ADHOC-TE-2016-0892 STATE of the ART - DIGITAL DATA COLLECTION

WEB PAGES

On the World Wide Web there are static and dynamic websites
Static Websites are characteristic for the beginning of the Web and consist of webpages (HTML files) hosted on Web servers and accessible through a browser.
The content and form (mark-up elements) are not separate. The structure of the pages is fixed and any change of the structure, content or form is done through editing the individual files or page templates (in some cases) or stylesheets. Due to better loading speeds, static websites have seen a resurgence in the past few years through many user-friendly site-makers / generators.
Dynamic Websites are generated by applications running on a server and connect to a database (hosted on the same server or another) where the content and configuration/structure are stored.
These websites are created and managed through WEB CONTENT MANAGEMENT SYSTEMS.
Individual pages that are accessed through a browser do not exist on a Web server, but are generated for each request using templates which combine HTML, scripting languages (PHP, JavaScript) and filling them in with content from the database, thus generating each page dynamically, adapted to the context.

DATA SOURCES

OPEN DATA (datasets made public or publicly accessible and useable without restrictions)
Open data are data which can be used freely, reused, redistributed by anyone – subjected at the most to attribution
DATA FROM DOCUMENTS (usually, tables in PDF)
DATA FROM WEBSITES
DATA FROM API INTERROGATION (Facebook, Twitter, Wikipedia etc.)
API (Application Programming Interface) – sets of functions which are made available for developers who want to integrate a system with another. To encourage the development of Web or mobile apps, many social media systems have (partially) open APIs
According to the type of data and their specific format, several automated data collection tactics may be used to gather large datasets

- OCR (Optical Character Recognition) and document conversion
- Web Scraping, Web Wrapping and Web Crawling
- API Interrogation

WEB SCRAPING

Web scraping / Screen scraping refers to different methods that can be used to collect data from Web pages, usually from dynamic web pages. It is also sometimes called Web data extraction, screen scraping or Web harvesting.
It is a form of DATA MINING and it is based on the identification of distinct patterns in HTML and CSS templates of dynamic pages to detect information of the same type and to automate the data collection process
The stages of a scraping project:

- **Choosing the target content**
- **Defining selectors / patterns**
- **Simulating navigation**
- **Configuring automation**

Tools: Chrome Web Scraper, Outwit Hub, Octoparse, Helium Scraper, Import.io, Screen Scraper

Choosing the target content

Search results, e-commerce websites, media sharing platforms, archives of news websites or blogs have linear structures and are usually displayed in a paged list (with navigation – `next` or `show more` buttons)
Forums or threaded comments have tree-like structures.
In dynamic websites, the same HTML tags and CSS styles are used for content of the same type.

Defining selectors / patterns

The semi-structured nature of webpages may be used.
HTML pages may be represented as tree structures of nested HTML element tags
Each level of the hierarchy represents a nesting level of the HTML elements. This representation is called DOM (Document Object Model)
For each HTML element in a document, a unique path can be defined in a similar way in which we call up the path of a file in a filesystem.
For example:
/html/head/title will refer to the title of the document
/html/body/div[2] will refer to the second div element in the body
Xpath specifications allow us to refer to the surrounding elements of a given element, conceptualized as a node with certain relations (ancestor, parent, sibling, child, descendant, preceding, following) in the document tree structure.

Types of navigation

- Liniar structures – Pagination (blogs/news)
- Tree structures – Multilevel (forums/comments)
- Network structures – Graph (wiki type sites)
- Tabular structures – sortable by variables (e-commerce)
- Relational structures – entities of several types (IMDB)

API INTERROGATION

API = Application Programming Interface is a software-to-software interface that allows data exchange between to applications
For example, Facebook offers the public API OpenGraph so that other applications created by developers to be able to integrate functionality/personalized information from Facebook – Facebook authentication, friends who liked a certain content etc.

Types of API

RESTful- (Representational State Transfer) – the most used form of API – communication through HTTP
SOAP (Simple Object Access Protocol) – data transfer in XML format

Types of HTTP requests

GET – to get data
PUT – to edit existing data
POST – to add new data
DELETE – to delete data

Most APIs require authentication.
HTTP Basic Access Authentication – user and password are transmitted in the header of the HTTP request
OAuth 1.0/2.0 – a unique token is generated for the user
A significant number of APIs will transmit responses as JSON (JavaScript Object Notation) - the most widely used data format for data interchange on the web

Facebook Graph API

Nodes – users, pages, posts, comments
Edges – connections between nodes – the comments of a certain post, the posts of a certain page
Fields – information about entities, objects – pages, users, posts

Each node has a unique Object ID

- Usually APIs limit the quantity of data that can be transmitted through a single response
- Paginated responses may allow the collection of large quantities of data
- In the aftermath of the Cambridge Analytica scandal, Facebook has introduced severe limitations in the access of data through its API (including data from public Facebook pages)

Tools: Facepager – usage outline:
Creating a database
Configuring an interrogation preset
Configuring data columns
Authentication and token
Starting the interrogation (actual data collection)
Exporting the data in CSV
Scenarios: Facebook, Twitter, YouTube, Wikipedia, generic APIs
Alternative tools: API interrogation with cURL command line tool, python libraries for interacting with APIs etc.

ADHOC-TE-2016-0892 STATE of the ART DIGITAL DATA ANALYSIS

DATA CLEANUP AND CONVERSION

Data collection from online sources using Web scrapers or API interrogation will usually mean data is initially in text format, character strings.
Data extracted through Web Scraping or API interrogation are semi-structured data.
Issues with saving or exporting data:
File format
Maximum field sizes (for each tool used)
Character encoding (special characters / diacritical marks may not be recognized by some analytical software)

Data types

- character
- string
- integer
- float
- boolean (TRUE/FALSE)
- Date / Time
- Coordinates (latitude/longitude)

File formats

Character Separated Values (comma, tab, semicolon)
Fixed width – each column takes up a number of characters on each row
Importing CSV files will require defining the field delimiter and any other characters used to define data fields
When importing text files into different analysis tools, after defining the data columns, fields are sometimes associated a certain data type. When converting data, maximal values that can be stored into each data type should be taken into consideration.
For data resulting from scraping several different websites (for example news sites) the date/time format will often be different and will need to be converted to a common format.

Some common string operations

CLEAN – deleting unprintable characters
TRIM – deleting extra spaces from the beginning or the end
UPPER/LOWER – conversion to upper or lower case
LEN- length of a character string
CONCAT / CONCATENATE / TEXTJOIN – joins two or more strings
LEFT / RIGHT – a certain number of characters from the beginning or end of a string
REPLACE / SUBSTITUTE – replaces a string with another
CONTAINS – returns TRUE if a string contains a given substring
Tools:*Tableau Data Prep* *– large dataset overview, filtering, sorting, pivoting, character/string replacing, calculated fields etc.*
ASAP Utilities – replacing accented characters, deleting extra spaces, recognizing dates, changing date/time formats

ELEMENTS OF COMPUTATIONAL LINGUISTICS

Stopwords – words in a language that are usually excluded before working with natural language processing tools – usually very frequent words or particles, deictics, pronouns, conjunctions, compound verb forms
POS tagging – Identification of parts of speech
Stemming – process through which the endings of words are trimmed in hopes of reducing words to their base form.
Lemmatization – process that uses a vocabulary and morphological analysis to identify and reduce all forms of a word to their base forms.
Jaccard Distance/ Jaccard similarity coefficient is used to compare elements of two sets with the purpose of observing which are common and which are distinct. It is a measure of similarity of two data sets or codes
Hierarchical clustering - data mining methods that group data into categories and subcategories according to a measure of distance / difference between the sets and their elements.
Co-occurrence analysis – a co-occurrence matrix describes in a binary way (presence/absence) an object with respect to a context (a word or a code in a document/text)
Co-occurrence network – a more visually intuitive representation based on a co-occurrence matrix where nodes represent objects (words/codes), and edges/lines between them represent their presence in the same context (co-occurrence). Nodes or edges may be weighed (sized or styled with according to their frequency or relative distance score)
TF - term frequency - (word or code)
TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).
IDF - inverse document frequency - - a measure of the importance of a term. While computing TF, all terms are considered equally important. However, it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:
IDF(t) = log_e(Total number of documents / Number of documents with term t in it).
Tools:
KH Coder – visual interface for natural language processing (based on R) for co-occurrence networks, hierarchical clustering of words or codes, heatmaps etc.
Sketch Engine – corpus linguistics, lexicography
Spacyr and Shiny – R libraries for visualizing co-occurrence networks based on spacy tokenization

KNOWLEDGE REPRESENTATION AND INFORMATION RETRIEVAL

Semantic networks are a logic-based formalism for knowledge representation. Semantic networks are graphs which are constructed from both a set of vertices (or nodes) and a set of directed and labeled edges. The vertices or nodes represent concepts, and the edges represent semantic relations between the concepts.
Knowledge about accepted meanings should be processed in adjacent regions of the semantic network. Therefore, semantic networks are often termed “associative networks.”
Formal Concept Analysis (FCA) is a method of knowledge representation introduced in the 1980s by Rudolf Wille, rooted in the pragmatic philosophy of Charles Sanders Peirce, based on a binary incidence relation, and building on applied lattice and order theory. It has applications in various fields and its advantage lies in the possibility to visualize and explore formal concepts in a formal context (a data table that represents binary relations between items in a set of objects and items in a set of attributes) as representations of complete lattices.
Tools: FCA Tools Bundle

MACHINE LEARNING and NATURAL LANGUAGE PROCESSING

Word embedding is one of the most popular representation of document vocabulary. It is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc.
The objective is to have words with similar context occupy close spatial positions. Mathematically, the cosine of the angle between such vectors should be close to 1, i.e. angle close to 0.
Word2Vec is a two-layer neural network trained to reconstruct linguistic contexts of words. It is a method to construct such an embedding. It uses large corpus inputs and produces a vector space of several hundred dimensions, each word being assigned a vector in the space. The word vectors’ positioning so that words which have common contexts in the corpus are located in proximity to one another.
NER – **Named Entity Recognition**, **Text Categorization**/Classification – require training language models or using pre-trained models.
Tools: Gensim (Word2Vec), Spacy (NLP), Prodigy (Annotation interface), PyTorch, LASER, scikit-learn