Citation extraction and modeling
Meen Chul Kim, Andrea Forte, Aaron Halfaker
History

2005 - Rebuilt Mediawiki with references as first class objects in the system.
   - it had a summary page and discussion for each reference!
   - it had a reverse citation index!
   - it had user tools to avoid duplication at the point of data insertion!

2006-2008 - Deployed in high schools and studied citation practices of science learners

2009-2017 - trying to get back home: lots of interview studies on motivations and meaning of citation to different Wikipedians, some efforts to extract and analyze references
Building a Dataset

The proposed dataset include the following constructs and attributes:

- **reference** - the text within an article that refers to a particular source
- **source** - an external resource that provides support for a statement in an article
- **article** - a concept being described

I think of *citation* as the activity.

References are our window for understanding sources.
A reference has the following fields

- **content** -- appears between the `<ref>` tags
  - **raw_content**
  - **templated** -- boolean, the entirety of the content a template call
  - **cite_template** -- citation templates are usually named something like `{{cite book|...}}` This field will contain the type the citation template if present.
- **urls** -- all the urls that appear in the content (domain)
- **identifiers** -- structured identifiers contained in the content (doi|isbn|arxiv|pubmed)

- **occurrences** -- the locations in the text that the `<ref>` tags appear
  - **section** -- the section # as defined by MediaWiki (starts at 0, split by headers)
  - **text_offset** -- the # of chars between the beginning of the article and the beginning of the `<ref>` tag
  - **preceding_text** -- 250 chars before the beginning of the `<ref>` tag.
  - **header_level** -- the level of the header immediately above the `<ref>` tag (0 for lead)
  - **header_text** -- the text of the closest header
  - **level_2_text** -- the text of the nearest level 2 header
  - **level_2_offset** -- the # of chars between the nearest level 2 header and the beginning of the `<ref>` tag

- **revid** -- revision ID associated with reference addition/deletion/edit
A source has the following fields

- **type (categorical)** -- books, journal articles, conference proceedings, magazines, mainstream newspapers, etc.
- **level (categorical)** -- primary, secondary, and tertiary
- **quality**
  - verifiability (see [Wikipedia Verification Check: A Chrome Browser Extension](#))
    - technical verifiability (boolean) -- the trustworthiness of a source' identifier (e.g. ISBN, DOI)
    - practical verifiability (boolean) -- the open accessibility to a source.
      - Has a URL
      - URL returns a 200
  - persistence (numerical) -- a continued duration of a source based on the history of reference revisions
- **relevance**
  - content similarity
    - raw distance (numerical) -- a source' textual proximity to titles, headers, specific sections, preceding text, etc.
    - semantic distance (numerical) -- semantic proximity referenced by WordNet
  - topological similarity (numerical) -- neighbor distance between sources, shortest paths, sigma of network diameter, etc.
- **importance**
  - cumulative measures (numeric) -- cited frequency, citation burst, etc.
  - topological measures (numeric) -- PageRank, HIT, betweenness centrality, sigma of modularity, network diameter, or clustering coefficient, etc.
  - possible measures: TBD
Reference Extraction Pipeline

**Wiki XML dumps**
All enwiki pages with full revision history

```xml
<page>
  <title>Wikipedia</title>
  <id>5043734</id>
  <text xml:space=..>
```

**Pattern extraction**
Island grammar pattern regex

```python
# 'ref' tag
r'\<ref\([^/]*\)?/\>|\<ref[^/]*\>[^\s]*\</ref>'

# section header
r'(^|\n)(=[^=]+=)|(==[^=]+=)|(===[^=]+===)'
```

**Parsing data**
MediaWiki parser & Apache Spark

```python
content: '{{cite web ..}}'
Url: 'http://www.niemanlab.org/2011/10/the-contrib..
 templated: True
level_2_text: 'History'
```

**Machine-learning citation quality**
Predictive models for grouping "good" and "bad" citations, i.e. citations that will be removed vs. those that stay
Other collaborative online encyclopedias were attempted before Wikipedia, but none were so successful.