



Textová data a dobývání znalostí

Obsah prezentace



- ❖ Co je to **dobývání znalostí z textových dat (TM: *text data mining*)** a proč je užitečné?
- ❖ Hlavní cíle a úlohy TM.
- ❖ Co je specifické pro práci s textovými daty?
- ❖ Textové dokumenty, jejich soubory a **volba vhodné reprezentace.**
- ❖ Postupy pro **provádění TM** v základních úlohách DM:
 - ❖ Klasifikace
 - ❖ Shlukování
 - ❖ ...
- ❖ Zajímavé aplikace TM

Co je to text mining (TM)?



“The objective of Text Mining is to exploit information contained in textual documents in various ways, including ...**discovery of patterns and trends in data, of associations among entities, of predictive rules**, etc.” (Grobelnik et al., 2001)

“Another way to view text data mining is as a process of exploratory data analysis that leads to **heretofore unknown information**, or to answers for questions for which the answer is not currently known.” (Hearst, 1999)

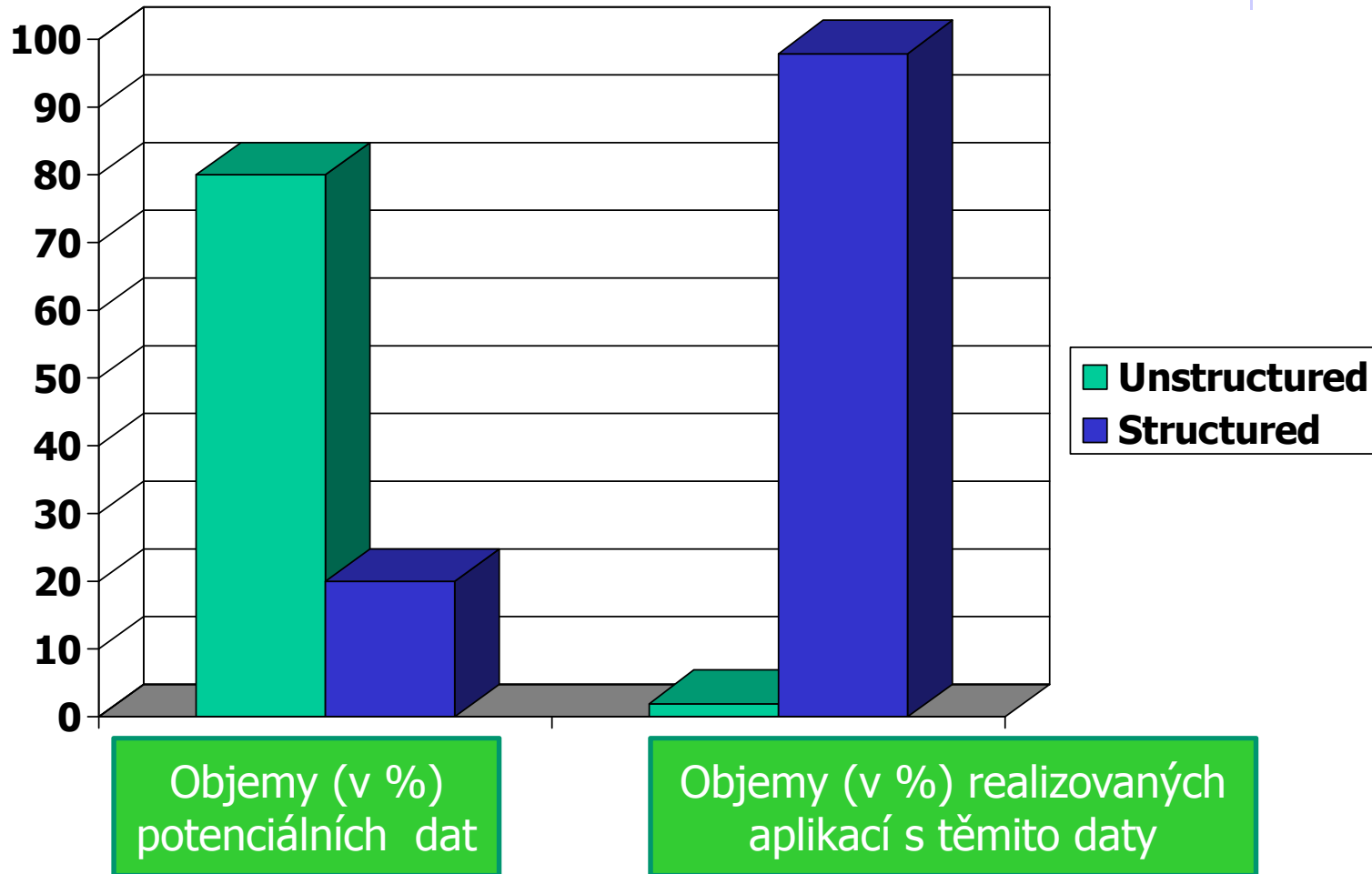
Kde je schováno „know-how“ různých firem či institucí?



Nejčastěji právě v psaných textech a dalších zdrojích, které nejsou vhodné pro použití metod standardního DM

- ❖ e-mailová komunikace
- ❖ Stížnosti zákazníků
- ❖ Žádosti o plnění při pojistných událostech
- ❖ Smlouvy
- ❖ Novinové články
- ❖ Technické dokumenty
- ❖ Webové stránky
- ❖ Přepisy telefon. rozhovorů se zákazníky
- ❖ Patentové přihlášky
- ❖ ...
- ❖ Vědecké články

— Příležitost pro DM, kterou představuje TM...



Výzvy, které představuje TM



- ❖ Základní data mají formu “běžného textového souboru (*free text*)”, což se velmi liší od tabulek, které jsou standardním vstupem pro základní algoritmy DM. **Problém nestrukturovaných dat.**
- ❖ Skutečný obsah je skryt kdesi uvnitř textových dokumentů: např. negace může být vyjádřena mnoha možnými způsoby. Při řešení řady úloh se neobejdeme bez **syntaktické analýzy!**
- ❖ V přirozeném jazyku se běžně setkáváme s **víceznačností** (*ambiguity*), kterou neřeší syntaktická analýza – nutné použít **sémantickou analýzu**. Příklady víceznačnosti na mnoha úrovních :
 - lexikální – víceznačnost slov (*diamond, Ford, ..*)
 - syntaktická – “*Read about the problem in newspapers !*”
 - referenční - “*The boy was passing a man with a dog. He stroked him.*”
Koho pohladil - psa nebo pána? Anaphora.

■ ...

Dobývání znalostí z textu (TM)



je problematika na pomezí více oborů:

- ❖ obecné dobývání znalostí z dat
- ❖ počítačová lingvistika
- ❖ vyhledávání informací

Obor řešící níže uvedené úlohy:

	Hledání častých vzorů nebo schémat	Hledání skrytých “pokladů”	
		Nových	Ne-nových
Netextová data	DM obecně	Explorat. analýza dat	Databáze dotazování
Textová data	Počítačová lingvistika		Vyhled. informací

Úlohy řešené TM a jejich příklady



❖ Exploratorní analýza dat

- ◆ Použití medicínských článků jako zdroj inspirace při návrhu hypotéz o příčinách onemocnění (Swanson and Smalheiser, 1997).

❖ Extrakce informací

- ◆ (Polo)automatická transformace textu do tvaru strukturované báze znalostí, na kterou už lze použít běžné nástroje DM.
 - ❖ Bootstrapping methods (Riloff and Jones, 1999).

❖ **Klasifikace textů**

- ◆ Používá se jako užitečný mezikrok při extrakci informací
 - ❖ Bootstrapping method using EM (Nigam et al., 2000).

Problémy explorace dat



- ❖ Jak lze nalézt platné vazby, aniž bychom se utopili v množství přípustných variant? Kombinatorická exploze!
 - ◆ Potřebujeme zlepšit modely pro popis lexikálních vztahů a pro formalizaci sémantických omezujících podmínek (velmi náročný úkol)
- ❖ V jaké formě by měla být používaná informace nabídnuta lidskému uživateli tak, aby se v ní mohl co nejnáze orientovat?

Extrakce informací (IE)



Nalezení potřebné problémově závislé (*domain-specific*) informace z podkladů v přirozeném jazyce s využitím slovníku vzorů a sémantického slovníku

- ❖ **Slovník vzorů** pro extrakci významných údajů (např. "cestoval do <x>" nebo "přeběhl <jméno_soupeře>"), kde výraz v lomených závorkách představuje hledanou informaci.

Způsoby konstrukce slovníku vzorů:

- ❖ **Ručně**
- ❖ automaticky s využitím metod strojového učení na ručně anotovaná data
- ❖ **Sémantický slovník** (slovník slov, kde u každého slova jsou uvedeny všechny jeho sémantické významy)
 - ❖ Konstruován **obvykle ručně!**

Otevřené problémy extrakce informací



- ❖ Jak zjednodušit použití metod strojového učení, když tyto metody obvykle předpokládají, že pracují v režimu „s učitelem“ (tj. potřebují klasifikované příklady)
- ❖ Jak nahradit ruční anotování textových dat, které je časově velmi náročné a drahé?
- ❖ Je třeba hledat a vyvíjet **nové algoritmy** vhodné
 - ❖ pro režim bez učitele!
 - ❖ kterým stačí menší soubor příkladů.
- ❖

Klasifikace textů (TC)



- ❖ Zařad' dokument do jedné z několika tříd, jejichž výběr je předem dán (např. beletrie, novinový článek, odborný článek)
 - ◆ *"Toto nevede k odhalení nových informací..."* (Hearst, 1999).
 - ◆ Velmi užitečné v praktických úlohách:
 - ❖ Seskupení dokumentů podle oblastí, kterých se týkají
 - ❖ Identifikace žánrové preference čtenářů
 - ❖ Třídění mailů
 - ❖ ...

Otevřené problémy při klasifikaci



- ❖ Úloha velmi příbuzná IE – v obou případech je třeba velké množství klasifikovaných příkladů
 - ◆ S použitím 1000 novinových článků z UseNetu, které byly ručně anotované, se systém naučil pravidla pro klasifikaci, která s úspěšností asi 50% nalézala články, které uživatel považoval za zajímavé.
 - ◆ Část práce se realizuje předem – nezdržuje reakci a dotaz uživatele.
- ❖ Hledání nových zdrojů informací, které umožní **snížit podíl ruční práce** při tvorbě klasifikovaných příkladů.

† Dokumenty a jejich soubory (*collection*)



- ❖ **Soubor dokumentů** = skupina dokumentů v přirozeném jazyce, která má buď *statický* nebo *dynamický* charakter (vzhledem ke změnám v čase).
 - ◆ e.g. **PubMed** "on-line repository of abstracts for > 13 million papers (about 40.000 new abstracts are added each month)"
- ❖ **Dokument** = samostatná jednotka vyskytující se v souboru dokumentů.
 - ◆ e.g. a business report, research paper, news story, e-mail, ..
- ❖ Tentýž dokument se může vyskytovat v několika různých souborech (např. články o „e-health“ jsou součástí jak souboru dokumentů se zdravotnickou tematikou, tak souboru věnovaném ICT)

Reprezentace dokumentů a efektivní vyhledávání



- ❖ **Jak je možné nalézt konkrétní článek** v databázi PubMed? *Klíčová slova* nejsou dostatečně diskriminativní:
 - ◆ Dotaz „*protein or gene*” → nalezeno víc než 3 miliony dokumentů
 - ◆ I velmi specifický termín *epidermal growth factor receptor* → 10.000 dok.!
- ❖ Jaké **příznaky** pro reprezentaci dokumentů je nejlépe použít, chceme-li pak efektivně použít metody DM ?
- ❖ V současné době se doporučuje pracovat s **příznaky, které odpovídají významu či obsahu textu!** Ovšem zpracování přirozeného jazyka (*natural language processing* NLP) a porozumění významu komplexního textu se stále považuje za jeden z nejnáročnějších cílů AI (Turingův test). V čem jsou problémy? *Víceznačnost, negace, interpretace zájmen, ...*

Kandidáti na příznaky



❖ **Písmena.** Jednotlivé symboly umožňují **odpovídat na některé morfologické otázky** (např. pro predikci následujícího textu)

- ◆ bigramy (trigramy) reprezentují posloupnosti 2 (či 3) symbolů

❖ **Slova** - často se setkáváme s pojmem **tokeny na úrovni slov** (*word-level tokens*)

- ◆ Takové tokeny mohou být ještě i ohodnocené (např. údajem o své gramatické kategorii – podstatné jméno, sloveso,..)
- ◆ reprez. pomocí **ranečku slov** (*bag-of-words*) ignoruje pořadí tokenů
- ◆ kmeny slov (word stem)

❖ **Termy** mohou reprezentovat samostatná slova nebo skupiny slov, např. "*White house*"

❖ **Koncepty** reprezentují větší celky potřebné pro řešení problémů se synonymy ...

- ◆ Např. podstatné jméno "*car*" může v textu odpovídat následujícím výrazům: *automobile, truck, Lightning McQueen*

Problémy s vysokou dimenzí



- ❖ Uvedené typy příznaků umožňují reprezentovat každý dokument jako **vektor slov** (či termů)
 - ◆ Každá složka vektoru odpovídá nějakému "kvantitativnímu údaji", který se vztahuje na příslušné slovo nebo term.

❖ Velmi užitečné zjednodušení, které má však řadu problémů:

- ◆ Použití slov či termů vede k zavedení **velkého množství příznaků**:
 - ❖ Small Reuters je soubor dokumentů, který obsahuje 15.000 článků. Vyskytuje se v něm 25.000 netriviálních příznaků (jedná se o kmeny slov)
 - ❖ Většina algoritmů nezvládá práci s daty o tak velkém množství příznaků → neobejdeme se bez použití technik na redukci příznaků!
- ◆ **Řídké příznaky**: Každý samostatný dokument obsahuje jen velmi omezenou část ze všech zavedených příznaků.

Nejobvyklejší úlohy TM



- ❖ **Extrakce příznaků** (*feature extraction*) hledá v dokumentu vhodnou množinu slov, která mohou dokument co nejlépe reprezentovat.
- ❖ **Klasifikace dokumentů** (*categorization*) – např. oblast žánr, předmětná oblast, ...
- ❖ **Vyhledávání informací** (*information retrieval*) - např. webové vyhledávače
- ❖ **Shlukování** či hledání vhodné organizace pro soubor dokumentů
- ❖ **Extrakce informací** (*information extraction*) – např. konkrétní data o aktuálních naměřených hodnotách v chorobopise psaném rukou
- ❖ ...

Extrakce příznaků: úloha



počet
slov v
textu

While more and more textual information is available on-line, effective retrieval is difficult without good indexing of text content.

20

Vynechání „vazebních“ slov

While-more-and-textual-information-available-online-effective-retrieval-difficult-without-good-indexing-text-content

15

Extrakce
příznaků

Text-information-online-retrieval-index

5

2 1 1 1 1

Frekvence vybraných výrazů ve výchozím textu

Kroky používané při postupném zjednodušování reprezentace dokumentu



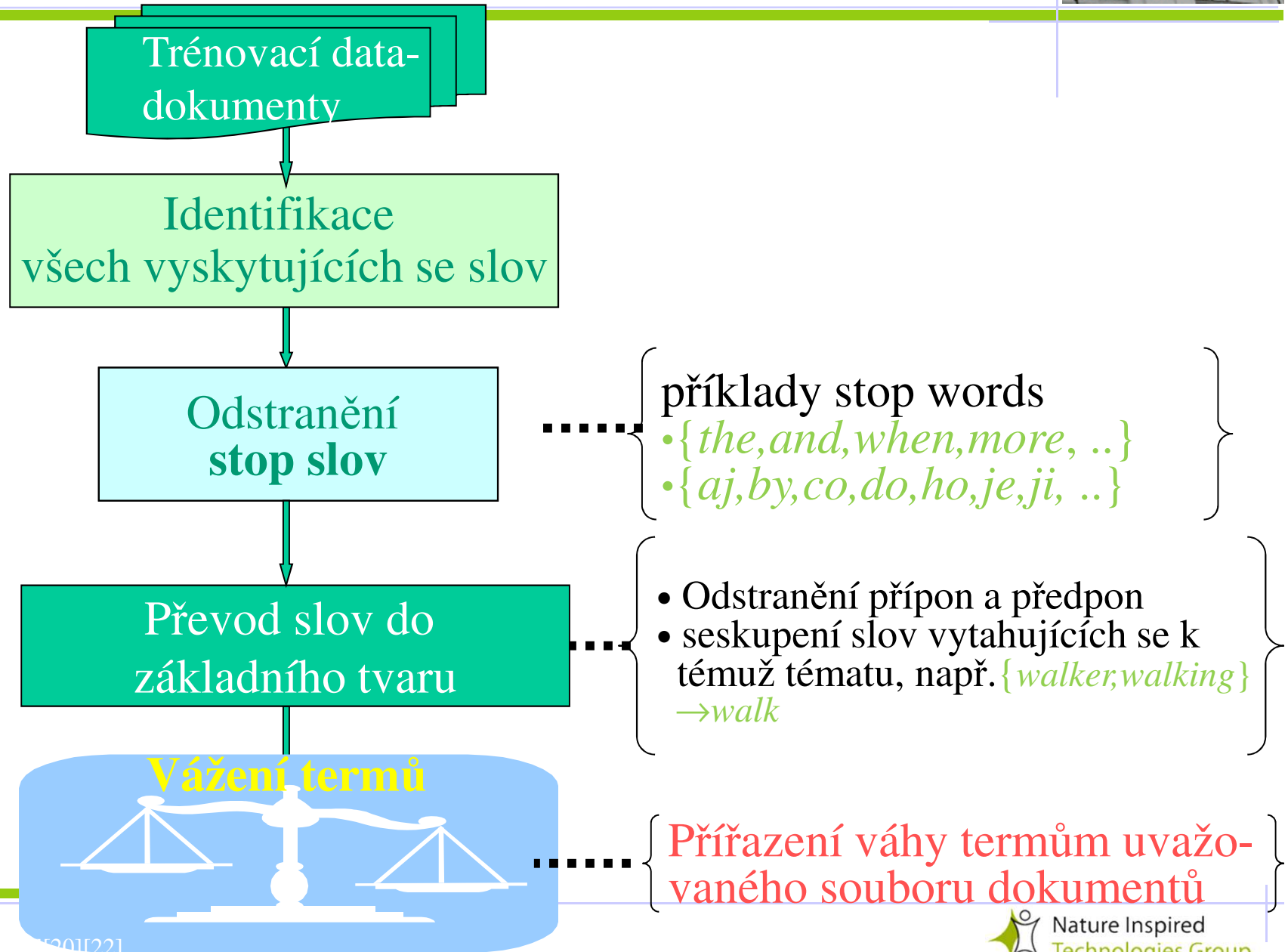
Vynechání nepodstatných slov

- ◆ Ne všechna slova jsou stejně důležitá, např. *the, is, and, to, he, she, it* nenesou obvykle nijak zásadní informaci (i když jsou situace, kdy mohou způsobit změnu interpretace)
- ◆ Můžeme vybrat různé úrovně, jak přísně budou vybírána slova k vynechání

obvykle postupuje takto

- ◆ Nejdřív jsou vynechána slova, která nenesou samy o sobě žádný význam a jsou označovaná jako **stop slova** (stopwords), např. spojky, spony (*je, ..*), ...
- ◆ Zbylým slovům je **přiřazena váha** podle jejich počtu výskytů (**Term Frequency - TF**) a podle toho, jak moc jejich přítomnost mění obsah (význam) dokumentu. Pro **indexování** se vybírají slova s nejvyšší vahou.
- ◆ Důležitá slova by měla získat vyšší váhu, méně důležitá naopak nižší. Oblíbenou měrou pro toto hodnocení je **TF_IDF**. Tato míra kombinuje informaci o frekvenci výskytu slova s údajem **IDF (Inverse Document Frequency)** , viz další strana. Slova s nejvyšší vahou jsou použita pro indexování

Extrakce příznaků (1): Indexování



Ext.příz.(2): Indexování s TF vážením slov



- Reprezentace dokumentu jako vektoru ve vektorovém prostoru

$$d=(w_1, w_2, \dots, w_t) \in \mathbf{R}^t$$

kde w_i je váha (počet výskytů) i - tého termu v dokumentu d .

- **míra tf** – vážení podle frekv. termů (*Term Freq. Weighting*)

$$w_{ij} = \text{Freq}_{ij}$$

$\text{Freq}_{ij} :=$ počet výskytů j tého termu v dokumentu d_i .

× **Nevýhody?** Hodnota TF nebere v úvahu důležitost slov a výsledkem je podobný obraz pro velmi rozlišné dokumenty:

$D1$ **A**BRT**S**A**Q**W
 AX**A**O

$D2$ **R**T**A**BB**A**X**A**
 QS**A**K

	A	B	K	O	Q	R	S	T	W	X
D1	4	1	0	1	1	1	1	1	1	1
D2	4	2	1	0	1	1	1	1	0	1

Ext.příznaků (3) s vážením podle IDF



tf versus idf : vážení podle Inverse Document Frequency

$$w_{ij} = \mathit{Freq}_{ij} * \log(N / \mathit{DocFreq}_j)$$

N := počet všech dokumentů v souboru použ. pro trénování.

$\mathit{DocFreq}_j$:= počet dokumentů, kde se vyskytuje j -tý term.

➔ **Výhoda:** Postup zohledňuje význam slova z hlediska možnosti rozlišit zpracovávané dokumenty.

Předpoklad: termy s nízkým **DocFreq** pro daný soubor (K, O, W) přispívají k rozlišení dokumentů v souboru daleko víc než termy velmi časté (tj. s vysokým **DocFreq**)

ABRTSAQWA
XAO

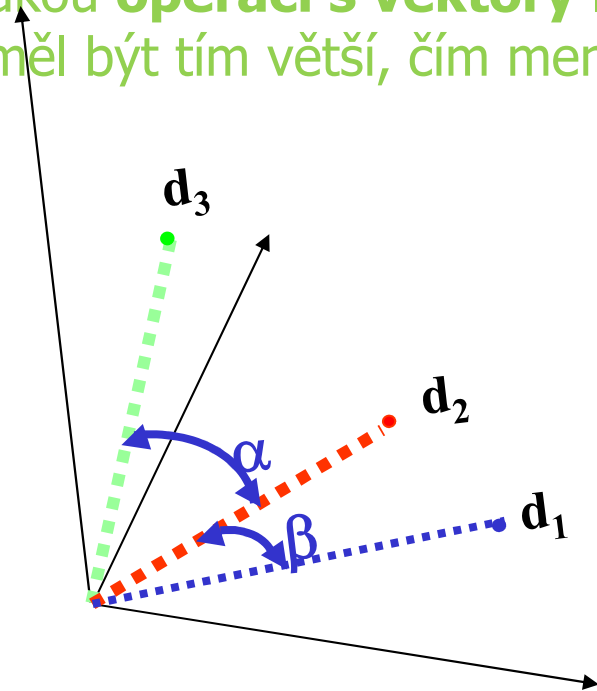
RTABBAXA
QSAK

	A	B	K	O	Q	R	S	T	W	X
D1	0	0	0	0.3	0	0	0	0	0.3	0
D2	0	0	0.3	0	0	0	0	0	0	0

Jak se definuje podobnost pro vektory dokumentů?



- ❖ docA "Java Programming Language" $\langle 0, 0, 1, 1, 1, 0, 0, 0 \rangle$
- ❖ docB "Barcelona beats Real Madrid" $\langle 1, 1, 0, 0, 0, 1, 1, 0 \rangle$
- ❖ docC "Barcelona beats Slavia" $\langle 1, 1, 0, 0, 0, 0, 0, 1 \rangle$
- ❖ Jakou **operaci s vektory** lze použít pro definici podobnosti? Výsledek by měl být tím větší, čím menší úhel oba vektory svírají: trigon. fce?

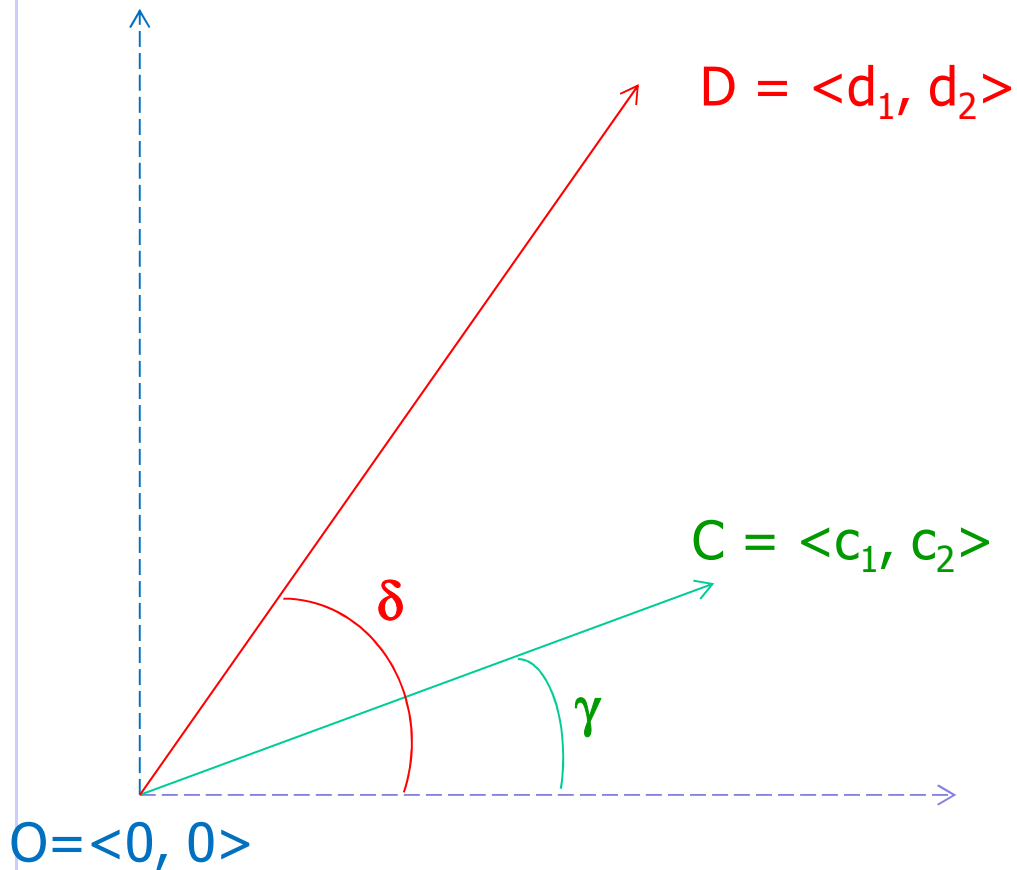


Pro $\alpha > \beta$ platí, že $\cos(\alpha) < \cos(\beta)$

Vektor d_2 je blíže k d_1 než k d_3

Podobnost definovaná kosinem
(*cosine-based similarity model*)
odpovídá uvedeným požadavkům.

Jak lze vypočítat vzdálenost 2 vektorů v_1 a v_2 v rovině?



Jak se počítá kosinová vzdálenost $\cos(\delta - \gamma)$ pro 2 vektory \mathbf{c} ($= \text{OC}$) a \mathbf{d} ($= \text{OD}$) v rovině?

Nechť $|\mathbf{c}|$ je délka vektoru \mathbf{c} :

$$|\mathbf{c}| = \sqrt{c_1^2 + c_2^2}$$

$$\begin{aligned}\cos(\delta - \gamma) &= \cos \delta \cdot \cos \gamma + \sin \delta \cdot \sin \gamma \\ &= c_1 \cdot d_1 / |\mathbf{c}| \cdot |\mathbf{d}| + c_2 \cdot d_2 / |\mathbf{c}| \cdot |\mathbf{d}| \\ &= (c_1 \cdot d_1 + c_2 \cdot d_2) / |\mathbf{c}| \cdot |\mathbf{d}| \\ &= (\text{skalární součin } \mathbf{c} \text{ a } \mathbf{d}) / |\mathbf{c}| \cdot |\mathbf{d}|\end{aligned}$$

Postupy používané pro klasifikaci dokumentů



❖ Zadání problému

- ◆ Mějme **soubor dokumentů**, z nichž každý je ohodnocen jménem *label* nějaké třídy z pevně dané množiny $C = \{c_1, c_2, \dots, c_j\}$ uvažovaných tříd.
- ◆ Dále mějme **klasifikátor**, který je naučen na těchto datech (trénovací množina).
- ◆ Pro libovolný nový, dosud nezpracovávaný dokument, by měl být klasifikátor schopný "predikovat" s vysokým stupněm přesnosti správné zařazení do třídy, kam patří

❖ Jaké postupy lze použít?

- ◆ **Rozhovací stromy**: Výhodné pro vektory dokumentů. Problémy se složitostí u velkých souborů.
- ◆ **Support Vector Machine** vhodná v případě 2 tříd
- ◆ **Bayessovská klasifikace**, Neural Networks, **Case-based reasoning**

Klasifikace dokumentů podle centroidů



Vstup:

- nový dokument $d = (w_1, w_2, \dots, w_n)$;
- Předem dané kategorie $C = \{c_1, c_2, \dots, c_l\}$, do nichž jsou zařazeny všechny dokumenty v trénovací množině

1. **Centroid** všech vektorů zařaz. do třídy i označ jako vektor c_i .

2. **Použij jako model podobnosti kosinovou funkci**

$$\text{Simil}(d_i, d_j) = \cos(d_i, d_j) = \frac{d_i \bullet d_j}{\|d_i\|_2 \times \|d_j\|_2} = \frac{\sum (w_{ik} \times w_{jk})}{\sqrt{\sum w_{ik}^2} \times \sqrt{\sum w_{jk}^2}}$$

a pro všechny c_i vypočti hodnotu

$$\text{Simil}(\vec{c}_i, d) = \cos(\vec{c}_i, d)$$

3.// **Výstup:** Dokument d zařad' do třídy **max** tak, aby pro všechna $i \leq l$ platilo

$$\text{Simil}(\vec{c}_i, d) \leq \text{Simil}(c_{\max}, d)$$

K čemu je TM užitečné?



- ❖ Shlukování dokumentů či termů
 - ◆ Najdi ve velkém souboru dokumentů ty, které jsou podobné.
- ❖ Klasifikace textů
 - ◆ Pro nový dokument zjisti, kterých tématům se věnuje
- ❖ Získávání informací
 - ◆ webové vyhledávače
- ❖ Extrakce Informací z textových dokumentů
 - ◆ Odpovídání na dotazy (*Question Answering*)

Vhodné dostupné nástroje, viz

<http://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics/> (e.g. <http://tm.r-forge.r-project.org/> **Text Mining in R**)

Aktualni typy aplikaci



- ❖ **Business/Competitive Intelligence** hleda ve vnejsim prostoru takove nove informace, ktere manazerum umozni vytvaret ucinnejsi strategii dalsiho smerovani firmy (udaje o konkurencnich produktech a firmach, o zakaznicich, novch technologiich, ...)
- ❖ **E-Discovery** (analyza el.dokumentu pro pravni analyzu, napr. jako soucast vysetrovani), **narodni bezpecnost**
- ❖ **Objevovani ve vede** (predevsim prirodni vedy)
- ❖ **Sentiment analysis** – napr. zjistovani politickych nalad ve spolecnosti na socialnich sitich
- ❖ **Monitorovani socialnich siti** (identifikace akt. temat, ..)

TM na webu



- ❖ WWW lze chápat jako obrovský orinetovaný graf, ve kterém webové stránky jsou uzly a odkazy na další stránky (hyperlinks) odpovídají orinetovaným hranám
- ❖ Tato grafová struktura obsahuje vedle samotného textu mnoho informací o tom, jak „užitečné“ jsou jednotlivé „uzly“
- ❖ Podobná situace nastává i ve společnosti
 - ◆ MUDr. A. a MUDr. K. mají stomatologickou ambulanci v téže ulici.
 - ◆ 10 náhodně vybraných chodců říká, že MUDr. A. je dobrý zubař
 - ◆ 5 uznávaných a vážených lékařů, z nichž někteří jsou stomatologové, považuje MUDr. K. za lepšího odborníka než je MUDr.A

Kterého zubaře byste si vybrali?

† Některá témata, která jsme vynechali



- ❖ Využití nezávislých (externích) slovníků, např. **WordNet**
- ❖ Využití technik, které jsou specifické pro zpracování přirozeného jazyka. Typickým příkladem jsou nástroje počítačové lingvistiky, např. využití gramatiky pro
 - ◆ pochopení smyslu dotazu v rámci scénáře pro the "získávání znalostí" (information retrieval) nebo
 - ◆ při implementaci systémů „odpovídajících na dotazy“
- ❖ Další zajímavá aplikace vyhledává články pro dané téma

<http://core.kmi.open.ac.uk/search>

- ❖ Někteří "puristé" nepovažují většinu současných aktivit v oblasti TM za skutečné dobývání znalostí z textů (real text mining) – směřují k něčemu opravdu inovativnímu!

† Další poznámky o budoucích možnostech TM



PubMed (<http://www.ncbi.nlm.nih.gov/pubmed/>) je „veřejně přístupný“ zdroj dat vytvořený a udržovaný Národním Centrem pro Biotechnologické Informace (NCBI) v US National Library of Medicine® (NLM).

PubMed obsahuje víc než 21 milionů článků a citací pro biomedínské publikace z MEDLINE: zdroj obsahující časopisy o živé přírodě (*life science journals*) a elektronické knihy.

RaJoLink či **CrossBee** (crossbee.ijs.si) jsou SW aplikace využívající PubMed pro hledání zajímavých „rozumných“ a velmi smělých hypotéz o tom, jak by bylo možné vysvětlit některé dosud ne dostatečně zdůvodněné fenomény pozorované v reálném životě. Významnou roli zde hrají „outliers“.