Launching Your Text Analysis Projects

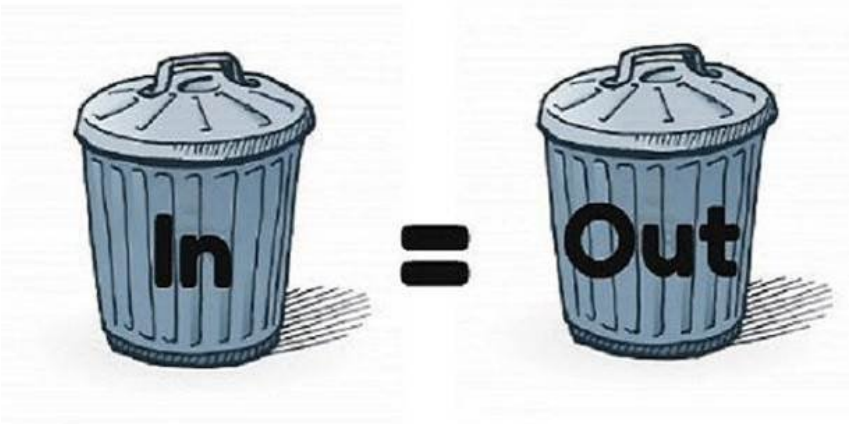
Institute for Qualitative and Multi-Method Research 2023

Zenobia Chan (Princeton) Will Lowe (Hertie)
June 30, 2023

Outline

Outline

Garbage in, Garbage out



Where to find the data I need?

1. Downloading from the source in a clean text or CSV file (in an ideal world...)
2. Ask the source
3. Scraping
4. Scanning + OCR
5. Typing (life can be hard...)

Some Common Databases for Downloading Text Data

Very limited selection of examples, due to my ignorance

- Official Document System of the United Nations
- *Foreign Relations of the United States* (*FRUS*)
- UCSB Presidency Project
- Newspaper databases
  - Factiva
  - WiseNews (Chinese newspapers)
  - ~~CNKI (Chinese newspapers)~~ (**2023 update:** Usable only if you are physically in Mainland China)
  - EastView (Russian newspapers)
- ~~Twitter API~~ (**2023 update:** Usable only if you pay some hefty fees...)
- National Diet Library of Japan
- British National Archives

Potential Biases in Text Data (and data in general)

- Who is generating the data? Why are they generating the data?
- What is omitted in this data?
  - What is not written down or said can be as important as what is written down or said
  - E.g. *FRUS* is a curated selection of archival materials!
- Why is this data available?

Scraping Websites

Let's go to R

Outline

Scriptio continua: Some languages have no space!

- Chinese
- Japanese
- Javanese
- Thai
- etc.

An *Actual* Newspaper Headline from Hong Kong

兒子生性病母倍感安慰

苦盡甘來

56歲的張婉貞，由於丈夫沉迷賭博欠債纍纍，01年時終難以忍受，與其離婚，因情緒受困，1年後更患上乳癌，面對不斷的

而丈夫報讀一些課程後，亦開始understand她，女兒見其狀況轉好，早前更向她表示明白她一直以來的痛苦，「咁已經好開心，好過佢界錢我用好多」

梁女士認為，家人與她每天共同生活，對其支持非常重要。張婉貞亦表示，現時與兒子關係如同朋友，希望天下母親亦能與子女有良好溝通，共度歡樂的母親節。

- 兒子 | 生性 | 病母 | 倍 | 感 | 安慰
  - The sick mother feels much comfort as the son is well-behaved
- 兒子 | 生 | 性病 | 母 | 倍 | 感 | 安慰
  - The mother feels much comfort as the son catches sexually transmitted infections
- ***Both*** ways of tokenization are grammatical!

Useful R Packages

Languages Quanteda can process

- English, German, Russian (quite well)
- Arabic, Hebrew, Chinese, Japanese (okay)
- More info: `https://tutorials.quanteda.io/multilingual/`

Useful R Packages

Languages Quanteda can process

- English, German, Russian (quite well)
- Arabic, Hebrew, Chinese, Japanese (okay)
- More info: `https://tutorials.quanteda.io/multilingual/`

Different variations of Chinese

- `jiebaR`
  - Developed by Baidu
  - Works better for simplified Chinese
- `songotsti`
  - Works best for Cantonese simplified Chinese
  - `devtools::install_github("justinchuntingho/songotsti")`

Useful R Packages

Languages Quanteda can process

- English, German, Russian (quite well)
- Arabic, Hebrew, Chinese, Japanese (okay)
- More info: `https://tutorials.quanteda.io/multilingual/`

Different variations of Chinese

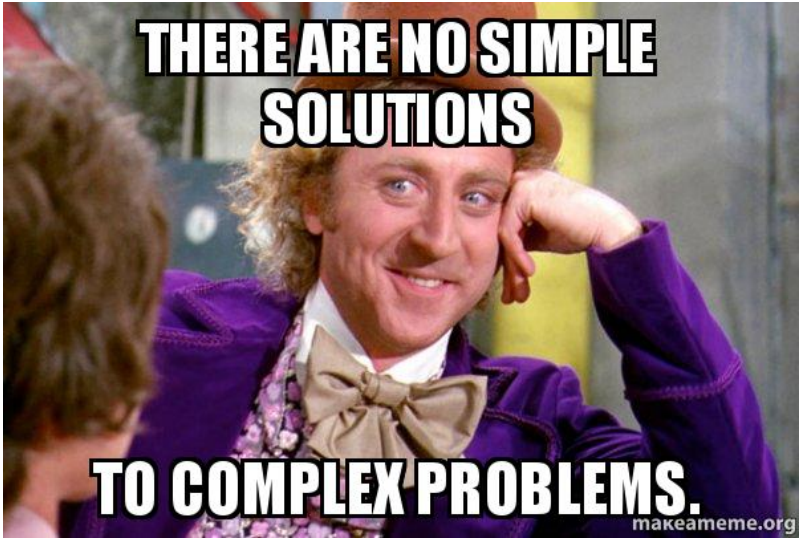
- `jiebaR`
  - Developed by Baidu
  - Works better for simplified Chinese
- `songotsti`
  - Works best for Cantonese simplified Chinese
  - `devtools::install_github("justinchuntingho/songotsti")`

Arabic

- Stanford Tokenizer
  - Developed by the Stanford Natural Language Processing Group
  - Works for Chinese too

Outline

Basically...


THERE ARE NO SIMPLE SOLUTIONS
TO COMPLEX PROBLEMS.
makeameme.org

Some Common Solutions

Machine translation to English

1. Translate everything to English
   - Google Translate API
   - ChatGPT API
   - Etc.
2. Merge the translated corpa (from different langauges) into one corpus (in English)
3. Analyze the merged corpus in English from Step 2

Some Common Solutions

Machine translation to English

1. Translate everything to English
   - Google Translate API
   - ChatGPT API
   - Etc.
2. Merge the translated corpa (from different langauges) into one corpus (in English)
3. Analyze the merged corpus in English from Step 2

Human coding

1. Design a coding scheme
2. Code the text data in different languages following the same coding scheme
3. Analyze the coded data

A Few Things to Consider...

1. How well is the machine translation in the language(s) of interest?
2. What is/are the purpose(s) of the analysis?
   - Something quick and dirty
   - Something subtler
   - Etc.
3. How much time and resources do you have and want to invest in?

Outline

Three Things to Takeaway

1. Text analysis is ***no*** magical solution
   - Requires researcher's (i.e. ***your***) knowledge about the text dataset

Three Things to Takeaway

1. Text analysis is ***no*** magical solution
   - Requires researcher's (i.e. ***your***) knowledge about the text dataset
2. Context matters!
   - Data generating process: Why is the data even there in the first place?
   - What are the potential biases? How will they affect your results?

Three Things to Takeaway

1. Text analysis is ***no*** magical solution
   - Requires researcher's (i.e. ***your***) knowledge about the text dataset
2. Context matters!
   - Data generating process: Why is the data even there in the first place?
   - What are the potential biases? How will they affect your results?
3. Validate, validate, validate...
   - You ***can*** discard nonsensical output