Extracting News Web Page Creation Time with DCTFinder

Xavier TANNIER
LIMSI-CNRS, Univ. Paris-Sud, Orsay, France,
firstname.lastname@limsi.fr

Motivation

1. Temporal parsing...

Temporal analysis of texts is often an essential component in a wide range of NLP and IR applications:
- **Question-Answering**
- Multidocument **summarization**
- **Timeline** building
- Medical **decision-making**

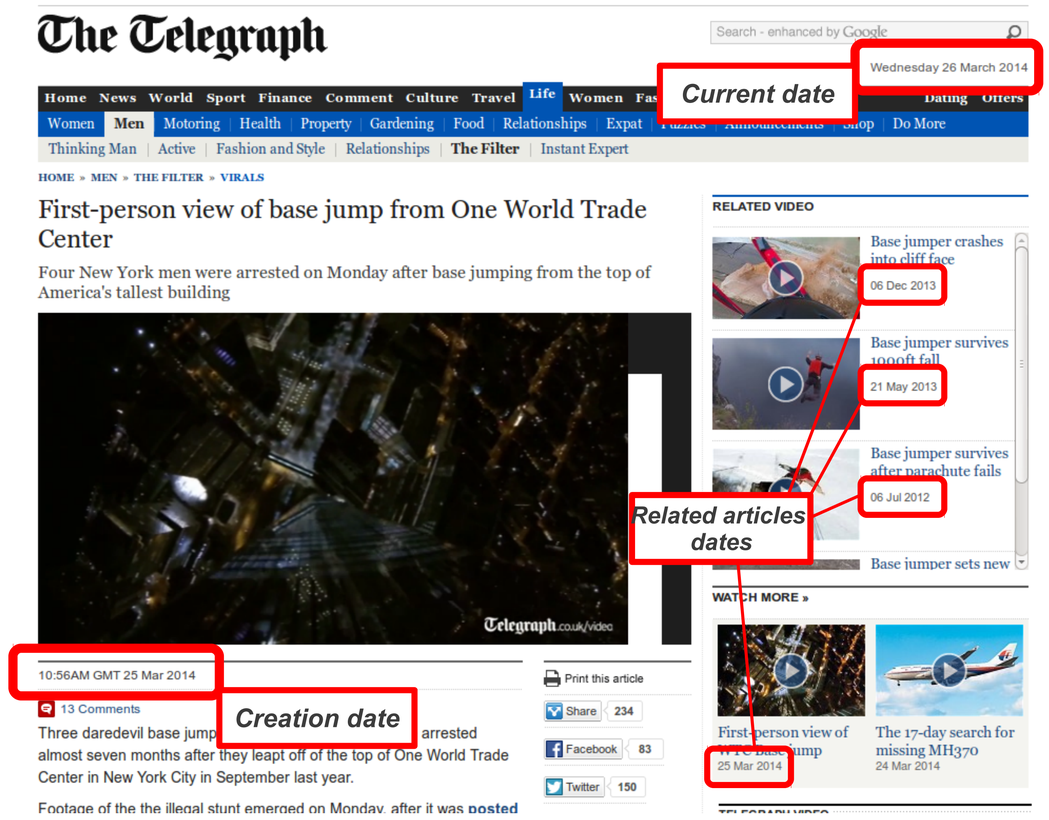
Tools like *Heideltime*, *SUTime*, *Timen*, *ManTIME*, etc. can be used to detect and normalize temporal expressions

2. ... of web pages...

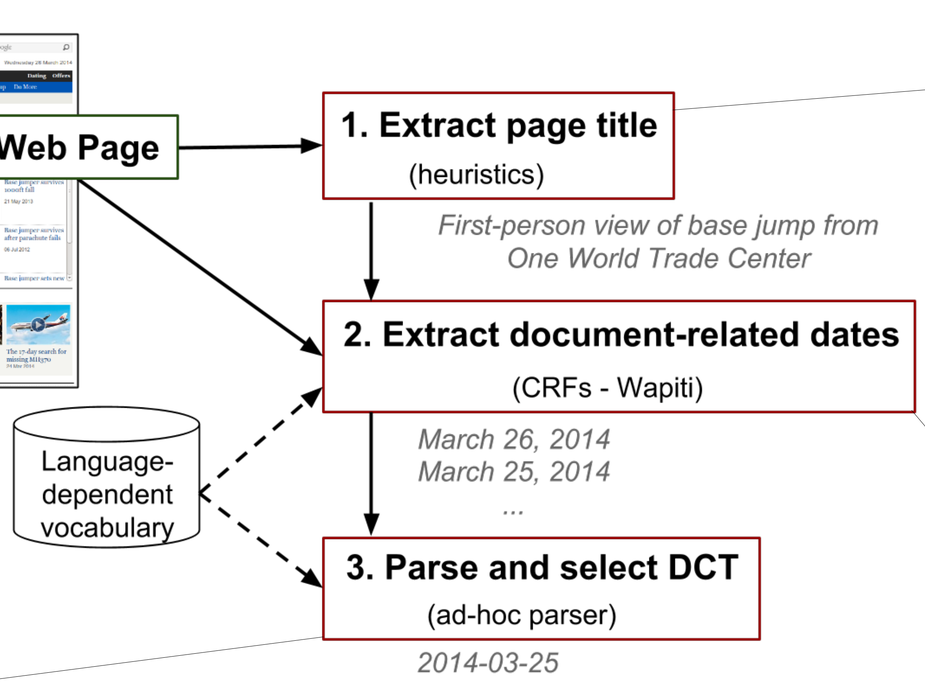
Two main issues make temporal parsing of web pages difficult:
- Web pages need to be **cleaned** before a proper analysis is performed on the text (textual content *vs.* menus, ads and non-informative content). This is addressed by cleaners such as *BodyTextExtraction*, *Boilerpipe*, *jusText*, *Readability*.
- There is **no reliable metadata providing the web page creation time**
  HTML5 `<pubdate>` is not used yet, all sites have a different way to insert the date in the HTML content.
  Server or RSS information are often wrong.

3. ... requires to extract the document creation time (DCT).

Almost all news web pages are time-stamped, but getting their creation date is not straightforward. A lot of dates occur in a web page, but only one is the creation date.



System Overview



1. Page title:
1. Content of tag `<h1>`, if only one `<h1>` is present in the document.
2. Content of any tag, if it is the longest string in the web page that is included in the HTML header `<title>` tag.
3. Content of tag `<h2>`, `<h3>` or `<h4>`, if only one such tag is present in the document.
4. Content of any tag, if the id or class attributes match language-dependent regular expressions (for example, ".*title.*", ".*headline.*").

Title proximity is an important clue for finding document dates

2. Document-related dates:
- Document creation date
- Last update date
- Current date ("now")

Context and structure of these three classes are similar and difficult to differentiate, so we don't try

We just want to discard "related articles" dates and dates inside the text

Extraction with Conditional Random Fields and Wapiti toolkit:
- **Lexical features** (language-dependent):
  - Date vocabulary and patterns (months, days, full dates, time zones, times)
  - Document date triggers (*"published"*, *"created"*, *"released"*...)
- **Structural features**:
  - Position in document, other dates around
  - Distance from title
  - Distance from triggers

Find full list of features and CRF templates in the LREC paper

3. Date extraction:
- The output of the CRF system is a list of tokens, where the tokens are tagged if they supposedly belong to document-related dates.
- Parsing dates from this output is straightforward (see "Language Dependence" for the exception).

Select DCT:
- Among document creation date, update date and "now", the DCT is always the oldest.

Also :
- If the URL is provided, try to extract the DCT from it
- If the download date is provided, avoid suggesting a DCT later than download date.

Language Dependence

US-English *vs.* other English:
- US-English often uses MM/DD/YYYY format, while others use rather DD/MM/YYYY, which can affect parsing when day ≤ 12.
- We use domain name extensions to (try to) handle this issue.
- But still, if you know if the page is US or non-US, you can specify it to the system to avoid confusion.

English *vs.* other languages:
- Applying models learned on English data to French leads to good results.
- This should be the same for most European languages.

Evaluation

Three corpora:
- Model learned on L3S-GN1 (from L3S), ~600 pages, 2007-2008 in English
- Tested on 100 more recent web pages in English
- Tested on 100 recent web pages in French

Dataset	Title Accuracy	DCT Accuracy
L3S-GN1 (cross-validation)	86.0%	92.4%
English recent dataset	94.0%	90.0%
French recent dataset	88.0%	87.0%

http://sourceforge.net/projects/dctfinder/