Distil: a Distributed Information Library

a document, bibliography, and research knowledge management system, built on top of Git

James Boyden

20th of April, 2011

Why I'm presenting this

- I find it useful to organise the papers I read.

  ⇒ It might be useful to you too.

- It's Open Source software.

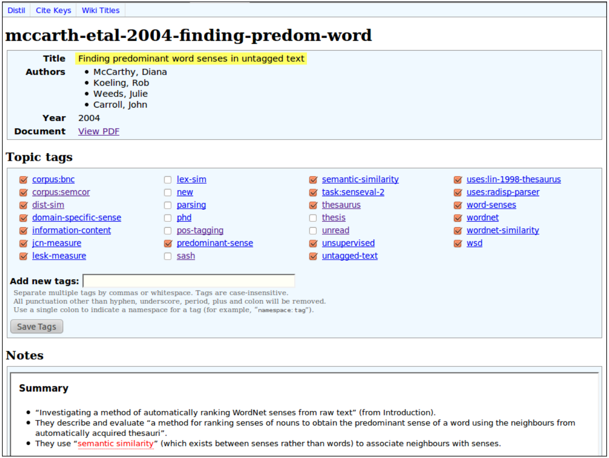
  ⇒ I welcome any additions or bugfixes.

  (Yes, it's written in Python.)

What Distil is

Two main components:

- a command-line tool (`distil`) to import new bibs & docs
- a Trac-inspired note-management wiki



mccarth-etal-2004-finding-predom-word

Title	Finding predominant word senses in untagged text
Authors	• McCarthy, Diana • Koeling, Rob • Weeds, Julie • Carroll, John
Year	2004
Document	View PDF

Topic tags

☑ corpus:bnc
☑ corpus:semcor
☑ dist-sim
☑ domain-specific-sense
☑ information-content
☑ jcn-measure
☑ lesk-measure
☐ lex-sim
☐ new
☐ parsing
☐ phd
☐ pos-tagging
☑ predominant-sense
☐ sash
☑ semantic-similarity
☑ task:senseval-2
☑ thesaurus
☐ thesis
☐ unread
☑ unsupervised
☑ untagged-text
☑ uses:lin-1998-thesaurus
☑ uses:radisp-parser
☑ word-senses
☑ wordnet
☑ wordnet-similarity
☑ wsd

Add new tags:

Separate multiple tags by commas or whitespace. Tags are case-insensitive.
All punctuation other than hyphen, underscore, period, plus and colon will be removed.
Use a single colon to indicate a namespace for a tag (for example, "namespace: tag").

Notes

Summary

- "Investigating a method of automatically ranking WordNet senses from raw text" (from Introduction).
- They describe and evaluate "a method for ranking senses of nouns to obtain the predominant sense of a word using the neighbours from automatically acquired thesauri".
- They use "semantic similarity" (which exists between senses rather than words) to associate neighbours with senses.

Why I built it

I wanted a system that...

- could store docs internally (PDF, DOC, PS.GZ, etc.)
- could synchronise easily
  - between home and uni
  - or to a backup location
  - asynchronously
- allowed me to take notes easily
  - in a wiki-like syntax
  - with inter-note wiki-links
  - using Vim keybindings (or Emacs!)
- allowed me to tag/label easily
- incorporated version control

Why I built it, part 2

I wanted a system that...

- could import new bibs & docs quickly and easily
- understood BibTeX
- could auto-generate consistent BibTeX cite-keys
    - which were also mnemonic
    - but not *too* long
- could work without a network connection
- allowed me to view/edit multiple items simultaneously
- was Unicode-friendly

Some disclaimers

- “Agile” development methodology
  ⇒ very feature-incomplete
- Definitely “alpha” software

Some assurances

- I use it every day
- Built upon the stable foundation of Git
  - If it fails halfway through an operation (rare)
    $\Rightarrow$ revert/complete the uncommitted changes
  - If it commits something wrong (never happened yet)
    $\Rightarrow$ revert the changeset
  - The Git wrapper functions are simple and well-tested
  - No database lockups/corruption!

The `distil` command

```
distil somefoo.bib otherbar.pdf whatever.abs
```

- Command-line arguments:
  1. a BibTeX **bibliography**
  2. the **document** (pdf, doc, ps.gz, etc.) *(optional)*
  3. a plain-text **abstract** *(optional)*

- What happens:
  1. Read the BibTeX file, auto-generate a cite-key
  2. Create a new directory in the repository
  3. Move the files into the new directory
  4. Rename the files
  5. Update the cite-key in the BibTeX file
  6. Git commit

The cite key

- Auto-generated from the BibTeX file:
  - (at most) **2 author** components
  - **year**
  - (at most) first **3 title** components
- For example:

  "Finding predominant word senses in untagged text"

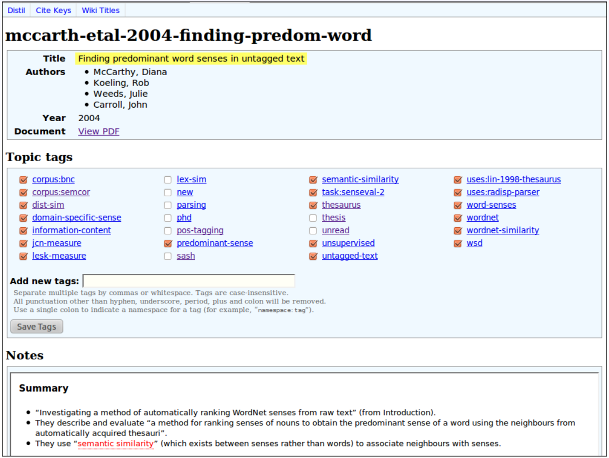
  by McCarthy et al (2004)

  becomes:

  `mccarth-etal-2004-finding-predom-word`

- Actual examples!
  - `mccarth-etal-2004-automat-acquir-predom`
  - `mccarth-etal-2004-automat-identif-infreq`
  - `mccarth-etal-2004-finding-predom-word`
  - `mccarth-etal-2004-ranking-wordnet-senses`

The wiki



mccarth-etal-2004-finding-predom-word

Title	Finding predominant word senses in untagged text
Authors	• McCarthy, Diana • Koeling, Rob • Weeds, Julie • Carroll, John
Year	2004
Document	View PDF

Topic tags

- ☑ corpus:bnc
- ☑ corpus:semcor
- ☑ dist-sim
- ☑ domain-specific-sense
- ☑ information-content
- ☑ jcn-measure
- ☑ lesk-measure
- □ lex-sim
- □ new
- □ parsing
- □ phd
- □ pos-tagging
- ☑ predominant-sense
- □ sash
- ☑ semantic-similarity
- ☑ task:senseval-2
- ☑ thesaurus
- □ thesis
- □ unread
- ☑ unsupervised
- ☑ untagged-text
- ☑ uses:lin-1998-thesaurus
- ☑ uses:radisp-parser
- ☑ word-senses
- ☑ wordnet
- ☑ wordnet-similarity
- ☑ wsd

Add new tags:

Separate multiple tags by commas or whitespace. Tags are case-insensitive.
All punctuation other than hyphen, underscore, period, plus and colon will be removed.
Use a single colon to indicate a namespace for a tag (for example, "namespace: tag").

Notes

Summary

- "Investigating a method of automatically ranking WordNet senses from raw text" (from Introduction).
- They describe and evaluate "a method for ranking senses of nouns to obtain the predominant sense of a word using the neighbours from automatically acquired thesauri".
- They use "semantic similarity" (which exists between senses rather than words) to associate neighbours with senses.

http://pc-4e32-1.it.usyd.edu.au:8888/

What we have today

- Importing bibs, docs, abstracts
- Auto-generating cite-keys
- Git wrapper functions
- Wiki for notes on bibs and topics
  - wiki-markup
  - inter-wiki links (including inter-bib “cite” links)
- Topic-tagging of bibs
- Indexes of tags and cite-keys

Some potential future features

- Topic-tagging wiki pages
- Attachments
- `/timeline`
- Creating/editing BibTeX bibliographies
- Better support for LateX in BibTeX
- Indexing/search
- Track-backs
  - Links to this page.
  - This paper was cited by *X*.
  - *X* said “Interesting comment” about this paper.
- `/authors/curran/j*`
- *Actually* being distributed

  (between people in a team, not just different computers)

Thank you

Any questions?