Data Collection and Preparation

DAPI . Information Description, Storage and Retrieval Course
MIEIC, 2020/21 Edition

Sérgio Nunes
DEI, FEUP, U.Porto

Work in progress

Plan for Today

- ➔ COVID-19 logistics
- ➔ Data Collection and Preparation
- ➔ Groups update
  - ➔ Present topic and data sources. Discuss and approve.
  - ➔ Check data exploration and characterization.
- ➔ Dataset characterization
- ➔ Group work

Logistics (ongoing)

➔ Managing "morning shifts" and DAPI.

  ➔ Start 30 min later?

  ➔ Try to reserve room for the morning?

  ➔ Implement shifts in DAPI?

➔ Other pending issues?

Milestone #1

➔ Goal: prepare and characterize the datasets.

➔ Depends heavily on the datasets (e.g. crawling, scraping, other).

➔ Checklist

- ➔ Properties of the datasets (media, formats, volume, structure, license, source authority, ...)
- ➔ Describe process to collect the dataset (sample, whole dataset, API, ...)
- ➔ Describe the data pipeline process (collecting, cleaning, alignment, integration, enrichment, ...)
- ➔ Present the conceptual model of the data (and of the domain if needed)
- ➔ Which are possible search tasks / ideias? What other works exist?

➔ https://web.fe.up.pt/~ssn/dokuwiki/teach/dapi/202021/delivery1/index

Data Pipelines

Data Workflow

➔ Data collection

➔ Data storing

➔ Data cleaning

➔ Data enrichment

➔ Data exploration and analysis

➔ Data presentation

Data Collection

➔ Data sources: data repositories, databases, APIs, web scrapping, files, ...

➔ Data formats: unstructured text, CSV, JSON, XML, Excel, PDF, ...

➔ Character encoding.

➔ Work with small samples.

Data Storage

➔ Flat files / local storage

➔ Databases: document oriented, relational database, key-value, ...

➔ Clusters (e.g. Hadoop)

➔ Cloud-storage (e.g. AWS, Azure)

Data Cleaning

- ➔ During data cleaning, data analysis is implicit.
- ➔ Automating the process is key — deal with changes in the input, document, repeat
- ➔ Expected tasks
  - ➔ Identify missing / invalid values
  - ➔ Normalize data (e.g. "PT" vs. "pt" vs. "Port." vs. "Portugal")
  - ➔ Format data
  - ➔ Find outliner and bad data
  - ➔ Find duplicate data

Data Enrichment

➔ Can be done by adding metadata (e.g. timestamps, author)

➔ Or by combining different datasets

  ➔ Finding key attributes for aligning different collections is key

Tools for Data { collection, preparation, exploration, characterization }

OpenRefine

➔ Open source tool for data exploration and cleaning (formerly Google Refine).

➔ Explore Data

➔ Clean and Transform Data

➔ Reconcile and Match Data

➔ https://openrefine.org/

➔ **Task**: view online tutorials / videos and experiment with your datasets.

Apache Tika

➔ Apache open-source tool to parse and extract text and metadata from multiple formats (e.g. PPT, XLS, PDF).

➔ https://tika.apache.org/

spaCy

➔ Natural Language Processing in Python

➔ Open source (MIT license)

➔ Pre-trained models (>50 languages)

➔ Very active project and community

➔ Features: tokenization, named entity recognition, part-of-speech tagging, similarity measures, ...

➔ https://spacy.io/

Other NLP tools

➔ NLTK (Natural Language Toolkit)

  ➔ Python. Open source tool with many resources available, including a freely available reference book — https://www.nltk.org/book/

  ➔ https://www.nltk.org

➔ Apache OpenNLP

  ➔ Java. Open source library.

  ➔ https://opennlp.apache.org

Web Data

- ➔ Scrapy
  - ➔ Open source Python tool for crawling and scraping
  - ➔ https://scrapy.org/

- ➔ Beautiful Soup
  - ➔ Python library designed for screen-scraping
  - ➔ https://www.crummy.com/software/BeautifulSoup/

Command Line tools

➔ Command Line is a powerful solution in many data processing stages.

- ➔ Agile (interactive, close to the file system)
- ➔ Extensible (integrates well with other technologies, language agnostic)
- ➔ Scalable (automatable with scripts, repeatable)
- ➔ Ubiquitous
- ➔ ! Core knowledge with wide impact in many areas

➔ https://www.datascienceatthecommandline.com/1e/

Data Exploration and Visualization

- R
  - Free software for statistical computing and graphics.
  - Many resources and documentation. Strong community
  - https://www.r-project.org/
- Pandas
  - Open source Python library for data analysis.
  - https://pandas.pydata.org/
- Excel

Conceptual Domain Modeling

You need to understand the fundamental concepts within your problem domain

➔ *... it is the task of discovering the entity types that represent the things and concepts, and their relationships, pertinent to your problem space.*

➔ *... depict your detailed understanding for the problem space for your system.*

➔ Various tools and techniques can be used for this task.

➔ We will use **UML Class Diagrams** and focus on the data dimension of the problem, i.e. consider what is included in the dataset.

Conceptual Data Modeling

➔ Iterative process, i.e. start simple and add complexity.

➔ Identify main concepts (i.e. things, entities) to define classes.

➔ Describe properties of the entities to define attributes.

➔ Use basic data types (e.g. number, text, date). Be specific if needed.

➔ Identify relationships (i.e. verbs) to define associations.

➔ Consider complex associations to better describe your domain, e.g. inheritance, dependency, compositions, association classes, etc.

Tasks

➔ Characterize the datasets

➔ Obtain the conceptual model of the domain

➔ Try available tools to work with datasets

➔ Discuss the storage of datasets

➔ Identify retrieval tasks using the datasets

➔ **Next week**: Moodle post with data pipeline diagram (thread with first message).

References

➔ Scott Ambler, The Object Primer, Chapter 8: Conceptual Domain Modeling, Cambridge University Press, 3rd Edition, 2004 (Section 8.4)

➔ Jacqueline Kazoo, and Katharine Jarmul. Data wrangling with python: tips and tools to make your life easier. O'Reilly, 2016.

➔ Garrett Grolemund, and Hadley Wickham. R for data science: import, tidy, transform, visualize, and model data. O'Reilly, 2016. https://r4ds.had.co.nz/