

TEXTOVÁ ANALYTIKA VŠETCI O NEJ POČULI, NIKTO JU NEVIDEL



KATARÍNA FRÜHWALDOVÁ

Námestie sv. Petra – Vatikán 2005

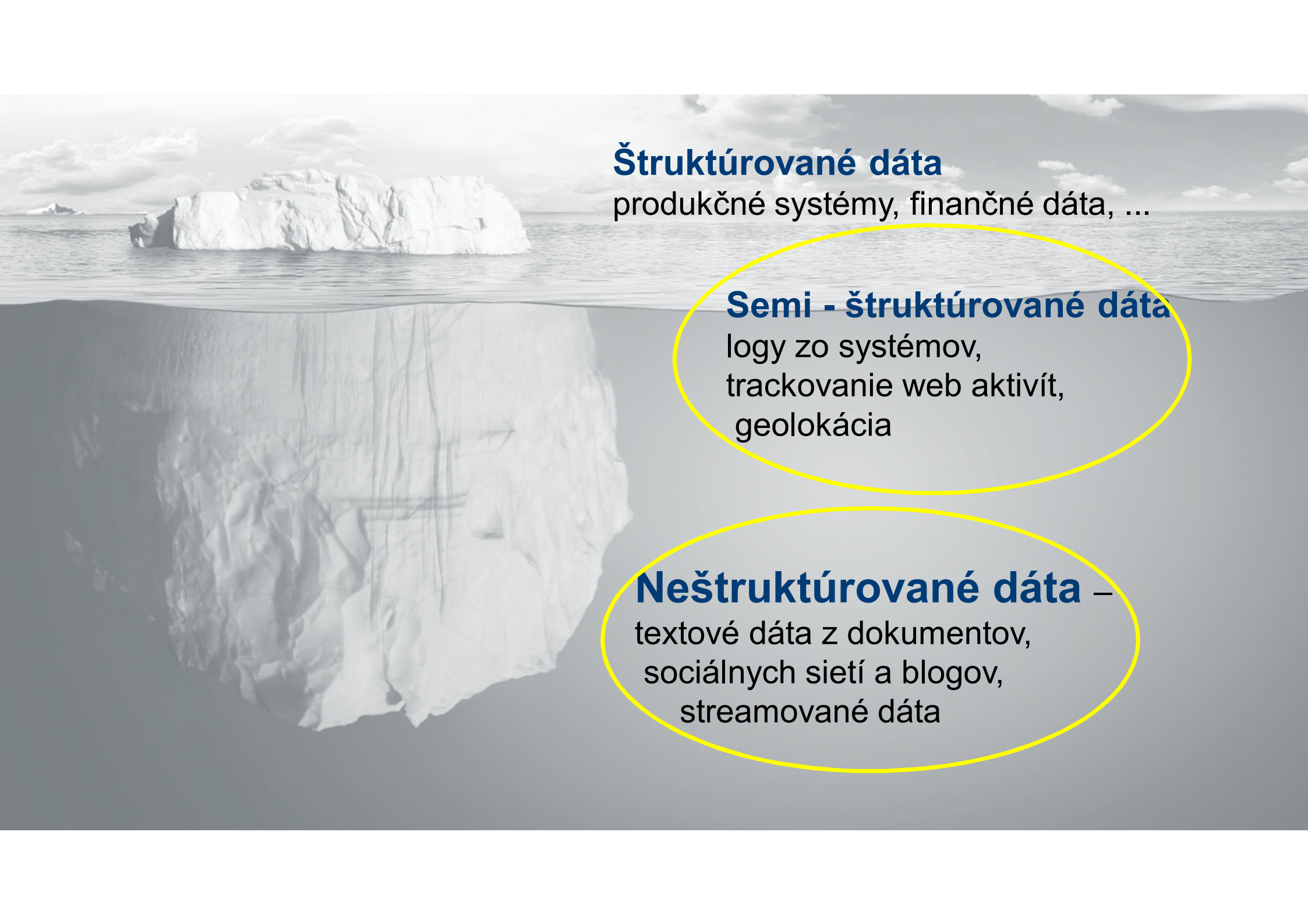


Námestie sv. Petra – Vatikán 2013





DÁTA

An iceberg floating in the ocean. The small tip above the water represents structured data, while the much larger submerged part represents semi-structured and unstructured data.

Štruktúrované dáta

produkčné systémy, finančné dáta, ...

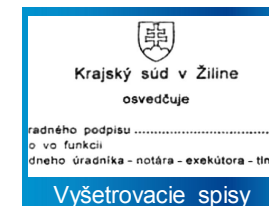
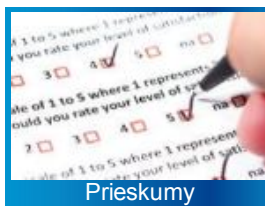
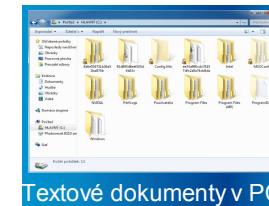
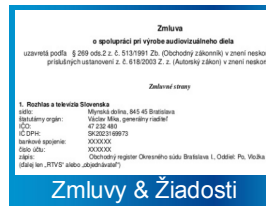
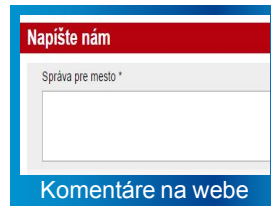
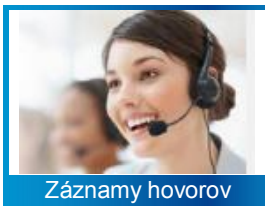
Semi - štruktúrované dáta

logy zo systémov,
trackovanie web aktivít,
geolokácia

Neštruktúrované dáta –

textové dáta z dokumentov,
sociálnych sietí a blogov,
streamované dáta

ZDROJE NEŠTRUKTÚROVANÝCH DÁT



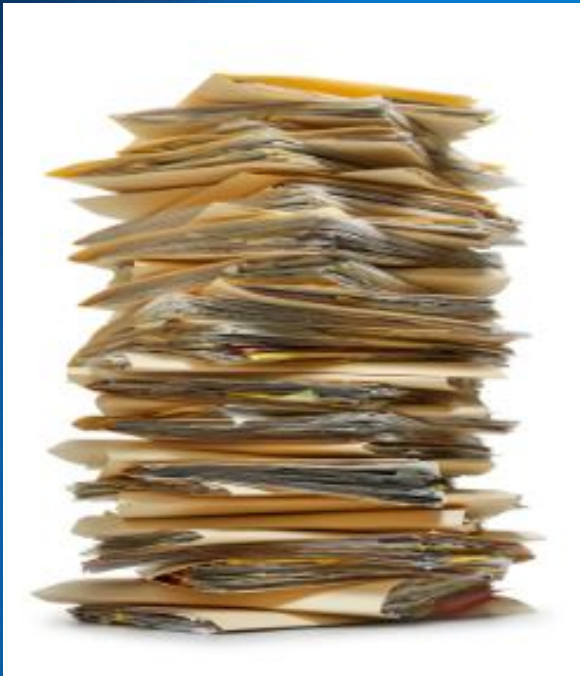
NIKTO JU NEVIDEL ...



POĎME JU POHLÁDATĚ..



TEXTOVÉ DÁTA



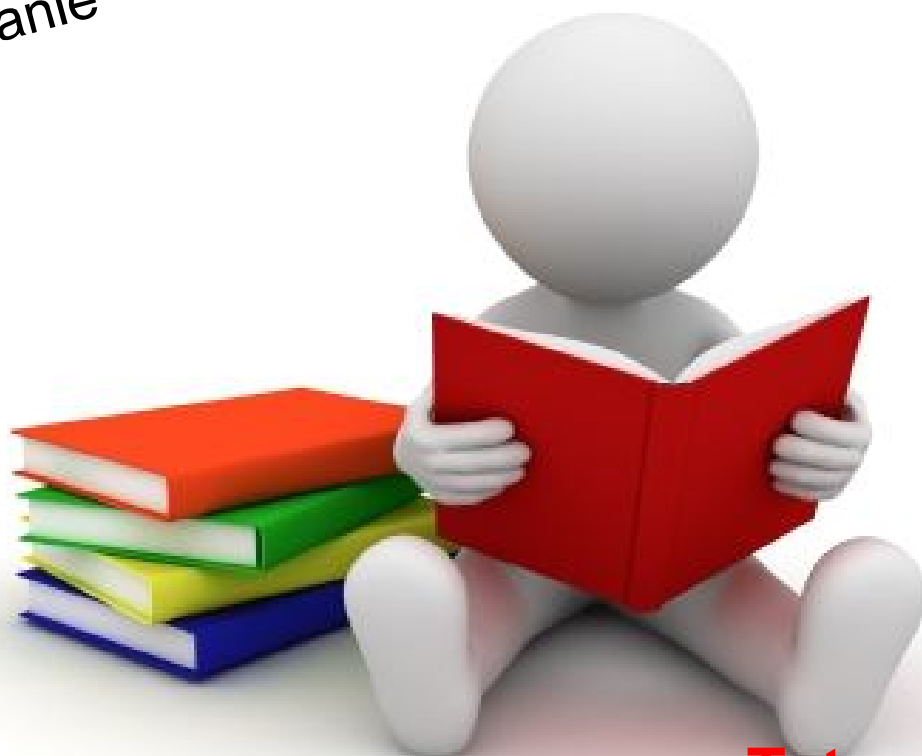
POMÓOC!

VÝZVA

- Jazyk je chaotický!
 - Preklepy, slang, skratky, diakritika
 - Veľké objemy dát
 - Rôznorodé zdroje dát a jazyky
- Nedostatok ľudských zdrojov
- Manuálne čítanie je subjektívne a časovo náročné a postup tvorby vzorky môže viesť k strate cenných informácií
- Ťažké odhaliť skryté vzory a trendy

VYHLÁDÁVANIE | PREDPOKLADOM JE ŽE VIEM ČO HLÁDAM

Vyhľadávanie – CTRL F

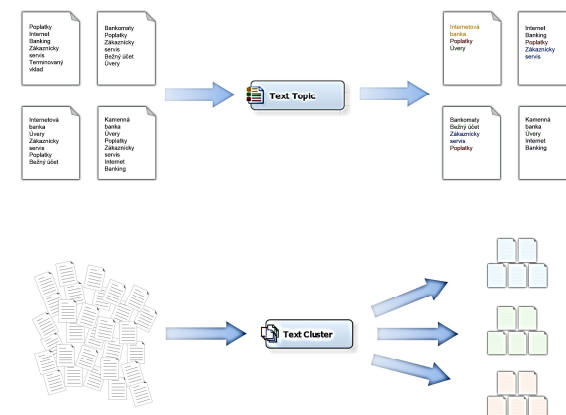
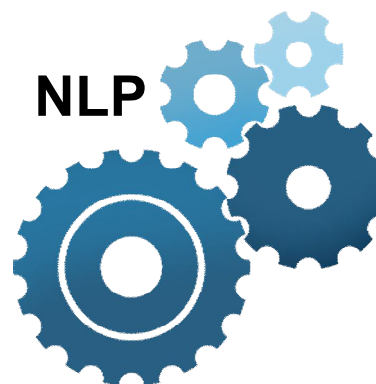


Toto ešte nie je textová analytika...

PROCES



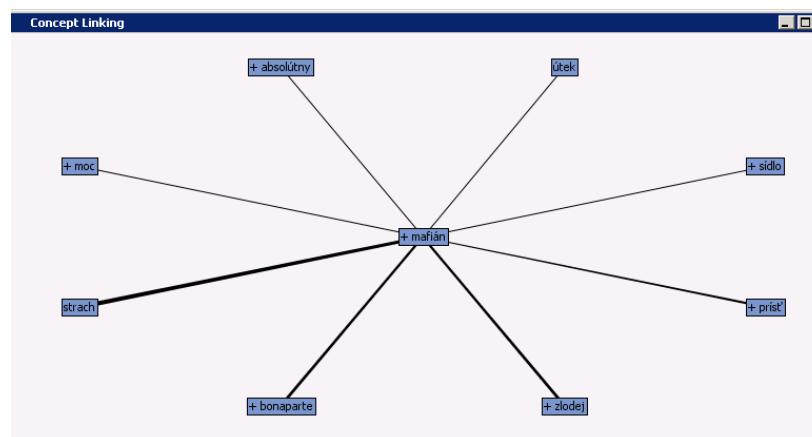
	ID	RECEZIA	HOODNENIE	ODPORICA
1	1	Rubina so 14 a 12 roč. detmi. Komplex veľkej palázy, moderné zariadenie pre...	100%	1
2	2	Kdeby hotel. krásne prostredie, transfer letiska cca 1. hod. Ubytovaní sm...	94%	1
3	5	Maja Beata rodná dovolenka v Turecku dopadne, ako si sam. Bál Sme u...	50%	0
4	7	Tento hotel je vhodný pre rodiny s deťmi. Ubytovanie je krásne, izby sú pre...	92%	1
5	8	Hotel s veľkým resortom je prekrásny, ročník čísla 11 super. Vďaka pohodliu, izby...	100%	1
6	9	Super dovolenka, stretnutie super, relaxačný výber, senzor parkovísk, veľa...	100%	1
7	11	Bývali sme v novom budove LT, takže super, každý deň upratovanie + žijúť...	86%	1
8	14	Začiatok by som rád povedal že toto bola moja sedmá dovolenka v Turecku...	42%	0
9	15	Hotel je vyhovujúce pre rodiny a malých detí do 12 rokov (ľahko odporúčam...	67%	0
10	17	Transfer z letiska cca 1,20hod. Po príchode do 15 min sme boli ubytovaní...	97%	1
11	18	Transfer z letiska 1. 5 hodiny, ubytovanie o jedinečnosť v noci bez nespánia...	81%	1
12	19	Hotel nedosahuje 5* ani náhodou, spájam tak máz 3, 5* pre izby, stretnutie a v...	42%	0
13	20	Fakt super dovolenka, veľký hotel, ale napriek tomu, miesto pít a v bazéne...	81%	1
14	21	Vďaka krásnemu resortu a relaxačným službám. Fíkale sme v noci a hneď rá...	79%	0
15	22	Avšak hotel je nádherný, bohatý ovet super, pláž obklopená 17 Turecku me...	81%	1
16	23	Tento hotel je úžasný! Bar bol pri hoteli, bazénoch a na pláži. Okrem toho m...	89%	1
17	24	Výborná dovolenka, izby boli čisté, ale nábytok stáli, pre 4 dieru rohu...	86%	1
18	25	Hotel pôsobí veľmi luxusne a určite si tu príde na svoje aj náročnejší kú...	94%	1
19	26	Veľmi pekný hotel, nádherný resort, vďaka krásnej krajine, strava. Bál sme a...	83%	1



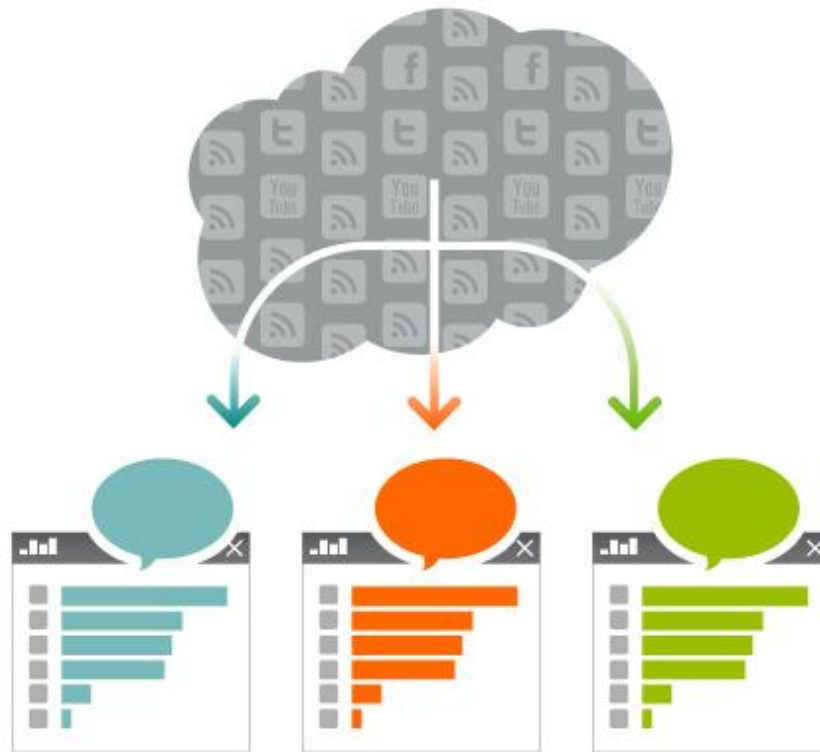
UKÁŽKA ANALÝZY ANALÝZA DISKUSNÝCH PRÍSPEVKOV

Terms					
	TERM	FREQ	# DOCS	KEEP ▼	WEIG
+	kaliňák	112	61	✓	
+	človek	50	43	✓	
+	smer	53	36	✓	
+	matovič	43	33	✓	
+	mafián	44	31	✓	
+	fico	40	29	✓	
+	milión	27	22	✓	
+	slovensko	29	22	✓	
+	dať	23	20	✓	
+	štát	20	18	✓	
+	bašternák	30	18	✓	
+	politik	25	17	✓	
+	minister	22	17	✓	
+	voliť	22	17	✓	
+	leško	26	16	✓	
+	vláda	28	16	✓	
+	začať	17	16	✓	

Clusters		
Cluster ID	Descriptive Terms	Frequency ▼
8	+kaliňák +smer +volič +bohatý +bašternák +počiatek ...	41
9	+príst' +zarobiť +mafián +kamarát +štát +firma ...	36
3	+občan +eú +pomáhať +slovensko +tisíc +banka ...	30
5	+matovič +kotleba +bolševik +vysvetliť +prevrat +voliť ...	30
7	+nejst' +politik +národ +basa +daňový +dať ...	25
11	+pirát +nájst' +novinár +červený +komunista +informácia...	25
4	+zodpovednosť +kiska +dovoliť +vláda +byť +pravda ...	23
2	+suma +tvrdiť +nemusieť +príjem +biznis +vyplatiť ...	17



KATEGORIZÁCIA TEXTU



.... Tvorba taxonómie....

Taxonómia je štruktúra priečinkov a priečinky s pod priečinkami sú pomenované ako kategórie resp. pod kategórie

... Automatizácia ...

... Objektivita ...

KATEGORIZÁCIA EXTRAKCIA KONCEPTU

- Extrakcia konceptu – vyhľadať hodnotné informácie v texte. Koncepty sú názvy entít (ľudí, adries, firiem, emailov...) a podstatných informácií (číslo účtu, ičo...)
- Využitie
 - Obohatenie dát
 - Využitie open / webových dát
 - Vyšetrovanie – zabavené PC, preskúmanie pošty...
 - Monitorovanie – konkurencia, vestníky, akékoľvek webové dáta

ZMLUVA č. MK – 78/09/M
o obsahoch, cieľoch a zabezpečení služieb verejnosti v oblasti rozhlasového vysielania
na roky 2010 – 2014

uzatvorená podľa § 21a zákona č. 619/2003 Z. z. o Slovenskom rozhlas v znení neskorších
predpisov

medzi

I. Slovenská republika
zastúpená: Ministerstvo kultúry Slovenskej republiky
sídlo: Námestie SNP 813 31 Bratislava
IČO: 00165182
bankové spojenie: Štátna pokladnica
výdavkový účet: 7000071652/8180
príjmový účet: 7000071599/8180
depozitný účet: 7000071687/8180
zastúpený: Marek Maďarič
minister kultúry
(ďalej len „štát“)

II. Slovenský rozhlas
zapsaný v Obchodnom registri Okresného súdu Bratislava I, oddiel: Po, vložka
číslo: 28/B,
sídlo: Mútna č. 1, 817 55 Bratislava

SENTIMENT

- Pomáha pochopiť a bližšie spoznať nálady a názory
- Využitie najmä v komunikácii s občanom
- Dôležitosť širšieho konceptu pri problémovej identifikácii:
sarkazmus („no pekne), negácia („nie je dobrý“), čertovsky dobrý...



... A PREDSA EXISTUJE ...





- Web – o čom sa hovorí, odčítanie nálady, povest' mena, ako sú vnímané významné udalosti/témy/kauzy
- Kontrola interných emailov – o čom sa píše, vynášanie informácií...
- Zabavené počítače / preskúmanie PC dokumentov
- Extrakcia konceptu / entít
- Kategorizácia dokumentov – správy, vyšetrovacie záznamy, podnety, komunikácia s občanom...

VĎAKA ZA POZORNOST

