Fingerprinting
případová studie

8. 10. 2014 Josef Šlerka, Socialbakers
Konference Big Data, Bratislava

Kdo jsem...

stručné představení...

Josef Šlerka

Head of R&D v Socialbakers

V minulosti šéf Ataxo Interactive

Stojím za projekty jako je Social Insider a další

Vedu Studia nových médiá na FF UK

Big Data

stručné vymezení pole

Big Data - 3V

1. Volume - velký objem dat

2. Variety - data mají různou povahu a formu, včetně obrázků či videa

3. Velocity - obsah dat se konstatně mění, přicházejí často ve streamu nebo bulkových kolekcí apod.

Small Data - Big Data

small Data - vetšinou designovaná pro konkrétní úkol, vysoce strukturovaná, z jedné domény, přicházejí z definovaných zdrojů, obvykle mají krátký časový cyklus zachování apod.

Big Data - není vždy jasné plné pole, čeká se nejmenší granularita, často nestrukturovaná, z vícero zdrojů

(zajímavé počtení: Jules J. Berman: Principles of Big Data)

Big Data - note

Big Data možná není dobré definovat velikostí, ale úplností. Nechceme uchovávat primárně agregace, ale všechna data. Pro výzkum nesamplujeme, ale pracujeme z celkem apod.

Big data je spíše pojmenování pro specifický typ datových problémů.

Social Insider a Listening

z pohledu “Big Data”

Social Insider & co.

Social Insider je social media monitoring pro český a slovenský trh, který ma světovou mutací Social Listening.

Technologicky využíváme Amazon, Elasticsearch, Redis, Ruby a další drobnosti.

Drobné problémy

Datové zdroje jsou zcela asynchronní (stream api v frekvenční api). Tlak je proměnlivý podle typu události.

Data jsou realativně velká (půl miliardy zmínek konstatně).

Není předem jasné použití. Nejsou tu jasné reporty.

Specifické požadavky

Asynchronost dat kupříkladu jinak postavený v interface, nelze kupříkladu stránkovat. (ukázka)

Drill-down menu je analytikou, nikoli navigací.

Není jasné kolik dat se bude exportovat, je proto třeba exportů na pozadí.

Specifický problém při aggregacích spojený s konstatním tlakem dat.

ATAXO SOCIAL INSIDER

Logged in as: Ataxo Social Insider

Mentions | Analytics | Settings | Exports

SOURCES

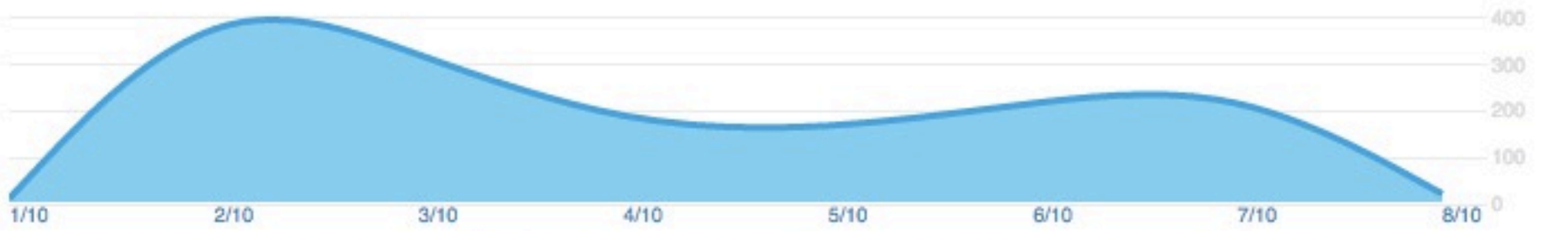
- All
- Blogs (4)
- Discussions (31,329)
- Facebook (5,544)
- Feeds (5)
- Forums (590)
- Google plus (217)
- Instagram (21)
- News (1,212)
- **Twitter** (1,636)
- Yelp (1)
- Youtube (10)
- Trash

TOPICS

- All
- Já (5)
- Univerzita (44)
- Ataxo (83)
- Konkurence (0)
- Instagram (393)
- TweetRoku (0)
- Karlín (11)
- Romové (201)
- PRISM (26)
- Strana zelených (4)
- Islam (108)
- Experiment (81)
- Vánoce (261)
- Kulturní souboj (9)

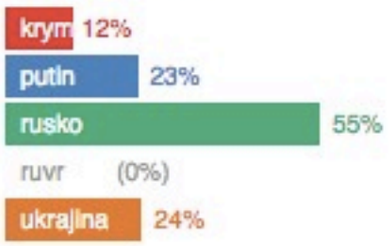
Search within currently displayed mentions

Show search tips



Sources

Top 5 Keywords



Sentiment 100.0% missing

Top 5 Authors

Zprávy : 170 (10.39%)

Hlas Ruska : 109 (6.66%)

News 24h Cze : 81 (4.95%)

Lukáš Rojko : 19 (1.16%)

Michael Roma... : 18 (1.10%)

Top 5 mentions

Auuu: "Uvěznili mě na 3 roky za nesouhlas se sov. invazí ČSSR. A CZ prezident chce, aby... (116)

Jde o putovní masový hrob? 1. Kolumbie před 4 lety: 2. Ukrajina dnes: (33)

Na ČT24 dali titulek "Krym, Rusko". To se mi jen zdá? Uznali jsme snad anexi? (28)

Český vývoz do Ruska za prvních šest měsíců tohoto roku dle MZV klesl (držte se!) o 0,1... (16)

Brutálnímu zásahu policie a vojska v Drážďanech přihlíží i agent sovětské tajné služby,... (15)

Top 5 hashtags

#ukrajina (35)

#putin (29)

#zprávy (15)

#sankc... (15)

#ukraine (14)

Hide Charts

Hide highlight

Displaying 25 of 1,636 records

Export displayed mentions to csv

Schedule export

RT @CZSecStateEU: Vít Kolář: Sankce nejsou cíl, ale prostř… read more

08 OCT 2014 10:40 — RT @CZSecStateEU: Vít Kolář: Sankce nejsou cíl, ale prostředek k tomu přimět **Rusko** k civilizovanému postoji. Debata k #sankce s @brestan …

08 OCT 2014 01:44 — Ladislav Jakl: Zastavme rozpínavost **Ukrajiny** ! http://t.co/yjBYESTs6l

AUTHOR: MARTIN PÁNEK | SOURCE: TWITTER | DOMICILE: CS,SK

FOLLOWERS: 913 | FRIENDS: 324 | KLOUT: 62

Human Rights Watch: Násilná mizení na **Krymu** : http://t.co/Ubhb6ZImps

08 OCT 2014 01:41 — Human Rights Watch: Násilná mizení na **Krymu** : http://t.co/Ubhb6ZImps

AUTHOR: NEWS 24H CZE | SOURCE: TWITTER | DOMICILE: CS,SK

FOLLOWERS: 192 | FRIENDS: 255 | KLOUT: 33

Udo Ulfkotte: „Německá a americká média zkouší přinést lidu… read more

08 OCT 2014 00:45 — Udo Ulfkotte: „Německá a americká média zkouší přinést lidu Evropy válku a vnést válku do **Ruska** . To je bod, za kterým už není návratu." #CIA

AUTHOR: STŘEDNÍ TŘÍDA V ČR | SOURCE: TWITTER | DOMICILE: CS,SK

FOLLOWERS: 333 | FRIENDS: 312 | KLOUT: 50

Loading...

Specifické požadavky

Asynchronost dat kupříkladu jinak postavený v interface, nelze kupříkladu stránkovat. (ukázka)

Drill-down menu je analytikou, nikoli navigací.

Není jasné kolik dat se bude exportovat, je proto třeba exportů na pozadí.

Specifický problém při aggregacích spojený s konstatním tlakem dat.

Nejvíc sdílený obsah

Etuda na téma podobné příspěvky aneb fingerprinting

Požadavek

najdi mi nebo posty, které jsou si hodně podobné:
“třeba tak na 90% nebo tak nějak”

příklad:

#Rusko poprvé soudí žoldáka za válčení v řadách separatistů na Ukrajině. Média ho označují za nacionalistu a fašistu. http://t.co/7JtHsc3YV0

RT @Aktualnecz: #Rusko poprvé soudí žoldáka za válčení v řadách separatistů na Ukrajině. Média ho označují za nacionalistu a fašistu. http:…

Klasika

Levenstheinova distance, NCD a další nearest neighbor methods

Levensthein & Co.

Hammingova distance - počet substitucí znaků, které je nutno změnít aby se jeden řetezec proměnil v druhý (předpokládá se stejná vzdálenost)

Levenstheinova distance - počet substitucí, vložení a smazání které je třeba pro změnu jednoho řetězce v druhý (řetězce mohou být různě dlouhé)

NCD

Astraktní měření vzdálenosti řetězců pomocí kompresí.

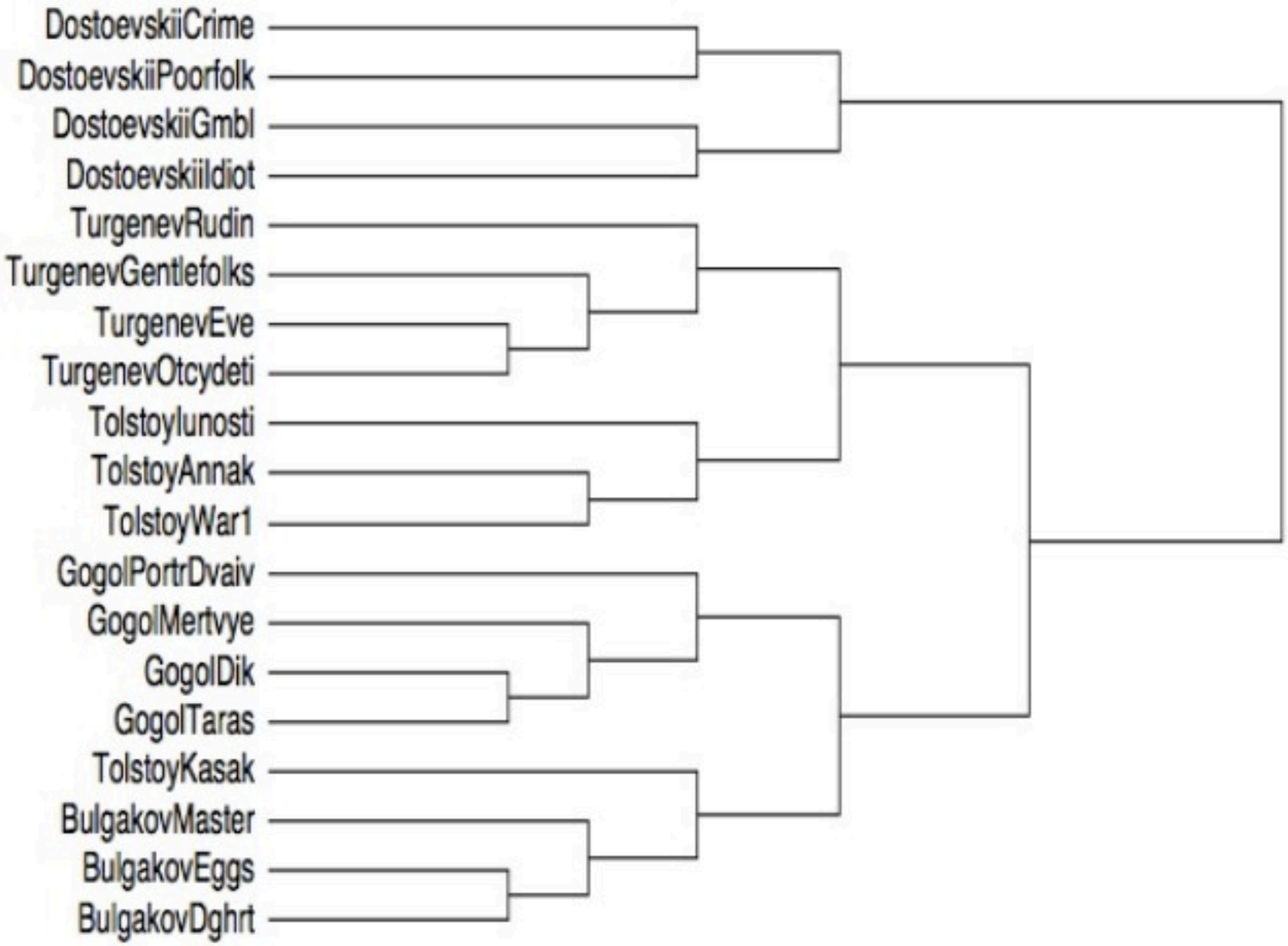
autory jsou Rudi Cilibrasi a Paul M. B. Vitanyi

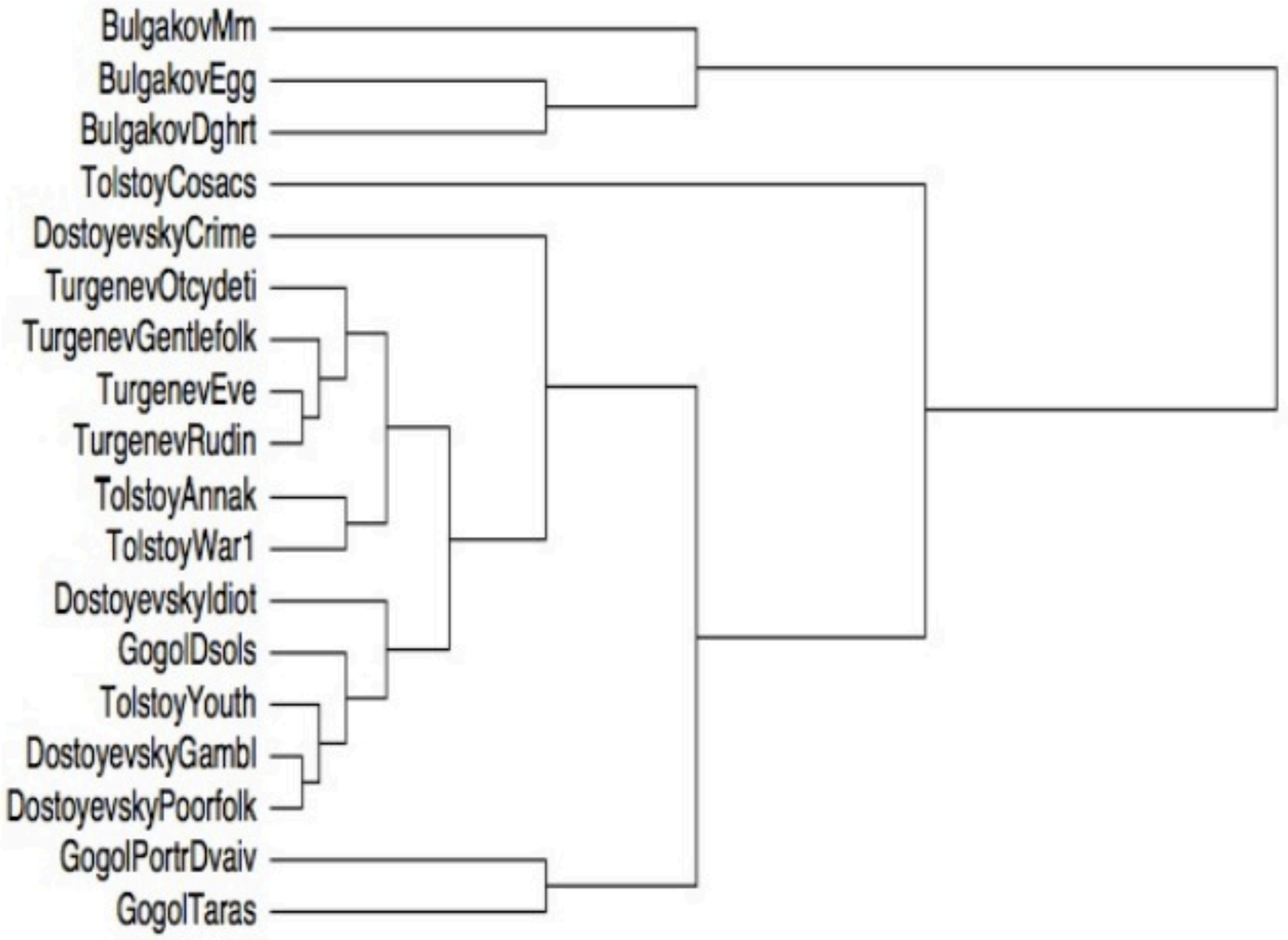
podobné věci sdílí stejné vlastnosti

dvě reprezentace jsou si tím podobnější, čím méně složitých změn je třeba k převodu jedné v druhou

NCD

$$NCD(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}.$$


DostoevskiiCrime
DostoevskiiPoorfolk
DostoevskiiGmbl
Dostoevskiildiot
TurgenevRudin
TurgenevGentlefolks
TurgenevEve
TurgenevOtcydeti
Tolstoylunosti
TolstoyAnnak
TolstoyWar1
GogolPortrDvaiv
GogolMertvye
GogolDik
GogolTaras
TolstoyKasak
BulgakovMaster
BulgakovEggs
BulgakovDghrt


BulgakovMm
BulgakovEgg
BulgakovDghrt
TolstoyCosacs
DostoyevskyCrime
TurgenevOtcydeti
TurgenevGentlefolk
TurgenevEve
TurgenevRudin
TolstoyAnnak
TolstoyWar1
DostoyevskyIdiot
GogolDsols
TolstoyYouth
DostoyevskyGambl
DostoyevskyPoorfolk
GogolPortrDvaiv
GogolTaras

Height

0.925 0.965

South.Park.S13E02.DSR.XviD-0TV.txt

South.Park.S13E05.DSR.XviD-0TV.txt

South.Park.S13E03.DSR.XviD-0TV.txt

South.Park.S13E06.HDTV.XVID-BAJSKORV.txt

how.i.met.your.mother.s01e01.txt

how.i.met.your.mother.s01e04.txt

how.i.met.your.mother.s01e02.txt

how.i.met.your.mother.s01e03.txt

1x03-The_Big_Bang_Theory-The_Fuzzy_Boots_Corollary.txt

1x01-The_Big_Bang_Theory-Pilot.txt

1x02-The_Big_Bang_Theory-The_Big_Bran_Hypotesis.txt

1x04-The_Big_Bang_Theory-The_Luminous_Fish_Effect.txt

Cluster Dendrogram

Pro a proti

rostoucí komplexita: n(n-1)/2

3000 řádku vyžaduje 4.5 million kalkulací

proměnlivá velikost textů

jde optimalizovat, ale POZOR máme tu konstatní tlak dat a případný clustering musíme dělat on-the-fly

umožňuje jemnější klastrování, někdy ale není jasné proč

Znovu a lépe

key collision methods
aneb fingerprinting (a taky perceptual hashing)

Základní idea

vytvořit takovou reprezentaci obsahu, která bude různé podoby obsahu normalizovat do jednoho klíče

de facto je to převrácená podoba hashování

místo co nejkratší reprezenace unikátního obsahu tu chcem mít reprezentaci, která postihne co největší variablitu obsahu

podobné příspěvky pak získáváme třeba agregací

Příklady

Fingerprint

N-Gram Fingerprint

findimagedupes.pl

Fingerprint

převést na běžnou ASCII reprezentac (zbabit je diakritiky)

odmazat všechny přebytečné prázdné mezery

všechny znaky změnit na malé

odmazat všechny interpunkční znaménka

rozdělit řetězec do samostatnách tokenů přes prázdné mezery

seřadit tokeny podle abecedy a odmazat duplikující se

spojit je všechny opět do jednoho řetězce

Detaily

řeší specifický problém s diakritikou a šumem znaků

pořadí slov v textu je pro něj nerelevatní

možnosti vylepšení: odstranění speficických slov (jak stopwords, tak RT či via), lemmatizace či stemming

možnostu ušteření místa pomocí klasického hash

N-Gram Fingerprint

převeď na ASCII reprezentaci

změň všechny znaky na malé

odmaž všechnu interpunkci a všechny speciální znaky

získej seznam n-gramů

spoje je dohromady

Detaily

příklad: “bratislava” ma 1-gram “abilrstv”

oproti předchozímu ve variantě s 1-gramy a 2-gramy

dokáže řešit specifické varianty: "Krzysztof", "Kryzysztof" a "Krzystof" mají stejný 1-gram fingerprint

findimagedupes.pl

standartizuje velikost na 160x160

převeď na škálu šedé

rozmazej trochu

normalizuje barevnou intenzitu

zvyš na maximum kontrast

znovu přesampluj na 16x16 a převeď na mono

vezmi 32 bytes obrazku a máš fingerprint

Pro a proti

výkon, výkon, výkon! podobný obsah se dostává pouhou agregací a šetrně

ale je to poměrně hrubá podobnost

lze však různě vylepšovat

Co to spojit?

Simahash - perceptualní hashing

Perceptual hash

vytvořit takový hash řetězce, aby podobné řetězce měli podobné hashe

příklad SimHash používaný Googlem a vyvinutý Mosesem Charikarem

mnoho paperů a článků, jeden z nich: http://matpalm.com/resemblance/simhash

Příklad z něj

```
irb(main):006:0> p1 = 'the cat sat on the mat'
irb(main):005:0> p2 = 'the cat sat on a mat'
irb(main):007:0> p3 = 'we all scream for ice cream'
irb(main):007:0> p1.hash
=> 415542861
irb(main):007:0> p2.hash
=> 668720516
irb(main):007:0> p3.hash
=> 767429688
```

klasicky nám dává hash různé hashe pro různé stringy

Příklad z něj

```
irb(main):006:0> p1 = 'the cat sat on the mat'
irb(main):005:0> p2 = 'the cat sat on a mat'
irb(main):007:0> p3 = 'we all scream for ice cream'
irb(main):007:0> p1.hash
=> 415542861
irb(main):007:0> p2.hash
=> 668720516
irb(main):007:0> p3.hash
=> 767429688
```

```
irb(main):003:0> p1.simhash
=> 851459198
00110010110000000011110001111110

irb(main):004:0> p2.simhash
=> 847263864
00110010100000000011100001111000

irb(main):002:0> p3.simhash
=> 984968088
00111010101101010110101110011000
```

simhash dává podobné hashe pro podobné řetězce

Příklad z něj

```
irb(main):006:0> p1 = 'the cat sat on the mat'
irb(main):005:0> p2 = 'the cat sat on a mat'
irb(main):007:0> p3 = 'we all scream for ice cream'

irb(main):003:0> p1.simhash
=> 851459198
00110010110000000011110001111110

irb(main):004:0> p2.simhash
=> 847263864
00110010100000000011100001111000

irb(main):002:0> p3.simhash
=> 984968088
00111010101101010110101110011000
```

hammingova vzdálenost podobného párů (p1,p2)=4

zatím co (p1,p3)=16 a (p2,p3)=12

XOR přítel

10101011100010001010000101111100

XOR 10101011100010011110000101111110

= 00000000000000001010000000000010

máme tak 3 změny z 32, neboli odhadovaný rozdíl 3/32 tedy 0.09375

nebo chcete-li 1 - 3 / 32 = 0,90625 (90%)

(viz třeba blogpost: http://www.titouangalopin.com/blog/articles/2014/05/simhash-or-the-way-to-compare-quickly-two-datasets)

Mimochodem

mimochodem u fingerprintu findimagedupes.pl funguje XOR stejně:-)

Integrace...

... zpět k technoligii

More like this...

Lucene based databáze podporují možnosti najdi mi podoné příspěvky jako je tento, které jsou v Levenstheinu vzdáleny o ...

V případě stejně dlouhý řetezců jako je binarní reprezentace simhash je to pak vlastně hammingova distance.

Děkuji za pozornost!

@josefslerka