NLP For Economists

Reading and Writing Text Files in different format

Sowmya Vajjala

Munich Graduate School of Economics - LMU Munich

Guest Course, October 2020

Goals for this session

- Why should we know how to work with various file formats?
- How do we extract text from different formats?
- What are some problems with existing solutions?

source material: Chapter 2 from `https://practicalnlp.ai`
note: I will leave code examples out of this, as I don't know what formats of text you eventually want to work with. I will point you to relevant references.

Why different formats?

- Generally, when we learn NLP, we work with existing data, already converted to plain text.
- However, real world scenarios are far from this.
- Especially in a field where NLP use is relatively new like yours, there is no guarantee of finding plain text data.
- That said, documents can come in many different formats. We better have at least a vague idea of extracting text from those formats!

Some example formats

- PDFs
- scanned texts/image files
- XML files
- HTML files such as web pages, forums etc.
- live Twitter stream
- already existing tweets stored in a database/.csv file etc.
- JSON
- Docx files
- **best format* txt files, other plain text files

... ...

Reading from HTML/XML

What we need: bs4 library

- We have to understand the structure of the document/tags used etc (e.g., inspect element in chrome browser) to be able to extract.
- What happens behind the scenes: The library "parses" HTML/XML formatted text and builds a tree like object of the format, so that we can query and extract what we want.

details: `https://pypi.org/project/beautifulsoup4/`

Reading from JSON

What we need: json library (just import json)

- JSON is a commonly used format to exchange data, including text.
- Good thing with this is: it is natively supported by Python, and it looks like a lot of dictionary objects when you see.
- We still have to figure out the structure to get what we want.

details: `https://realpython.com/python-json/`

Reading from PDFs

What we need: pypdf2 library is a good start.

- There are many such libraries for pdf parsing, each good at a certain kind of data.
- camelot/tabula/excalibur-py can be used to extract tabular data from pdfs.
- Google/Amazon are now offering some tools in their web-based (paid) services.
- Yet, I did not come across a perfect solution so far.
- Very challenging format to process.

useful link: `https://realpython.com/pdf-python/`

Reading from twitter stream

What we need: tweepy

- We need a twitter account, and some authentication tokens for using twitter through a program
- This is a good source of streaming, latest data
- However, be aware of Twitter's terms of use, don't store identifying information, and remember: many NLP tools don't work well on tweets. So, look for custom variants (they exist).

details: `https://www.earthdatascience.org/courses/use-data-open-source-python/intro-to-apis/twitter-data-in-python/`

Reading from scanned images

What we need: OCR (Optical Character Recognition)

```
from PIL import Image
from pytesseract import image_to_string
filename = "somefile.png"
text = image_to_string(Image.open(filename))
print(text)
```

details: `https://pypi.org/project/pytesseract/`

Conclusion

- Many different file formats
- Many different libraries for each
- They may not be perfect - but it could meet your needs, depending on what level of NLP you want/need
- What to do with what we managed to extract? - next video!

... ... `https://nostarch.com/automatestuff2` - this free ebook has a lot of code examples on extracting data from different formats.