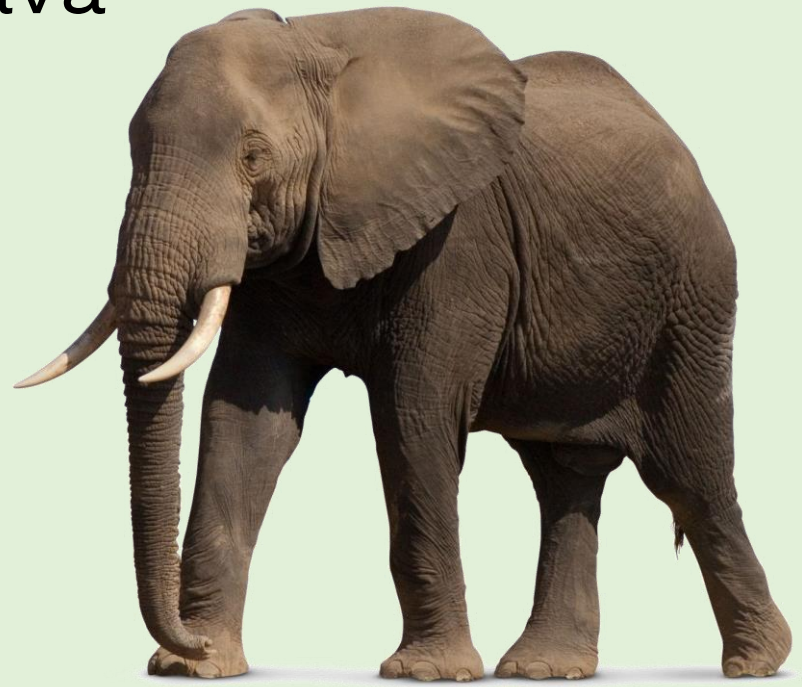
Vyhledávání informací

Modely IR systémů

28. 10. 2021

Rozehřívací vyhledávačka

Jak končí svět podle díla, v němž se na místě motta (epigrafu) vyskytuje kusá informace o tom, že centrální postava románu Josepha Conrada *Srdce temnoty*, postava posedlá slonovinou, je mrtva?



Základní cíl IR systému

Vybrat na základě dotazu *(q)* relevantní dokumenty *(d)* z kolekce dokumentů.

Indexování dokumentů

- *Shakespearovo sebrané dílo*
- není to malé, ale není to kolekce webových dokumentů
- hledáme díla, kde se vyskytuje slovo Brutus a Caesar, ale ne Cézarova žena Calpurnia
- Brutus AND Caesar NOT Calpurnia

Indexování dokumentů

- *Shakespearovo sebrané dílo*
- hledáme díla, kde se vyskytuje slovo Brutus a Caesar, ale ne Cézarova žena Calpurnia
- Brutus AND Caesar NOT Calpurnia

- *grepping* – lineární prohledávání
- trvá to, neřadí to, logika?...
- řešením je [redacted]

Indexování dokumentů

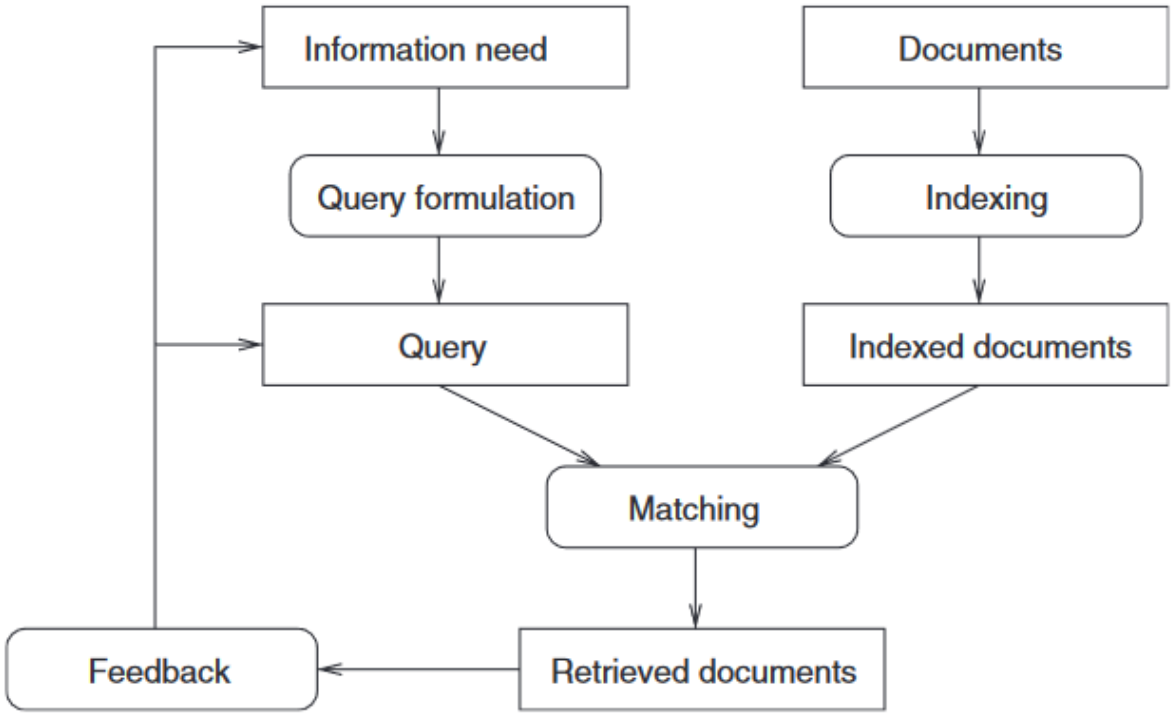
- *Shakespearovo sebrané dílo*
- hledáme díla, kde se vyskytuje slovo Brutus a Caesar, ale ne Cézarova žena Calpurnia
- Brutus AND Caesar NOT Calpurnia

- *grepping* – lineární prohledávání
- trvá to, neřadí to, logika?...
- řešením je indexovat dokumenty předem

Indexování dokumentů

Tři základní cíle systému IR:

- reprezentace dokumentů (indexování)
- reprezentace dotazu
- porovnání obou



Indexování dokumentů

matice dokument-termín

binární zaznamenání

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	…
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
…							

Index

Sestavte index pro tyto čtyři dokumenty:

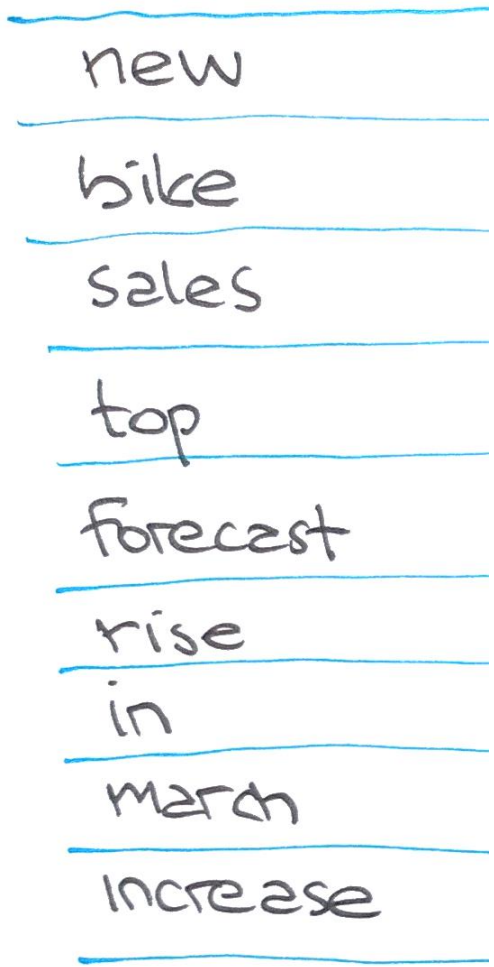
New bike sales! Top forecast!	*Bike sales rise in March*	Increase in bike sales in March!	**March:** new bike sales rise…

Index

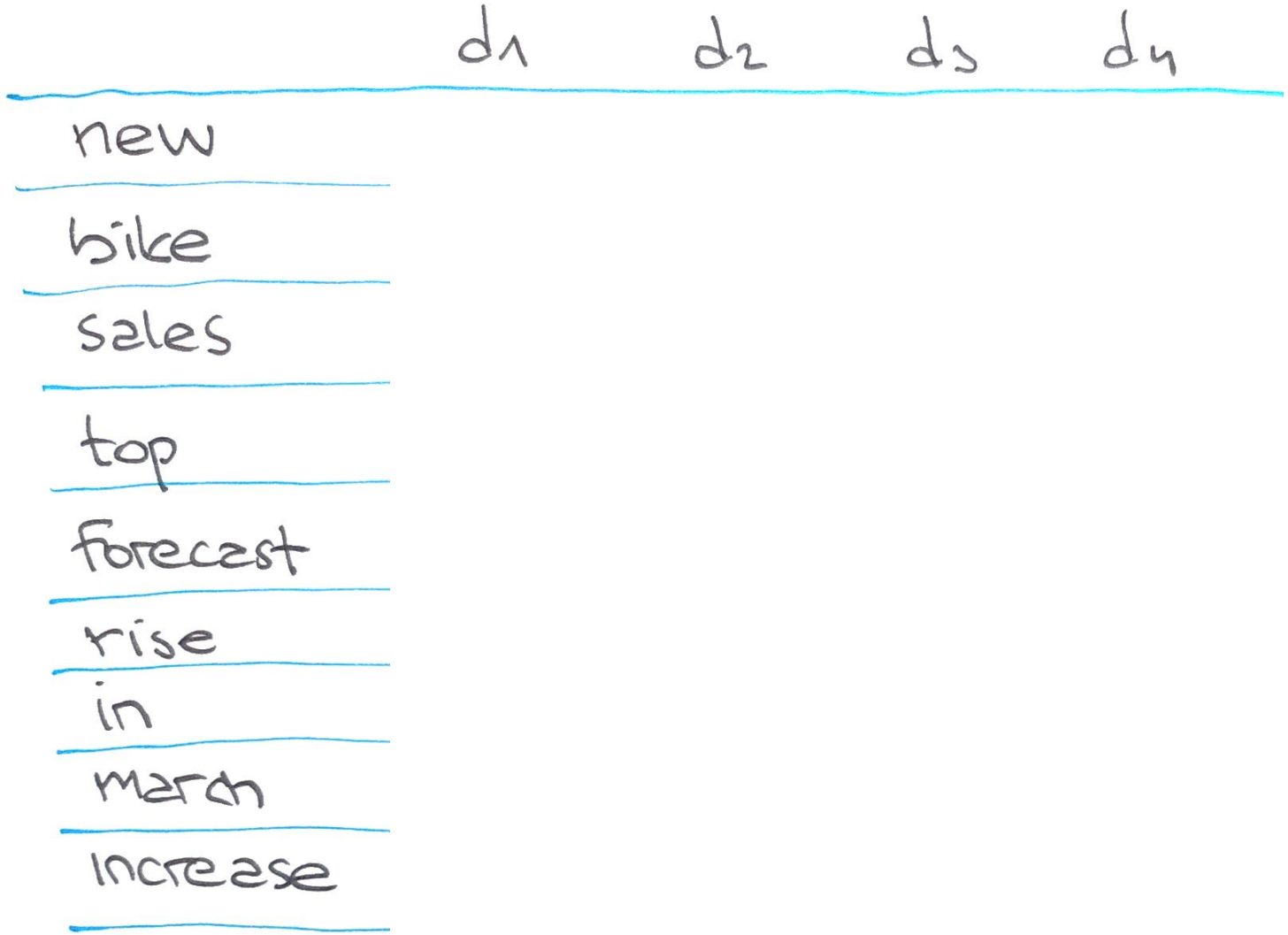
Sestavte index pro tyto čtyři dokumenty:

d1: new bike sales top forecast
d2: bike sales rise in march
d3: increase in bike sales in march
d4: march new bike sales rise

Co budeme potřebovat, abychom dokázali sestavit index podobně jako před chvílí u Shakespeara?



d1: new bike sales top forecast
d2: bike sales rise in march
d3: increase in bike sales in march
d4: march new bike sales rise



d1: new bike sales top forecast
d2: bike sales rise in march
d3: increase in bike sales in march
d4: march new bike sales rise

	d_1	d_2	d_3	d_4
new	1	0	0	1
bike	1	1	1	1
sales	1	1	1	1
top	1	0	0	0
forecast	1	0	0	0
rise	0	1	0	1
in	0	1	1	0
march	0	1	1	1
increase	0	0	1	0

d1: new bike sales top forecast
d2: bike sales rise in march
d3: increase in bike sales in march
d4: march new bike sales rise

	d_1	d_2	d_3	d_4
new	1	0	0	1
bike	1	1	1	1
sales	1	1	1	1
top	1	0	0	0
forecast	1	0	0	0
rise	0	1	0	1
in	0	1	1	0
march	0	1	1	1
increase	0	0	1	0

$d_4 = 1, 1, 1, 0, 0, 1, 0, 1, 0;$

d1: new bike sales top forecast
d2: bike sales rise in march
d3: increase in bike sales in march
d4: march new bike sales rise

$\mathbf{d_1}$ = 1, 1, 1, 0, 0, 1, 0, 1, 0;

Jaký problém bude mít tento jednoduchý index v případě reálných dokumentů?

$\mathbf{d_1}$ = 1, 1, 1, 0, 0, 1, 0, 1, 0;

$\mathbf{t_1}$ = d_1, d_4;

q = query = dotaz
d = dokument

Převrácený index

- efektivnější k účelům IR
- díváme se na to z pohledu termínů
- *neznáme to odněkud?*

Sestavte převrácený index pro tyto čtyři dokumenty:

d1: new bike sales top forecast

d2: bike sales rise in march

d3: increase in bike sales in march

d4: march new bike sales rise

	d_1	d_2	d_3	d_4
new	1	0	0	1
bike	1	1	1	1
sales	1	1	1	1
top	1	0	0	0
forecast	1	0	0	0
rise	0	1	0	1
in	0	1	1	0
march	0	1	1	1
increase	0	0	1	0

$d_4 = 1, 1, 1, 0, 0, 1, 0, 1, 0;$

d1: new bike sales top forecast
d2: bike sales rise in march
d3: increase in bike sales in march
d4: march new bike sales rise

	d_1	d_2	d_3	d_4	
new	1	0	0	1	$= d_1, d_4;$
bike	1	1	1	1	$= d_1, d_2, d_3, d_4;$
sales	1	1	1	1	$= \ldots$
top	1	0	0	0	
forecast	1	0	0	0	
rise	0	1	0	1	
in	0	1	1	0	
march	0	1	1	1	
increase	0	0	1	0	

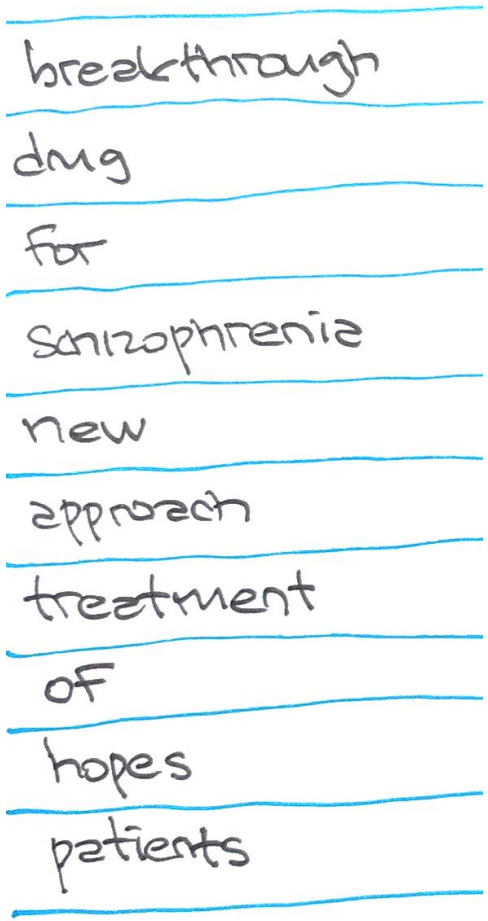
$d_4 = 1,1,1,0,0,1,0,1,0;$

d1: new bike sales top forecast
d2: bike sales rise in march
d3: increase in bike sales in march
d4: march new bike sales rise

Index a převrácený index

- sestavte matici dokument-termín
- sestavte převrácený index

d1: breakthrough drug for schizophrenia
d2: new schizophrenia drug
d3: new approach for treatment of schizophrenia
d4: new hopes for schizophrenia patients



d1: breakthrough drug for schizophrenia
d2: new schizophrenia drug
d3: new approach for treatment of schizophrenia
d4: new hopes for schizophrenia patients

	d_1	d_2	d_3	d_4
breakthrough	1	0	0	0
drug	1	1	0	0
for	1	0	1	1
schizophrenia	1	1	1	1
new	0	1	1	1
approach	0	0	1	0
treatment	0	0	1	0
of	0	0	1	0
hopes	0	0	0	1
patients	0	0	0	1

d1: breakthrough drug for schizophrenia
d2: new schizophrenia drug
d3: new approach for treatment of schizophrenia
d4: new hopes for schizophrenia patients

	d_1	d_2	d_3	d_4	
breakthrough	1	0	0	0	d_1;
drug	1	1	0	0	d_1, d_2;
for	1	0	1	1	d_1, d_3, d_4;
schizophrenia	1	1	1	1	d_1, d_2, d_3, d_4;
new	0	1	1	1	d_2, d_3, d_4;
approach	0	0	1	0	d_3;
treatment	0	0	1	0	d_3;
of	0	0	1	0	d_3;
hopes	0	0	0	1	d_4;
patients	0	0	0	1	d_4;

d1: breakthrough drug for schizophrenia
d2: new schizophrenia drug
d3: new approach for treatment of schizophrenia
d4: new hopes for schizophrenia patients

Indexování a vyhledávání

- subject indexing – *o čem je dokument?*
- v knihovnách je to jasné…
- je to výzkumná oblast

- jenže počet dokumentů roste
- potřebujeme automatickou indexaci
- modely a postupy se vyvíjejí
- nástup webu
- počítačová lingvistika; AI, NS, computer vision,…

Indexování a vyhledávání

- velká oblast výzkumu
- specifické systémy IR
- např. lékařská dokumentace
- specifika různých druhů dokumentů

Modely IR

- existuje nespočet modelů
- booleovský model (50. léta)
- vektorový model (70. léta)
- pravděpodobnostní modely (70. léta)
- modely založené na zpracování jazyka (90. léta)
- Google PageRank (1998)
- *desítky a stovky dalších…*

Booleovský model

- booleovská logika
- první, nejrozšířenější a široce aplikovaný
- dokument i dotaz jsou pojímány jako soubor výrazů
- vyhledání je založeno na výskytu výrazů z dotazu
- vyhledávání výrazu „jablko“ jednoduše vrací množinu dokumentů, kde je výraz „jablko“
- pomocí logických operátorů booleovské logiky lze vytvářet nové množiny dokumentů odpovídající vyhledávacímu dotazu

Booleovský model

- Brutus AND Caesar
- najdi v převráceném indexu výskyt Brutus
- natáhni si ID dokumentů
- najdi v převráceném indexu výskyt Caesar
- natáhni si ID dokumentů
- najdi shodu

- *příklad se schizofrenií* – q: schizophrenia AND drug

	d_1	d_2	d_3	d_4	
breakthrough	1	0	0	0	d_1;
drug	1	1	0	0	d_1, d_2;
for	1	0	1	1	d_1, d_3, d_4;
schizophrenia	1	1	1	1	d_1, d_2, d_3, d_4;
new	0	1	1	1	d_2, d_3, d_4;
approach	0	0	1	0	d_3;
treatment	0	0	1	0	d_3;
of	0	0	1	0	d_3;
hopes	0	0	0	1	d_4;
patients	0	0	0	1	d_4;

d1: breakthrough drug for schizophrenia
d2: new schizophrenia drug
d3: new approach for treatment of schizophrenia
d4: new hopes for schizophrenia patients

Booleovský model

- zároveň nejkritizovanější
- výsledky nejsou hodnoceny a řazeny
- model dokument prostě najde nebo ne

- „jablko AND sad AND štrůdl“
- nenajde „letadlo AND pilot AND létání“,
- nenajde „jablko AND štrůdl“ - očividně užitečnější, ale model nemá jak určit, že tomu tak je…

	d_1	d_2	d_3	d_4	
breakthrough	1	0	0	0	d_1;
drug	1	1	0	0	d_1, d_2;
for	1	0	1	1	d_1, d_3, d_4;
schizophrenia	1	1	1	1	d_1, d_2, d_3, d_4
new	0	1	1	1	d_2, d_3, d_4;
approach	0	0	1	0	d_3;
treatment	0	0	1	0	d_3;
of	0	0	1	0	d_3;
hopes	0	0	0	1	d_4;
patients	0	0	0	1	d_4;

jak to můžu vyřešit?

d1: breakthrough drug for schizophrenia
d2: new schizophrenia drug
d3: new approach for treatment of schizophrenia
d4: new hopes for schizophrenia patients

Regionový model

- rozšíření boolovského modelu
- regiony – tj. model pro jasně strukturované dokumenty
- *co to může být?*
- dva základní operátory CONTAINING, CONTAINEDBY

(<LINE> CONTAINING farewell) CONTAINEDBY
(<SPEECH>CONTAINING (<SPEAKER> CONTAINING Hamlet))

Modely přesné shody

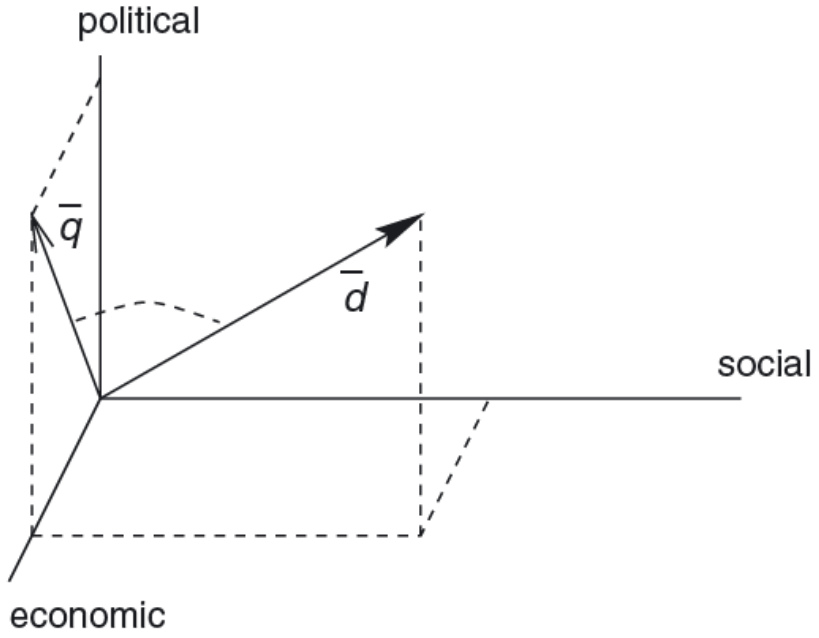
- booleovský, regionový
- *WYSIWYG*
- nemožnost řadit dokumenty
- existují rozšíření
- např. *rozšířený booleovský model*

Vektorové modely

- nutnost řazení výsledků jako vstupní bod
- Hans Peter Luhn (1957) – *statistický přístup*
- k prohledání kolekce dokumentů by měl uživatel připravit dokument podobný k těm, co potřebuje
- míra podobnosti mezi reprezentací připraveného dokumentu a reprezentací dokumentů v kolekci
- první krok: spočítat počet prvků které dotaz a indexovaná reprezentace dokumentů sdílejí

Vektorové modely

- Gerard Salton (1983)
- každý termín má přidělenu svou dimenzi v prostoru
- podobnost se měří jako cosinus úhlu který odděluje vektor dokumentu a vektor dotazu
- výzkumy oblasti seskupování dokumentů (clustering), automatické kategorizace textu



podle čeho to vynesu?

Vektorové modely

d1: the man walked the dog

d2: the man took the dog to the park

d3: the dog went to the park

Zdroje příkladů: https://www.dissertations.se/dissertation/245456a998/

https://blog.seznam.cz/2011/08/semanticka-analyza-textu-1/

Vektorové modely

d1: the man walked the dog
d2: the man took the dog to the park
d3: the dog went to the park

[dog, man, park, the, to, took, walked, went]

Vektorové modely

d1: the man walked the dog
d2: the man took the dog to the park
d3: the dog went to the park

[dog, man, park, the, to, took, walked, went]

d1 [1, 1, 0, 1, 0, 0, 1, 0]
d2 [1, 1, 1, 1, 1, 1, 0, 0]
d3 [1, 0, 1, 1, 1, 0, 0, 1]

Vektorové modely

d1: the man walked the dog
d2: the man took the dog to the park
d3: the dog went to the park

[dog, man, park, the, to, took, walked, went]

d1 [1, 1, 0, 1, 0, 0, 1, 0]
d2 [1, 1, 1, 1, 1, 1, 0, 0]
d3 [1, 0, 1, 1, 1, 0, 0, 1]

Jak zvážit, které termíny jsou důležité a které ne?

Vektorové modely

d1: the man walked the dog
d2: the man took the dog to the park
d3: the dog went to the park

[dog, man, park, the, to, took, walked, went]

d1 [1, 1, 0, 2, 0, 0, 1, 0]
d2 [1, 1, 1, 3, 1, 1, 0, 0]
d3 [1, 0, 1, 2, 1, 0, 0, 1]

TF-IDF

- term frequency–inverse document frequency
- statistické vyjádření důležitosti výrazu v dokumentu
- nejvyužívanější metoda vážení výrazů

d1 [1, 1, 0, 2, 0, 0, 1, 0]

d2 [1, 1, 1, 3, 1, 1, 0, 0]

d3 [1, 0, 1, 2, 1, 0, 0, 1]

$$\mathrm{idf_i} = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

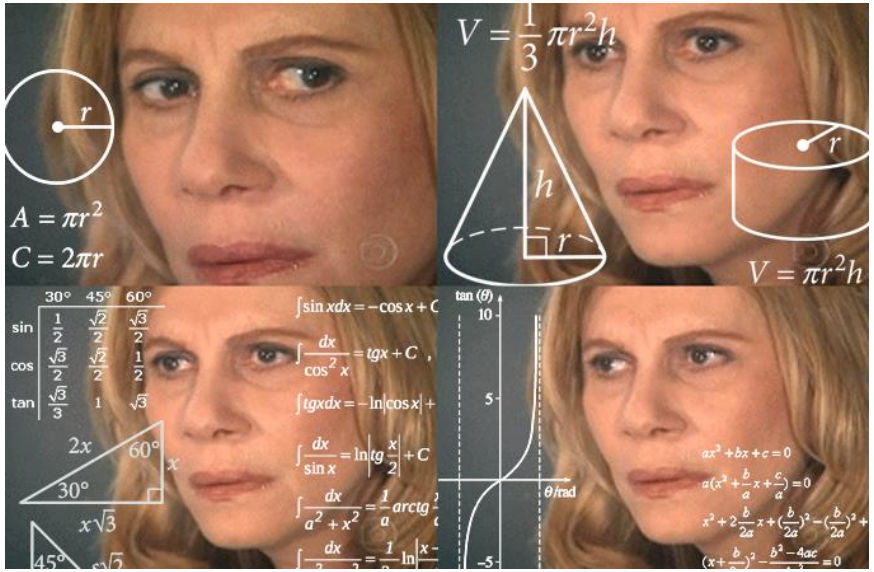
TF-IDF

- term frequency–inverse document frequency
- statistické vyjádření důležitosti výrazu v dokumentu
- nejvyužívanější metoda vážení výrazů

d1 [0, 0.18, 0, 0, 0, 0, 0.48, 0]

d2 [0, 0.18, 0.18, 0, 0.18, 0.48, 0, 0]

d3 [0, 0, 0.18, 0, 0.18, 0, 0, 0.48]



Vektorové modely

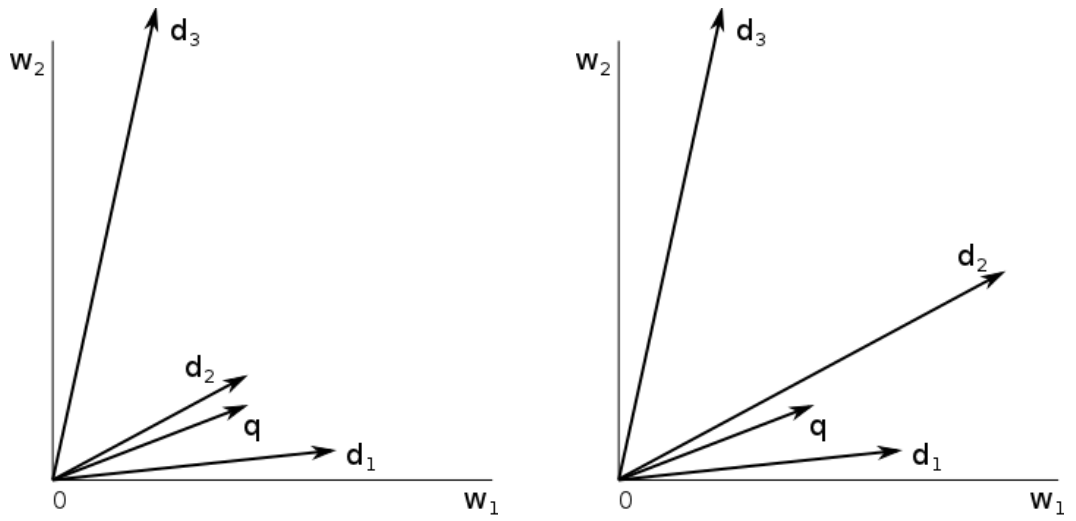
1. *the man walked the dog*
2. *the man took the dog to the park*
3. *the dog went to the park*

[dog, man, park, the, to, took, walked, went]

d1 [0, 0.18, 0, 0, 0, 0, 0.48, 0]
d2 [0, 0.18, 0.18, 0, 0.18, 0.48, 0, 0]
d3 [0, 0, 0.18, 0, 0.18, 0, 0, 0.48]

Vektorové modely

- relevance = vektor *q* leží nejblíže vektoru *d*
- TF problém: *the man walked the dog, and they walked and walked and walked…*
- neměříme vzdálenost
- měříme úhel
- kosinová podobnost
- EN: *cosine similarity*



Obrázek: https://blog.seznam.cz/2011/08/semanticka-analyza-textu-1/

Příliš mnoho dimenzí

- redukce dimenzionality
- omezení slov, které bereme v potaz
- výběr příznaků (stop slova)
- extrakce příznaků (lemmatizace, stemmatizace)
- desambiguace (letech – lemma let a léto), [redacted]
- oko – řada významů, stejné lemma
- stomatolog, zubař – jeden význam, různé lemma

Příliš mnoho dimenzí

- redukce dimenzionality
- omezení slov, které bereme v potaz
- výběr příznaků (stop slova)
- extrakce příznaků (lemmatizace, stemmatizace)
- desambiguace (letech – lemma let a léto), *kontext*
- oko – řada významů, stejné lemma
- stomatolog, zubař – jeden význam, různé lemma

Latentní sémantická analýza

- automatická identifikace kontextu
- pracuje se např. se společným výskytem slov
- v několika úrovních *(fotbal – tenis | sport)*

d1: the man walked the dog

d2: the man took the dog to the park

d3: the dog went to the park

q: „walked“ – co bude ve výsledcích?

NLP

- zpracování přirozeného jazyka (NLP)
- topic modelling / topic classification
- lat. sémantická analýza je klasickým přístupem NLP
- https://monkeylearn.com/topic-analysis/

- *např. automatická klasifikace emailů*
- automatická indexace dotazu
- https://www.jstor.org/analyze/

Pravděpodobnostní modely

- aplikace pravděpodobnosti do IR
- IR systém má pouze *q*
- nejisté porozumění informační potřebě
- pravděpodobnost pracuje s nejistotou…
- pravděpodobnost nám může pomoci
- *aktuálně exponované*
- *mnoho modelů a přístupů*

Fuzzy logika v IR

- obecný model fuzzy IR systému bere v potaz vágnost
- nejistota skrze nepřesné hodnoty a neurčité dotazy
- „datum vydání je nedávné“
- řešení neurčitosti dotazů
- řešení nejistot v databázích

information retrieval

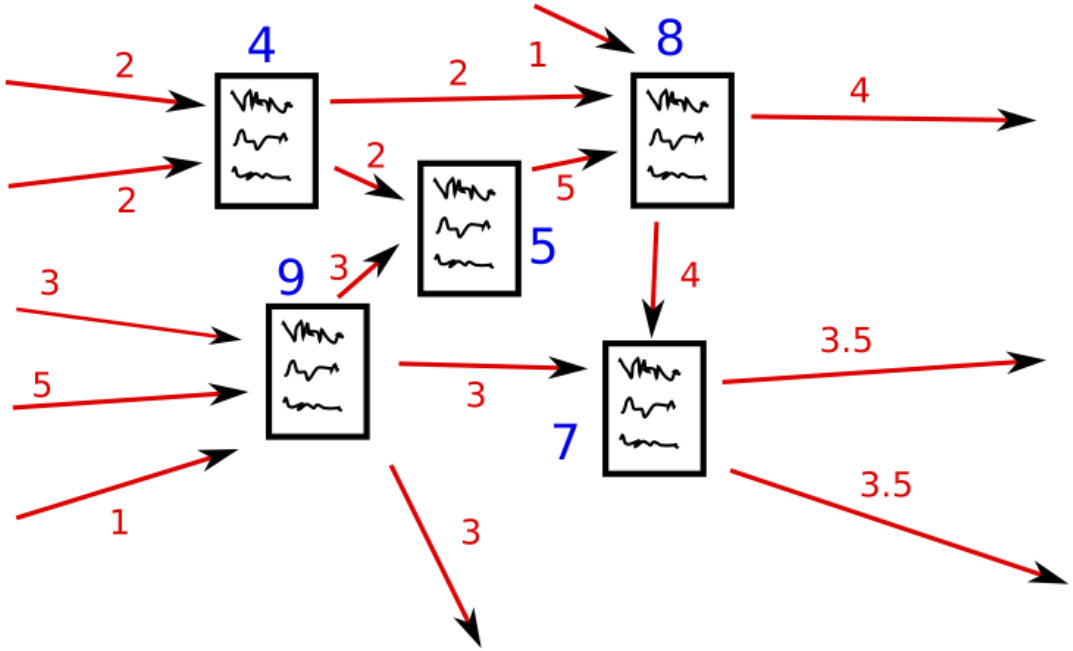
Indexování webu

- stránek na webu jsou miliardy
- indexování a vyhledávání na webu je obtížené
- *jak se indexuje web?*

Indexování webu

- stránek na webu jsou miliardy
- indexování a vyhledávání na webu je obtížené
- *jak se indexuje web?*

PageRank



- předchozí modely vracejí *d* na základě shody s *q*
- *Larry Page a Sergey Brin*
- cíl PR = vracet kvalitní a důvěryhodné dokumenty
- využití architektury hyperlinku
- vyhodnocení kvality dokumentu (stránky)
- více odkazů = větší kvalita

$$R'(a) = c\left(\sum_{u \in B_a} \frac{R'(u)}{N_u}\right) + (1-c)R'(a)$$

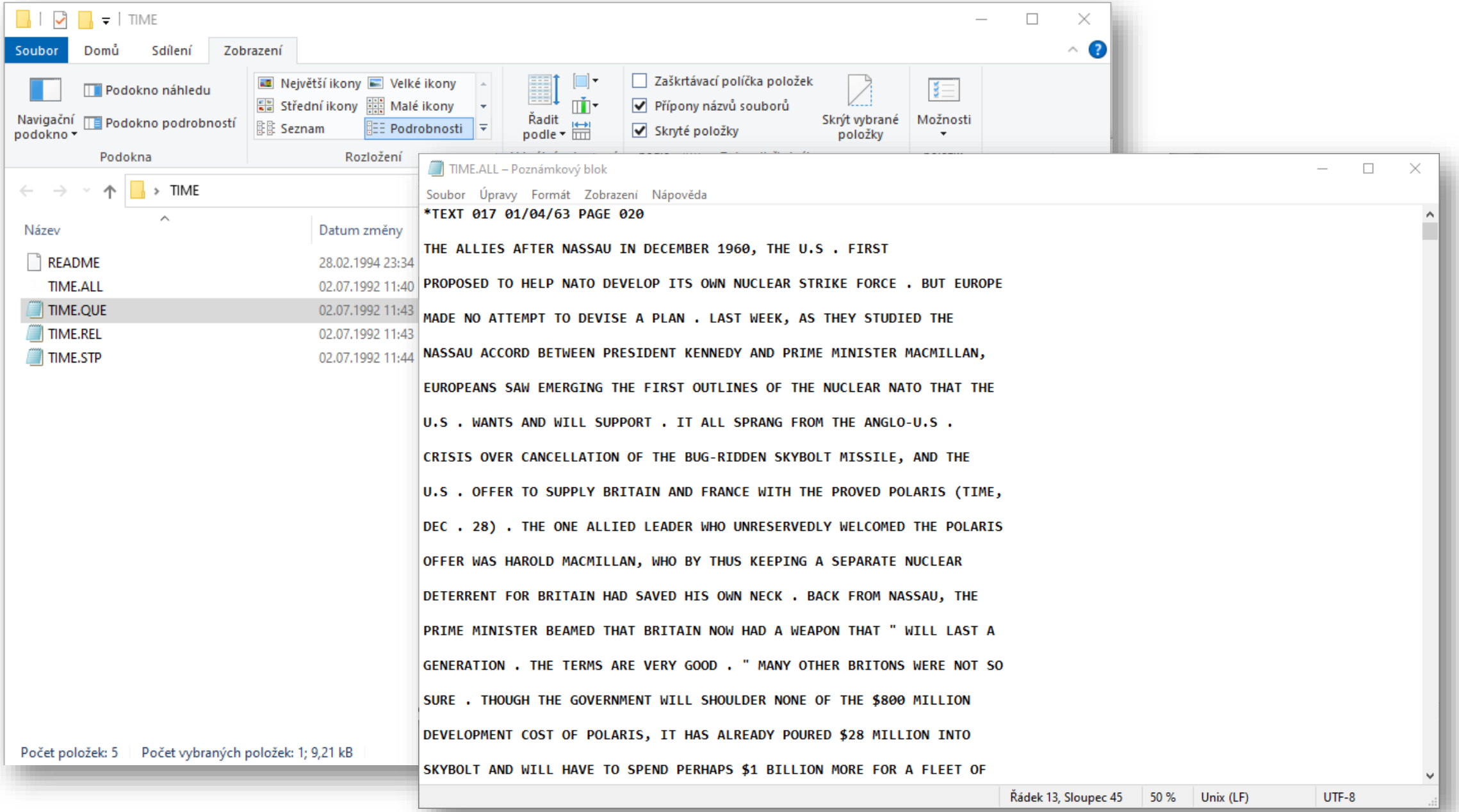
Hodnocení modelů

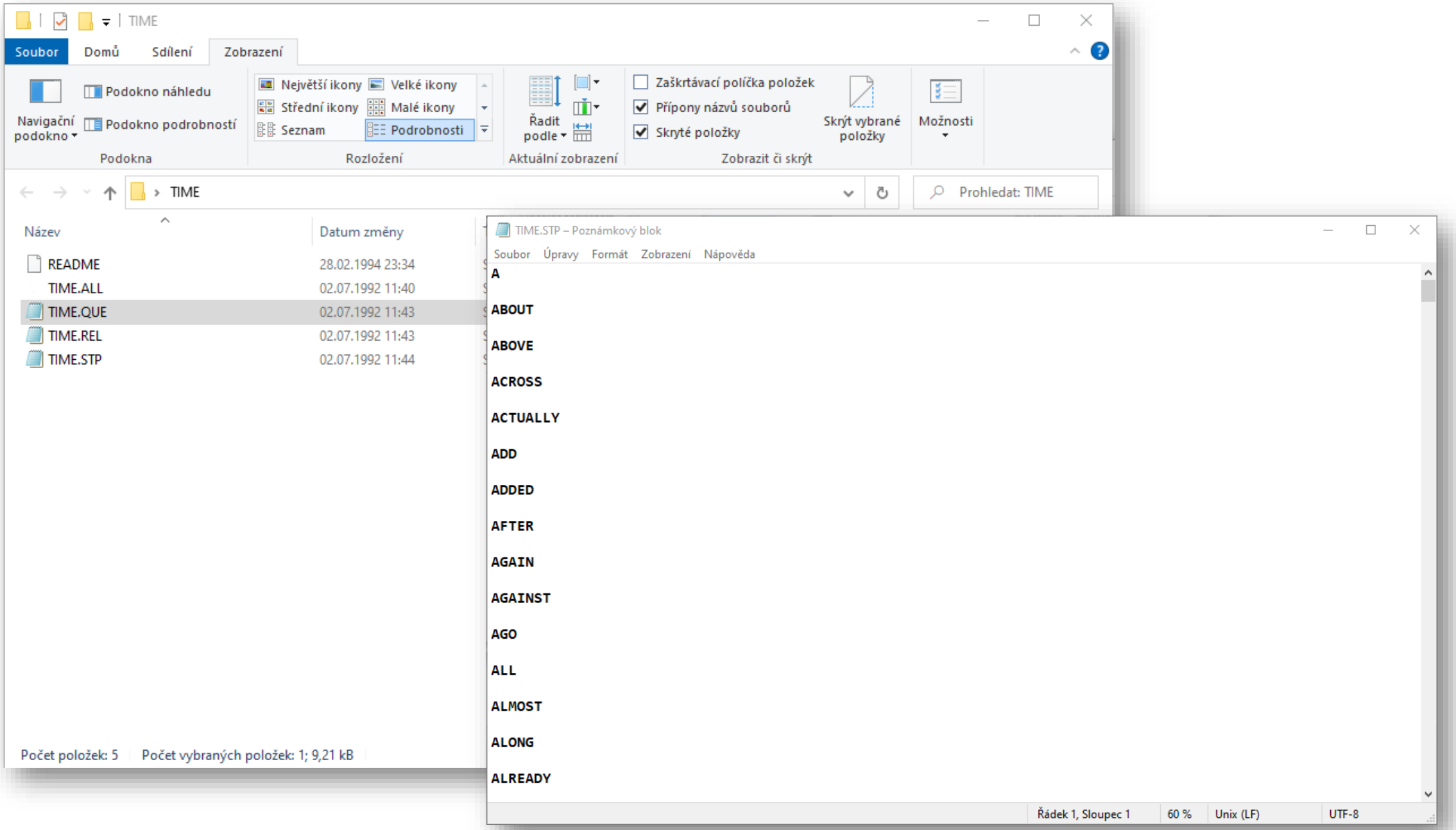
- modelů je celá řada
- jak zjistíme, co funguje a co ne?
- IR je značně empirické odvětví
- přínos modelu je třeba dokázat

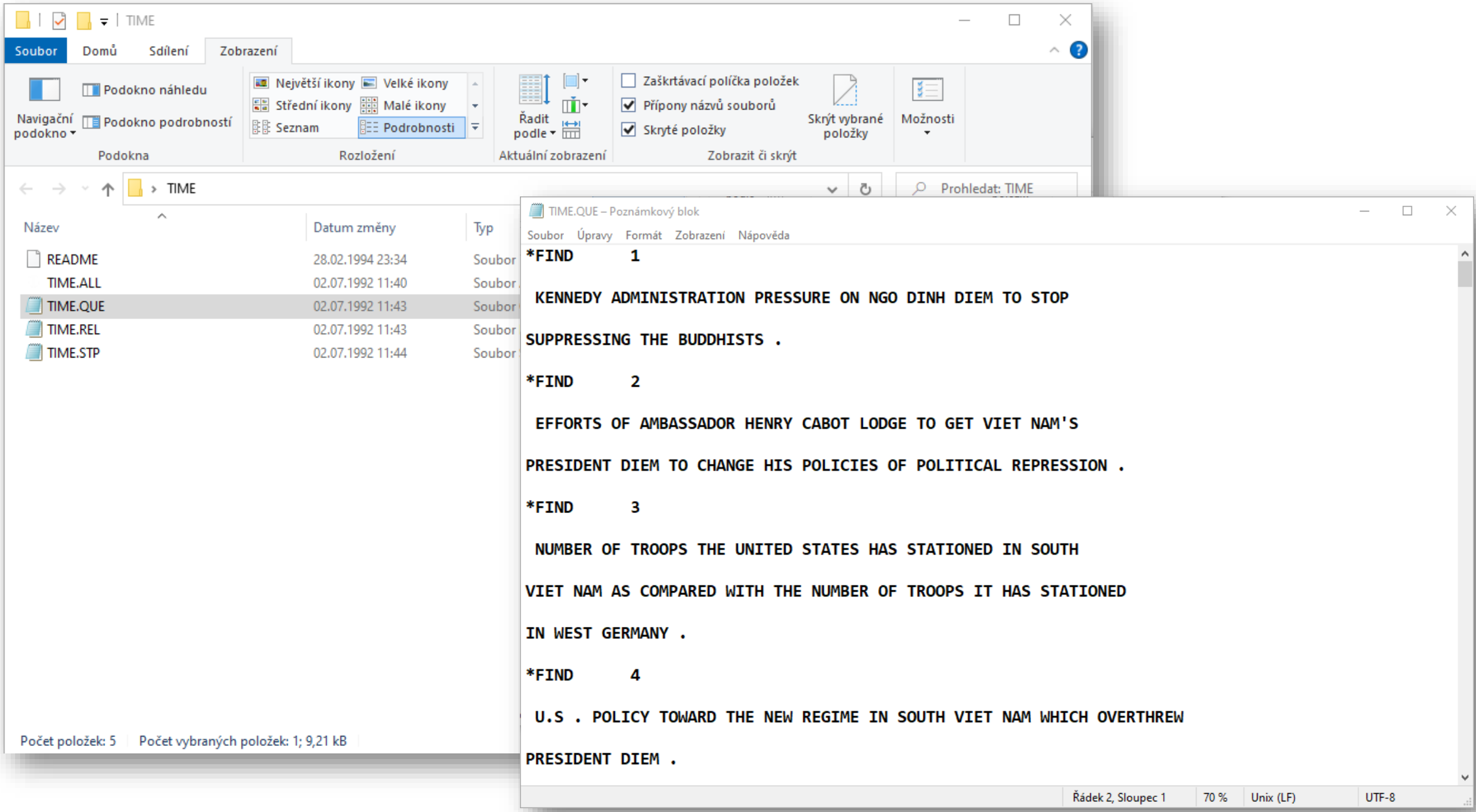
Jak můžeme posoudit, který model je lepší než jiný model?

Hodnocení modelů

- ustálené kolekce dokumentů
- přiřazené otázky a informační potřeby
- http://web.eecs.utk.edu/research/lsi/corpa.html
- *jak vypadá takový testovací korpus? - TIME*
- http://www.statmt.org/europarl/


TIME
Soubor
Domů
Sdílení
Zobrazení
Podokno náhledu
Podokno podrobností
Navigační podokno
Podokna
Největší ikony
Velké ikony
Střední ikony
Malé ikony
Seznam
Podrobnosti
Rozložení
Řadit podle
Zaškrtávací políčka položek
Přípony názvů souborů
Skryté položky
Skrýt vybrané položky
Možnosti
TIME
Název
Datum změny
README
28.02.1994 23:34
TIME.ALL
02.07.1992 11:40
TIME.QUE
02.07.1992 11:43
TIME.REL
02.07.1992 11:43
TIME.STP
02.07.1992 11:44
Počet položek: 5
Počet vybraných položek: 1; 9,21 kB
TIME.ALL – Poznámkový blok
Soubor Úpravy Formát Zobrazení Nápověda
*TEXT 017 01/04/63 PAGE 020
THE ALLIES AFTER NASSAU IN DECEMBER 1960, THE U.S . FIRST
PROPOSED TO HELP NATO DEVELOP ITS OWN NUCLEAR STRIKE FORCE . BUT EUROPE
MADE NO ATTEMPT TO DEVISE A PLAN . LAST WEEK, AS THEY STUDIED THE
NASSAU ACCORD BETWEEN PRESIDENT KENNEDY AND PRIME MINISTER MACMILLAN,
EUROPEANS SAW EMERGING THE FIRST OUTLINES OF THE NUCLEAR NATO THAT THE
U.S . WANTS AND WILL SUPPORT . IT ALL SPRANG FROM THE ANGLO-U.S .
CRISIS OVER CANCELLATION OF THE BUG-RIDDEN SKYBOLT MISSILE, AND THE
U.S . OFFER TO SUPPLY BRITAIN AND FRANCE WITH THE PROVED POLARIS (TIME,
DEC . 28) . THE ONE ALLIED LEADER WHO UNRESERVEDLY WELCOMED THE POLARIS
OFFER WAS HAROLD MACMILLAN, WHO BY THUS KEEPING A SEPARATE NUCLEAR
DETERRENT FOR BRITAIN HAD SAVED HIS OWN NECK . BACK FROM NASSAU, THE
PRIME MINISTER BEAMED THAT BRITAIN NOW HAD A WEAPON THAT " WILL LAST A
GENERATION . THE TERMS ARE VERY GOOD . " MANY OTHER BRITONS WERE NOT SO
SURE . THOUGH THE GOVERNMENT WILL SHOULDER NONE OF THE $800 MILLION
DEVELOPMENT COST OF POLARIS, IT HAS ALREADY POURED $28 MILLION INTO
SKYBOLT AND WILL HAVE TO SPEND PERHAPS $1 BILLION MORE FOR A FLEET OF
Řádek 13, Sloupec 45
50 %
Unix (LF)
UTF-8


TIME
Soubor
Domů
Sdílení
Zobrazení
Podokno náhledu
Podokno podrobností
Navigační podokno
Podokna
Největší ikony
Velké ikony
Střední ikony
Malé ikony
Seznam
Podrobnosti
Rozložení
Řadit podle
Aktuální zobrazení
Zaškrtávací políčka položek
Přípony názvů souborů
Skryté položky
Skrýt vybrané položky
Zobrazit či skrýt
Možnosti
TIME
Prohledat: TIME
Název
Datum změny
README 28.02.1994 23:34
TIME.ALL 02.07.1992 11:40
TIME.QUE 02.07.1992 11:43
TIME.REL 02.07.1992 11:43
TIME.STP 02.07.1992 11:44
Počet položek: 5
Počet vybraných položek: 1; 9,21 kB
TIME.STP – Poznámkový blok
Soubor Úpravy Formát Zobrazení Nápověda
A
ABOUT
ABOVE
ACROSS
ACTUALLY
ADD
ADDED
AFTER
AGAIN
AGAINST
AGO
ALL
ALMOST
ALONG
ALREADY
Řádek 1, Sloupec 1
60 %
Unix (LF)
UTF-8


TIME
Soubor
Domů
Sdílení
Zobrazení
Podokno náhledu
Podokno podrobností
Navigační podokno
Podokna
Největší ikony
Velké ikony
Střední ikony
Malé ikony
Seznam
Podrobnosti
Rozložení
Řadit podle
Aktuální zobrazení
Zaškrtávací políčka položek
Přípony názvů souborů
Skryté položky
Skrýt vybrané položky
Zobrazit či skrýt
Možnosti
TIME
Prohledat: TIME
Název
Datum změny
Typ
README 28.02.1994 23:34 Soubor
TIME.ALL 02.07.1992 11:40 Soubor
TIME.QUE 02.07.1992 11:43 Soubor
TIME.REL 02.07.1992 11:43 Soubor
TIME.STP 02.07.1992 11:44 Soubor
Počet položek: 5
Počet vybraných položek: 1; 9,21 kB
TIME.QUE – Poznámkový blok
Soubor Úpravy Formát Zobrazení Nápověda
*FIND 1
KENNEDY ADMINISTRATION PRESSURE ON NGO DINH DIEM TO STOP
SUPPRESSING THE BUDDHISTS .
*FIND 2
EFFORTS OF AMBASSADOR HENRY CABOT LODGE TO GET VIET NAM'S
PRESIDENT DIEM TO CHANGE HIS POLICIES OF POLITICAL REPRESSION .
*FIND 3
NUMBER OF TROOPS THE UNITED STATES HAS STATIONED IN SOUTH
VIET NAM AS COMPARED WITH THE NUMBER OF TROOPS IT HAS STATIONED
IN WEST GERMANY .
*FIND 4
U.S . POLICY TOWARD THE NEW REGIME IN SOUTH VIET NAM WHICH OVERTHREW
PRESIDENT DIEM .
Řádek 2, Sloupec 1
70 %
Unix (LF)
UTF-8

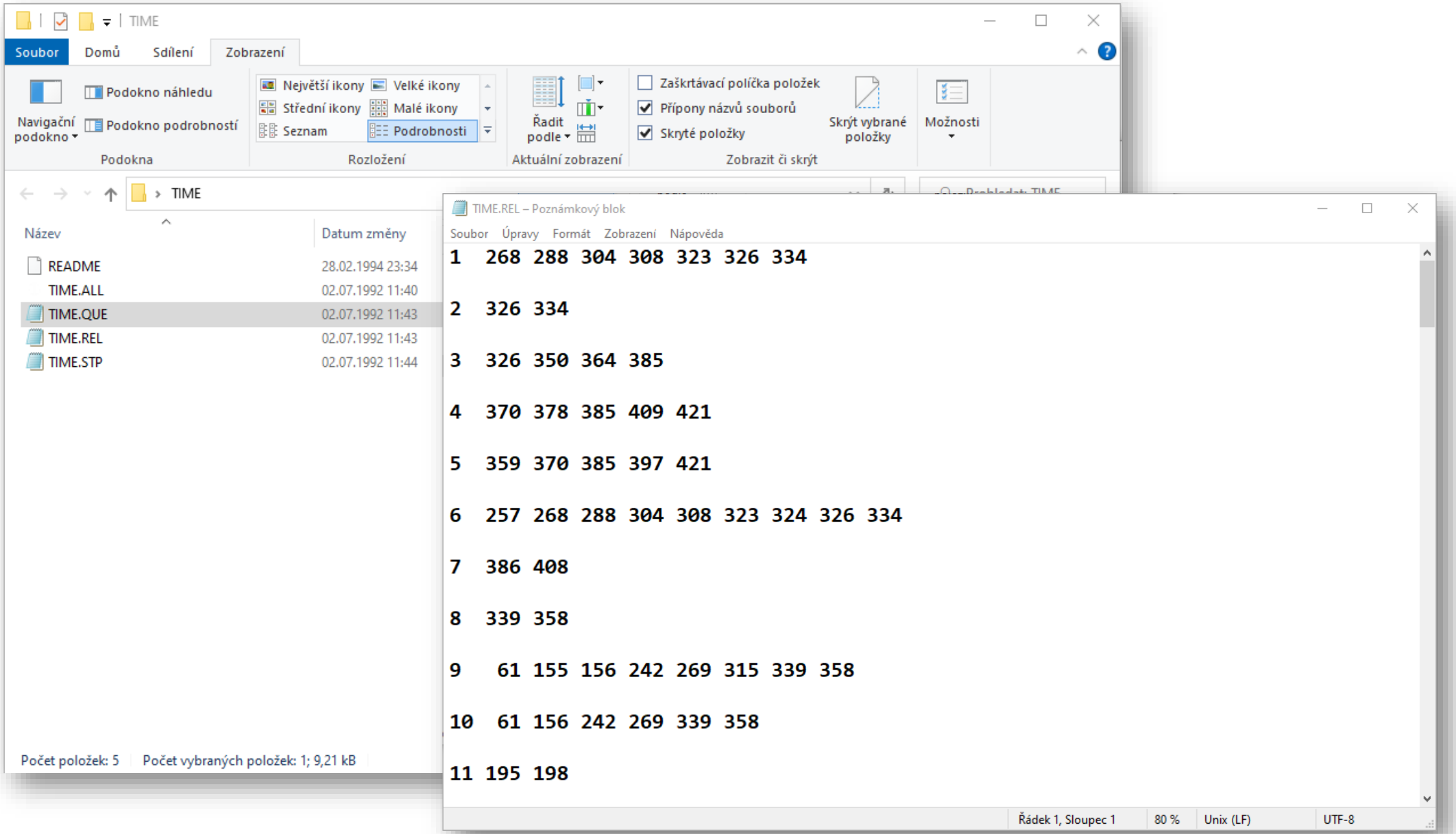

TIME
Soubor
Domů
Sdílení
Zobrazení
Podokno náhledu
Podokno podrobností
Navigační podokno
Podokna
Největší ikony
Velké ikony
Střední ikony
Malé ikony
Seznam
Podrobnosti
Rozložení
Řadit podle
Aktuální zobrazení
Zaškrtávací políčka položek
Přípony názvů souborů
Skryté položky
Skrýt vybrané položky
Zobrazit či skrýt
Možnosti
TIME
Název
Datum změny
README 28.02.1994 23:34
TIME.ALL 02.07.1992 11:40
TIME.QUE 02.07.1992 11:43
TIME.REL 02.07.1992 11:43
TIME.STP 02.07.1992 11:44
Počet položek: 5
Počet vybraných položek: 1; 9,21 kB
TIME.REL – Poznámkový blok
Soubor Úpravy Formát Zobrazení Nápověda
1 268 288 304 308 323 326 334
2 326 334
3 326 350 364 385
4 370 378 385 409 421
5 359 370 385 397 421
6 257 268 288 304 308 323 324 326 334
7 386 408
8 339 358
9 61 155 156 242 269 315 339 358
10 61 156 242 269 339 358
11 195 198
Řádek 1, Sloupec 1
80 %
Unix (LF)
UTF-8

Table 1.2: SMART test collections

Collection	Year	#Docs	#Questions	why built
Cran-2*	1964	1398	225	compare indexing methods
Cran-1*	1964	200	42	Cranfield subset of 42 questions
Cran424	1970	424	155	Cornell "reduced" subset
IRE-3	1965	780	34	indexing/dictionary experiments
ADI	1965	82	35	doc length experiments
ISPRA	1967	1268	48	multiple relevance judgements
ISPRA	1968	1095/468	48	CLIR English/German
MED273	1967	273	18	comparison to Lancaster
MED450	1970	450	30	"corrected" Medlars
MEDLARS	1970	1033	30	larger medical collection
OPHTH.	1970	853	30	specific medical domain
TIME	1970	425	83	full text articles
NPL*	1970	11429	93	indexing experiments
INSPEC*	1982	12684	84	indexing, Boolean
CACM	1982	3204	52	additional metadata
ISI/CISI	1982	1460	76	co-citations

Uklidňovací vyhledávačka

Z jakých měst pocházeli komici,
kteří inspirovali dvojici kocourů
v animovaném filmu, kde se poprvé
objevil také slavný ptáček Tweety?

Příště...

- hodnocení výsledků vyhledávání
- hodnocení z pohledu uživatele
- hodnocení informací obecně
- prostě budeme hodnotit...

Mgr. Tomáš Marek
marek@kisk.cz