Internationalizing Quantitative Text Analysis

27 June 2018

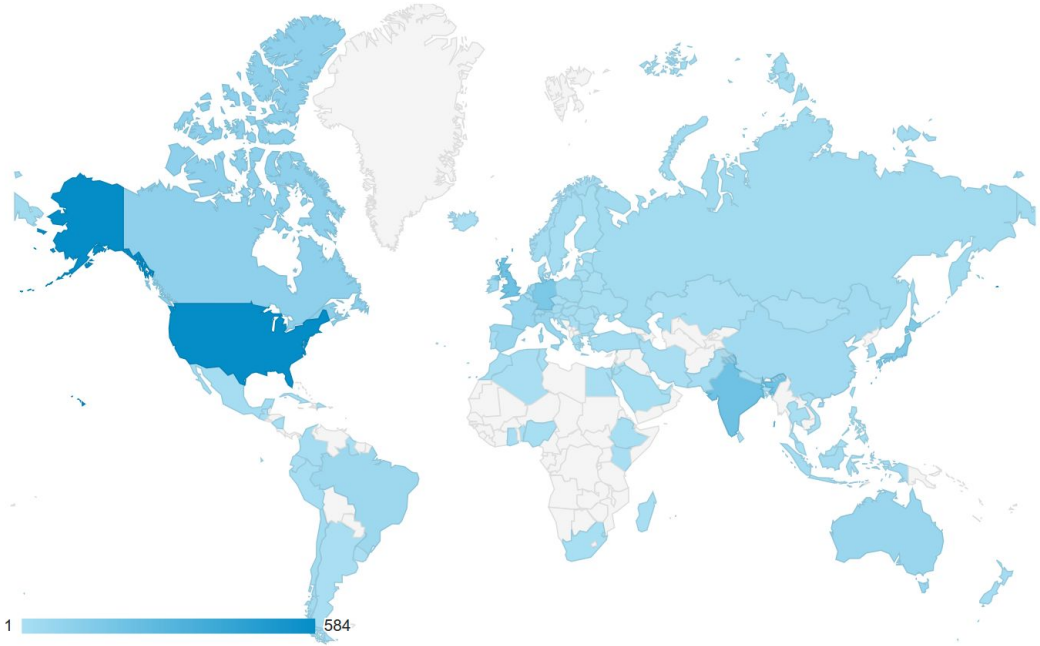
Kohei Watanabe (WIAS/LSE)
Oul Han (GESIS/SNU)

LSE SEDS Social and Economic Data Science

QI quanteda initiative

Visitor to docs.quanteda.io

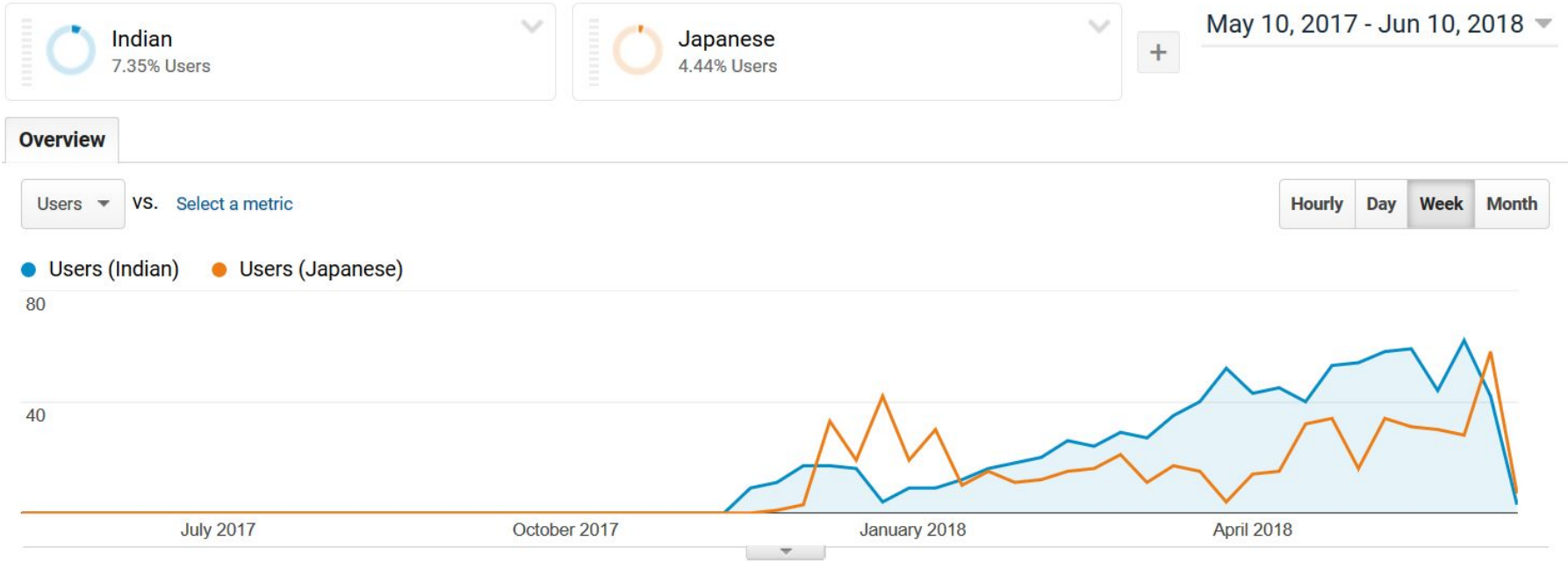
Quanteda's users are concentrated in the US and Europe



Country	Acquisition
	Users ↓
	2,433 % of Total: 100.00% (2,433)
1. United States	584 (23.64%)
2. United Kingdom	209 (8.46%)
3. India	208 (8.42%)
4. Germany	163 (6.60%)
5. Japan	148 (5.99%)
6. Canada	98 (3.97%)
7. Netherlands	78 (3.16%)
8. France	68 (2.75%)
9. Spain	59 (2.39%)
10. Italy	59 (2.39%)

Indian and Japanese visitors

However, we are getting more Asian users.



Keys to internationalizing text analysis

They are mainly about tools but not necessarily.

- Tools (software packages, morphological analysis tools)
- Data (corpus, dictionary)
- Skills (knowledge of statistics and programing)
- Literature (text books, papers)

quanteda
Quantitative Analysis of Textual Data

Tools

quanteda
Quantitative Analysis of Textual Data

Text analysis tools

There aren't a lot of text analysis packages that can handle Asian languages.

- KH Coder (MySQL + Java + R + perl + Mecab)
  - Officially supports Japanese, Chinese, Russian, Korean
  - Widely used in Japan but not in the overseas
- Quanteda (R)
  - Supports all languages in International Component for Unicode (ICU)
  - Use in the US and Europe, but less so in Asia
- Homegrown tools (Python + morphological analysis tool)
  - Python is not so good with Unicode
  - Usually heavily system dependent

quanteda
Quantitative Analysis of Textual Data

Morphological analysis

表層形	品詞	品詞細分類1	品詞細分類2	品詞細分類3	活用型	活用形	原形	読み	発音
○	記号	一般	*	*	*	*	○	○	○
宮本	名詞	固有名詞	人名	姓	*	*	宮本	ミヤモト	ミヤモト
（	記号	括弧開	*	*	*	*	（	（	（
徹	名詞	固有名詞	人名	名	*	*	徹	トオル	トール
）	記号	括弧閉	*	*	*	*	）	）	）
委員	名詞	一般	*	*	*	*	委員	イイン	イイン
	記号	空白	*	*	*	*			
内容	名詞	一般	*	*	*	*	内容	ナイヨウ	ナイヨー
について	助詞	格助詞	連語	*	*	*	について	ニツイテ	ニツイテ
差し控え	動詞	自立	*	*	一段	未然形	差し控える	サシヒカエ	サシヒカエ
させ	動詞	接尾	*	*	一段	連用形	させる	サセ	サセ
て	助詞	接続助詞	*	*	*	*	て	テ	テ
じゃ	助詞	副助詞	*	*	*	*	じゃ	ジャ	ジャ
なく	助動詞	*	*	*	特殊・ナイ	連用テ接続	ない	ナク	ナク
て	助詞	接続助詞	*	*	*	*	て	テ	テ

Morphological analysis tools (1)

Morphological analysis tools are used in tokenization of Japanese

- Mecab (Japanese, Korean and Chinese)
  - RMecab (not on CRAN)
    - Documentation is only available in Japanese
  - RcppMecab (on CRAN)
    - Internationalization of RMecabKo
    - Supports both Korean, Japanese and Chinese
    - Documented in English
  - mecab-python
    - No longer maintained
- Rakuten MA (Japanese and Chinese)
  - javascript library

quanteda
Quantitative Analysis of Textual Data

Morphological analysis tools (2)

Morphological analysis tools are used in tokenization of Chinese and Korean

- HanNanum, Kkma, Komoran, Twitter Korean Text, ... (Korean)
  - Java libraries, different performances, choose according to aim
  - Available all-in-one via KoNLP (on CRAN), KoNLPy (Python)
- jieba (Chinese)
  - Python module
  - jiebaR (on CRAN)

quanteda
Quantitative Analysis of Textual Data

Challenges in tokenizing Korean

- We can “stem” Japanese but not Korean:

하다	行う	do
하였다	行った	did
했다	行った	did (shorter form)
했었다	行った	had done

- Two possible approaches:
  - Apply morphological analysis tools to identify lemma
  - Tokenize only nouns using noun extraction tools (HanNanum)

quanteda
Quantitative Analysis of Textual Data

Part-of-speech (POS) tagging

doc_id	sentence_id	token_id	token	lemma	pos	entity
text1	1	1	When	when	ADV	
text1	1	2	I	-PRON-	PRON	
text1	1	3	presented	present	VERB	
text1	1	4	the	the	DET	
text1	1	5	supplementary	supplementary	ADJ	
text1	1	6	budget	budget	NOUN	
text1	1	7	to	to	ADP	
text1	1	8	this	this	DET	
text1	1	9	House	house	PROPN	ORG_B
text1	1	10	last	last	ADJ	DATE_B
text1	1	11	April	april	PROPN	DATE_I
text1	1	12	,	,	PUNCT	
text1	1	13	I	-PRON-	PRON	
text1	1	14	said	say	VERB	
text1	1	15	we	-PRON-	PRON	
text1	1	16	could	could	VERB	

Application of POS tools

Part-of-speech tagging tools can be adopted for Asian languages, but usually not as good as specialist tools

- StanfordNLP (Chinese)
  - Java tool for part-of-speech tagging
- Spacy (Chinese, Japanese, Thai, Vietnamese etc.)
  - Python tool for part-of-speech tagging
  - Currently only supports tokenization
- Universal Dependencies (60 languages including CJK)
  - udpipe (on CRAN)

quanteda
Quantitative Analysis of Textual Data

Dictionary-based tokenization

Use of external tools is not necessary

- International Component for Unicode (ICU)
  - ICU is the IT infrastructure
    - Adobe, Amazon, Apple, DELL, Google, IBM etc.
    - Also part of Python and Java libraries
  - ICU defines word boundaries using dictionary
    - Dictionary for Chinese and Japanese are developed based on existing lexicon
      - IPA for Japanese
      - Libtabe for Chinese
    - However, dictionary is missing for Korea
  - The string package provides interface to ICU in R

quanteda
Quantitative Analysis of Textual Data

Tokenization by ICU

Japanese segmentation without morphological analysis

```
> txt_jp <- "政治とは社会に対して全体的な影響を及ぼし、社会で生きるひとりひとりの人の人
生にも様々な影響を及ぼす複雑な領域である。"
> quanteda::tokens(txt_jp)
tokens from 1 document.
text1 :
 [1] "政治"         "と"           "は"           "社会"         "に対して"
 [6] "全体"         "的"           "な"           "影響"         "を"
[11] "及"           "ぼ"           "し"           "、"           "社会"
[16] "で"           "生きる"       "ひとりひとり" "の"           "人"
[21] "の"           "人生"         "に"           "も"           "様々"
[26] "な"           "影響"         "を"           "及ぼす"       "複雑"
[31] "な"           "領域"         "で"           "ある"         "。"
```

quanteda
Quantitative Analysis of Textual Data

Tokenization by ICU

Chinese segmentation without morphological analysis

```
> txt_cn <- "政治是各种团體进行集体决策的一个过程，也是各种团體或个人为了各自的領域所结
成的特定关系，尤指對於某一政治實體的統治，例如統治一個國家，亦指對於一國內外事務之監督
與管制。"
> quanteda::tokens(txt_cn)
tokens from 1 document.
text1 :
 [1] "政治"   "是"     "各种"   "团"     "體"     "进行"   "集体"   "决策"
 [9] "的"     "一个"   "过程"   "，"     "也是"   "各种"   "团"     "體"
[17] "或"     "个人"   "为了"   "各自"   "的"     "領域"   "所"     "结成"
[25] "的"     "特定"   "关系"   "，"     "尤"     "指"     "對於"   "某一"
[33] "政治"   "實體"   "的"     "統治"   "，"     "例如"   "統治"   "一個"
[41] "國家"   "，"     "亦"     "指"     "對於"   "一"     "國內外" "事務"
[49] "之"     "監督"   "與"     "管制"   "。"
```

quanteda
Quantitative Analysis of Textual Data

Data

quanteda
Quantitative Analysis of Textual Data

Sources of textual data (Japanese)

- Politics
  - Japanese National Diet Minutes API (http://kokkai.ndl.go.jp/api.html)
    - Accessible via the kaigiroku package (https://github.com/amatsuo/kaigiroku/)
  - Japanese Local Political Corpus (http://local-politics.jp/)
    - No API to download raw text yet
- Online forum
  - Yahoo Japan News
- Mass media
  - Asahi Shimbun (Kikuzo database)
  - Yomiuri Shimbun (Yomidas database)
  - Full newspaper corpus (http://www.nichigai.co.jp/sales/corpus.html)
- More...

quanteda
Quantitative Analysis of Textual Data

Sources of textual data (Korean)

- Politics
  - Politics in Korea API (http://en.pokr.kr/main, http://data.popong.com/)
- Online forum
  - Naver News
- Mass media
  - KINDS (https://www.kinds.or.kr/)
  - Dow Jones Factiva
- More...

quanteda
Quantitative Analysis of Textual Data

Skills

quanteda
Quantitative Analysis of Textual Data

Necessary skills and knowledge

- Data collection
  - API
    - Knowledge of machine readable formats (XML, JSON etc.)
  - Scraping
    - Advanced skill in programing (R or Python)
    - Knowledge of HTML, javascript and Selenium
- Analysis
  - Basic knowledge of
    - Descriptive and inferential statistics (chi-square, t-test, regression analysis)
    - network analysis
  - Basic skill in programing R or Python
    - However, KH Coder does not require programming skill
- Research design
  - Broad knowledge of social scientific text analysis is essential

quanteda
Quantitative Analysis of Textual Data

Literature

quanteda
Quantitative Analysis of Textual Data

Textbook

Many people need a textbook to teach themselves, but there aren't any.

- There is no good textbook on the application of text analysis
  - Textbooks on text analysis are usually about computer programming
    - Ken Benoit should be writing a social scientific textbook
  - Textbooks in computer science are about complex algorithms
- CJK languages require special handling
  - Language specific textbooks are needed for the analysis of CJK languages
    - We have created a section for Japanese (https://tutorials.quanteda.io/language-specific/japanese/)
    - Add pages for Korean and Chinese

Research paper

More papers using text analysis should be available, but there are obstacles.

- Papers on CJK texts might be difficult to publish in English journals
  - Weak incentives to use text analysis to study Asian countries
- Asian journals might be more conservative about methodology
  - Methodologically innovative papers are more difficult to publish
  - Comparative analysis in Asian languages and English is necessary
- There is no international and interdisciplinary venue for text analysis
  - Kyushu's Multidisciplinary Text-mining Colloquium is too domestic
  - Asia PolMeth is good, but too narrowly focuses on political science

Summary (1)

- Tools
  - Tokenization of CJK text is becoming easier but not without problems
    - Cross-lingual morphological analysis tools (e.g. rcppMecab)
    - Tokenization without morphological analysis (e.g. ICU)
  - Requires international effort and demand stimulation to improve support of CJK languages
  - For the moment, there are pragmatic workarounds (e.g. noun stemming in Korean)
- Data
  - Political texts are becoming increasingly accessible in Asia
    - Comprehensive list of sources is needed
  - Media text is still difficult to collect
    - Scrapers is required for large scale data collection

Summary (2)

- Skill
  - Advanced quantitative text analysis requires programming skills
    - Waseda University’s Global Education Center will offer a statistics course using R from the next term
    - However, there is no course about programming itself
  -
- Literature
  - Needs more opportunities and incentives to publish analysis of CJK languages

Conclusions

Actions to promote text analysis in CJK languages:

1. Develop more educational materials for CJK languages
    - Develop techniques to make Asian and European languages comparable
    - Translate materials into local languages
2. Improve accessibility to tools and datasets
    - Create a centralized list of tools (e.g. scrapers) and datasets
3. Expose students and researchers to the latest projects
    - Invite leading figures from overseas for research seminars
    - Fund participation to text conferences (Text-as-data, Manifesto Corpus, PolText etc.)
4. Establish interdisciplinary venue or medium for CJK text analysis
    - International and multidisciplinary events on text analysis should be organized
    - Launch an open-access journal on Asian text analysis