Open Data Processing Tools

Dr Sepi Chakaveh

20 February 2017
Data-Information-Knowledge-Wisdom

Wisdom
Knowledge
Information
Data
Data Collection
Data Processing

- Collect
- Clean
- Visualise
- Analyse
Data processing as a series of steps

- Acquire
- Parse
- Filter
- Mine
- Represent
- Refine
- Interact
Value Proposition

Analysed Data → Predictions → Application
Open Data Cleaning Data
Introduction

• This process focusses on using tools to validate, clean and explore data sets.

• A big problem with publicly available datasets is the number of errors within them.

• These problems vary from simple spelling errors, to the more complex problems involving misuse of units.

• We use Google Open Refine as an example of a typical tool.
Problems

• Date Validation
  – One of the most common problems in data is mixed date formats, this can be particularly troublesome when you have British and American date formats e.g. (7/12/2012 and 12/31/2012)

• Multiple Representations
  – Most common in datasets containing abbreviations, for example in location data or role based data. It is common that abbreviations will change and even be present in fully expanded form (e.g. Vice-President Marketing and VP Marketing)
Problems

• **Summation Records**
  
  – When data has been extracted from a spreadsheet application, it is common to be left with both columns and rows of data containing the sums (or other formula) of the other data. While not an error, it is inconvenient when re-processing the data.

• **Duplicate Record Detection**
  
  – Duplicated records are common place both at the point of entry (by a human) but also a common occurrence when exporting a huge amount of data from multiple systems.
  
  – It is often the case that the data has been duplicated in order to speed up searching across multiple domains where the data is applicable in both.
Problems

• **Mixed use of numerical scales**
  
  – A common, but critical, failure in data that can lead to audit failure. E.g. outliers are often clear to see as one record may contain a figure multiple factors bigger than any other.

• **Redundant Data**
  
  – Redundant data is not required, thus it is common that errors are made when entering it.
Problems

• **Numeric Ranges**
  – Numeric ranges, often used to anonymise data, cause problems when wanting to explore and visualise the data.

• **Spelling Errors**
  – While not critical in all cases, spelling errors can lead to awkwardness when querying and visualising data.
Importing Data

• In order to carry out this exercise three datasets are required.
  – Datasets are genuine,
  – You need to access them directly

• Dataset 1 – Department for Environment, Food & Rural Affair
  https://data.gov.uk/dataset/agricultural_market_reports

• Dataset 2 – UK GP Earnings
  https://data.gov.uk/dataset/gp_earnings_and_expenses
This exercise focusses on using tools to process, visualise, publish & analyse data sets.

– **Data Processing Tools**
  - Open Refine
  - Open Data Soft

– **Data Publishing Tools**
  - GitHub
  - CKAN

– **Data Analytic Tools**
  - Hadoop
Data Processing Tools (Validation & Cleaning)
Importing into Google OpenRefine

- We need to use the Google OpenRefine tool to process our datasets.
- You may download Google OpenRefine from openrefine.org.
- OpenRefine is an application that runs locally on your machine.
- There is an additional benefit of the data to remain private.
- Once installed and running it should open a browser window on the refine home screen.
OpenRefine

• Interactive Data Transformation (IDT) Tool

• An extremely powerful IDT for data visualisation & manipulation.

• Data in the following format is supported by
  
  – OpenRefine
  – TSV, CSV, *SV, Excel(.xls and .xlsx)
  – JSON,
  – XML,
  – RDF as XML, Wiki markup, and
  – Google Data documents are all supported
OpenRefine
OpenRefine (Limitations)

• It is not a good tool for creating new datasets
• Not useful for huge datasets
• The amount of data that can be processed depends on the memory size
• For a file size of 512MB, a simple operation can take more than 15 minutes which is not accepted in real-time application
• Use OpenRefine as an offline data cleaning tool
Data Publishing Tools
GitHub

- GitHub is a web-based Git repository hosting service.
- Git is a VC System keeping developers up to date for all versions (non-beta) release using a distributed revision control and source code.
- GitHub has a Source Code Management (SCM) functionality of Git as well as adding its own features.
- The success of GitHub is due to the fact that GitHub managed to create a developers community where likeminded people can store & share their projects on a network.
Sometimes you just need a little help.

Common Issues

- Why are my contributions not showing up on my profile?
- Why is Git always asking for my password?
- Dealing with non-fast-forward errors
- Error: Repository not found
- Do you have custom plans?
- HTTPS cloning errors
- What is my disk quota?
- What are the limits for viewing content and diffs in my repository?
CKAN

• The **Comprehensive Knowledge Archive Network (CKAN)** is a web-based open source data management system for the storage and distribution of data,

• **CKAN** has recently managed to be the preferred tool for publishing, sharing, finding and using data for a number of Governments, including the UK.

• **CKAN** like many other data processing tools first considers the metadata aspects of the data-sets.
**Feature Tour**

CKAN is a fully-featured, mature, open source data management solution. CKAN provides a streamlined way to make your data discoverable and presentable. Each dataset is given its own page with a rich collection of metadata, making it a valuable and easily searchable resource. Check out our live interactive demo!

**Publish & find datasets**
Publish datasets via import or through a web interface. Search by keyword or filter by tags. See dataset information at a glance. Full change history lets you easily undo changes or view old versions.

**Store & manage data**
Store the raw data and metadata. Visualize structured data with interactive tables, graphs and maps. Get statistics and usage metrics for your datasets. Search geospatial data on a map by area.

**Engage with users & others**
Federate networks with other CKAN nodes. Theme with CSS or integrate with a CMS. Build a community with extensions that allow users to comment on and follow datasets.

**Customise & extend**
Use the API's rich programming interface, and benefit from over 60 extensions including link checking, comments, and analytics. CKAN's Open Source licence allows you to download and run it for free.
Data Analytics Tools
Hadoop

- **Hadoop** is an open source collection of tools for manipulating large & distributed data sets.

- **Hadoop** is licensed by Apache guaranteeing no particular company controls the direction of Hadoop.

- **Hadoop** is primarily used for massive amounts of structured and unstructured data and be processed.
Hadoop Ecosystem

- **Ambari** A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters. **Avro** A data serialization system.

- **Cassandra** A scalable multi-master database with no single points of failure.

- **Chukwa** A data collection system for managing large distributed systems.

- **HBase** A scalable, distributed database that supports structured data storage for large tables.
Hadoop Ecosystem

- **Mahout** A Scalable machine learning and data mining library.

- **Pig** A high-level data-flow language and execution framework for parallel computation.

- **Spark** A fast and general compute engine for Hadoop data.

- **Tez** A generalized data-flow programming framework,
Hadoop

Hadoop Ecosystem

- Hive
- Flume
- HBase
- Avro
- Hue
Hadoop

Who uses Hadoop?

[Logos of various companies including IBM, Samsung, SK telecom, Yahoo!, The New York Times, Facebook, VISA, Intel, Netflix, Ning, Autodesk, contextWeb, Twitter, Amazon, Groupon, JPMorgan Chase & Co., NAVTEQ, Rackspace Hosting, visibleMEASURES, eBay, University of Phoenix.]