Voyant Tools for Japanese Text Analysis

Molly Des Jardin
University of Pennsylvania
@mdesjardin / mollydesjardin.com
Text Analysis?

"Computer-assisted reading" of texts (using Natural Language Processing)

Need to have machine-readable text – depends on your analysis

- Plain text
- Marked up in TEI, XML, or other formats
- Tagged with parts of speech or other linguistic information

The more semantic, the more difficult

- Tokenizing pretty easy; POS tagging not too hard; representing concepts and relationships really hard!

Can be "distant" or "close" reading or in-between

Requires human interpretation
Real-World NLP Techniques & Applications

Topic Modeling
- Identify possible themes and areas of interest in huge bodies of text
- Ex: Wikileaks cables, a corpus of 19th-century novels

Sentiment Analysis
- Often used on things like product reviews or Twitter data in business context

Automatic Summarization and Text Generation
- Used to automatically write news articles about relatively simple topics

Concordance and Keyword-in-Context
- For looking at language use in small or large corpora
- "Closer" reading than a human would want to do or could do
Why Is Japanese Hard?

TOKENIZATION

Most existing tools expect whitespace-separated words like English. Linguists could argue about where to separate "words" in Japanese. Several programs available for tokenizing, specialize in different types of language:

- Tiny Segmenter, MeCab IPADic -> Contemporary
- MeCab with UniDic for 近代, 近世, 中世, etc. -> Historical
- MeCab with UniDic for Internet Language -> Specialized
- Voyant Tools uses its own tokenizer

ENCODINGS – Shift-JIS, Unicode
Voyant Tools

Now with Japanese out of the box! Includes:

• Tokenization
• Beginnings of (quirky) Japanese stopword list
• Japanese interface

Various visualizations of basic analyses – nothing semantic or syntactic

Word cloud, concordance, word frequency, keyword in context, uniqueness of words, word counts

A way to begin exploring your texts
How to Use Voyant Tools

Go to Voyant's website or run on your computer (see docs for info)
Choose a sample corpus, upload files, or copy-paste URLs or text
Choose whether to tokenize Japanese in the upload page options
Customize your stop word list
  • These are words you want to ignore in the word cloud, etc.
  • Common words or fragments, or things you just don't care about
  • Changes depending on your corpus and purpose!
Don't use Safari; works fine on Firefox; haven't tested other browsers
Go!

Voyant at http://voyant-tools.org
Get some files of your own – ex. Aozora Bunko or even news articles
Customize your stopword list per your interests and corpus
Compare pre-tokenized vs. Voyant-tokenized files
Try out pasting news articles or content from Aozora Bunko
See about further paths for exploration from your initial results!
Feel free to contact developers with questions and feedback!