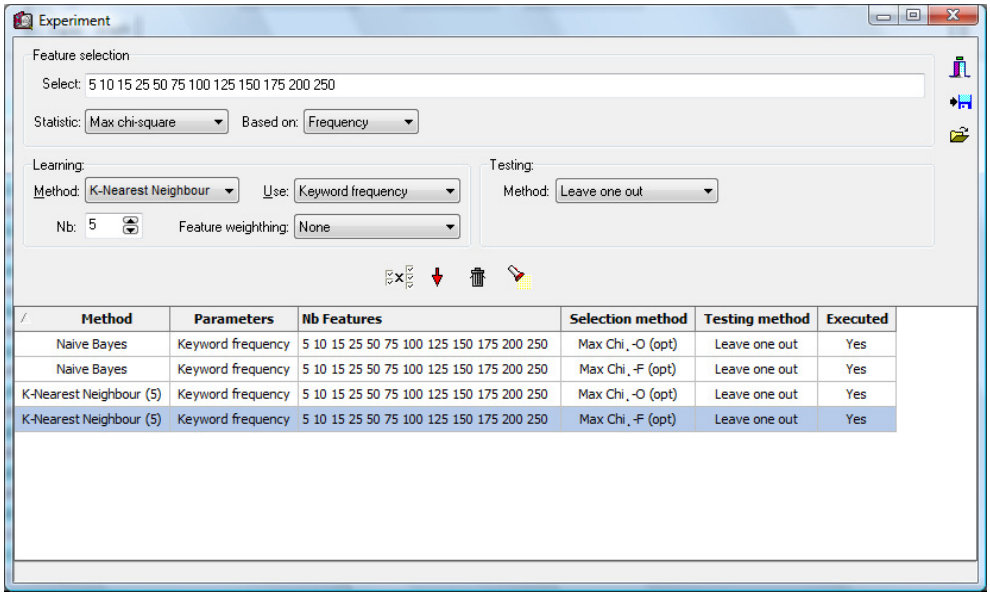
Obr. 33 - Dendogram shlukování testovací sady

Similarity index je pro jednotlivé skupiny uveden na obrázku (Obr. 34).

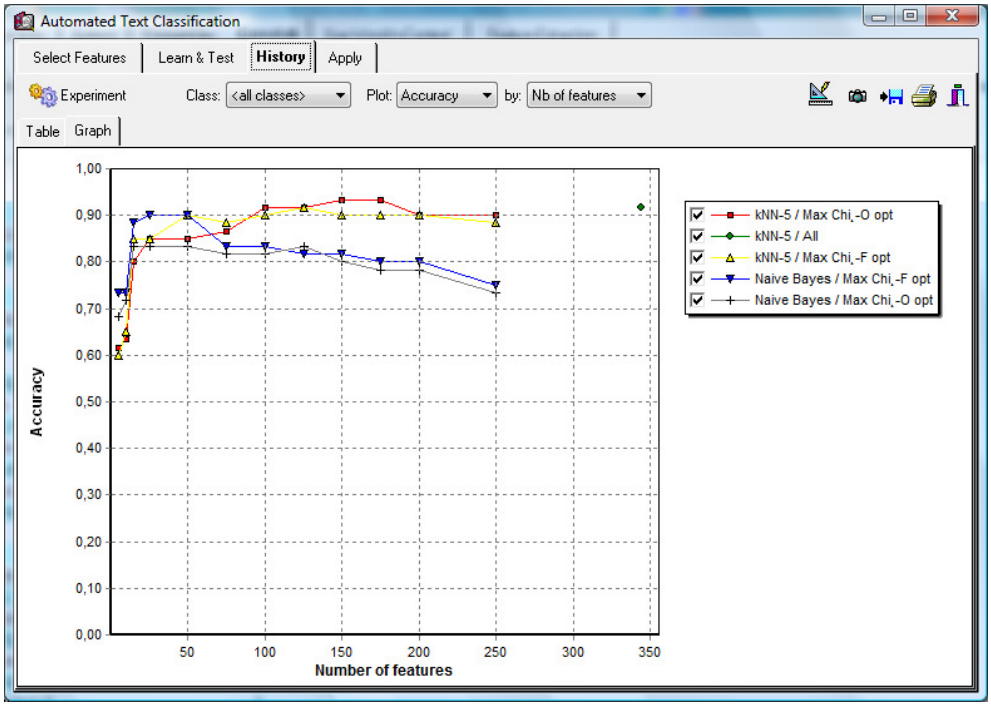


Obr. 34 - Similarity index skupin testovací sady

Pro kategorizaci byly provedeny experimenty jak pomocí metody nejbližšího souseda, tak pomocí metody Naive Bayes. Bylo nutné otestovat optimální nastavení parametrů, to bylo uskutečněno pomocí zadání jednotlivých experimentů, které je na obrázku (Obr. 35) a výsledky experimentu jsou uvedeny v grafu na obrázku (Obr. 36).

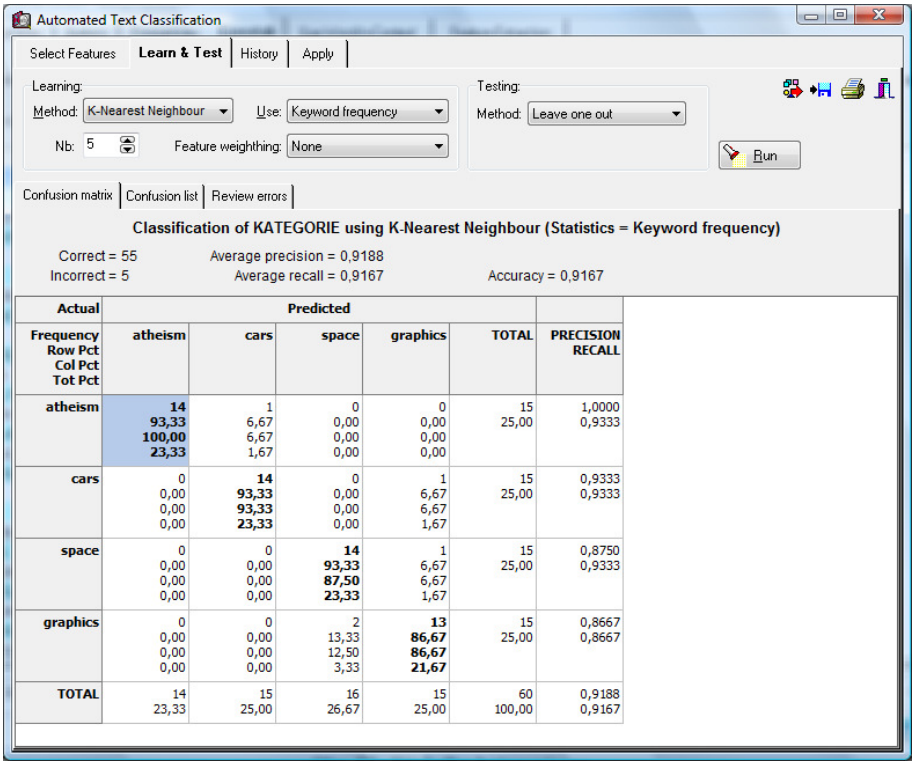


Obr. 35 - Experimenty pro zjištění vhodných parametrů



Obr. 36 - Výsledky experimentů kategorizace

Na základě výsledků testů byla zvolena metoda Nejbližšího souseda s parametry uvedenými na obrázku (Obr. 37), kde je zároveň provedeno i učení a testování kategorizace dokumentů na testovací datové sadě. Úspěšnost této metody učení byla 91,67%.



Actual	Predicted					
Frequency Row Pct Col Pct Tot Pct	atheism	cars	space	graphics	TOTAL	PRECISION RECALL
atheism	**14** **93,33** **100,00** **23,33**	1 6,67 6,67 1,67	0 0,00 0,00 0,00	0 0,00 0,00 0,00	15 25,00	1,0000 0,9333
cars	0 0,00 0,00 0,00	**14** **93,33** **93,33** **23,33**	0 0,00 0,00 0,00	1 6,67 6,67 1,67	15 25,00	0,9333 0,9333
space	0 0,00 0,00 0,00	0 0,00 0,00 0,00	**14** **93,33** **87,50** **23,33**	1 6,67 6,67 1,67	15 25,00	0,8750 0,9333
graphics	0 0,00 0,00 0,00	0 0,00 0,00 0,00	2 13,33 12,50 3,33	**13** **86,67** **86,67** **21,67**	15 25,00	0,8667 0,8667
TOTAL	14 23,33	15 25,00	16 26,67	15 25,00	60 100,00	0,9188 0,9167

Obr. 37 - Výsledky kategorizace

5.2.Testování dat v programu T-LAB

Datová sada projevů – Bush

Tato datová sada je kompilací 22 anglicky psaných projevů, vystoupení a vyjádření prezidenta Bushe z roku 2001 v rozsahu od 167 po 3250 slov v jednom textu. Některé z textů jsou v sadě vůči originálu kráceny. Soubory nejsou blíže kategorizovány, obsahují všechna hlavní témata politických projevů charakteristických pro autora jako terorismus, náboženství, finance, ekonomiku, svobodu a další. Sada je jež připravená jako korpus pro použití v programu T-Lab. Testování tvorby korpusu tedy proběhlo na opětovném vytvoření tohoto souboru ze samostatných textů.

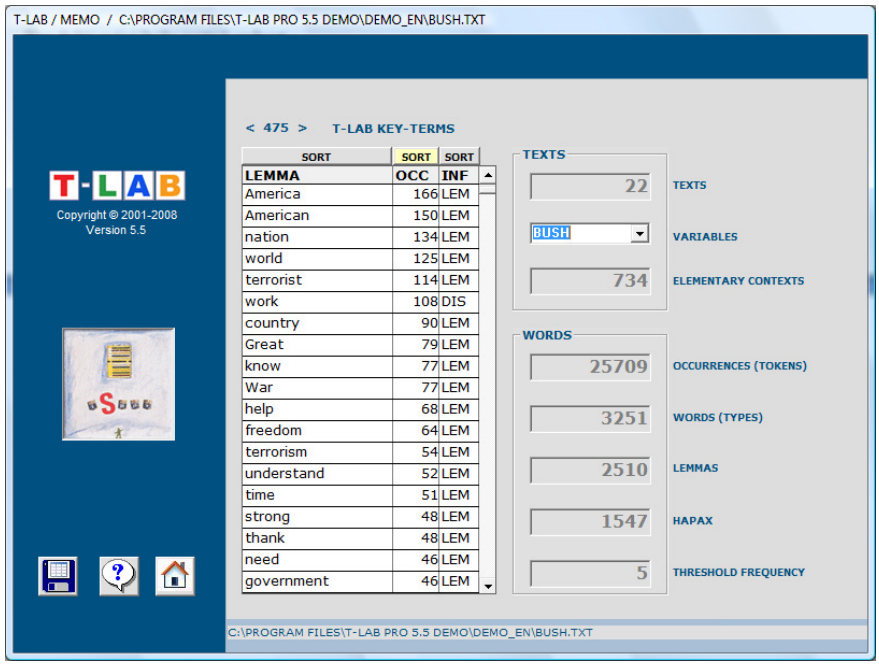
Práce s programem

Před počátkem práce s programem je doporučeno vytvoření pracovní složky, úprava souborů obsahujících text a jejich přesun do samostatné složky, dále je doporučeno rozvržení variables a categories.

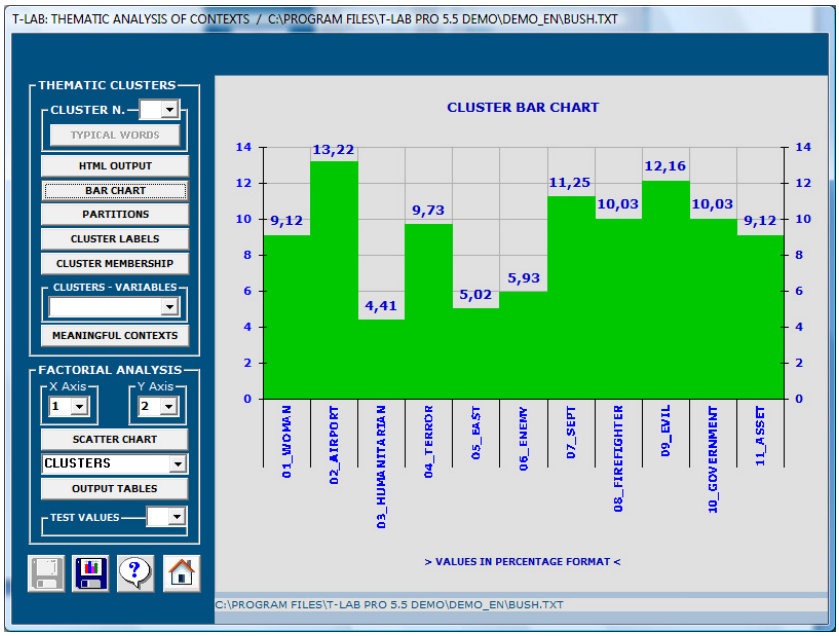
Tvorba korpusu je možná pouze ze souborů ve formátu txt neobsahujících vnitřní dělení pomocí vynechaných řádků či sekvencemi znaků, které jsou dále v korpusu používány pro dělení jednotlivých textů. Pokud jsou odstraněny tyto potíže, pak program načítá všechny soubory v uživatelem určené složce, v opačném případě je proces zpracování zastaven na prvním chybném souboru a program dále nepokračuje.

Pro testování byla připravená datová sada Bush načtena do prostředí programu s využitím anglické lemmatizace obsažené v programu a použita základní verze Stop-word check a Multi-word check. Zpracováním byl vytvořen seznam 475 lemma. Další statistické údaje zpracovaného korpusu jsou na obrázku (Obr. 38).

Tematická analýza pracuje s lemma či dokumenty. Hodnoty analýzy pro lemma ve výsledcích jsou však uváděny jak pro ně, tak pro elementární slova. K analýze byl zvolen maximální počet jedenácti shluků, který lze po výpočtu měnit (ovšem pouze nižší počet). Nabízeno je též automatické pojmenování shluků dle nejvíce frekventovaného lemma či možnost vlastních názvů na základě obsažených slov ve shluku. Výsledek tematické shlukovací analýzy elementárních slov s jedenácti automaticky pojmenovanými shluky je na obrázku (Obr. 39).

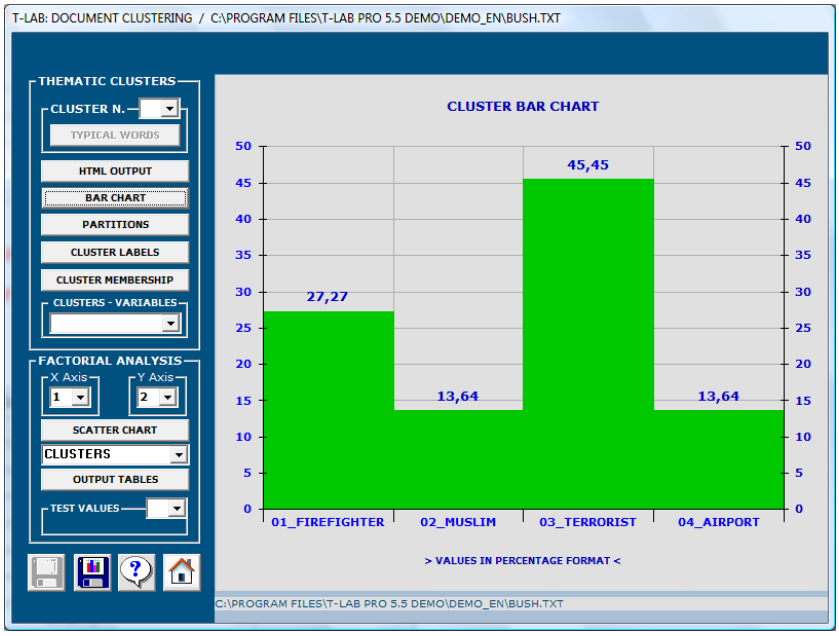


Obr. 38 - Statistický přehled datové sady Bush



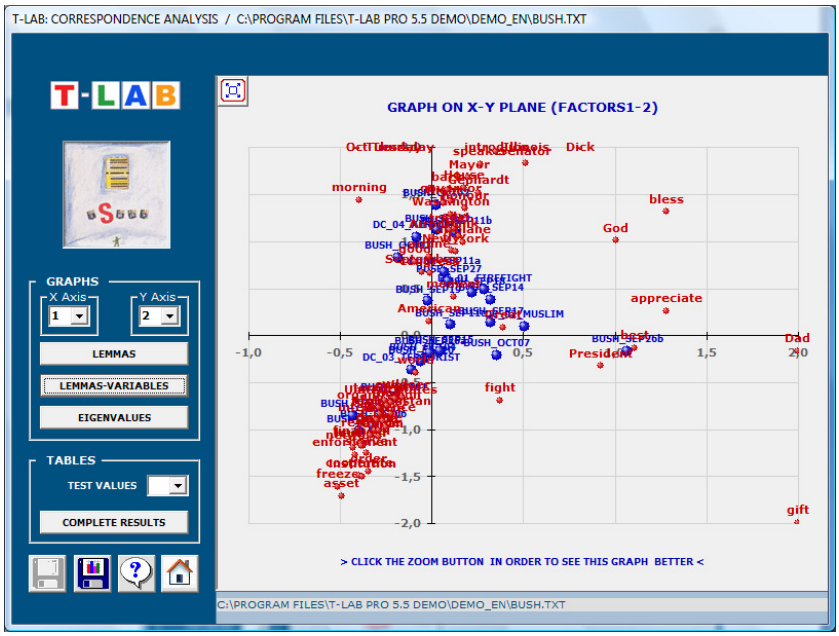
Obr. 39 - Výsledek tématické analýzy elementárních slov

Dále byla tematická analýza zaměřena na shlukování dokumentů, zde program paradoxně nabídl maximální počet shluků 10, avšak po jeho potvrzení zredukoval výčet možných shluků pouze na tři či čtyři. Výsledek tematické shlukové analýzy dokumentů je na obrázku (Obr. 40). Názvy kategorií byly opět použity z nabídky přiřazených názvů programem po provedení shlukování.



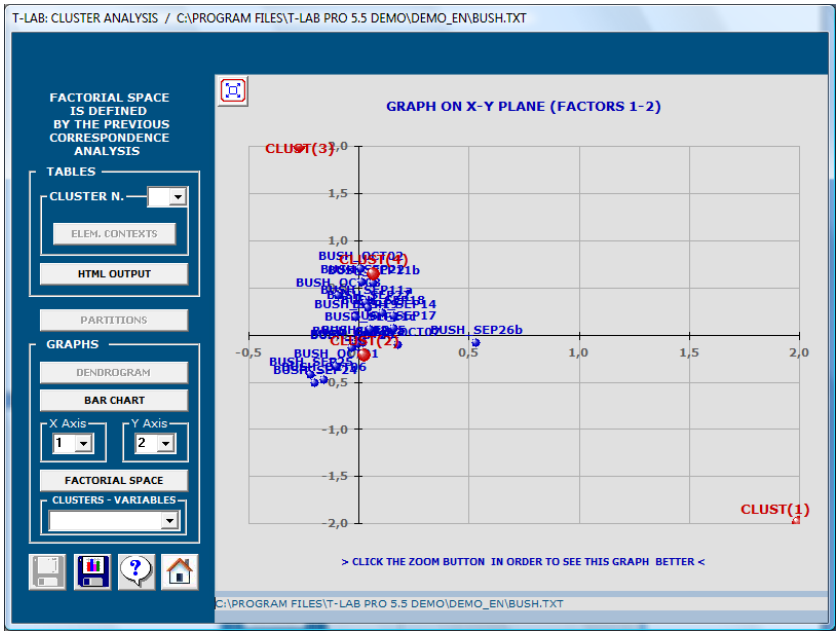
Obr. 40 - Výsledek tématické analýzy dokumentů

Ve srovnávací analýze byly otestovány vztahy mezi variables a lemmas, výsledky analýzy však ztrácely přehlednost v grafickém podání, což je zobrazeno na obrázku (Obr. 41),



Obr. 41 - Výsledek srovnávací analýzy pro lemmas a variables

Tato analýza podává rychlý přehled o základních vztazích rozložení dokumentů a jejich přibližných obsahů, dalším výstupem pak je obsáhlá tabulka se všemy vypočtenými hodnotami. Použití shlukové analýzy a testování vytvoření shluků již bylo problematické vzhledem k chybám v programu při sledu více testů. Cílem bylo vytvořit čtyři shluky pomocí metody K-means, výsledek se zobrazením shluků a variables je na obrázku (Obr. 42).



Obr. 42 - Výsledek shlukování metodou K-means

5.3.Datová sada elektronické korespondence

Pro testování funkčnosti byla vytvořena datová sada z poskytnuté firemní e-mailové korespondence z let 2004 – 2008. Tato data byla očištěna a převedena do textových souborů. Jednotlivé e-maily byly s ohledem na jejich obsah rozřazeny do skupin

- české e-maily důležité,
- české e-maily nedůležité,
- anglické e-maily důležité,
- anglické e-maily nedůležité.

Celkem bylo z korespondence vytvořeno 218 souborů spadajících do těchto čtyř skupin. Při vytváření byly splněny podmínky zachování obchodního tajemství a ochrany osobních dat. Použitá datová sada není volně šiřitelná.

5.3.1. Proč datová sada z elektronické korespondence

Na základě podnětu společnosti, která poskytnula data pro testování byla provedena analýza struktury příchozí pošty v závislosti na čase v průběhu pěti let. Tento požadavek vznikl na základě tvrzení společnosti o zahlcování elektronické pošty SPAMem (nevyžádanou poštou) natolik, že se tato komunikace díky času potřebnému k jejímu vyřízení stává neefektivní. K potvrzení či vyvrácení této domněnky byla data roztříděna do základních skupin a sledovány změny velikosti těchto skupin v čase. Výsledky jsou zobrazeny v tabulce (Tab. 5 a Tab. 6).

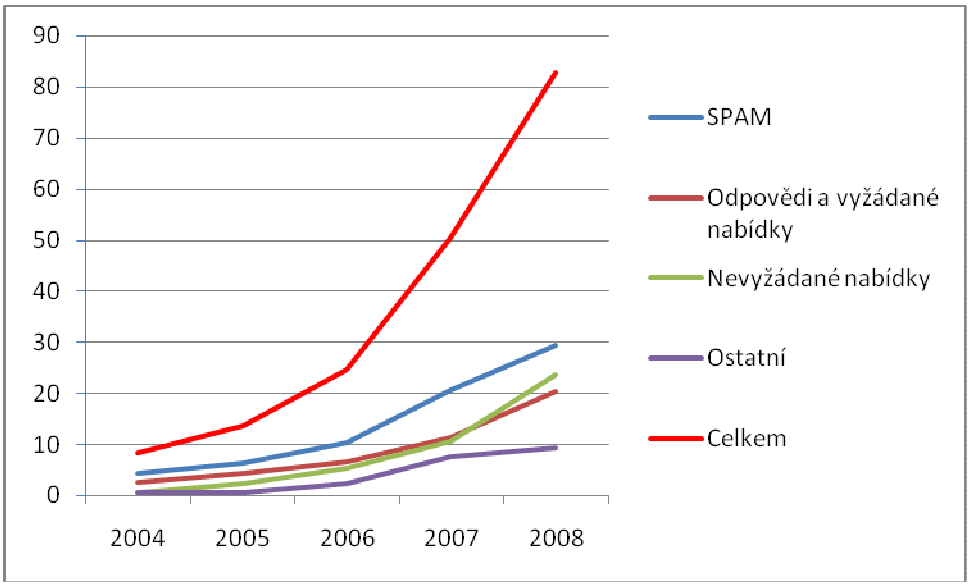
Tab. 5 - Průměrný počet přijatých skupin dat za den

	2004	2005	2006	2007	2008
SPAM	4,3	6,3	10,3	20,7	29,3
Odpovědi a vyžádané nabídky	2,7	4,3	6,7	11,3	20,3
Nevyžádané nabídky	0,7	2,3	5,3	10,7	23,7
Ostatní	0,7	0,7	2,3	7,7	9,3
Celkem	8,3	13,7	24,7	50,3	82,7

Tab. 6 - Procentuální vyjádření přijatých skupin dat za den

	2004	2005	2006	2007	2008
SPAM	52%	46%	42%	41%	35%
Odpovědi a vyžádané nabídky	32%	32%	27%	23%	25%
Nevyžádané nabídky	8%	17%	22%	21%	29%
Ostatní	8%	5%	9%	15%	11%

Z dat je patrný výrazný nárůst korespondence, v grafu na obrázku (Obr. 43) je vidět téměř exponenciální tvar křivky celkového množství přijatých dat.



Obr. 43 - Průměrný počet přijatých skupin dat za den

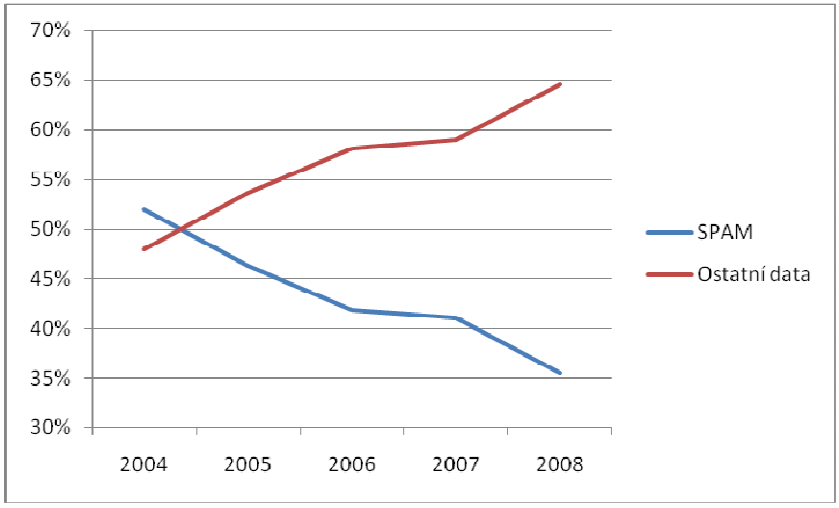
Třídění dat elektronické korespondence je v širším měřítku omezeno až do současné doby prakticky pouze na filtrování nevyžádané pošty (SPAM). Toto očišťování je již relativně rozšířené a probíhá v několika možných úrovních, které mohou působit i současně,

- třídění dat na úrovni poskytovatele internetu či doménového registrátora,
- třídění dat na úrovni lokálního serveru (toto řešení je zpravidla použito pouze ve firemním prostředí se SMTP serverem),
- třídění dat na úrovni uživatelských účtů na lokálních PC zajišťované software pro práci s elektronickou poštou,
- třídění dat na úrovni uživatelských účtů na lokálních PC zajišťované software třetích stran, doplňujících software pro práci s elektronickou poštou.

Takové rozdělení dat je zobrazeno v tabulce (Tab. 7) a grafu na obrázku (Obr. 44), data skupin jsou vyjádřena v procentech.

Tab. 7 - Procentuální vyjádření přijatých skupin dat za den

	2004	2005	2006	2007	2008
SPAM	52%	46%	42%	41%	35%
Ostatní data	48%	54%	58%	59%	65%



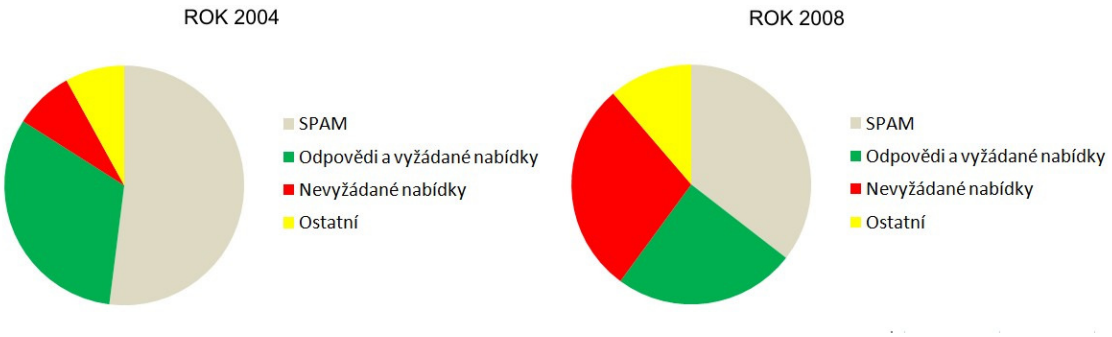
Obr. 44 - Procentuální podíl SPAMu a součtu ostatních dat

V průběhu času je patrný slábnoucí podíl SPAMu na celkovém počtu dat. Z toho však plyne i ztráta efektivnosti takového třídění dat, neboť skupina dat, kterou je nutné ručně zpracovat velmi rychle roste.

Pro data nespadající do kategorie SPAM však existují pouze velmi omezené možnosti třídění založené zpravidla pouze na základních jednoduchých nástrojích obsažených v softwaru pro zpracování elektronické komunikace, tzv. pravidel, kdy jsou data tříděna dle určitých a předem

definovaných slov obsažených v různých částech zprávy, jako odesílatel, předmět, tělo zprávy a dalších. Vytváření těchto pravidel je uživatelsky velmi zdlouhavé, s omezenou funkcionalitou například použitím či nepoužitím diakritiky a bez možnosti jakéhokoliv učení či rozsáhlejší automatizace.

Vývoj struktury příchozích dat v roce 2004 a v roce 2008 je v grafu na obrázku (Obr. 45).



Obr. 45 - Vývoj struktury dat v čase

Největší nárůst dat je v oblasti nevyžádaných nabídek. Využití těchto dat se dle odhadu pohybuje jen mezi 5 – 10%, avšak jejich ekonomický přínos je prokazatelný. Díky jejich množství jsou však tato data příjemci buď ignorována a mazána, nebo označována jako spam. Použití elektronické korespondence by se stalo efektivnější, pokud by bylo možné využít například metod text miningu a rozřazovat data do více různých kategorií.

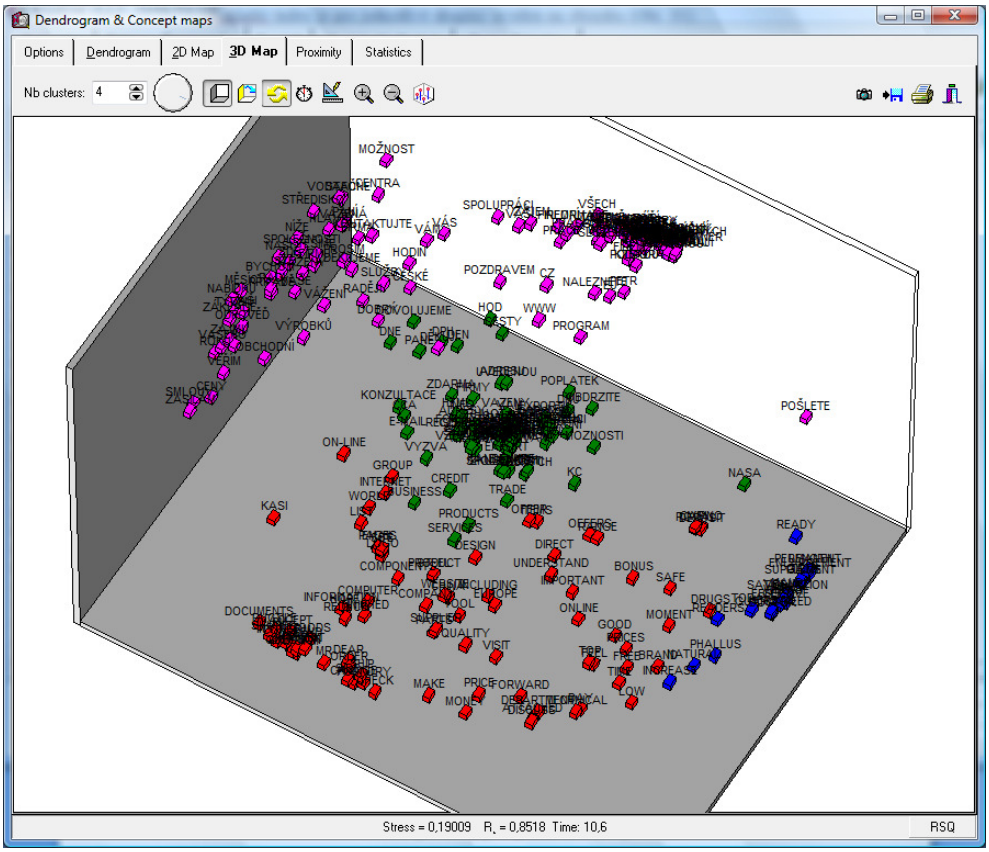
5.3.2. Kategorizace dat v programu Wordstat

Cílem práce s datovou sadou elektronické korespondence je ověření funkčnosti kategorizace pro texty odpovídající současné obchodní korespondenci ve více jazycích (v této datové sadě anglické a německé texty). Tato sada se vyznačuje oproti jiným sadám především ne příliš dlouhými texty jednotlivých souborů, což je dáno jejich původem.

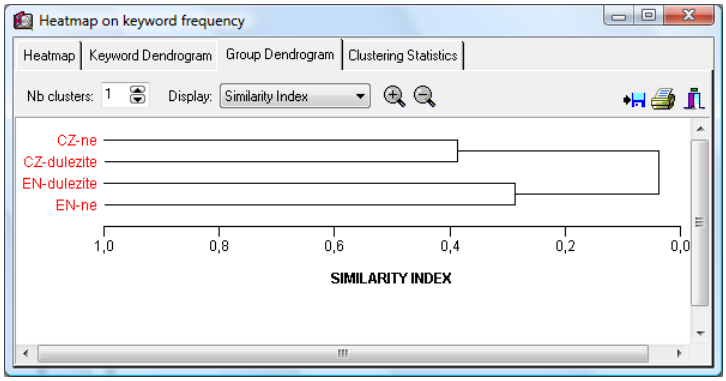
Testovací datová sada obsahuje 218 souborů ve čtyřech skupinách. Z této sady byla vytvořena trénovací sada obsahující 59 náhodně vybraných souborů, každá ze čtyřech skupin byla zastoupena minimálně 11 soubory.

K další práci s programem je určena pouze trénovací sada, testovací sada slouží pro vyhodnocení kvality kategorizace programem na sadě pro něj neznámých dat. Výběr vhodné metody lze určit pomocí výsledků experimentů na trénovací sadě.

Pro textovou analýzu byl použit slovník vyjmutých slov skládající se z českých i anglických výrazů. Termy byly experimentálně rozřazeny do čtyř shluků na obrázku (Obr. 46). Hierarchické shlukování a similarity index je na obrázku (Obr. 47).

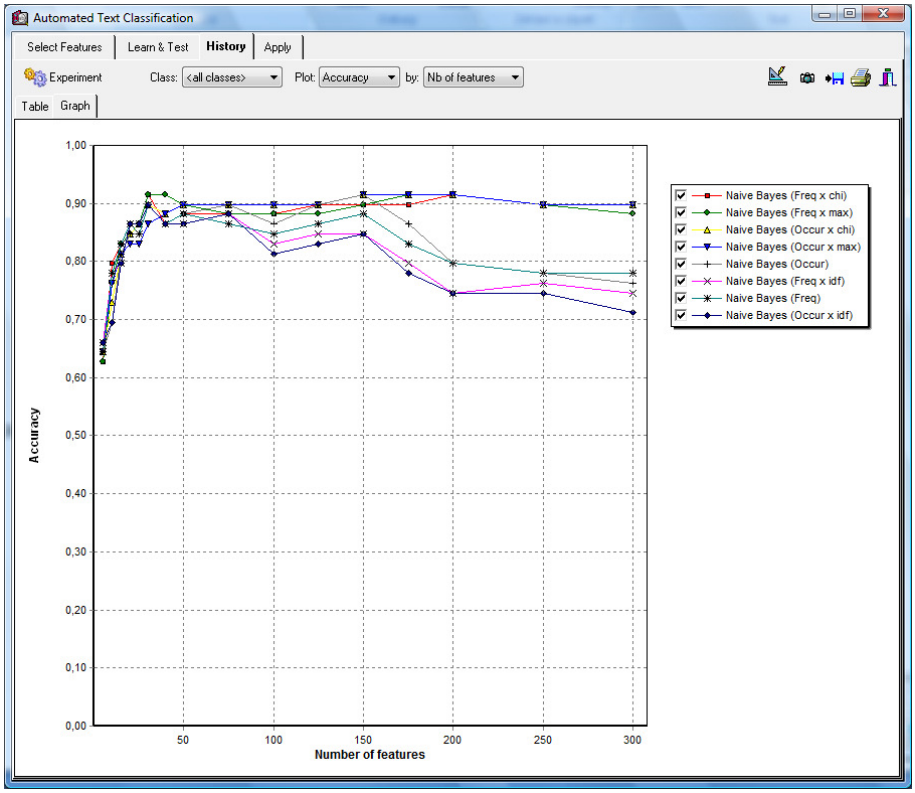


Obr. 46 - Shlukování termů

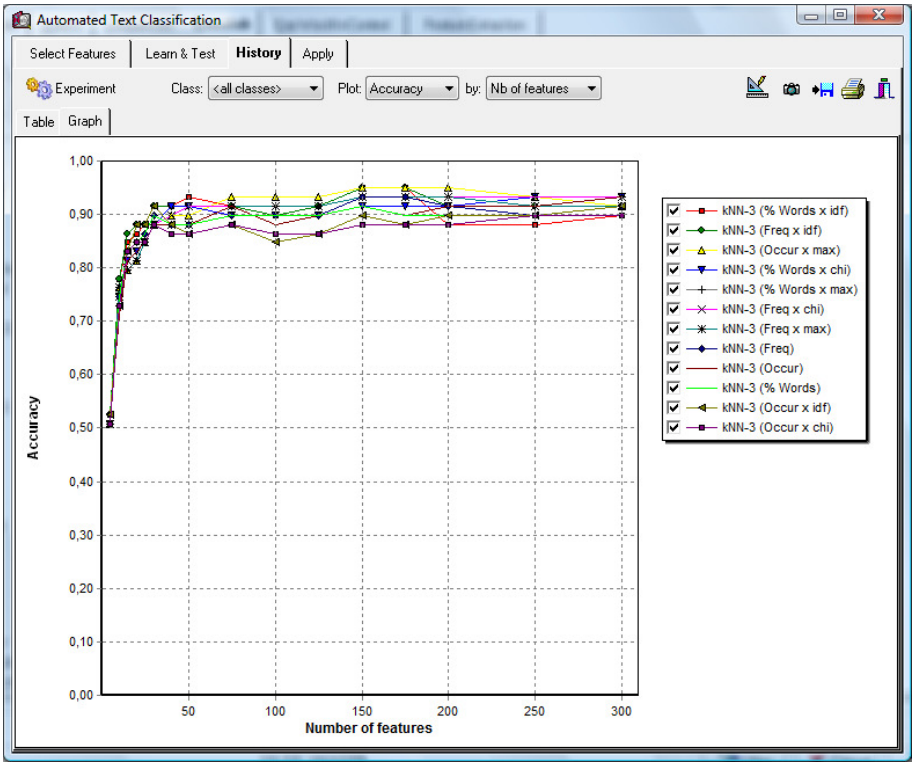


Obr. 47 - Similarity index

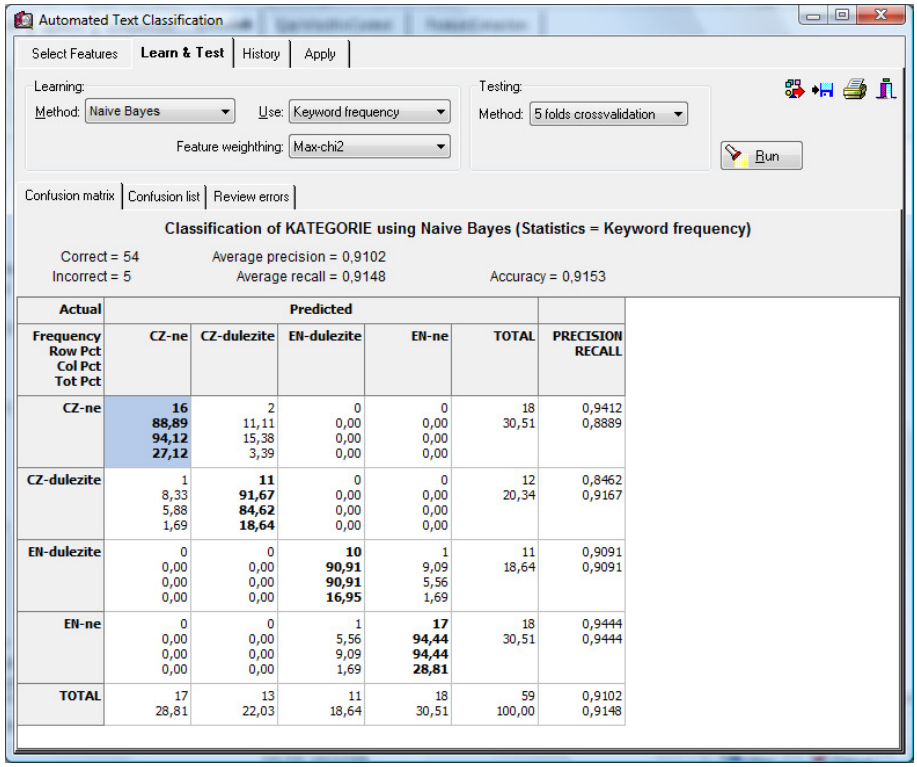
Pro kategorizaci je důležitý především výběr metody, program Wordstat umožňuje testovat různá nastavení, pro přehlednost jsou rozdělena do dvou grafů s metodou Naive Bayes na obrázku (Obr. 48) a metodu nejbližšího souseda na obrázku (Obr. 49).



Obr. 48 - Experimenty na bázi Naive Bayes



Obr. 49 - Experimenty na bázi Metoda nejbližšího souseda



Actual	Predicted					
Frequency Row Pct Col Pct Tot Pct	CZ-ne	CZ-dulezite	EN-dulezite	EN-ne	TOTAL	PRECISION RECALL
CZ-ne	**16** **88,89** **94,12** **27,12**	2 11,11 15,38 3,39	0 0,00 0,00 0,00	0 0,00 0,00 0,00	18 30,51	0,9412 0,8889
CZ-dulezite	1 8,33 5,88 1,69	**11** **91,67** **84,62** **18,64**	0 0,00 0,00 0,00	0 0,00 0,00 0,00	12 20,34	0,8462 0,9167
EN-dulezite	0 0,00 0,00 0,00	0 0,00 0,00 0,00	**10** **90,91** **90,91** **16,95**	1 9,09 5,56 1,69	11 18,64	0,9091 0,9091
EN-ne	0 0,00 0,00 0,00	0 0,00 0,00 0,00	1 5,56 9,09 1,69	**17** **94,44** **94,44** **28,81**	18 30,51	0,9444 0,9444
TOTAL	17 28,81	13 22,03	11 18,64	18 30,51	59 100,00	0,9102 0,9148

Obr. 50 - Vybraná metoda učení

Při fázi učení a testování s pomocí trénovací sady dat byla vybrána metoda Naive Bayes využívající Keyword frequency a dosažená úspěšnost kategorizace byla 91,53%. Tabulka výsledků zatřídění dat do skupin pro trénovací sadu a nastavení použité metody je uvedeno na obrázku (Obr. 50). Pro ověření funkčnosti kategorizace byla použita celá testovací sada a výsledky kategorizace jsou uvedeny v tabulce (Tab. 8). Úspěšnost kategorizace testovací sady elektronické korespondence byla vypočtena z těchto výsledků a její hodnota činí 86,7%.

Tab. 8 - Výsledky kategorizace elektronické korespondence

Kategorie	Přiřazeno do:			
	České e-maily důležité	České e-maily nedůležité	Anglické e-maily důležité	Anglické e-maily nedůležité
České e-maily důležité	21	2		
České e-maily nedůležité	14	75		
Anglické e-maily důležité			20	2
Anglické e-maily nedůležité		4	7	73

Analýza těchto výsledků naznačuje velký potenciál v možnostech zpracování tohoto typu dat a v případě použití v elektronické korespondenci by text mining dokázal především díky výsledkům ve sloupcích českých a anglických důležitých e-mailů v tabulce (Tab. 8) velmi zefektivnit činnost práce (v případě této datové sady v managerských pozicích firmy).

6. Hodnocení použitých nástrojů

6.1. Wordstat a QDA Miner

Hodnocené programy tvoří celek, který je určen pro zpracování úloh text miningu. Program QDA Miner, ve kterém uživatel připravuje textová data pro další zpracování nabízí po spuštění základní menu pro otevření či vytvoření souboru projektu, pod kterým jsou zastřešena všechna další nastavení i data. Soubor projektu se ukládá průběžně a automaticky, to má své výhody i nevýhody, nicméně při jakémkoliv nechtěném zásahu, zpravidla smazáním části dat či dokumentu, uživatelem nelze toto vrátit zpět.

Při vytváření nového souboru projektu jsou přehledně nabídnuty všechny možnosti včetně nadstandardní funkce (v této cenové kategorii) získání dat přímo z běžící databáze. Pro výběr souborů do nového projektu je schopen program zpracovat základní formáty textových souborů včetně doc, rtf či htm(l), což velmi usnadňuje práci s převáděním dokumentů před jejich použitím, naopak program neumí zpracovat soubory bez přípony i pokud obsahují známou vnitřní strukturu dat (například txt soubory bez přípony). U dokumentů s mnoha netextovými objekty není program při načítání stabilní a dochází k pádu programu. Pro tyto dokumenty však program obsahuje možnost samostatného Průvodce převodem formátů pro jejich pozdější využití. Poslední možností vytvoření nového projektu je načtení databázového souboru (offline) opět v několika možných formátech. Ovšem opět i zde může dojít k pádu programu v případě nekorektnosti dat.

Program QDA Miner má přehledné rozložení pracovních oken, kde dominuje náhled aktuálního souboru a umožňuje intuitivně provádět všechny základní úpravy textu včetně formátování pomocí standardizovaných ikon, úpravy velmi napomáhají v další orientaci s textem. Codes se vkládají do libovolně dlouhých úseků a jsou trvale provázané a editovatelné, práce s nimi je velmi jednoduchá. Proměnné spojené s text miningem jsou v levém sloupci a jejich přidávání a třídění do skupin je přehledné. Velkým nedostatkem při zpracování dat je však nemožnosti hromadného přidání dokumentů do vytvořených kategorií, neboť program neumí označit více než jeden dokument. To velmi ztěžuje práci především s většími rozsahy souborů.

Analytické nástroje programu umožňují rozsáhlou práci se slovy či spojeními a rozsáhlé možnosti vyhledávání v textech dle Codes, Při zpracování textů pak slouží jako výstupy statistické tabulky, grafy a přehledy, všechny s možností exportu dat do často používaných formátů souborů. Vlastnosti grafů je možné editovat, což pomáhá k jejich přehlednosti. Dále je možné ukládat též nastavení parametrů hledání. Program při práci se slovy i codes v projektu pracuje bezchybně a uživatelsky přívětivě.

Program Wordstat je spouštěn jako součást programu QDA Miner, avšak ve vlastním okně se záložkami nastavení a analýz. Program obsahuje též lemmatizaci, slovníky vyjmutých slov, či

kategorizaci, nikoliv však pro češtinu. Tyto funkčnosti programu jsou neoddělitelné a bohužel nelze například vynecháním lemmatizace snížit pořizovací cenu produktu. Naopak formát slovníku vyjmutých slov či kategorií je sice s vlastními příponami exc a cat, nicméně mají strukturu textového souboru a lze je tudíž velmi jednoduše vytvářet i pro český jazyk. Tyto lze navíc editovat i přímo v programu. Většina analýz je programem spouštěna již při kliknutí na příslušnou záložku, aniž by mohl uživatel změnit nastavení pro výpočet. V programu toto není jednotně upraveno, některé záložky mají tlačítko spuštění výpočtu, některé se aktualizují pouze přechodem na ně. Především při výpočtu kontingenční tabulky je tato vlastnost velmi nepříjemná, neboť základní nastavení programu počítá implicitně nejrozsáhlejší tabulku term x term, což vede na některých konfiguracích hardware k pádu programu. Všechny výstupy programu, ať již grafické či tabulkové je možné ukládat do externích souborů, to velmi usnadňuje další využití výstupů dat. Program nabízí širokou škálu formátů výstupů a ty ukládá korektně. Při vyhledávání termů v kontextu program umožňuje vstup přímo do vybrané části textu a jeho další editaci, aniž by bylo nutné opouštět Wordstat. Při práci s Wordstatem nelze bez jeho přejít zpět do programu QDA Miner a jakákoliv činnost v něm je tedy blokována. Program v několika případech nabízí jako jednu z možností výstupu zobrazení dendogramu. Práce s ním je však omezena a chybí některé funkce. V případě větších celků především možnost omezení zobrazení termů či možnost interaktivních vláken, která je možno označit, popřípadě barevně odlišit. Při více než několika desítkách termů obsažených v dendogramu tak graf ztrácí přehlednost a díky nutnosti použití rolovací lišty není možné se v zobrazení orientovat. Funkce lupy obsažená v programu se jeví jako nedostatečná, neboť s oddálením se termy stávají nečitelné. Výhodou naopak je při výběru počtu shluků barevné oddělení slov v nich a ponechání větví stromu nad tímto počtem čárkovanou čarou. Funkci Heatmap pak v programu chybí možnosti úprav zobrazení pouze shluků, ne jednotlivých termů a tato funkce je při vyšším než malém počtu termů velmi obtížně použitelná vzhledem ke způsobu zobrazení. Kategorizace je v programu zpracována přehledně a možnost hromadných experimentů s výpočty je velmi vítaná vzhledem k rychlému a efektivnímu dosáhnutí výsledků. Nastavení výpočtů jak pro učení, tak pro experimenty jsou přehledná a program umožňuje ukládat nejen výsledky, ale i nastavení testů. Bohužel není vazba mezi výsledky experimentů a nastavením učení, takže i po nalezení nejvhodnějšího postupu musí být tento opět manuálně nastaven na sousední záložce. Výstup učení a testování je v kategorizaci maticový souhrný, výčtový i tabulkový, pro experimenty jsou pak dostupné tabulky hodnot všech testů a graf s přehledem všech testů, ty lze navíc jednotlivě vypínat. Kategorizace dokumentů pak probíhá v samostatné záložce v návaznosti na učení a testování dle vybrané metody a parametrů. K dispozici je načtení jednoho dokumentu, seznamu dokumentů, aktuální databáze (z projetku) či externí databáze. Výstupy jsou pak ve formě tabulek s možností uložení do externích souborů.

Programy QDA Miner a Wordstat byly s danou testovací sadou stabilní a práci s nimi je možné i přes drobné nedostatky doporučit.

6.2.T-Lab

Program je spouštěn v jediném okně, má vlastní výraznou grafickou úpravu a pro práci s ním musí uživatel tento koncept přijmout. Program sám se snaží uživateli pomáhat v nastavení a automatizaci některých kroků. To je v případě hledání přesného cíle v některých analýzách kontraproduktivní.

Pro práci s programem je nutné vytvořit z analyzovaných dat korpus, s kterým pak program dále pracuje. Korpus v tomto programu má tvar textového souboru s definovaným obsahem. Tento soubor je sice možné ručně měnit, avšak doplňkové informace ke každému textu jsou velmi nesnadno zapsatelné manuálně. Program potřebuje mít pro svoji činnost data připravená v textových souborech formátu txt uložených v samostatné složce. U souborů s delšími texty nesmí být tyto s mezerami mezi odstavci, neboť to program vyhodnocuje jako chybná vstupní data. Nevýhodou je nemožnost této analýzy pro všechny soubory naráz před jejich další úpravou, neboť program skončí s vytvářením korpusu při nalezení prvního nevhodně formátovaného souboru. Vzhledem k další práci s programem je pak velkou nevýhodou nemožnost jakékoliv úpravy již vytvořeného korpusu. Pro každou změnu tedy musí být vygenerován nový korpus. Pro něj je možnost nechat vygenerovat automaticky Variable nebo zvolit vlastní nastavení a doplnit Variable a k nim příslušné Labels. Dalším krokem je přiřazení dokumentů tvořících korpus do příslušných kategorií. I v tomto programu lze přiřazovat pouze po jednom souboru, tudíž tento krok je problematický při velkém rozsahu dat.

Načtení korpusu nabízí základní nastavení pro další práci. Tím je lemmatizace, ovšem pouze pro naprogramované jazyky, dále co-occurrences a volby Stop-word a Multi-word seznamu, ty je možnost upravovat během zpracování korpusu. Po vytvoření korpusu a jeho načtení je programem vyžadováno potvrzení seznamu lemmas, které budou dále používány programem. Ten opět uživateli nabízí pomoc v podobě automatického seznamu, kdy při jeho výběru pouze uživateli nabídne export seznamu lemmas do externího souboru, či v druhé volbě je možné z nabídnutých lemma vyřadit uživatelem definované. Výsledky uživatelem definovaných analýz jsou zobrazeny zpravidla v grafických náhledech a tabulkách, je možné exportovat do formátu xls. V závislostní analýze program pracuje spolehlivě, výpočty je možné upravovat pouze po jejich prvním spuštění.

Program je však značně problematický ve srovnávací analýze. Shluková analýza vede nezřídka k pádu programu, vyskytnou-li se chybové hlášky, pak jsou tyto v italštině a program se chová ne příliš korektně. Pro možnost zpracování různých typů analýz bylo zapotřebí několikrát manuálně vymazat předchozí pomocná data vytvořená programem. Vzhledem k použití

demonstračních dat dodaných výrobcem však toto nepůsobí příliš dobrým dojmem. Užití Kohonenových map se po volbě jednoho ze sedmi přednastavených rozměrů omezuje pouze na automatické vygenerování html tabulky.

Program T-Lab je tedy určen pro zpracování textů malého objemu a analýzy týkající se především termů v nich použitých a práce s nimi. Funkce jako shluková analýza nejsou silnou stránkou tohoto programu. Nástroje pro klasifikaci tento program nepodporuje.

7. Závěr

Tato práce měla kromě jiného komparaci dvou nástrojů Wordstat od firmy Provalis Research a T-LAB od stejnojmenné firmy určených pro text mining a patřících přibližně do stejné cenové kategorie.

Popis těchto nástrojů jejich výrobci pak uváděl tyto nástrojé jako velmi podobné. Nicméně na trhu neexistují dva stejné softwarové nástroje a po úvodním seznámení se s dvěma vybranými pro další analýzu byl u programu T-Lab zjištěn výrazný posun funkcionality do oblasti sémantické, program se zaměřuje především na zpracování slov, slovních spojení a práce s textovými dokumenty není hlavní oblastí tohoto programu. Díky tomuto zjištění bylo od zakoupení akademické licence ustoupeno a tento program byl hodnocen pouze v demoverzi s příslušným omezením spočívajícím v nemožnosti jej otestovat standardní datovou sadou. Testování funkčnosti tak proběhlo na výrobcem dodané sadě, kde byl její obsah analyzován a popsán. Tento program navíc neobsahuje kategorizaci, což jej činí nevhodným pro možnost zpracování vlastních dokumentů dle definovaných tříd.

Program Wordstat byl testován společně s programem QDA Miner od stejné společnosti, neboť jemu samotnému chybí možnost importu a preprocessingu dat. Oba tyto programy byly testovány s využitím akademické licence zakoupené fakultou ekonomicko-správní Univerzity Pardubice. Tato licence umožňuje používat oba programy v rámci licenční smlouvy bez dalšího omezení. Program byl testován na výběru ze sady 20 Newsgroups, což je světově uznávaná sada anglických textů určená pro práci s text miningovými nástroji. Použití této sady v programu Wordstat přineslo uspokojivé výsledky především v kategorizaci, při shlukování dokumentů se projevila skladba textů datové sady, kde mnoho dokumentů má shodnou hlavičku, a shlukování těchto dokumentů pak proběhlo optimálně až při vyšším počtu shluků.

Naopak použití vlastní datové sady s česko-anglickými texty pocházejícími z obchodní korespondence přineslo velmi dobré výsledky v kategorizaci pomocí programu Wordstat. Tento program dokázal korektně zpracovávat i českou část datové sady a jeho schopnost kategorizace se potvrdila i na nezávislých datech.

Prověřením těchto dvou text miningových programů byly splněny cíle stanovené v zadání práce a především u programu Wordstat tak byla potvrzena vhodnost použití text miningu při zpracování dokumentů. Text mining by tak v budoucnu mohl být velmi přínosný i pro každodenní použití na základě jeho schopnosti pokročilé analýzy dat a v oblasti komunikace by mohl velmi zefektivnit práci s daty.

Použitá literatura

[1] AAS, K., EIKVIL., L. *Text categorisation*: A surfy. Oslo, Norway: Norwegian Computing Center, 1999. 37 s.

[2] ALDENDERFER, M., BLASHFIELD, R. *Cluster Analysis*. 1st edition. [s.1.] : Sage Publications, Newbury Park, Cal, 1984. 88 s. ISBN 978-0803923768.

[3] BERKA, Petr. *Dobývání znalostí z databází.* Vyd. 1. Praha : Academia, 2003. 366 s. ISBN 80-200-1062-9.

[4] CIMIANO, P., HOTHO, A., STAAB, S. *Comparing conceptual, divisive and agglomerative clustering for leasing taxonomies from text.* In Proceedings of the European Conference on Artificial Intelligence (ECAI), 2004.

[5] *Členění data miningových úloh* [online]. 2002 [cit. 2008-08-18]. Dostupný z WWW: <http://datamining.xf.cz/view.php?cisloclanku=2002102801>.

[6] GRIFFITHS, A. et al. Hierarchic Agglomerative Clustering Methods for Automatic Document Classification, The Journal of Documentation, vol. 40, No. 3, 1984. s. 175-205.

[7] *Hodnocení kvality sumarizátorů textů* [online]. Západočeská Univerzita v Plzni, 2004 [cit. 2008-08-23]. Dostupný z WWW: <http://textmining.zcu.cz/publications/znalosti.pdf>.

[8] *Intersteno* [online]. 2007 [cit. 2008-08-23]. Dostupný z WWW: <http://www.intersteno.cz/>.

[9] KASKI, S., *Data exploration using self-organizing maps*. Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82, Espoo 1997, 57 pp. ISBN 952-5148-13-0.

[10] *LISp-miner* [online]. VŠE Praha. 2008 [cit. 2008-08-23]. Dostupný z WWW: <http://lispminer.vse.cz/index.html>.

[11] MacQUEEN, J. B. Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symphosium on Mathematical Statistics and Probability, Berkeley*: University of California Press, vol. 1, 1967. s. 281-297.

[12] NAHM, Un Yong, MOONEY, RAYMOND J. *Text mining with information extraction*. [s.l.] : The University of Texas at Austin, 2004. 218 s. ISBN 0-496-01283-5.

[13] NOVÁČEK, Libor. Google pracuje lépe s češtinou, zavedl stemming.. *Lupa.cz* [online]. 2007 [cit. 2008-08-18].

[14] PETR, Pavel. *Data Mining : Díl 1.*. 1. vyd. Pardubice : Univerzita Pardubice, 2006. 144 s. ISBN 80-7194-886-1.

[15] *SAS Institute Inc.* [online]. 2008 [cit. 2008-08-23]. Dostupný z WWW: <http://www.sas.com/>.

[16] SKLENÁK, Vilém a kol. *Data, informace, znalosti a Internet*. Vyd. 1. V Praze : C.H. Beck, 2001. xvii, 507 s. ISBN 80-7179-409-0.

[17] *T-LAB User\'s Manual* [online]. 2008 [cit. 2008-08-15]. Dostupný z WWW: <http://www.mytlab.com/Manual_en.zip>.

[18] TAN, Ah-Hwee: *Text mining: the state of the art and the challenges.* In: Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases-PAKDD'99, Beijing, April (1999).

[19] *Text Retrieval and Mining* [online]. 2004 [cit. 2008-08-19]. Dostupný z WWW: http://www.stanford.edu/class/cs276a/

[20] *TextSTAT* [online]. 2008 [cit. 2008-08-23]. Dostupný z WWW: <http://www.niederlandistik.fu-berlin.de/textstat/>.

[21] WEISS, Sholom M., et al. *Predictive Methods for Analyzing Unstructured Information*: Springer, 2005. 236 s. ISBN 978-0-387-95433-2.

[22] *Wordstat manual* [online]. c2005 [cit. 2008-08-15]. Dostupný z WWW: <http://www.provalisresearch.com/Download/Manuals.html>.