TC3 Information Retrieval

Master 1 AI, Upsay
T4, March to April

Kim Gerdes
Lisn

Last time

- Extensive review of the previous week's notebook
- BM25
- Sense2Vec

Planning

1. 10/3 gentle introduction
2. 17/3 big dataset, binary evaluation
3. 24/3 improvements: embeddings
4. **28/3 project presentation, technical terms, semantic structure**
5. 31/3 many absent: how many? remote class?
   exploring existing tools
   NO 7/4!
6. 14/4 work on projects, discussions
7. 21/4 project presentation

Today

- Kaggle
- terms
- patents
- semantic relations
- term detection

technical terminology

- ***term***:
  - Merriam-Webster: a word or expression that has a precise meaning in some uses or is peculiar to a science, art, profession, or subject
  - Wiktionary:
    - From Middle English terme, borrowed from Old French terme, from Latin terminus (“a bound, boundary, limit, end; in Medieval Latin, also a time, period, word, covenant, etc.”).
    - word or phrase, especially one from a specialised area of knowledge
- https://en.wikipedia.org/wiki/Terminology_extraction
  - Terminology extraction (also known as term extraction, glossary extraction, term recognition, or terminology mining) is a subtask of information extraction. The goal of terminology extraction is to automatically extract relevant terms from a given corpus.
  - Keywords and term extraction:
    - keywords (often single word items, often from a fixed list, for dispatching documents)
    - terms (multiword items).

why are terms interesting?

- for learning about a domain (humans and machines)
  - information extraction: what does the text say?
  - semantic structure: who does what to whom?
  - NLG: controlled generation, detect hallucinations
  - trend detection, diachronic analysis
- for translation (bilingual terminology lists)
- for manual control of classification
- Search engine optimization (SEO)
- taxonomy / ontology extraction (competitions)

How to get a list of terms for a domain?

- frequent ngrams in a technology corpus?
  - find out in the notebook…

Term structure

- In each language, a term can have a different format.
- In most situations, the requirement is for a term to be a **noun phrase (NP)**.
  - For example, a term in English can be composed of
    - nouns (N),
    - adjectives (J)
    - and also prepositions

    so the phrase should match one of these patterns N+N, N of N, J+N, J+J+N, J+N of N, J+N of J+N etc. while preposition + article + adjective is unlikely to be considered a term.

Term frequency

- If we analyse texts from tabloid newspapers and texts from books on accounting, we are likely to find *income tax* and *best way* in both of them.
- Both phrases match the structure of a term in English (N+N and J+N respectively), however, the frequencies are likely to differ.
- While the frequency of *best way* is likely to be similar in both texts, the frequency of *income tax* is likely to be much higher in texts on accounting.
- This is how the system can automatically tell a frequent phrase from a term and will identify *income tax* as a term.

patent texts

- classification
  - CPC ~300k classes
  - IPC ~70k classes
  - A,B,C,D,E,F,G,H,Y
  - A01, A02, …
  - IPC4 : G06F

G06F

- https://www.wipo.int/classifications/ipc/en/ITsupport/Version20170101/transformations/ipc/20170101/en/htm/G06F.htm
- ELECTRIC DIGITAL DATA PROCESSING (computer systems based on specific computational models G06N)
  - Subclass indexes
    - DATA PROCESSING 7/00, 15/00-19/00
    - INPUT, OUTPUT; INTERCONNECTIONS BETWEEN FUNCTIONAL ELEMENTS 3/00, 13/00
    - ADDRESSING OR ALLOCATION 12/00
    - CONVERSION; PROGRAMME CONTROL; ERROR DETECTION, MONITORING 5/00, 9/00, 11/00
    - DETAILS 1/00
    - SECURITY ARRANGEMENTS 21/00
- AI
  - https://www.wipo.int/tech_trends/en/artificial_intelligence/patentscope.html
  - https://patentscope.wipo.int/search/en/result.jsf?query=AItechniqueMachineLearning&sortOption=Relevance

patent example

https://patents.google.com/patent/EP3502925A1/en?oq=3502925

patent example

COMPUTER SYSTEM AND METHOD FOR EXTRACTING DYNAMIC CONTENT FROM WEBSITES
_____3502925_____487783840_____EP3500000.txt_____G06F_____G06F16/951:G06F16/986
Computer system (100), computer-implemented method and computer program product are provided for extracting dynamic content data (221) from a website (220) in a machine-readable format. The system has an interface (110) to access configuration data (250) reflecting the structure of the website (220). The configuration data includes at least a website specific scraping script and one or more website specific XPath statements. Further, the interface receives a data retrieval request (210) specifying the website (220) and corresponding dynamic content data (221) to be retrieved. A scraper module (120) provides the scraping script (2050) for execution wherein the scraping script is configured to perform one or more parameterized navigation steps on the website (220) to access the dynamic content data (221). A script module (140) triggers execution of the scraping script and receives HTML/XML data associated with the dynamic content data from the website (220) in response to the scraping script execution. An XPath extraction module (150) extracts machine-readable content data (222) from the HTML/XML data wherein the XPath extraction module is pre-configured with the website specific XPath statements in accordance with the structure of the website (220).
_____d:
Technical Field
The present invention generally relates to systems for data retrieval, and more particularly, relates to methods, computer program products and systems for extracting dynamic content from websites in machine-readable format.
Background
Web scraping or web data extraction methods are known in the art. Web scraping is used to access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. Web scraping typically refers to automated processes implemented using a bot or web crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.
Web scraping a web page involves retrieving a predefined Hypertext Markup Language (HTML) page and extracting data from it. Fetching is the downloading of a page which is stored under a static Web address typically specified by a Uniform Resource Locator (URL). Once the page is fetched from where it had been stored, extraction can take place. The content of a page may then be parsed, searched, reformatted, etc. Web scrapers typically extract certain parts of a page to make use of it for another purpose. An example is to find and copy names and phone numbers, or companies and their URLs, to a list (so-called contact scraping).
Prior art Web scraping tools can retrieve web page content from pages which are stored as predefined HTML data. Such content is referred to as static content herein because it relates to content provided by static web pages. However, current web technology allows to dynamically generate web pages on a web server in response to requests which may be received from a user or a computer system. As a consequence, data shown on websites can continuously change. A web page containing respective data can change its layout and new data fields may be introduced at any time. The content of such dynamic web pages (dynamic content) typically depends on the navigation history through a website. In other words, it depends on where the user currently is and which information and requests have been sent previously. Current Web scraping tools fail to scrape dynamic content data from such dynamically generated web pages and provide respective content data in a machine-readable format so that the content can be further processed by other computer systems provided with the extracted data.
Summary
Hence, there is a need for providing improved methods and systems to enable web scraping for dynamic content on dynamic web pages.
This technical problem is solved by a computer system, a computer-implemented method and a computer program product as disclosed in the independent claims. The disclosed embodiments define a screen-scraping framework which addresses the above problem by automatically connecting to a target website and extracting dynamic data from said target website.
In one embodiment, a computer system is provided for extracting dynamic content data from a website in a machine-readable form
…

file format

- \n\n between patents

COMPUTER SYSTEM AND METHOD FOR EXTRACTING DYNAMIC CONTENT FROM WEBSITES
_____3502925_____487783840_____EP3500000.txt_____G06F_____G06F16/951:G06F16/986
Computer system (100), computer-implemented method and computer program product are provided for extracting dynamic content data (221) from a website (220) in a machine-readable format. The system has an interface (110) to access configuration data (250) reflecting the structure of the website (220). The configuration data includes at least a website specific scraping script and one or more website specific XPath statements. Further, the interface receives a data retrieval request (210) specifying the website (220) and corresponding dynamic content data (221) to be retrieved. A scraper module (120) provides the scraping script (2050) for execution wherein the scraping script is configured to perform one or more parameterized navigation steps on the website (220) to access the dynamic content data (221). A script module (140) triggers execution of the scraping script and receives HTML/XML data associated with the dynamic content data from the website (220) in response to the scraping script execution. An XPath extraction module (150) extracts machine-readable content data (222) from the HTML/XML data wherein the XPath extraction module is pre-configured with the website specific XPath statements in accordance with the structure of the website (220).

file format

_____d:
Technical Field
The present invention generally relates to systems for data retrieval, and more particularly, relates to methods, computer program products and systems for extracting dynamic content from websites in machine-readable format.
Background
Web scraping or web data extraction methods are known in the art. Web scraping is used to access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. Web scraping typically refers to automated processes implemented using a bot or web crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

claims

_____c:

1. An apparatus for compression of untyped data comprising: a graphical processing unit (GPU) including a data compression pipeline, the data compression pipeline including: a data port coupled with one or more shader cores, wherein the data port is to allow transfer of untyped data without format conversion, and a 3D compression/decompression unit to provide for compression of untyped data to be stored to a memory subsystem and decompression of untyped data from the memory subsystem.
2. The apparatus of claim 1, wherein the apparatus is to utilize fixed sequential blocks for the storage of the untyped data.
3. The apparatus of claim 1, wherein the apparatus is to convert a buffer for the untyped data to a stateful buffer, the stateful buffer to identify untyped data for compression.
4. The apparatus of claim 1, wherein the memory allocation of the untyped data is to be determined by software, and wherein a GPU driver is to determine whether the untyped data is to be compressed.
5. The apparatus of claim 4, wherein the apparatus is to pass one or more hints regarding data compression of the untyped data to the GPU driver, wherein the one or more hints include one or more of whether compression should be enabled for a buffer and a native data size that maps to the buffer.
6. The apparatus of claim 5, wherein the GPU driver is to allocate an auxiliary buffer to store compression metadata.

Finding terms here

- using frequency?
- using an external dictionary?
  - with frequency?
    - how to count multi-word matches in a big corpus?
      - regex?
      - Trie
- using numbering in patent texts?
  - not everywhere

many terms

- A ready-made list:
  - manyterms.lower.txt contains all these potential terms from these categories and subcategories (and some heuristics such as minimum frequency, some no-go Categories, ...): ['Category:Physics', 'Category:Technology', 'Category:Technology_by_type', 'Category:Techniques']
  - 49166 categories, 2227891 pages
  - 700k potential multi-word terms
- problem:
  - terms may also consist of one token only!

Project

On a patent domain:

1. detection of technical terms
2. semantic relations between technical terms

Project

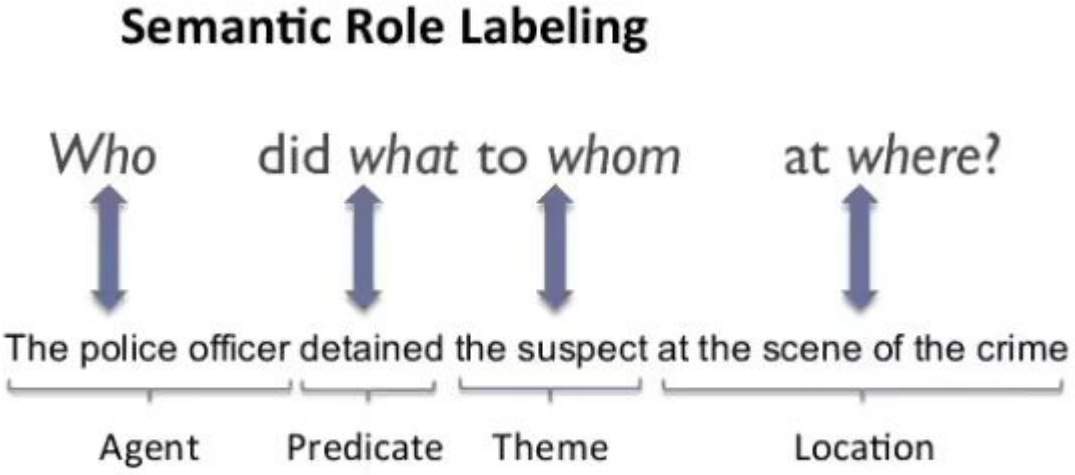
- Choose your partner
- choose your technical domain
  - beyond computer science?
- enter your names at the project table (see ecampus)

Project:
1. detection of technical terms

1. build gold
   a. describe what you call a term, criteria
2. build baseline
3. train term detector
4. evaluate on your gold and on SciIE

Project:
2. semantic relations between technical terms

Semantic parsing is the task of converting a natural language utterance to a logical form



different from syntax:
subject, object, modifier

the cable had been carelessly disconnected from the processor

Project:
2. semantic relations between technical terms

1. Each term should be related to at least one other term
   = each term should appear in one relationship
2. Let's look at
   a. http://nlp.cs.washington.edu/sciIE/
   b. https://prodi.gy/docs/api-interfaces#relations
   c. SQL queries https://guide.allennlp.org/semantic-parsing-seq2seq
   d. recent research on LLM https://arxiv.org/pdf/2209.15003.pdf
3. We want
   a. Synonyms, hyperonyms (is a type of), meronyms (contains, is part of)
   b. all other semantic relations as the verb form

TBA

- submission approximately 10 days after last class, until May 2 (?)
- details of required work
  - baselines, comparisons?
  - minimal set of relations?
  - compare to existing work?

notebook

- experimental session
- finding multi-word terminology by statistics
  could do better: mutual information, specificity, entropy
- matching terminology from a list
- training a NER for terminology

next time

- chunking, syntactic analysis
- NER: active learning
- tools for semantic analysis

let’s take a break

- grab the notebook on https://ecampus.paris-saclay.fr/course/view.php?id=28064#section-8
  - start the imports and download right away if you haven’t done so…
- today:
  - experimental session
  - finding multi-word terminology by statistics
  - matching terminology from a list
  - training a NER for terminology

Merci de votre

- attention
- considération
- intérêt
- écoute
- présence
- curiosité
- question

!