

Obtaining, Cleaning, and Exploring Corpora

Elliott Ash

September 6, 2018

Publicly Available Corpora

- ▶ There is already a vast amount of data out there that has already been compiled (e.g. CourtListener, Twitter, New York Times, Reuters, Google, Wikipedia).
- ▶ Chris Bail curates a list of these datasets:
 - ▶ <https://docs.google.com/spreadsheets/d/1I7cvuCBQxosQK2evTcdL3qtglaEPc0WFEs6rZMx-xiE/edit>
- ▶ Some interesting corpora described in NLTK Book Chapter 2.
- ▶ Many proprietary corpora are becoming available for research:
 - ▶ Lexis
 - ▶ Web of Science

Screen Scraping

- ▶ A screen scraper is a computer program that:
 - ▶ loads/reads in a web page
 - ▶ finds some information on it
 - ▶ grabs the information
 - ▶ stores it in a dataset
- ▶ Once upon a time you could collect virtually any piece of information from the internet by screen scraping.
 - ▶ But now web sites make it difficult with restrictive terms of use, bot-blockers, javascript, etc.
 - ▶ Still, a little creativity goes a long way.

What a web site looks like to us



[Main page](#)

[Contents](#)

[Featured content](#)

[Current events](#)

[Random article](#)

[Donate to Wikipedia](#)

[Wikipedia store](#)

Interaction

[Help](#)

[About Wikipedia](#)

[Community portal](#)

[Recent changes](#)

[Contact page](#)

Tools

[What links here](#)

[Create account](#) [Log in](#)

Article [Talk](#)

Read

[Edit](#)

[View history](#)



World Health Organization ranking of health systems in 2000

From Wikipedia, the free encyclopedia

The **World Health Organization (WHO) ranked the health systems** of its 191 member states in its [World Health Report](#)^[1] 2000. It provided a framework and measurement approach to examine and compare aspects of [health systems](#) around the world.^[2] It developed a series of performance indicators to assess the overall level and distribution of [health](#) in the populations, and the responsiveness and financing of [health care](#) services. It was the organization's first ever analysis of the world's health systems.^[3]

Contents [\[hide\]](#)

[1 Ranking](#)

[2 Methodology](#)

[3 Criticism](#)

[4 See also](#)

[5 References](#)

What a web site looks like to a computer

```
1 <!DOCTYPE html>
2 <html lang="en" dir="ltr" class="client-nojs">
3 <head>
4 <meta charset="UTF-8" />
5 <title>World Health Organization ranking of health systems in 2000 - Wikipedia, the free encyclopedia</title>
6 <meta name="generator" content="MediaWiki 1.26wmf10" />
7 <link rel="alternate" href="android-
  app://org.wikipedia/http/en.m.wikipedia.org/wiki/World_Health_Organization_ranking_of_health_systems_in_2000"
  />
8 <link rel="alternate" type="application/x-wiki" title="Edit this page" href="/w/index.php?
  title=World_Health_Organization_ranking_of_health_systems_in_2000&action=edit" />
9 <link rel="edit" title="Edit this page" href="/w/index.php?
  title=World_Health_Organization_ranking_of_health_systems_in_2000&action=edit" />
10 <link rel="apple-touch-icon" href="/static/apple-touch/wikipedia.png" />
11 <link rel="shortcut icon" href="/static/favicon/wikipedia.ico" />
12 <link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikipedia
  (en)" />
13 <link rel="EditURI" type="application/rsd+xml" href="//en.wikipedia.org/w/api.php?action=rsd" />
14 <link rel="alternate" hreflang="x-default"
  href="/wiki/World_Health_Organization_ranking_of_health_systems_in_2000" />
15 <link rel="copyright" href="//creativecommons.org/licenses/by-sa/3.0/" />
16 <link rel="alternate" type="application/atom+xml" title="Wikipedia Atom feed" href="/w/index.php?
  title=Special:RecentChanges&feed=atom" />
17 <link rel="canonical"
  href="https://en.wikipedia.org/wiki/World_Health_Organization_ranking_of_health_systems_in_2000" />
18 <link rel="stylesheet" href="//en.wikipedia.org/w/load.php?
  debug=false&lang=en&modules=ext.uls.nojs%7Cext.visualEditor.viewPageTarget.noscript%7Cext.wikihihero%7C
  mediawiki.legacy.commonPrint%2Cshared%7Cmediawiki.sectionAnchor%7Cmediawiki.skinning.interface%7Cmediawiki.ui.
  button%7Cskins.vector.styles%7Cwikibase.client.init&only=styles&skin=vector&*" />
19 <meta name="ResourceLoaderDynamicStyles" content="" />
20 <link rel="stylesheet" href="//en.wikipedia.org/w/load.php?
  debug=false&lang=en&modules=sites&only=styles&skin=vector&skin=vector&*" />
https://en.wikipedia.org/w/index.php?title=World_Health_Organization&a:lang(mzn),a:lang(ps),a:lang(ur){text-decoration:none}
```

Screen Scraping in Python

```
# package to access web pages
from urllib.request import urlopen
url = 'https://goo.gl/VRF8Xs' # shortened URL
page = urlopen(url) # open page

html = page.read() # read page as string
print(html[:400]) # print first 400 characters
print(html[-400:]) # print last 400 characters
print(len(html)) # print length of string
```

Browser Automation

- ▶ Many web sites are designed to be difficult to scrape.
- ▶ Python has many solutions for simulating a human browser:
 - ▶ robobrowser
 - ▶ selenium (chromedriver, phantomjs)
- ▶ Other solutions if all else fails:
 - ▶ DownThemAll! plug-in for Firefox
 - ▶ Hire mechanical turkers to manually download data.

API's

- ▶ API = Application Programming Interface
 - ▶ These are developer-oriented tools that provide access to cleaner data.
- ▶ Chris Bail's list of API's that could be interesting for research:
 - ▶ <https://docs.google.com/spreadsheets/d/1ZEr3okdlb0zctmX0MZKo-gZKPsq5WGn1nJ0xPV7a1-Q/edit>
- ▶ The example data set was obtained from the CourtListener API (courtlistener.com/api).

Other Languages

- ▶ All of the tools that we discuss in this class are available in many languages.
- ▶ spaCy has full functionality in English, German, Spanish, Portuguese, French, Italian, and Dutch.
 - ▶ beta functionality in dozens of other languages including Chinese and Arabic
 - ▶ See <https://spacy.io/usage/models>.

The machine learning models are language-independent.

Translation

- ▶ Can also translate languages before running an analysis

```
from googletrans import Translator
translator = Translator()
lang = translator.detect(korean_string).lang
eng = translator.translate(korean_string ,
                           src=lang ,
                           dest='en')

eng.text
```

HTML Parsing

```
from bs4 import BeautifulSoup # HTML parser
soup = BeautifulSoup(html, 'lxml') # parse page
print(soup.title)

# extract text
text = soup.get_text() # remove HTML markup
lines = text.splitlines() # split by line
print(len(lines)) # print number of lines

# drop empty lines
lines = [line for line in lines if line != '']
print(len(lines))
print(lines[:20]) # print first 20 lines
```

Character Encoding

The screenshot shows the macOS 'Characters' application window. The 'Unicode' tab is selected in the left sidebar. The main pane displays the 'Latin Extended-B' character set, which includes characters from U+0180 to U+024F. The characters are arranged in a grid with columns for different character sets (e.g., Basic Latin, IPA Extensions, Spacing Modifier Letters) and rows for different character sets (e.g., Latin, IPA Extensions, Spacing Modifier Letters). The character 'b' is highlighted in the first row, first column. The right sidebar shows the 'Font Variation' section, displaying the character 'b' in various font styles (e.g., Latin, IPA Extensions, Spacing Modifier Letters).

Unicode	Title	Category
00000000	Basic Latin	Latin
00000080	Latin-1 Supplement	Latin
00000100	Latin Extended-A	Latin
00000180	Latin Extended-B	Latin
00000250	IPA Extensions	Latin
00000280	Spacing Modifier Letters	Modifier Letters
00000300	Combining Diacritical Marks	Combining Marks
00000320	Greek and Coptic	Greek

Latin Extended-B

Unicode	Title	Category
0180	À à	Latin
0181	Á á	Latin
0182	Â â	Latin
0183	Ã ã	Latin
0184	Ä ä	Latin
0185	Å å	Latin
0186	Æ æ	Latin
0187	Ç ç	Latin
0188	Ð ð	Latin
0189	Ñ ñ	Latin
018A	Ò ò	Latin
018B	Ó ó	Latin
018C	Ô ô	Latin
018D	Õ õ	Latin
018E	Ö ö	Latin
018F	Ø ø	Latin
0190	Ù ù	Latin
0191	Ú ú	Latin
0192	Û û	Latin
0193	Ü ü	Latin
0194	Ý ý	Latin
0195	ÿ	Latin
0196	Ž ž	Latin
0197	Š š	Latin
0198	Š š	Latin
0199	š š	Latin
019A	š š	Latin
019B	š š	Latin
019C	š š	Latin
019D	š š	Latin
019E	š š	Latin
019F	š š	Latin
01A0	Š š	Latin
01A1	š š	Latin
01A2	š š	Latin
01A3	š š	Latin
01A4	š š	Latin
01A5	š š	Latin
01A6	š š	Latin
01A7	š š	Latin
01A8	š š	Latin
01A9	š š	Latin
01AA	š š	Latin
01AB	š š	Latin
01AC	š š	Latin
01AD	š š	Latin
01AE	š š	Latin
01AF	š š	Latin
01B0	š š	Latin
01B1	š š	Latin
01B2	š š	Latin
01B3	š š	Latin
01B4	š š	Latin
01B5	š š	Latin
01B6	š š	Latin
01B7	š š	Latin
01B8	š š	Latin
01B9	š š	Latin
01BA	š š	Latin
01BB	š š	Latin
01BC	š š	Latin
01BD	š š	Latin
01BE	š š	Latin
01BF	š š	Latin
01C0	Š š	Latin
01C1	š š	Latin
01C2	š š	Latin
01C3	š š	Latin
01C4	š š	Latin
01C5	š š	Latin
01C6	š š	Latin
01C7	š š	Latin
01C8	š š	Latin
01C9	š š	Latin
01CA	š š	Latin
01CB	š š	Latin
01CC	š š	Latin
01CD	š š	Latin
01CE	š š	Latin
01CF	š š	Latin
01D0	š š	Latin
01D1	š š	Latin
01D2	š š	Latin
01D3	š š	Latin
01D4	š š	Latin
01D5	š š	Latin
01D6	š š	Latin
01D7	š š	Latin
01D8	š š	Latin
01D9	š š	Latin
01DA	š š	Latin
01DB	š š	Latin
01DC	š š	Latin
01DD	š š	Latin
01DE	š š	Latin
01DF	š š	Latin
01E0	š š	Latin
01E1	š š	Latin
01E2	š š	Latin
01E3	š š	Latin
01E4	š š	Latin
01E5	š š	Latin
01E6	š š	Latin
01E7	š š	Latin
01E8	š š	Latin
01E9	š š	Latin
01EA	š š	Latin
01EB	š š	Latin
01EC	š š	Latin
01ED	š š	Latin
01EE	š š	Latin
01EF	š š	Latin
01F0	š š	Latin
01F1	š š	Latin
01F2	š š	Latin
01F3	š š	Latin
01F4	š š	Latin
01F5	š š	Latin
01F6	š š	Latin
01F7	š š	Latin
01F8	š š	Latin
01F9	š š	Latin
01FA	š š	Latin
01FB	š š	Latin
01FC	š š	Latin
01FD	š š	Latin
01FE	š š	Latin
01FF	š š	Latin

IPA Extensions

Unicode	Title	Category
0250	ɐ	IPA Extensions
0251	ɑ	IPA Extensions
0252	ɒ	IPA Extensions
0253	ɔ	IPA Extensions
0254	ɔ̃	IPA Extensions
0255	ɔ̄	IPA Extensions
0256	ɔ̅	IPA Extensions
0257	ɔ̆	IPA Extensions
0258	ɔ̇	IPA Extensions
0259	ɔ̈	IPA Extensions
025A	ɔ̉	IPA Extensions
025B	ɔ̊	IPA Extensions
025C	ɔ̋	IPA Extensions
025D	ɔ̌	IPA Extensions
025E	ɔ̍	IPA Extensions
025F	ɔ̎	IPA Extensions
0260	ɔ̏	IPA Extensions
0261	ɔ̐	IPA Extensions
0262	ɔ̑	IPA Extensions
0263	ɔ̒	IPA Extensions
0264	ɔ̓	IPA Extensions
0265	ɔ̔	IPA Extensions
0266	ɔ̕	IPA Extensions
0267	ɔ̖	IPA Extensions
0268	ɔ̗	IPA Extensions
0269	ɔ̘	IPA Extensions
026A	ɔ̙	IPA Extensions
026B	ɔ̚	IPA Extensions
026C	ɔ̛	IPA Extensions
026D	ɔ̜	IPA Extensions
026E	ɔ̝	IPA Extensions
026F	ɔ̞	IPA Extensions
0270	ɔ̟	IPA Extensions
0271	ɔ̠	IPA Extensions
0272	ɔ̡	IPA Extensions
0273	ɔ̢	IPA Extensions
0274	ɔ̣	IPA Extensions
0275	ɔ̤	IPA Extensions
0276	ɔ̥	IPA Extensions
0277	ɔ̦	IPA Extensions
0278	ɔ̧	IPA Extensions
0279	ɔ̨	IPA Extensions
027A	ɔ̩	IPA Extensions
027B	ɔ̪	IPA Extensions
027C	ɔ̫	IPA Extensions
027D	ɔ̬	IPA Extensions
027E	ɔ̭	IPA Extensions
027F	ɔ̮	IPA Extensions
0280	ɔ̯	IPA Extensions
0281	ɔ̰	IPA Extensions
0282	ɔ̱	IPA Extensions
0283	ɔ̲	IPA Extensions
0284	ɔ̳	IPA Extensions
0285	ɔ̴	IPA Extensions
0286	ɔ̵	IPA Extensions
0287	ɔ̶	IPA Extensions
0288	ɔ̷	IPA Extensions
0289	ɔ̸	IPA Extensions
028A	ɔ̹	IPA Extensions
028B	ɔ̺	IPA Extensions
028C	ɔ̻	IPA Extensions
028D	ɔ̼	IPA Extensions
028E	ɔ̽	IPA Extensions
028F	ɔ̾	IPA Extensions
0290	ɔ̿	IPA Extensions
0291	ɔ̀	IPA Extensions
0292	ɔ́	IPA Extensions
0293	ɔ̂	IPA Extensions
0294	ɔ̃	IPA Extensions
0295	ɔ̄	IPA Extensions
0296	ɔ̅	IPA Extensions
0297	ɔ̆	IPA Extensions
0298	ɔ̇	IPA Extensions
0299	ɔ̈	IPA Extensions
029A	ɔ̉	IPA Extensions
029B	ɔ̊	IPA Extensions
029C	ɔ̋	IPA Extensions
029D	ɔ̌	IPA Extensions
029E	ɔ̍	IPA Extensions
029F	ɔ̎	IPA Extensions
02A0	ɔ̏	IPA Extensions
02A1	ɔ̐	IPA Extensions
02A2	ɔ̑	IPA Extensions
02A3	ɔ̒	IPA Extensions
02A4	ɔ̓	IPA Extensions
02A5	ɔ̔	IPA Extensions
02A6	ɔ̕	IPA Extensions
02A7	ɔ̖	IPA Extensions
02A8	ɔ̗	IPA Extensions
02A9	ɔ̘	IPA Extensions
02AA	ɔ̙	IPA Extensions
02AB	ɔ̚	IPA Extensions
02AC	ɔ̛	IPA Extensions
02AD	ɔ̜	IPA Extensions
02AE	ɔ̝	IPA Extensions
02AF	ɔ̞	IPA Extensions
02B0	ɔ̟	IPA Extensions
02B1	ɔ̠	IPA Extensions
02B2	ɔ̡	IPA Extensions
02B3	ɔ̢	IPA Extensions
02B4	ɔ̣	IPA Extensions
02B5	ɔ̤	IPA Extensions
02B6	ɔ̥	IPA Extensions
02B7	ɔ̦	IPA Extensions
02B8	ɔ̧	IPA Extensions
02B9	ɔ̨	IPA Extensions
02BA	ɔ̩	IPA Extensions
02BB	ɔ̪	IPA Extensions
02BC	ɔ̫	IPA Extensions
02BD	ɔ̬	IPA Extensions
02BE	ɔ̭	IPA Extensions
02BF	ɔ̮	IPA Extensions
02C0	ɔ̯	IPA Extensions
02C1	ɔ̰	IPA Extensions
02C2	ɔ̱	IPA Extensions
02C3	ɔ̲	IPA Extensions
02C4	ɔ̳	IPA Extensions
02C5	ɔ̴	IPA Extensions
02C6	ɔ̵	IPA Extensions
02C7	ɔ̶	IPA Extensions
02C8	ɔ̷	IPA Extensions
02C9	ɔ̸	IPA Extensions
02CA	ɔ̹	IPA Extensions
02CB	ɔ̺	IPA Extensions
02CC	ɔ̻	IPA Extensions
02CD	ɔ̼	IPA Extensions
02CE	ɔ̽	IPA Extensions
02CF	ɔ̾	IPA Extensions
02D0	ɔ̿	IPA Extensions
02D1	ɔ̀	IPA Extensions
02D2	ɔ́	IPA Extensions
02D3	ɔ̂	IPA Extensions
02D4	ɔ̃	IPA Extensions
02D5	ɔ̄	IPA Extensions
02D6	ɔ̅	IPA Extensions
02D7	ɔ̆	IPA Extensions
02D8	ɔ̇	IPA Extensions
02D9	ɔ̈	IPA Extensions
02DA	ɔ̉	IPA Extensions
02DB	ɔ̊	IPA Extensions
02DC	ɔ̋	IPA Extensions
02DD	ɔ̌	IPA Extensions
02DE	ɔ̍	IPA Extensions
02DF	ɔ̎	IPA Extensions
02E0	ɔ̏	IPA Extensions
02E1	ɔ̐	IPA Extensions
02E2	ɔ̑	IPA Extensions
02E3	ɔ̒	IPA Extensions
02E4	ɔ̓	IPA Extensions
02E5	ɔ̔	IPA Extensions
02E6	ɔ̕	IPA Extensions
02E7	ɔ̖	IPA Extensions
02E8	ɔ̗	IPA Extensions
02E9	ɔ̘	IPA Extensions
02EA	ɔ̙	IPA Extensions
02EB	ɔ̚	IPA Extensions
02EC	ɔ̛	IPA Extensions
02ED	ɔ̜	IPA Extensions
02EE	ɔ̝	IPA Extensions
02EF	ɔ̞	IPA Extensions
02F0	ɔ̟	IPA Extensions
02F1	ɔ̠	IPA Extensions
02F2	ɔ̡	IPA Extensions
02F3	ɔ̢	IPA Extensions
02F4	ɔ̣	IPA Extensions
02F5	ɔ̤	IPA Extensions
02F6	ɔ̥	IPA Extensions
02F7	ɔ̦	IPA Extensions
02F8	ɔ̧	IPA Extensions
02F9	ɔ̨	IPA Extensions
02FA	ɔ̩	IPA Extensions
02FB	ɔ̪	IPA Extensions
02FC	ɔ̫	IPA Extensions
02FD	ɔ̬	IPA Extensions
02FE	ɔ̭	IPA Extensions
02FF	ɔ̮	IPA Extensions

Spacing Modifier Letters

Unicode	Title	Category
02B0	ɐ̘	Spacing Modifier Letters
02B1	ɐ̙	Spacing Modifier Letters
02B2	ɐ̚	Spacing Modifier Letters
02B3	ɐ̛	Spacing Modifier Letters
02B4	ɐ̜	Spacing Modifier Letters
02B5	ɐ̝	Spacing Modifier Letters
02B6	ɐ̞	Spacing Modifier Letters
02B7	ɐ̟	Spacing Modifier Letters
02B8	ɐ̠	Spacing Modifier Letters
02B9	ɐ̡	Spacing Modifier Letters
02BA	ɐ̢	Spacing Modifier Letters
02BB	ɐ̣	Spacing Modifier Letters
02BC	ɐ̤	Spacing Modifier Letters
02BD	ɐ̥	Spacing Modifier Letters
02BE	ɐ̦	Spacing Modifier Letters
02BF	ɐ̧	Spacing Modifier Letters
02C0	ɐ̨	Spacing Modifier Letters
02C1	ɐ̩	Spacing Modifier Letters
02C2	ɐ̪	Spacing Modifier Letters
02C3	ɐ̫	Spacing Modifier Letters
02C4	ɐ̬	Spacing Modifier Letters
02C5	ɐ̭	Spacing Modifier Letters
02C6	ɐ̮	Spacing Modifier Letters
02C7	ɐ̯	Spacing Modifier Letters
02C8	ɐ̰	Spacing Modifier Letters
02C9	ɐ̱	Spacing Modifier Letters
02CA	ɐ̲	Spacing Modifier Letters
02CB	ɐ̳	Spacing Modifier Letters
02CC	ɐ̴	Spacing Modifier Letters
02CD	ɐ̵	Spacing Modifier Letters
02CE	ɐ̶	Spacing Modifier Letters
02CF	ɐ̷	Spacing Modifier Letters
02D0	ɐ̸	Spacing Modifier Letters
02D1	ɐ̹	Spacing Modifier Letters
02D2	ɐ̺	Spacing Modifier Letters
02D3	ɐ̻	Spacing Modifier Letters
02D4	ɐ̼	Spacing Modifier Letters
02D5	ɐ̽	Spacing Modifier Letters
02D6	ɐ̾	Spacing Modifier Letters
02D7	ɐ̿	Spacing Modifier Letters
02D8	ɐ̀	Spacing Modifier Letters
02D9	ɐ́	Spacing Modifier Letters
02DA	ɐ̂	Spacing Modifier Letters
02DB	ɐ̃	Spacing Modifier Letters
02DC	ɐ̄	Spacing Modifier Letters
02DD	ɐ̅	Spacing Modifier Letters
02DE	ɐ̆	Spacing Modifier Letters
02DF	ɐ̇	Spacing Modifier Letters
02E0	ɐ̈	Spacing Modifier Letters
02E1	ɐ̉	Spacing Modifier Letters
02E2	ɐ̊	Spacing Modifier Letters
02E3	ɐ̋	Spacing Modifier Letters
02E4	ɐ̌	Spacing Modifier Letters
02E5	ɐ̍	Spacing Modifier Letters
02E6	ɐ̎	Spacing Modifier Letters
02E7	ɐ̏	Spacing Modifier Letters
02E8	ɐ̐	Spacing Modifier Letters
02E9	ɐ̑	Spacing Modifier Letters
02EA	ɐ̒	Spacing Modifier Letters
02EB	ɐ̓	Spacing Modifier Letters
02EC	ɐ̔	Spacing Modifier Letters
02ED	ɐ̕	Spacing Modifier Letters
02EE	ɐ̖	Spacing Modifier Letters
02EF	ɐ̗	Spacing Modifier Letters
02F0	ɐ̘	Spacing Modifier Letters
02F1	ɐ̙	Spacing Modifier Letters
02F2	ɐ̚	Spacing Modifier Letters
02F3	ɐ̛	Spacing Modifier Letters
02F4	ɐ̜	Spacing Modifier Letters
02F5	ɐ̝	Spacing Modifier Letters
02F6	ɐ̞	Spacing Modifier Letters
02F7	ɐ̟	Spacing Modifier Letters
02F8	ɐ̠	Spacing Modifier Letters
02F9	ɐ̡	Spacing Modifier Letters
02FA	ɐ̢	Spacing Modifier Letters
02FB	ɐ̣	Spacing Modifier Letters
02FC	ɐ̤	Spacing Modifier Letters
02FD	ɐ̥	Spacing Modifier Letters
02FE	ɐ̦	Spacing Modifier Letters
02FF	ɐ̧	Spacing Modifier Letters

Removing unicode characters

```
# package for removing unicode  
from unicodecode import unicodecode  
fixed = unicodecode('Visualizations\xa0')  
print(fixed) # print cleaned string
```

Corpus cleaning

- ▶ What we've already done:
 - ▶ removed HTML markup, extra white space, and unicode
- ▶ But HTML markup is often valuable:
 - ▶ HTML markup for section header names.
 - ▶ Legal database web sites often have HTML tags for citations to other cases.
- ▶ Other cleaning steps:
 - ▶ page numbers
 - ▶ hyphenations at line breaks
 - ▶ table of contents, indexes, etc.
- ▶ These are all corpus-specific, so inspect ahead of time.

Regular Expressions

- ▶ Regular Expressions, implemented in the Python package **re**, provide a powerful string matching tool.
 - ▶ A systematic string matching protocol – can match arbitrary string patterns
 - ▶ e.g., use `utilit*` to match `utility`, `utilities`, `utilitarian`, ...
 - ▶ Important for identifying speaker names (in political documents) section headers (in statutes), citations (in judicial opinions), etc.
- ▶ Also quite tedious, so we will not cover it here.
 - ▶ See NLTK book Chapter 3.4-3.5 for an introduction.

OCR (Optical Character Recognition)

- ▶ Your data might be in PDF's or images. Needs to be converted to text
- ▶ The best solution (that I know of) is ABBYY FineReader, which is expensive but might be available at your university library.
- ▶ My colleague Joe Sutherland at Columbia has a nice open-source package for OCR:
 - ▶ <https://github.com/jlsutherland/doc2text>

Should you run a spell checker?

- ▶ The short answer is no:
 - ▶ Most corpora have important specialized vocabulary that would be flagged by standard spell-checkers.
 - ▶ They are also very slow to run on large corpora.
 - ▶ In most empirical contexts, it's safe to assume that spelling errors (especially OCR errors) are uncorrelated with treatment assignment.
- ▶ Better solutions:
 - ▶ drop short (one or two letters) and long words (over 12 letters).
 - ▶ get doc frequencies for each word and filter out rare words
 - ▶ or use word embeddings (Lecture 13) and trust that misspellings will be nearby the true word.
- ▶ But:
 - ▶ There are cases where spelling errors could be correlated with treatment (for example, increasing legislator salaries might change both policy priorities and spelling error rates)
 - ▶ Check out the **enchant** module in Python.

Collect Key Metrics

```
df1 = df1[['state', 'snippet']]
# Number of documents
len(df1['snippet'])
# Number of label categories (e.g. states)
df1['state'].describe()
# Number of samples per class
counts_per_class = df1.groupby('state').count()
counts_per_class.head()
# Words per sample
def get_words_per_sample(txt):
    return len(txt.split())
df1['num_words'] = df1['snippet'].apply(get_words_per_sample)
df1['num_words'].describe()
# Frequency distribution over words
from collections import Counter
freqs = Counter()
for i, row in df1.iterrows():
    freqs.update(row['snippet'].lower().split())
freqs.most_common()[:20]
# (Number of samples) / number of words per sample
len(df1['snippet']) / df1['num_words'].mean()
```