Zápisky z akce

The Big Clean

Obsah

Úvodní informace

19.3.2011 - NTK

Oficiální web: http://bigclean.org/praha/
IRC: http://bit.ly/okfn-irc
Doplňkový GDoc (otázky, diskuze): http://bit.ly/otevrenadata

Získávání a čištění dat

Screen-scraping

Jindřich Mynarz (NTK) - ScraperWiki

O scrapování
Scraper - převádí webové stránky na data.

Kroky:

- Získat data (stažení)
- Parsování
- Extrakce
  - HTML
    - HTMLTidy (úprava HTML)
    - DOM
  - text
    - regulární výrazy

Zodpovědné scrapování
Don't be evil :) - ať na vás nikdo nepřijde.

Jak na to

- Omezte počet HTTP požadavků.
- Časově rozložte požadavky.
- Používat cache.

Podmínky scrapování

- Jsme oprávněni získat ty informace?
- Respektujte licence.
- Respektujte robots.txt.

Nástroje

- Needlebase
  - umožní používat stránku jako by měla API
- Yahoo Query Language (YQL)

    - umožní dotazovat se jako při SQL
- Google Spreadsheet
    - importHTML()
- ScraperWiki
    - i hosting, verzování, historie, hostovaná databáze (SQLite), programovací jazyky Python, Ruby, PHP
    - pracuje s různými formáty: HTML, CSV, XLS, PDF
    - ke stažení jako CSV, XML, JSON, atp. (přístupné i přes API)
    - intervaly lze nastavovat (default = 1 denně)
    - k nekomerčnímu použití zdarma / komerční dle domluvy
    - umí i vizualizaci dat
    - Zajímavosti:
        - podporován z fondů Channel 4 -> určená i pro novináře
        - v beta už asi 3 roky
        - existuje trh scraperů (poptávka - odměny)

Práce se Scraperwiki

Tutoriály - http://scraperwiki.com/help/tutorials/
Webové rozhraní pro přímé spouštění/ladění skriptů

Ukázkový příklad *Znečištění ovzduší*

- Zdroj
    - http://www.chmi.cz/files/portal/docs/uoco/web_generator/actual_hour_data_CZ.html
- Scraper
    - http://scraperwiki.com/scrapers/air_pollution/

Více odpoledne (hackování). Bodou se scrapovat datasety veřejné správy.

Čištění dat

Jakub Nešetřil (GoodData) - Google Refine

- volně dostupný nástroj (open-sourced by Google)
- nástroj pro čištění dat
- ideální pro osobní použití
- není potřeba programovat (snadné jako Excel :)
- podívat se na výuková videa

Prakticky

Dataset:

- http://aplikace.mvcr.cz/auta/stazeni.html
    a. Neplatné doklady
    b. Odcizená vozidla

Facety - pro filtrování (nejdůležitější prvek).

Možnost rozdělit sloupec podle dalšího oddělovače (u celého souboru to byl znak |, u konkrétního sloupce mezera).
Cluster - automaticky najde příbuznost slov - spojí je do skupin

- dobré pro shlukování dat plných překlepů

Pokročilejší práce s hodnotami - Google Refine Expression Language + reg. výrazy (Python, Clojure...)

Např. výraz value.split(‘ ‘)[0] vybere první slovo (po mezeru)

Histogram hodnot (zobrazit pomocí *Facet by choice counts*)

Hodnoty jsou vlastně pole - pomocí value[3] přistoupíme ke 4. znaku.
Stahování dat přes URL (doplňování sloupců).

Štefan Urbánek - čištění a kvalita dat

// @Stiivi, stefan.urbanek@gmail.com

Background

- Otvorený vestník verejného obstarávania
- Brewery - framework pro zpracování toku dat (inspirace v SPS Clementine)
- Cubes - analytický framework

Znaky kvalitních dat:

- kompletnost (základní ukazatel)
  - kontrolovat správnost parseru
  - snadno automaticky měřitelné
- přesnost (odpovídají skutečnému světu, ověřovat nesmysly)
  - hůře automaticky měřitelné
- důvěryhodnost
  - nakolik se datům dá věřit (např. cena rohlíku 1 Kč vs. 1 EUR)
  - špatně automaticky měřitelné
- aktuálnost
- konzistence
- integrita

Co jsou to kvalitní data? - např. 85 % dle znaků výše - nutno měřit.
Doporučení používat sondy - ověřování kvality dat v čase - upozornění na změny ve zdrojích => nutná změna algoritmu.

Používat reporty kvality dat.

Extrakce a Transformace
HTML dokumenty

- často narušená struktura (i u tabulek), ošklivé věci jako třídy a

in-line styly (např. uppercase) mající sémantický význam.

Spreadsheety - problémy

- Obrázky a nadpisy
- Opakování skupin sloupců
- Redundance
- Vizuální formátovnání (barvy místo příznaků)
- (Víceřádkové) buňky pomocí rámečků

Nástroje vybírat dle povahy dat a měřit kvalitu dat.

Otevřená data ve veřejné správě

Diskuze

Martin Nečaský

Odkaz na dokument z místosti, kde se konala diskuse.

Využití dat

"Analýza" dat

Josef Šlerka (STUNOME, Ataxo)

Google

site:mvcr.cz ext:xls - všechny excely na stránkách MVČR

Google Docs

ImportXML()
ImportHTML()
ImportFeed()

Podrobnosti viz:
Google Spreadsheets: Funkce -> Další -> Google
Functions: Functions for external data
Functions: Function list

Dotazovací jazyk je XPath

YQL

- Dotazovací jazyk (ala SQL) - dialekt pro dotazování nad webovými

stránkami
- výstupy v XML nebo JSONu
- v doplní části je odkaz pro volání dotazu pomocí URL
- parsuje mikroformáty (např. stánka profilu na twitteru)
- mapování dalších připravených datasetů:
  - Social Graph
  - last.fm
  - salesforce
  - atd...

Dotazy

Jednoduchý dotaz:

```
select * from html where url = “http://www.novinky.cz”
```

taky se kombinuje s XPath:

```
select * from html where url="http://suchosch.net" and xpath='//li/a'
```

zpřesnění obsahu:

```
select content  from html where url="http://gug.cz" and xpath='//li/a'
```

umí i joiny:

```
select * from search.web where query in (select content  from html where url="http://codeasi.net" and xpath='//li/a')
```

parsování míst z textu (vrátí objekt s adresou)

```
select * from geo.placefinder where text="Hlavní město je třeba Praha"
```

možnost získávat konkrétní entity

```
select p.content, span.content from html where url = “http://seznam.cz”
```

práce s mikroformáty

```
select * from microformats where url=”http:/twitter.com/bigcleancz”
```

Yahoo Pipes

Vizuální prostředí pro neprogramátorskou tvorbu mashupů.

Př. 1) spojení RSS novinky.cz a idnes.cz
- seřazení podle času

Př. 2) twitter CNN a @BreakingNews
- vizualizace na mapu

Další možnosti
- celé použít jako RSS, JSON, gadget, zasílat emailem
- taky podporuje reg. výrazy - např. se do mixu RSS dá vepsat název zdroje do titulku apod.
- hotové “trubky” se dají použít jako součást dalšího projektu

Google Fusion Tables

- “hovadsky a zvířecky” zpracuje až 100MB .csv
- upload souboru nebo import z Google Spreadsheets
- spousta zajímavých veřejných dat (Public tables)
- umí SQLite dotazy do hostované tabulky

- Vizualizace
  - Mapa (i intenzity)
  - Tabulka
  - Grafy
  - Storyline (chronologické řazení)

Př.: Pokutované pumpy a Google Fusion Tables

Social Network Analysis

Př.: Magistrát hl. m. Prahy - společné hlasování
Další zajímavé příklady: http://tlampac.webnode.cz/ (a má to i Facebook!)

Využití dat v žurnalistice

Adam Javůrek (Next Big)

Data - Filter - Visualize - **Story**

Ukázky

- Zdroj: bit.ly/odkazy-datazurba

Mapa světa dle průměrné délky penisu

- http://www.targetmap.com/viewer.aspx?reportId=3073
- "bulvární téma" - velký zájem, ale pochybná zdrojová data (některé údaje získány měřením, některé dotazníkem apod.)

Campaign Finances

- http://elections.nytimes.com/2008/president/campaign-finance/map.html
- zajímavé, ale moc toho neříká

Oakland crimespotting

- http://oakland.crimespotting.org/
- Zobrazení různých druhů zločinu, dle času, na mapě. Možnost nechat zasílat upozornění na email.

Spiegel - iraq war logs

- http://www.spiegel.de/international/world/0,1518,724000,00.html
- Z Wikileaks - velké množství špatně uchopitelných dat. Spiegel vizualizoval na mapě, s možností filtrace

Deadly day in baghdad

- http://www.nytimes.com/interactive/2010/10/24/world/1024-surge-graphic.html
- smrtelné konflikty: bagdad / jeden den

CNN War Causalities - Home and Away

- http://edition.cnn.com/SPECIALS/war.casualties/index.html
- Spojení bydliště zemřelého vojáka s místem jeho smrti (z dat vznikne příběh)

Snake oil? - Co říkají vědecké důkazy o zdravotních doplňcích

- www.informationisbeautiful.net/play/snake-oil-supplements/
- Velmi náročné na tvorbu
- Čím více je doplněk stravy nahoře, tím více opravdu pomáhá dle vědeckých studií. Vpravo možno zvolit konkrétní zdravotní problém.

Vizualizace projevu

- http://projects.datajamming.com/fun/sotu_2010_v_2011.jpg
- Vizualizace slov, z projevu Baraka Obamy

Jak hlasoval západ a východ v Eurovizi

- http://www.svd.se/multimedia/archive/00470/S__r_stade_Europa_i_470153a.swf
- Vazby při hlasování, vychází z podezření autorů, že obyvatelé východních zemí vždy hlasují pro svoje "umělce"

Vývoj rozpočtu vs realita

- http://www.nytimes.com/interactive/2010/02/02/us/politics/20100201-budget-porcupine-graphic.html
- Amanda Cox - specialistka NY Times na vizualizace

Jaká je nezaměstnanost ve vaší "třídě"?

- http://www.nytimes.com/interactive/2009/11/06/business/economy/unemployment-lines.html
- Vizualizace umožňující čtenáři hrát si s filtry (hledat sám sebe - vymezovat se dle pohlaví, věku, barvy pleti)

Jak vypadá váš blok baráků?

- http://projects.nytimes.com/census/2010/explorer?ref=us
- Zobrazení barevných teček dle barvy pleti obyvatel (kde je jaká rasa nejvíce zastoupená)

Jak vypadají naši mediální strašáci

- http://www.informationisbeautiful.net/play/mountains-out-of-molehills
- Propojení s Google Trends
- Související TED - David McCandless: The beauty of data visualization

Jak trávíme den

- http://www.nytimes.com/interactive/2009/07/31/business/20080801-metrics-graphic.html
- Stream graf

Z čeho se skládá inflace (Budget puzzle)

- http://www.nytimes.com/interactive/2008/05/03/business/20080403_SPENDING_GRAPHIC.html
- transformace z běžného stavebnicového grafu

Budget puzzle: dejte rozpočet do latě

- http://www.nytimes.com/interactive/2010/11/13/weekinreview/deficits-graphic.html

Public data explorer

- http://www.google.com/publicdata/directory
- Založeno na nástroji Gapminder (který koupil Google) největšího populátora statistiky Hanse Roslinga

Ukázka motion v google spreadsheet

- https://spreadsheets.google.com/ccc?key=0AiyV1MGwO21EdGMzb2xSNWs4M1YwUW5WcHNsaHk1c3c&hl=cs&authkey=CK-g5i8

Every Block

- http://chicago.everyblock.com/

Výdaje poslanců v Británii

- http://mps-expenses.guardian.co.uk/
- Crowdsourcing v praxi - analýza datasetu (kopie účtenek britských poslanců, lidé označují o co v účtence jde)

Cena za trávu

- http://www.priceofweed.com/
- crowdsourcovaná data

Koupit si dům, nebo pronajmout?

- http://www.nytimes.com/interactive/business/buy-rent-calculator.html
- Pomáhá lidem v rozhodování

Nástroje

- Many Eyes (free nástroj od IBM, je tím např. vizualizována velikost světových novin)
- OpenHeatMap
- Tableau Public
- Protovis
- Polymaps

Kde vzít data?

- DataMarket

- Get the Data
    - “fórum” kde vzít jaký dataset (UK)
- Datastore - World Government Data (Guardian)

Hrozby

- nutné porovnávat podobné hodnoty / za stejných podmínek (např. k počtu kilometrů při srovnávání potrubí)
- souvislosti - podloženost přímosti
- chyby (změna na straně “poskytovatele” dat)
- použitelnost pro uživatele - nabízet vodítko (návod viz NY Times)
    - Takhle ne! (Aktuálně.cz)
- pochopitelnost/přesnost (steam graph)
- přidaná hodnota
- zajímavost (Přidávat příběh, jinak můžeme vizualizovat telefonní seznam)

Datový žurnalista

- novinář
- programátor
- statistik
- ? (grafik?)
- ...v jednom

Co nejvíce dat!

Příklad vizualizace dat z minulých dob - mapa, odkud pocházejí lidé nakaženi cholerou -> nejvíce lidí bydlelo u pumpy -> zdroj nákazy pumpa.

Video - červená karkulka očima infografika

Jan Boček (OS Brnění)

Trocha teorie

Informace z klasické žurnalistiky s časem mizí... Datová žurnalistika by to měla vyřešit, rozložit blok textu do dále zpracovatelných dat.

Výhody datové žurnalistiky

- samonosná data
- relativně tvrdá
- wow efekt
- více příběhů najednou
- není potřeba trigger (nemusí se čekat na údernou zprávu), stačí upozorňovat na trend
- ajťák?

Zkušenosti z ČR

Přístup státu

- nástroj: stošestka (Zákon 106/1999 Sb. o svobodném přístupu k informacím)
- žádost, do 15 dnů musí odpovědět (lepší ale dát stížnost a odvolání najednou, protože stejně nereagují), pak podat žalobu
- "mimořádně" složitá data mohou být zpoplatněna (definováno)
- zaslané údaje musí být zveřejněny i na webu

Česká média

- zpoždění: nové profese a postupy?
- málo peněz
- neznalost technologií
- data: co s tím?

Ajťáci

- příležitost:
    - vizualizace
    - přerod v datového novináře
    - efektivní aktivismus

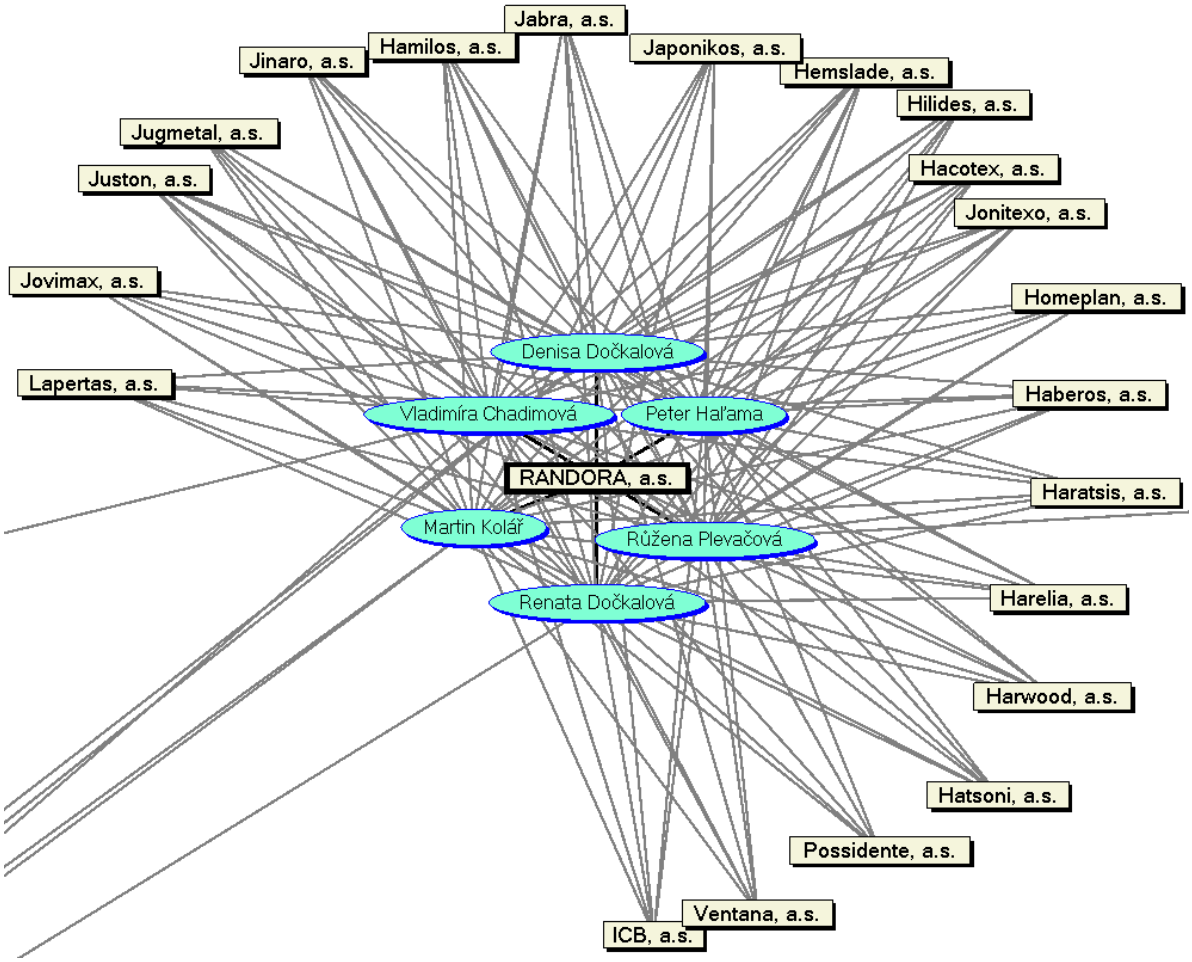
Praktická aplikace

Mapa heren

- http://www.osbrneni.cz/mapa-heren/
- mapa se rozšířila do médií - reakce politiků
    - zmizely automaty

Komu patří Jižní centrum

- http://www.osbrneni.cz/komu-patri-jizni-centrum/komu-patri-jizni-centrum
- přesun nádraží
- vizualizace katastru nemovitostí
- ORNetFace pro vizualizaci obchodního rejstříku
- hledání vztahů u majitelů nemovitostí okolo nádraží
    - RANDORA, a.s.



Odpolední hackování

Scrapování a čištění dat

Štefan Urbánek a Jindřich Mynarz

Odkaz na dokument z místosti, kde se konalo scrapování a čištění dat.

Shrnutí diskuze a workshopu

Dopolední sekce

Nějaké projekty jsou rozjeté a dokonce se kryjí. Nutná větší informovanost a spolupráce. Výstupy v dokumentu z místosti.

Odpolední sekce

Scrapery budou zveřejněny.

Účastníci rozdělení do skupin dle prog. jazyků:

- Python
  - seznam insolvenčních správců
- Ruby
  - seznam amnestií
- PHP
  - seznam veřejných sbírek
  - meteočidla
  - seznam církví
  - seznam příjemcí dotací

Více v dokumentu Hackovací odpoledne.

Další užitečné odkazy

- Data journalism developer studio
- http://datajournalism.stanford.edu/

Autoři

@pre_mysl - Přemysl Brýl (appsatori.eu, vse.gug.cz) - idea, formátování, zápisky
@codeas - Ivan Kutil (appsatori.eu, vse.gug.cz) - zápisky
@AnnaCeskova Anna Češková (snm.gug.cz) - zápisky
@codeas - Ivan Kutil (appsatori.eu, vse.gug.cz) - zápisky
@lindash - Linda Hlaváčová (vse.gug.cz) - zápisky
@suchosch - Jiří Suchomel (snm.gug.cz) - zápisky, mr. odkaz