**Technology for Text Analytics**

1. **Introduction: Tools for Handling Text Data**
2. Data acquisition: Collecting and converting text data for analysis (e.g., downloading documents)
3. Data preprocessing: Restructuring the collected data for analysis (e.g., tagging data)
4. Data analysis: Getting insights from data (e.g., identifying features/patterns, extracting useful information, ~~finding relationships and differences~~, and visualizing insights)
5. **Tool types**
6. **For all users**
7. Stand-alone tools (單機版工具)
8. Online tools (線上系統)
9. **For users with programming knowledge (when current tools do not provide features you need)**
10. Libraries (函式庫)
11. Application Programming Interfaces (APIs, 應用程式介面)
12. **Useful Tools**

**Tools for PDF-to-text Conversion**

|  |
| --- |
| **Examples**  File 1: 香港行政長官施政報告2017（中文版）  <https://www.policyaddress.gov.hk/2017/chi/pdf/PA2017.pdf>  File 2: The Chief Executive’s 2017 Policy Address (English Version)  <https://www.policyaddress.gov.hk/2017/eng/pdf/PA2017.pdf> |

1. Word [Demonstration: Open your pdf file with Word 2013+; Save it as Plain Text]
2. Google Docs [Demonstration: Upload your pdf file and open it with Google Docs; Download as Plain Text]
3. Lotapps (<http://www.lotapps.com/>) [Demonstration: Convert both Chinese and English files at the same time]

**Tools for Image-to-text Conversion: Optical Character Recognition**

|  |
| --- |
| **Examples**  File 3: 藥材與重金屬（中文版）  <http://www.cmd.gov.hk/html/b5/health_info/pamphlet/pdf/Herbal_Medicines_and_Heavy_Metals_tc_pp.pdf>  File 4: Herbal Medicines and Heavy Metals (English Version)  <http://www.cmd.gov.hk/html/eng/health_info/pamphlet/pdf/Herbal_Medicines_and_Heavy_Metals_en_pp.pdf>  Files 5a and 5b: Sample files (in Japanese)  [**http://dl.ndl.go.jp/info:ndljp/pid/953214**](http://dl.ndl.go.jp/info:ndljp/pid/953214) **(Select: 12)**  [Output files: JPEG and PDF (see files 5a and 5b)] |

1. OneNote [Demonstration: Use “Snipping tool” (剪取工具) and OneNote for the first two files.]
2. Google Drive (image, including images in PDF) [Demonstration: Open PDF/images with Google Docs (all the four files)]
3. Convertio (<https://convertio.co/ocr/japanese/>) [Demonstration: JPEG file in Japanese; note that the layout is kept.]

**Tools for Speech-to-text Conversion: Speech Recognition**

|  |
| --- |
| **Examples**  File 6: Weekly Address by President Trump  File 7: Chinese News |

1. Temi (<https://www.temi.com/>)
2. IBM Watson Speech to Text (<https://speech-to-text-demo.ng.bluemix.net/>) [Demonstration: Chinese (Use file 7), English and Japanese (Use the demo audio files)]

Other options (Login is required for trial)

* Transcribe (<https://transcribe.wreally.com/>)
* Vocalmatic (<https://vocalmatic.com/>)
* Sonix (<https://sonix.ai/>)
* VoiceBase (<https://app.voicebase.com/login>)

**Tools for Text Cleaning**

|  |
| --- |
| **Example**  File 8: Policy Address 2017 (Chinese) from pdf-to-text Conversion |

1. Text Cleaner (<https://approsto.com/text-cleaner/>) [Demonstration: remove empty lines; combine paragraphs]
2. TextCleanr (<http://www.textcleanr.com/>) [Demonstration: remove spaces]
3. Text Tool (<https://www.computerhope.com/text-tool/>)

**Tools for Text Comparison and Alignment**

|  |
| --- |
| **Examples**  Files 9 and 10: English Text Comparison  Files 11 and 12: Chinese PDF Comparison  Files 9 and 13: Bilingual Text Alignment |

1. Diffchecker (<https://www.diffchecker.com/diff>) [Demonstration: English only for free online version]
2. Draftable (<https://draftable.com/compare>) [Demonstration: Chinese pdf files]
3. Alignment (<https://www.tmxmall.com/aligner>)

**Tools for Segmentation and Tagging**

|  |
| --- |
| **Examples**  File 14: English News  File 15: Chinese News |

1. Part of Speech Tagging (<http://cogcomp.cs.illinois.edu/page/demo_view/pos>)
2. PHP簡易中文分詞 (<http://www.xunsearch.com/scws/demo/v4.php>) (Check the box 標註詞性) (Details: <http://www.xunsearch.com/scws/docs.php#attr>)
3. Stanford Parser (<http://nlp.stanford.edu:8080/parser/>)
4. TextAnalysisOnline (<http://textanalysisonline.com/>)

**Tools for Automatic Summarization**

|  |
| --- |
| **Examples**  File 16: English Text |

1. Automatic Text Summarizer (<http://autosummarizer.com/index.php>)
2. Free Summarizer (<http://freesummarizer.com/#summarizecontainer>)
3. Smmry (<http://smmry.com/>)
4. Text Compactor (<https://www.textcompactor.com/>)
5. Text Summarization (<http://textsummarization.net/text-summarizer>)

**Tools for Text Statistics, Keyword Extraction and Other Analysis**

|  |
| --- |
| **Examples**  File 16: See above  File 17: Chinese Text  File 18: Japanese Text  File 19: Video Indexer (Play the video while it is being uploaded) |

1. Text Analyzer (<https://www.online-utility.org/text/analyzer.jsp>)
2. Lexicool/Textalyser (<https://www.lexicool.com/text_analyzer.asp>)
3. Termbox (<http://termbox.lingosail.com/>)
4. Intencheck (<https://app.intencheck.com/analyze/>)
5. MeaningCloud (<https://www.meaningcloud.com/demos>)
6. TextRazor (<https://www.textrazor.com/demo>)
7. Tencent NLP (<http://nlp.qq.com/index.cgi>)
8. Baidu Yun (<https://cloud.baidu.com/product/nlp.html>)
9. Microsoft Text Analytics (<https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/?v=18.05>)
10. Voyant (<https://voyant-tools.org/>) [Another similar tool: Text Analysis API (<https://developer.aylien.com/text-api-demo?text=&language=en&tab=classify-taxonomy>)]
11. Microsoft Video Indexer (<https://vi.microsoft.com/en-us/>)
12. **Programming for Text Analytics**
13. There are scenarios in which you need to write your own programs for text analytics: You need to handle a huge number of special documents (instead of a small corpus). Or you want to have workflow or tool integration. Or you need features that are not offered by current tools.
14. Sample scenario: Downloading Chinese and English documents from <https://www.gov.hk/tc/about/abouthk/factsheets/> and automatic alignment
15. Python and R could be good starting points.
16. Get familiar with the following: (1) basic data types (e.g., numbers, strings, lists, dictionaries and functions), (2) syntax (e.g., loops, if statements, file input and output), (3) libraries (e.g., built-in (内建), standard (標準) and third-party (第三方)), and (4) API.

|  |
| --- |
| Other interesting tools  <https://rewordify.com/>  <http://www.hemingwayapp.com/>  Google: text tools (many interesting applications)  <https://guides.library.upenn.edu/japanesetext> [Very useful website]  <http://www.karsdorp.io/python-course/> [Very useful course]  <http://ctext.org/plugins/texttools/> [A collection of interesting applications that can be further developed] |

|  |
| --- |
| **Practice**  <https://www.gov.hk/tc/about/abouthk/docs/2017HK_in_brief.pdf>  <https://www.gov.hk/en/about/abouthk/docs/2017HK_in_brief.pdf>  **More information**  Clean text:  <https://machinelearningmastery.com/clean-text-machine-learning-python/>  <https://www.analyticsvidhya.com/blog/2014/11/text-data-cleaning-steps-python/>  <https://chrisalbon.com/python/basics/cleaning_text/> |