MUNI
SCI

Bi8700 Vybrané kapitoly ze zpracování, analýzy a vizualizace dat

Strojové zpracování a analýza textových dat

Jakub Ščavnický

Obsah

- Textové dáta
- Web scraping
- Modelový príklad 1
- Modelový príklad 2
- Samostatná práca

MUNI SCI

Text

- Textové dáta v dnešnej dobe internetu sú všade
  - google
  - wikipedia
  - sociálne siete
  - články
  - fóra
  - …
- Typ zdroja dát
  - štruktúrovaný (tabuľka, databáza)
  - neštruktúrovaný (kus textu)

MUNI SCI

Textové vyhľadávanie

Chcem nájsť knihy, ktoré napísal Tolkien.

- Štruktúrovaný zdroj
  - **v DB dotaz:**
    "SELECT * FROM books WHERE author_name = 'J.R.R. Tolkien'"
  - **v Exceli filter**

  - výsledky exaktné podľa povahy DB

- Neštruktúrovaný zdroj
  - **funkcia Find**
  - **na webe:** google search "*J.R.R. Tolkien books*"

  - problémy: jazyk, sémantika, preklepy, interpunkcia, diakritika, ...

MUNI SCI

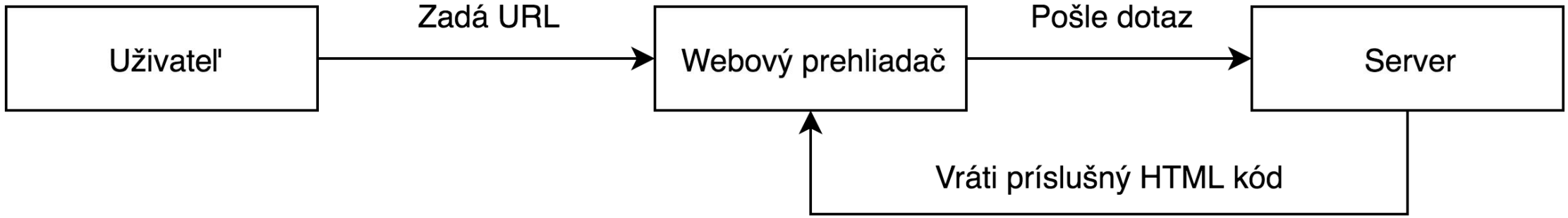
Získavanie textu z online zdrojov

- Kopírovanie a vloženie
- Stiahnutie (rôzne formáty)
- Využitie aplikačného rozhrania pre sťahovanie dát (API)
  - Facebook, Twitter, NASA, reddit, ...
- HTML (Hypertext Markup Language)
  - štandard pre tvorbu webových stránok a aplikácií
  - obsahuje dáta vo forme textu → dokument

MUNI SCI

HTML

- Hypertextový značkovací jazyk
- V pozadí webových stránok je vždy nejaká forma HTML
- Komunikácia s webovým serverom



MUNI SCI

HTML

- Webový prehliadač prekladá HTML do vizuálnej podoby pomocou HTML tagov
- *<head>, <title>, <body>, <h1> až <h6>, <p>, <a>, …*
- Portál Matematická Biologie

*https://en.wikipedia.org/wiki/HTML

MUNI SCI

HTML

- HTML dokument je štruktúrovaný
- Text v HTML tagoch nemusí byť štruktúrovaný

Vedie na strojové spracovanie textu z online zdrojov s cieľom získať štruktúrovanú podobu dát.

MUNI SCI

Web scraping

- Strojová technika extrahovania informácií z webových stránok prostredníctvom HTML kódu
- Vhodný pre:
  - veľké množstvo dát
  - dáta z viacerých stránok
  - opakovanie v čase

MUNI SCI

Scraping vs. crawling

- Web scraper
  - program, ktorý pomocou dotazov na server získa dáta a následne si ich uloží bokom
  - bez prehliadača
  - vedie na **parsing** a **text mining**

*https://en.wikipedia.org/wiki/Web_scraping

- Web crawler
  - program, ktorý systematicky prechádza množinu webových stránok typicky za účelom indexovania
  - vedie na **skúmanie štruktúry webu** a **mapovanie obsahu**

*https://en.wikipedia.org/wiki/Web_crawler

MUNI SCI

Právne normy a etika

- Scraping - všeobecne negatívny pohľad
  - šedá zóna, no nie je nelegálny – dotýkajú sa ho rôzne zákony
  - stránky verejné ⟶ dáta verejné ?
  - business, súkromné účely, výskum, edukačné aktivity, ...
- Pravidlá
  - rešpektovať podmienky použitia a tzv. “robots.txt” súbor (príklad 1, príklad 2)
  - rešpektovať autorské práva (*copyright infringement*)
  - s dátami nakladať legálne
  - nezaťažovať server (*trespass to chattels*)
  - vždy sa môžete opýtať na povolenie autorov webu
  - v prípade zverejnenia výstupu – agregovať, neduplikovať, nepredávať a uviesť zdroj
- Etika
  - nepriame odhalenie niekoho osobnej identity
  - nepriame odhalenie niekoho know-how
  - znižovanie výnosov stránok z reklám
  - znehodnotenie štatistík google analytics

*https://www.researchgate.net/publication/324907302_Legality_and_Ethics_of_Web_Scraping

MUNI SCI

Google

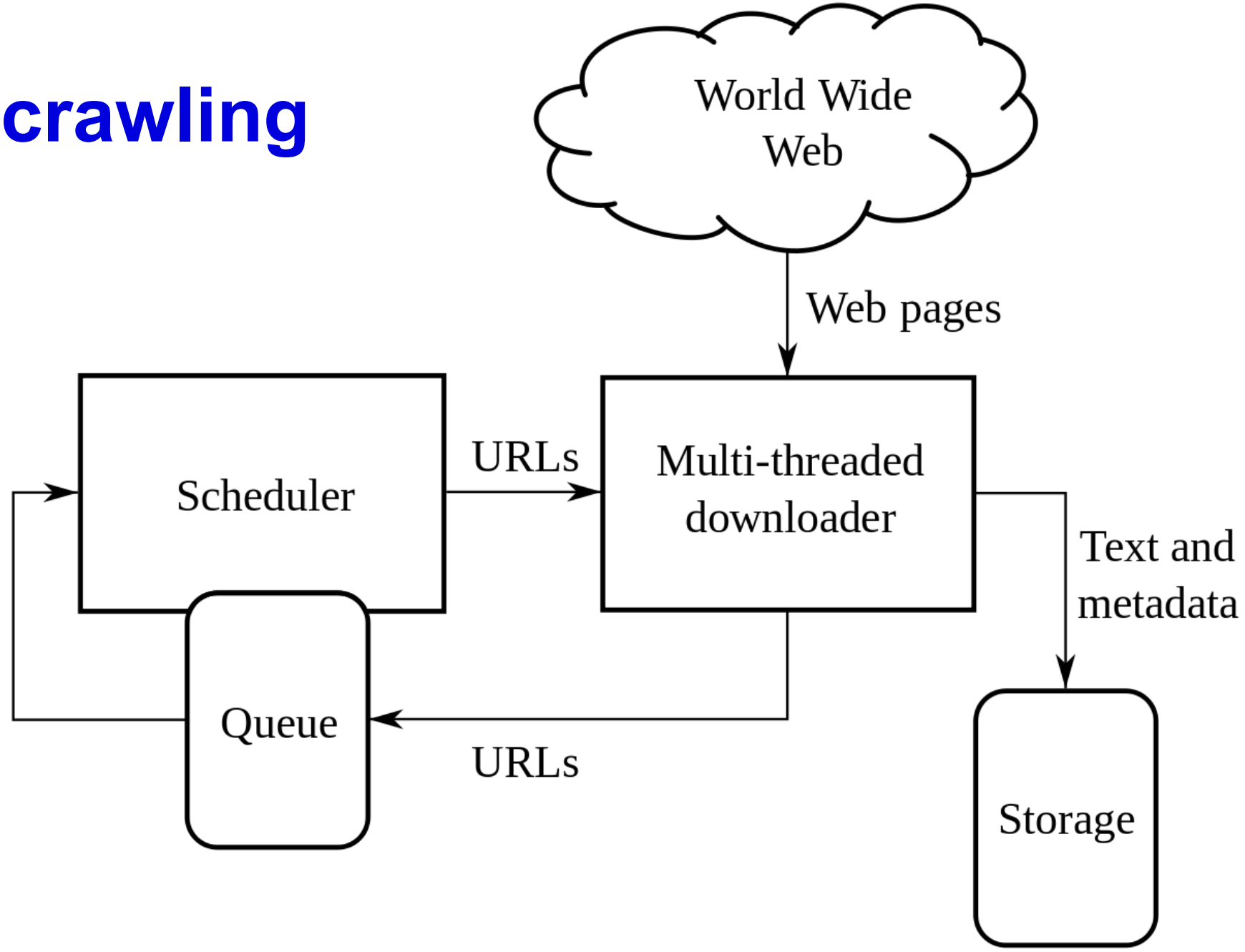
Crawling, scraping, spracovanie, indexácia, vypisovanie vykopírovaných informácií ...

- Googlebot – crawler, ktorý neustále prechádza weby
- Má cieľový URL list, ktorý sa v čase mení
- Obsah sťahuje a predkladá ďalším službám na spracovanie a indexáciu
- Rešpektuje *robots.txt* a využíva aj *sitemap.xml*
- Je to etické a v súlade so zákonom?

*https://www.gdata.in/blog/seo/how-google-indexes-your-website-using-googlebot

MUNI SCI

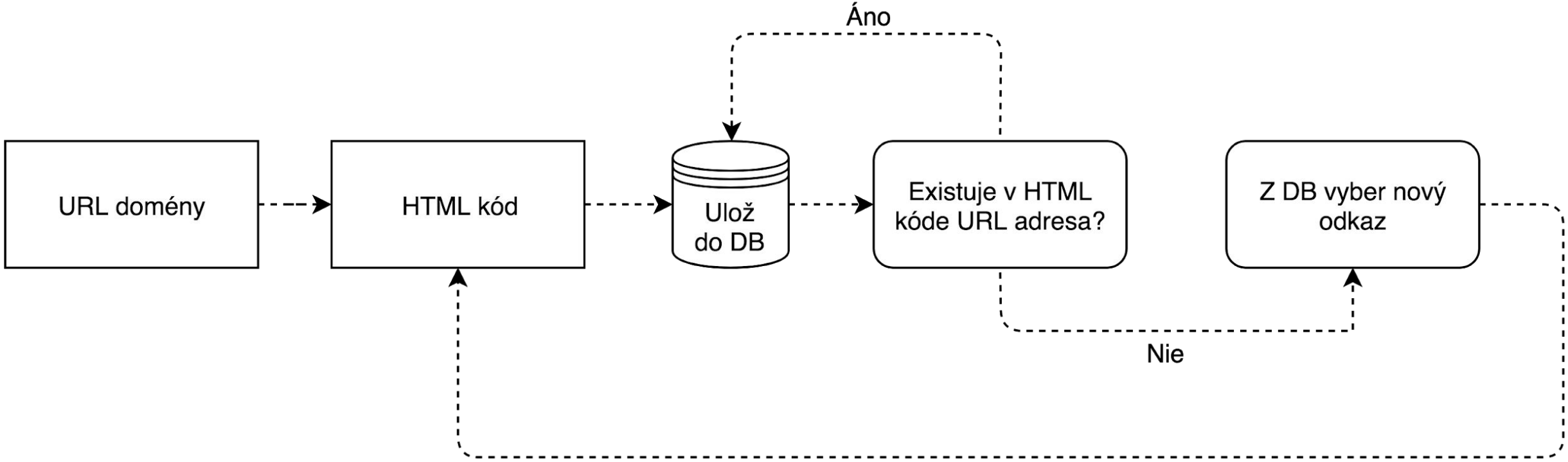
Web crawling



*https://en.wikipedia.org/wiki/Web_crawler

MUNI SCI

Web crawling



MUNI
SCI

Aplikácie

- Prieskum trhu
- Dáta zo sociálnych sietí
- Komentáre z fóra
- Výskumné účely
- Edukačné účely
- How a Math Genius Hacked OkCupid to Find True Love
- *"I think that what I did is just a slightly more algorithmic, large-scale, and machine-learning-based version of what everyone does on the site," McKinlay says. Everyone tries to create an optimal profile—he just had the data to engineer one.*

MUNI SCI

Technológie

- Jazyk
  - python, R, PHP, JavaScript, ...
- Knižnice
  - python – urllib, re, BeautifulSoup, nltk
  - R – rvest, tm, stringr, tidytext

MUNI SCI

Text mining

- Proces získavania nových informácií z textových dát
  - Natural Language Processing
  - Semantical Analysis
  - Frequency Analysis
  - Similarity Analysis
  - …
    - Stemming
    - Tokenization
    - Lemmatization
    - Parts of speech identification
    - Stop words list
    - …
- Vizualizácie
  - wordcloud, barchart, bubblechart, network charts, dendrogramy, ...

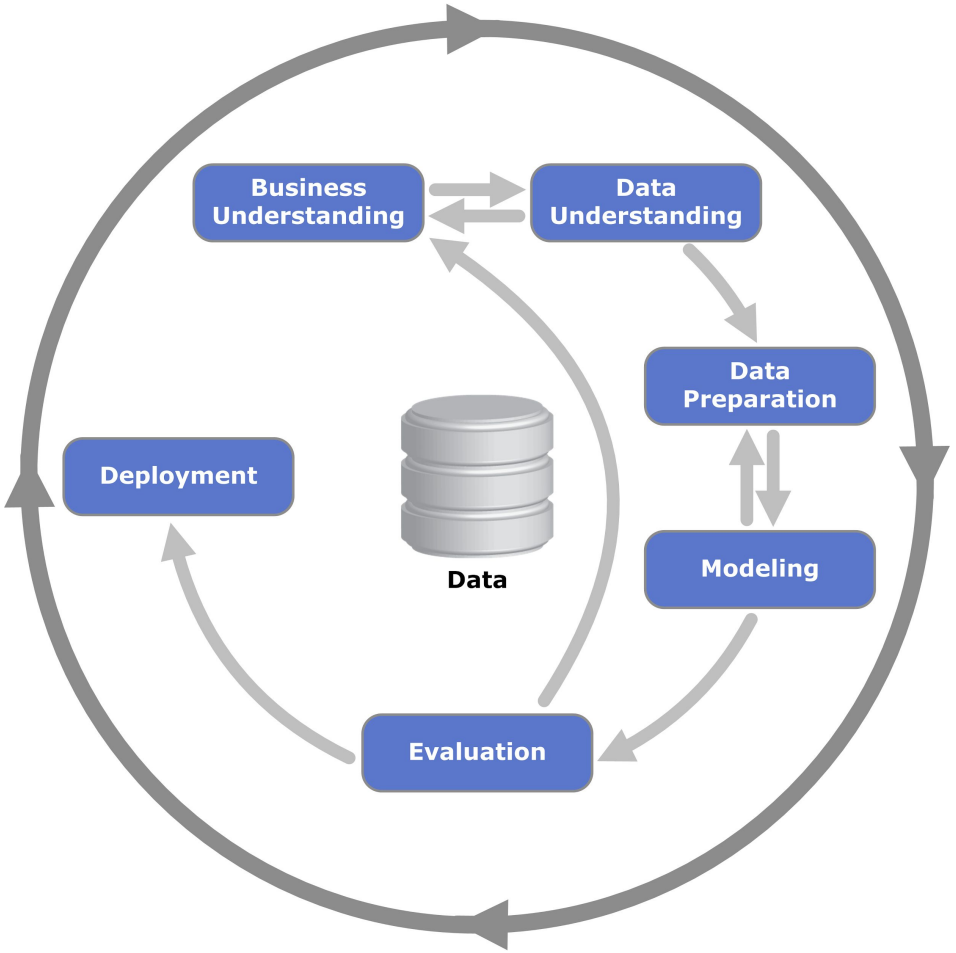
MUNI SCI

Text Mining

Text mining involves a series of activities to be performed in order to efficiently mine the information. These activities are:

*https://www.fosteropenscience.eu/content/text-mining-101

CRISP-DM



*https://www.dataproduct.at/methodology.html

MUNI SCI

MUNI
SCI

Prestávka

Modelový príklad 1

Aké je zloženie zamestnancov vo FN Brno v areáli Bohunice v závislosti na type oddelení, pozícii či pohlaví zamestnanca?

- Aké dáta potrebujem získať?
- Kde dáta nájdem?
- Obsahujú dáta všetky potrebné informácie?
- Môžem dáta stiahnuť a použiť?

MUNI SCI

Modelový príklad 1 – postup

Ako by ste postupovali?

1. Web scraping
2. Čistenie dát
3. Štrukturalizácia dát
4. Agregácia, vizualizácia
5. Kontrola, evaluácia

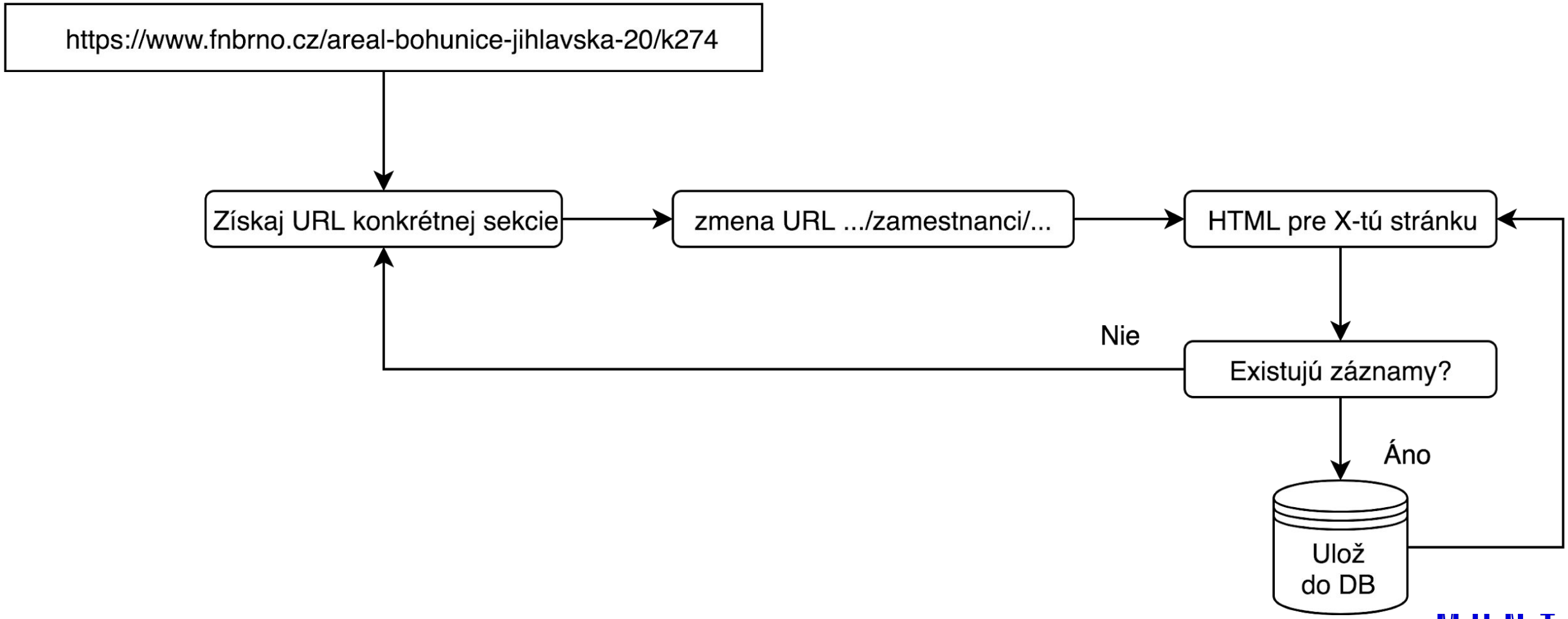
MUNI SCI

Modelový príklad 1 – dáta

www.fnbrno.cz

MUNI SCI

Modelový príklad 1 – scraping



MUNI
SCI

Modelový príklad 1 – technológie

- **python** (knižnice urllib3, BeautifulSoup, re)
- **SQLite** (knižnica sqlite3)
- **jupyter-notebook** (knižnice pandas, plotly)

MUNI SCI

MUNI
SCI

Prestávka

Analýza textu online

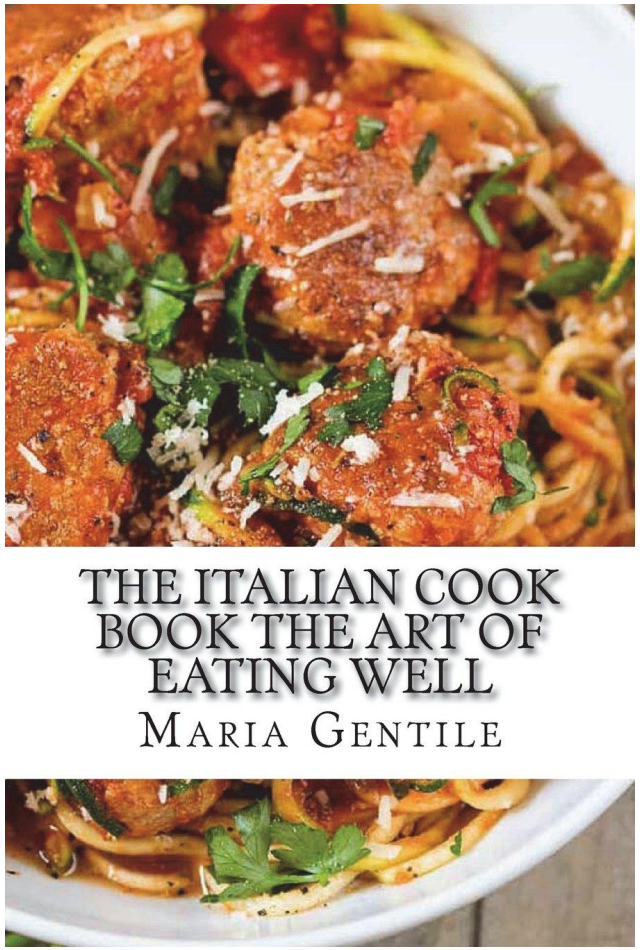
- Nástroj Voyant Tools
- Online, open-source, školský projekt
- UI vo viacerých jazykoch
- Analýzy sú jazykovo nezávislé (až na niektoré procedúry)
- ***Vstup*** – textové dokumenty (1 a viac...)
- ***Výstup*** – interaktívny dashboard s vizualizáciami, tabuľkami, porovnaniami, ...

*https://voyant-tools.org/docs/#!/guide/about

MUNI SCI

Modelový príklad 2

- eKnihy zdarma – Project Gutenberg
  - The Italian Cook Book The Art of Eating Well

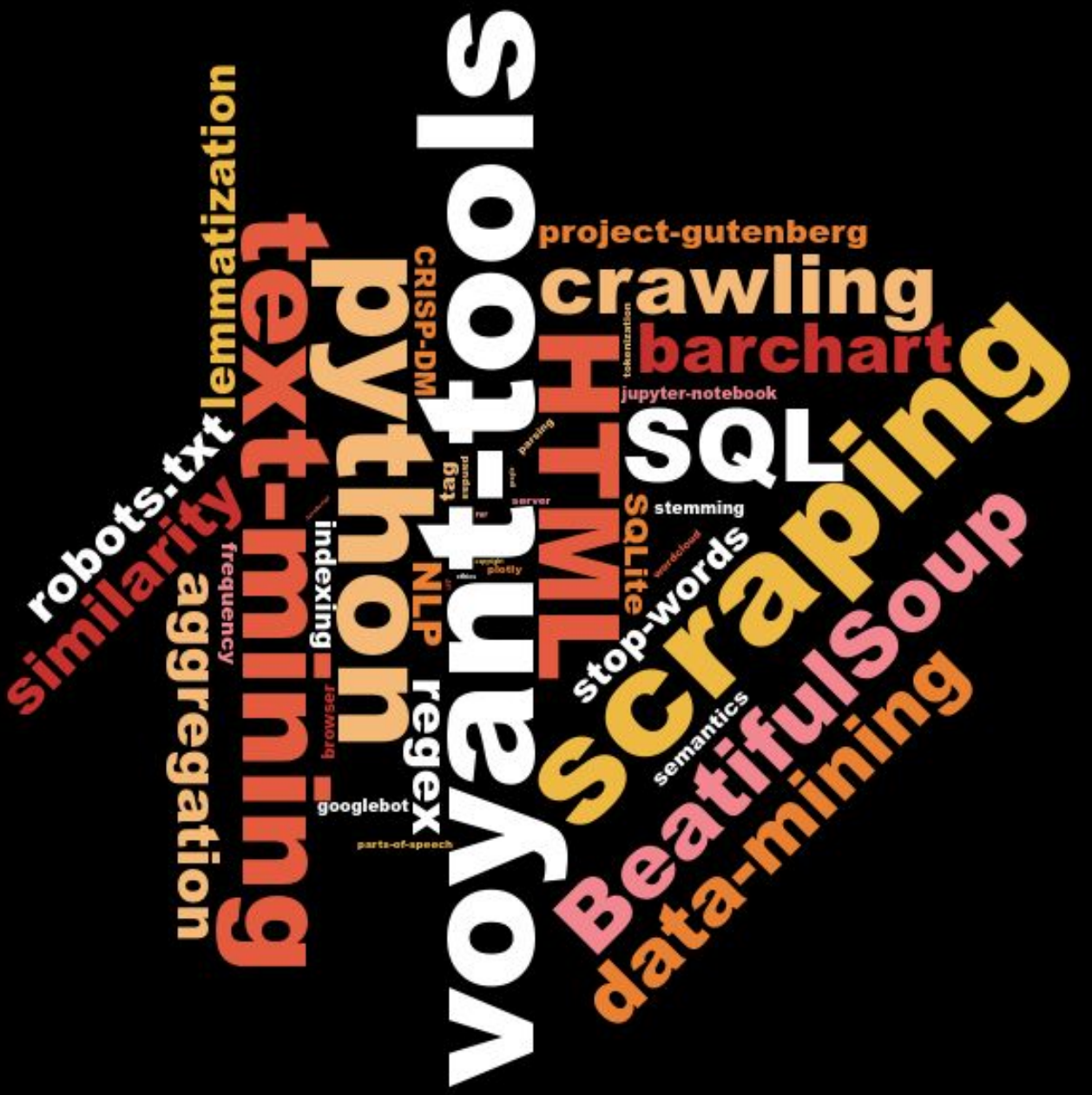


MUNI SCI

Samostatná práca

1. Vytvorte 3 skupiny
2. Vyberte si zdroj dát:
   - Harry Potter and the Philosopher's Stone (EN book)
   - The Lord of the Rings: The Fellowship of the Ring (EN book)
   - Star Wars IV, V, VI (EN subtitles)
   - Game of Thrones – Book One of A Song of Ice and Fire (EN book)
   - Forrest Gump (EN book)
3. Pomocou online nástroja Voyant Tools "**vydolujte**" zaujímavé informácie a na základe výstupov "**prerozprávajte**" príbeh vybraného diela

MUNI SCI


voyant-tools
scraping
text-mining
python
HTML
crawling
SQL
barchart
project-gutenberg
BeatifulSoup
data-mining
similarity
robots.txt
lemmatization
aggregation
stop-words
stemming
SQLite
NLP
regex
indexing
googlebot
semantics
CRISP-DM
jupyter-notebook
parts-of-speech