NLP without Annotated Dataset

Where are these datasets coming from?

Sowmya Vajjala

Seminar für Sprachwissenschaft, University of Tübingen, Germany

13 January 2021

Class Outline

- Quick recap of last class
- Corpus Collection
- Text Extraction
- Corpus Exploration

Last Class: Quick Recap

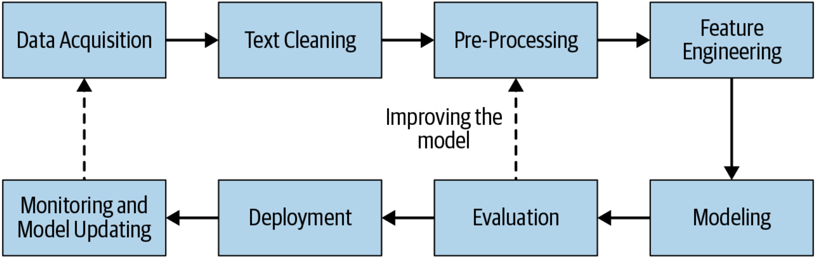
- NLP Pipeline: Various steps involved
- A real world NLP pipeline: Uber's COTA
- A code example for some of the steps
- General comments on building NLP systems
- Text representation for NLP

- Any questions on these so far?

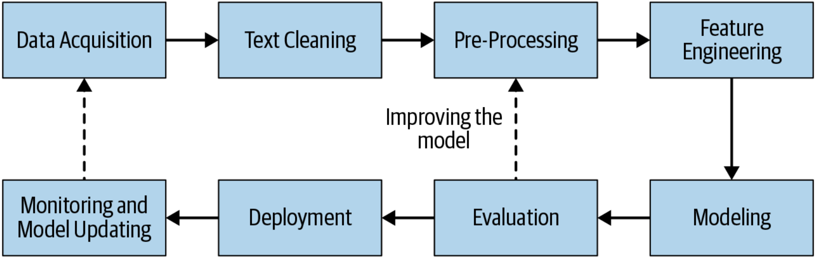
Housekeeping

- After enquiries/interest in BERT, I decided to do one class (mostly next week, just before your group discussion sessions begin) focusing on an intuitive understanding, how to use in code, languages it supports, fine-tuning etc.
- Deadlines: Decide your team (13th -Today!) and paper (15th - Friday).
- I will choose teams and papers and post on Forum on Friday, for those who did not choose. (based on what you wrote in the questionnaire, if you did it).

NLP Pipeline



NLP Pipeline



-the first step is "data acquisition". Why do we need it at all

Data Acquisition/Corpus Collection

note: I am going to use corpus and data/dataset interchangeably.

Why do we need data at all?

my one line answer: to teach the machine!

Why do we need data at all?

my one line answer: to teach the machine!

- Modern NLP is heavily machine learning driven and machine learning approaches typically require lots and lots of examples to "train" on and learn a task.

Why do we need data at all?

my one line answer: to teach the machine!

- Modern NLP is heavily machine learning driven and machine learning approaches typically require lots and lots of examples to "train" on and learn a task.
- Assuming we are "engineering" everything manually, we still need some kind of curated data to evaluate our approach for its accuracy and coverage.

Why do we need data at all?

my one line answer: to teach the machine!

- Modern NLP is heavily machine learning driven and machine learning approaches typically require lots and lots of examples to "train" on and learn a task.
- Assuming we are "engineering" everything manually, we still need some kind of curated data to evaluate our approach for its accuracy and coverage.

So, good quality datasets are very (very) important for building any NLP system.

Why collect our own data?

- Clearly, building and evaluating NLP systems requires some corpus or the other.
- When we are working on developing new methods/algorithms, we often work with standard/pre-existing datasets to compare among different approaches.

Why collect our own data?

- Clearly, building and evaluating NLP systems requires some corpus or the other.
- When we are working on developing new methods/algorithms, we often work with standard/pre-existing datasets to compare among different approaches.
- However, when we are working on using NLP for a specific problem scenario, we won't often have such datasets that meet our exact needs.
- Data can come in various forms, but in most cases, we need some form of "labeled" data. (What's that?)

What kind of data do we need for NLP? -1

- Huge collections of text (language modeling, topic modeling etc)

What kind of data do we need for NLP? -1

- Huge collections of text (language modeling, topic modeling etc)
- Lot of examples of expected input $->$ expected output pairs. E.g.,
  - sentence-translated sentence pairs (machine translation)
  - spam/non-spam emails (spam classification)
  - question-answer pairs
  - sentence $->$ names of entities in it, relations between them etc (information extraction)

.... Data can come in various forms, but in most cases, we need some form of "labeled" data (i.e., input $->$ output pairs).

What kind of data do we need? - 2

- Quantity: Typically, "learning" methods are data hungry. The more, the better, although it may plateau at some point. (What is large?)

What kind of data do we need? - 2

- Quantity: Typically, "learning" methods are data hungry. The more, the better, although it may plateau at some point. (What is large?)
- Quality: Garbage in $->$ Garbage out. We can't take **anything** we can lay hands on. (Why?)

What kind of data do we need? - 2

- Quantity: Typically, "learning" methods are data hungry. The more, the better, although it may plateau at some point. (What is large?)
- Quality: Garbage in $->$ Garbage out. We can't take **anything** we can lay hands on. (Why?)
- Data without ethical concerns such as using data without consent, keeping personally identifiable information, racial/gender bias in training examples etc. (Why is this important?)

What kind of data do we need? - 2

- Quantity: Typically, "learning" methods are data hungry. The more, the better, although it may plateau at some point. (What is large?)
- Quality: Garbage in $->$ Garbage out. We can't take **anything** we can lay hands on. (Why?)
- Data without ethical concerns such as using data without consent, keeping personally identifiable information, racial/gender bias in training examples etc. (Why is this important?)
- (Ideally) Variety: spoken language, social media, literary texts, dialect variation, legal docs, non-native language, different topics/subjects etc (Why?)

How do we collect a corpus?

1. Use available data

There are many openly (most of them are free) accessible NLP datasets.

- Quantum Stats NLP database
- Hugging Face NLP datasets
- Linguistic Data Consortium
- Researchers and organizations sometimes release their datasets for public (check research papers/company blog posts etc)

An Exercise

In the breakout rooms: spend 15 minutes and look for publicly available datasets in any two languages (first language: English or German, second language - choose something that is not widely spoken or if you don't know if NLP resources exist for that language).
Make a note on how big is this dataset, and any details (if mentioned) about what are its contents (i.e., what kind of text is it?), where is it useful etc. I will ask each room to do a 2 minute summary of their findings afterwards.

Discussion about findings

How do we collect a corpus?

2. Collect your own data

- Use existing data on the web
  - scraping websites (forums, newspapers, wikipedia etc)
  - collecting social media content (blog posts, tweets etc)
  - newspapers, wikipedia etc.
  - public archives (copyright free books, parliament proceedings, others)
- Collect your own (crowd sourcing, user studies, surveys etc)

Let us see some examples for each.

How do we collect a corpus?

2. Collect your own data

Scraping websites

- In a recently reported research work, researchers in USA collected a dataset of COVID-19 FAQs through webscraping (Process here)

How do we collect a corpus?

2. Collect your own data

Scraping websites

- In a recently reported research work, researchers in USA collected a dataset of COVID-19 FAQs through webscraping (Process here)
- This project describes how a government body in Mexico city used web scraping and NLP for improving job search experience.

How do we collect a corpus?

2. Collect your own data

Scraping websites

- In a recently reported research work, researchers in USA collected a dataset of COVID-19 FAQs through webscraping (Process here)
- This project describes how a government body in Mexico city used web scraping and NLP for improving job search experience.
- personal experiences: While working at SfS, I and Prof. Meurers collected data for training machine learning models using several websites (Weekly Reader, BBC Bitesize, Geo-Geolino (German), Time- Time for Kids etc)

How do we collect a corpus?

2. Collect your own data

Scraping websites

- In a recently reported research work, researchers in USA collected a dataset of COVID-19 FAQs through webscraping (Process here)
- This project describes how a government body in Mexico city used web scraping and NLP for improving job search experience.
- personal experiences: While working at SfS, I and Prof. Meurers collected data for training machine learning models using several websites (Weekly Reader, BBC Bitesize, Geo-Geolino (German), Time- Time for Kids etc)
- While working at a company long back, I collected question-answer pairs from a lot of forum websites, to build a search engine.

How do we collect a corpus?

2. Collect your own data

public archives, newspapers, Wikipedia etc

- Parliament proceedings from European Parliament (EUROPARL - several languages), Hansard in Canada (English-French), Nunavut Hansard (English-Inuktitut) etc. are commonly used as training corpora for machine translation.
- Crawled and annotated corpora from WSJ, NYT etc are frequently used as training data in NLP (you may have noticed if you browsed LDC).

How do we collect a corpus?

2. Collect your own data

public archives, newspapers, Wikipedia etc

- Parliament proceedings from European Parliament (EUROPARL - several languages), Hansard in Canada (English-French), Nunavut Hansard (English-Inuktitut) etc. are commonly used as training corpora for machine translation.
- Crawled and annotated corpora from WSJ, NYT etc are frequently used as training data in NLP (you may have noticed if you browsed LDC).

How do we collect a corpus?

2. Collect your own data

public archives, newspapers, Wikipedia etc

- Parliament proceedings from European Parliament (EUROPARL - several languages), Hansard in Canada (English-French), Nunavut Hansard (English-Inuktitut) etc. are commonly used as training corpora for machine translation.
- Crawled and annotated corpora from WSJ, NYT etc are frequently used as training data in NLP (you may have noticed if you browsed LDC).
- personal experiences: While working at a company, I scraped Canadian supreme court websites for case summaries.

How do we collect a corpus?

2. Collect your own data

public archives, newspapers, Wikipedia etc

- Parliament proceedings from European Parliament (EUROPARL - several languages), Hansard in Canada (English-French), Nunavut Hansard (English-Inuktitut) etc. are commonly used as training corpora for machine translation.
- Crawled and annotated corpora from WSJ, NYT etc are frequently used as training data in NLP (you may have noticed if you browsed LDC).
- personal experiences: While working at a company, I scraped Canadian supreme court websites for case summaries.
- I also crawled Wikipedia to build a labeled dataset of various categories of articles related to legal domain (e.g., finance law, human rights law etc) to develop a document tagger.

How do we collect a corpus?

2. Collect your own data

Social media content

- Twitter is a common source of data to study various issues such as trending topics, fake news, offensive text detection, sentiment analysis etc.
- It is also useful in industry scenarios for customer support, product reviews etc.
- Other social media websites such as Facebook, Reddit etc are also regularly used to collect data in NLP.
- This link shows how common it is to use such data for NLP Research!

How do we collect a corpus?

2. Collect your own data

crowd sourcing, user studies etc

- Crowdsourced datasets from https://msropendata.com/datasets?term=crowdsourcing and other such large organizations (Google etc)

How do we collect a corpus?

2. Collect your own data

crowd sourcing, user studies etc

- Crowdsourced datasets from https://msropendata.com/datasets?term=crowdsourcing and other such large organizations (Google etc)
- Crowdsourcing for compamy's internal use: Google collects translation data through user contributions

How do we collect a corpus?

2. Collect your own data

crowd sourcing, user studies etc

- Crowdsourced datasets from https://msropendata.com/datasets?term=crowdsourcing and other such large organizations (Google etc)
- Crowdsourcing for compamy's internal use: Google collects translation data through user contributions
- personal experiences: In 2016, Prof Meurers and I collected eye-tracking experiment data to study an NLP problem, by collaborating with cognitve science researchers at IWM-Tuebingen.
- in 2018, I and my student collected data through a user study where university students read texts and answered questions about them.

How do we collect a corpus?

2. Collect your own data

crowd sourcing, user studies etc

- Crowdsourced datasets from https://msropendata.com/datasets?term=crowdsourcing and other such large organizations (Google etc)
- Crowdsourcing for compamy's internal use: Google collects translation data through user contributions
- personal experiences: In 2016, Prof Meurers and I collected eye-tracking experiment data to study an NLP problem, by collaborating with cognitve science researchers at IWM-Tuebingen.
- in 2018, I and my student collected data through a user study where university students read texts and answered questions about them.
- Our own data can also come from internal organizational data sources like search logs, customer support data etc.

How do we collect a corpus?

"Annotated" Data?

- In some of these examples (scraping q&a, crawling Wikipedia with categories), we see data that does not need additional annotations.

How do we collect a corpus?

"Annotated" Data?

- In some of these examples (scraping q&a, crawling Wikipedia with categories), we see data that does not need additional annotations.
- In others, we wonder where the labels (e.g., positive vs negative sentiments, fake news vs normal), where does the labeled/annotated data come from?

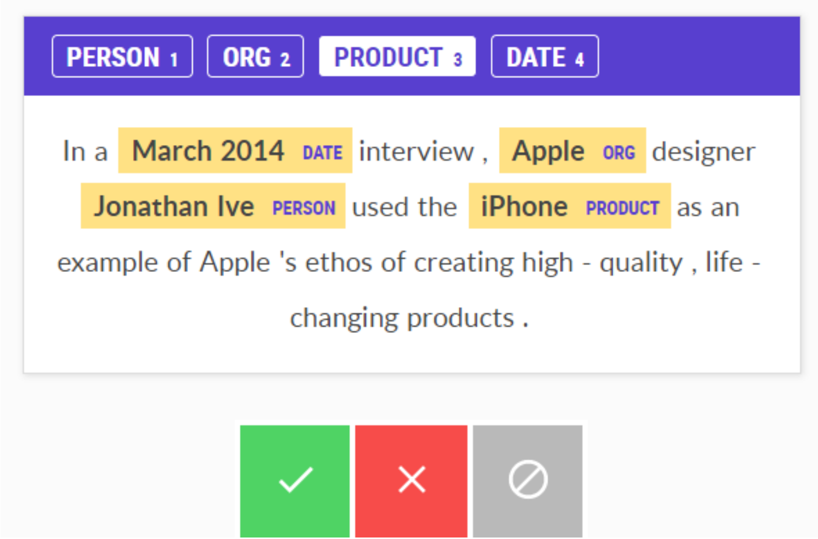
How do we collect a corpus?

"Annotated" Data?

- In some of these examples (scraping q&a, crawling Wikipedia with categories), we see data that does not need additional annotations.
- In others, we wonder where the labels (e.g., positive vs negative sentiments, fake news vs normal), where does the labeled/annotated data come from?
- crowd sourcing or having a more structured annotation experiments are two common ways of getting them.

How do we collect a corpus?

Active Learning



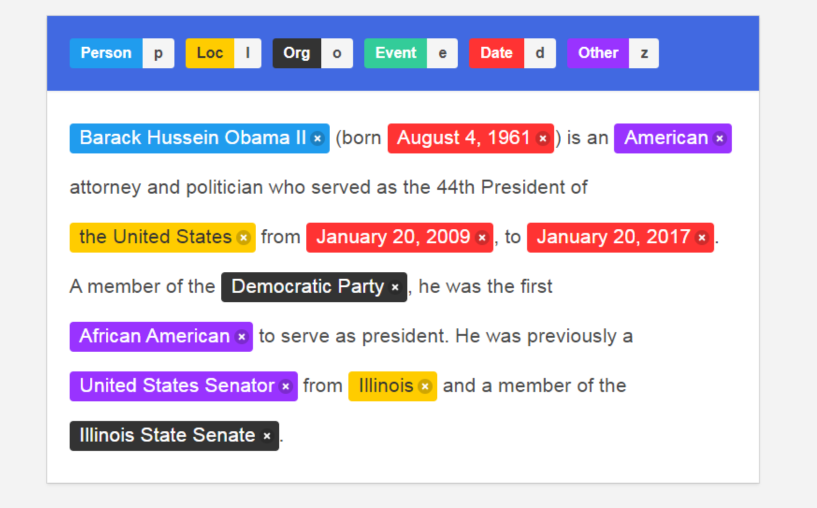
In a March 2014 DATE interview , Apple ORG designer Jonathan Ive PERSON used the iPhone PRODUCT as an example of Apple 's ethos of creating high - quality , life - changing products .

Source

(I briefly talked about this in the last class)

Annotation Tools

- Several tools exist, each serving its own purpose.
- Recent, open-source tool of interest, especially for small projects: Doccano



Barack Hussein Obama II × (born August 4, 1961 ×) is an American × attorney and politician who served as the 44th President of the United States × from January 20, 2009 ×, to January 20, 2017 ×. A member of the Democratic Party ×, he was the first African American × to serve as president. He was previously a United States Senator × from Illinois × and a member of the Illinois State Senate ×.

An Exercise

In the breakout rooms (15 minutes): go to https://datasets.quantumstat.com/ and look for datasets collected using various ways we just saw. You don't have to cover each and every possibility - just be ready to discuss about say 3 datasets collected from different sources (i.e., one from social media, one from crowd sourcing, one from news/wiki,), and comment on how big they are, what language they are in etc.
Each group should summarize their findings after 15 minutes.

An Exercise

Summary of your breakout room discussions.

How do we collect a corpus?

3. "Generate" your own data

- Data labeling: Look for patterns in the data, and generate labels using string matching, regular expressions etc.
- Data Augmentation: Generate synthetic data to add to what is already there through various strategies like - replacing words with synonyms, back translation, replacing names etc.
- Active Learning: interactively labeling the text to teach the learning algorithm

How do we collect a corpus?

3. ”Generate” your own data

Data Labeling

- e.g., using labeling functions in Snorkel: noisy, programmatic rules and heuristics that assign labels to unlabeled training data.

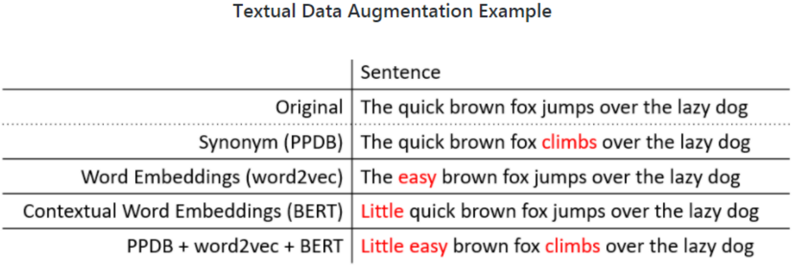
```
from snorkel.labeling import labeling_function

@labeling_function()
def lf_contains_link(x):
    # Return a label of SPAM if "http" in comment text, otherwise ABSTAIN
    return SPAM if "http" in x.text.lower() else ABSTAIN

@labeling_function()
def check_out(x):
    return SPAM if "check out" in x.text.lower() else ABSTAIN
```

More at: https://www.snorkel.org/use-cases/
(Topic for next class)

How do we collect a corpus?

3. "Generate" your own data

Data Augmentation: Replacing words with synonyms (of different kinds)



Textual Data Augmentation Example

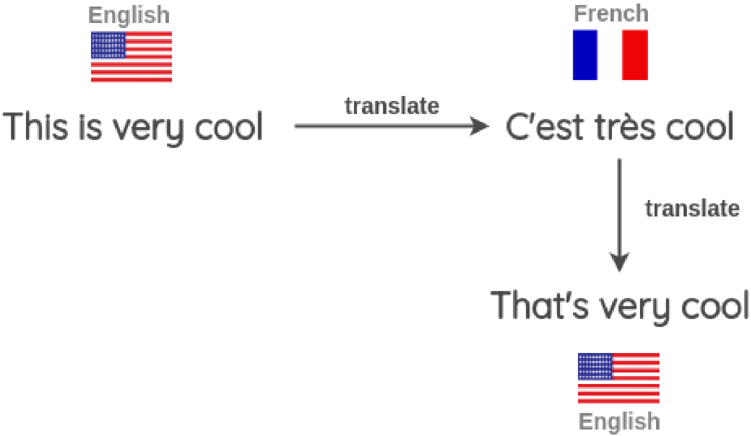
	Sentence
Original	The quick brown fox jumps over the lazy dog
Synonym (PPDB)	The quick brown fox climbs over the lazy dog
Word Embeddings (word2vec)	The easy brown fox jumps over the lazy dog
Contextual Word Embeddings (BERT)	Little quick brown fox jumps over the lazy dog
PPDB + word2vec + BERT	Little easy brown fox climbs over the lazy dog

source

How do we collect a corpus?

3. "Generate" your own data

Data Augmentation: Back Translation



Source
(more on this on Monday)

A fun Exercise

In the breakout rooms (10 minutes): Try some back translation using Google or Bing translate (i.e., type a sentence in one language to translate to another. Now, take the translated sentence, translate it back to original language). Share your observations after trying a few examples. Did you notice something interesting? Something funny? How can this be a useful data collection method according to you? Each group should summarize their findings after 10 minutes.

Summary of your breakout room discussions.

Different Ways of acquiring data: a summary

- Using existing public datasets
- Scraping data from websites, social media, public archives etc.
- Collecting our own data (crowd sourcing, user studies etc), setting up annotation experiments, Active learning etc
- Data Labeling
- Data augmentation

Some practical issues with data collection

- Clearly, there are benefits with scraping data from websites of various kinds.
- What are some potential risks?

Some practical issues with data collection

- Clearly, there are benefits with scraping data from websites of various kinds.
- What are some potential risks?
  1. user created, publicly available data can pose privacy risks for individuals who created it.

Some practical issues with data collection

- Clearly, there are benefits with scraping data from websites of various kinds.
- What are some potential risks?
  1. user created, publicly available data can pose privacy risks for individuals who created it. (sometimes, datasets are released after anonymization).
  2. you may be collecting the data without asking for permissions (ask: is everything freely visible on the web essentially free for such use?)
  3. copyrights/terms of service conflicts may come up.

Some practical issues with data collection

- Clearly, there are benefits with scraping data from websites of various kinds.
- What are some potential risks?
  1. user created, publicly available data can pose privacy risks for individuals who created it. (sometimes, datasets are released after anonymization).
  2. you may be collecting the data without asking for permissions (ask: is everything freely visible on the web essentially free for such use?)
  3. copyrights/terms of service conflicts may come up.
  4. Often, websites' policies allow free browsing, but not crawling.

Diesner, J., & Chin, C. L. (2016, May). Gratis, libre, or something else? Regulations and misassumptions related to working with publicly available text data. In Actes du Workshop on Ethics In Corpus Collection, Annotation & Application (ETHI-CA2), LREC, Portoroz, Slovénie.

Datasheets for Datasets-1

Ask the following while creating/using a dataset

1. Why was the dataset created? Who funded its creation? How and when was it created?
2. Who were involved in the collection? (students, crowdworkers etc.) How were they compensated?
3. What preprocessing/cleaning was done?
4. How is the dataset released/distributed?
5. Will the dataset be updated? How often? by whom?

Source: "Datasheets for Datasets", Gebru et.al., 2018

Datasheets for Datasets-2

Legal and Ethical considerations

- If it relates to people, were they told what the dataset would be used for and did they consent?
- If it relates to people, could this dataset expose people to harm or legal action?
- If it relates to people, does it unfairly advantage or dis-advantage a particular social group?
- Does the dataset comply with the EU General Data Protection Regulation (GDPR)?

Full list of questions in the original paper linked in the previous slide. Bender & Friedman, 2018 is another good paper on the topic of corpora creation/use (it is in our syllabus' readings section).

Corpus Collection: Summary

- There are many ways to get data, and label it.
- There are some concerns to be addressed while doing it too.
- I hope this part of today's class gave you a broad overview of all things related to data in NLP.

Corpus Collection: Summary

- There are many ways to get data, and label it.
- There are some concerns to be addressed while doing it too.
- I hope this part of today's class gave you a broad overview of all things related to data in NLP.
- Some questions:
  - What are some sources of data for your native language? (if it is not English/German).
  - Do you need annotated data? what sort of annotations are needed? How do you get them? - think about these questions.

Class Outline

- Quick recap of last class
- Corpus Collection
- **Text Extraction**
- Corpus Exploration

We can take a short 5 min break if you want.

Text Extraction - some questions

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

source material: Chapter 2 from https://practicalnlp.ai

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.
- 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
- Data stored in the cloud
- **best format* txt files, other plain text files

... ...

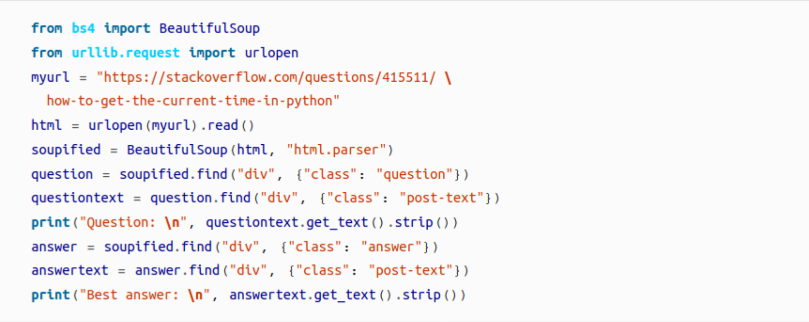
Reading from HTML/XML

What we need: bs4 library in Python

- 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/

bs4 - an example



```python
from bs4 import BeautifulSoup
from urllib.request import urlopen
myurl = "https://stackoverflow.com/questions/415511/ \
  how-to-get-the-current-time-in-python"
html = urlopen(myurl).read()
soupified = BeautifulSoup(html, "html.parser")
question = soupified.find("div", {"class": "question"})
questiontext = question.find("div", {"class": "post-text"})
print("Question: \n", questiontext.get_text().strip())
answer = soupified.find("div", {"class": "answer"})
answertext = answer.find("div", {"class": "post-text"})
print("Best answer: \n", answertext.get_text().strip())
```

note: current version of stackoverflow does not work with this code, as they changed their layout.

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/

Text Extraction: Summary

- 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
- However, some formats such as pdfs or images can never give you a perfect solution.
- What to do with what we managed to extract?

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

Class Outline

- Quick recap of last class
- Corpus Collection
- Text Extraction
- **Corpus Exploration**

Corpus Exploration - Goals

- Understand what corpus analysis is and why it is needed
- Do some basic analyses
- source: Chapter 1-2 in nltk.org/book, and a few linked blog posts

What is corpus exploration?

- Understanding what's in the corpus:
  - Looking at frequently used words/ngrams in the corpus
  - What words go together? (collocations)
  - Other properties of the corpus such as lexical diversity, linguistic coverage etc.
  - What categories are more frequent/less frequent etc.
- How can we visualize a corpus quickly?

.. ...

Why?

- Understand what the texts are about (generally speaking)
- What can be some useful features to use in an NLP model for this corpus
- Identify potential noise in the data
- Understand the limitations (may be something you want is not represented in the corpus? may be the data is heavily unbalanced and only some kinds of text are over represented?)

How?

Some basic analyses

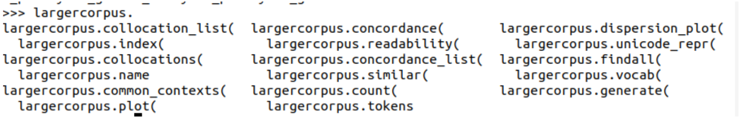
- Length of documents, average sentence length etc.
- Specific measure such as lexical diversity etc may be useful in some cases.
- For text classification, for example, it is useful to look at the distribution of various categories in the corpus, most common words per category etc.

On the entire corpus

I am taking the example texts that come with nltk in this example:

```
from nltk.book import * #loads some sample texts
texts() #lists the texts.
list(text1) #shows a text as a list of words.

largercorpus = Text(list(text1) + list(text2) + list(text3))
```

What we can do with this corpus:

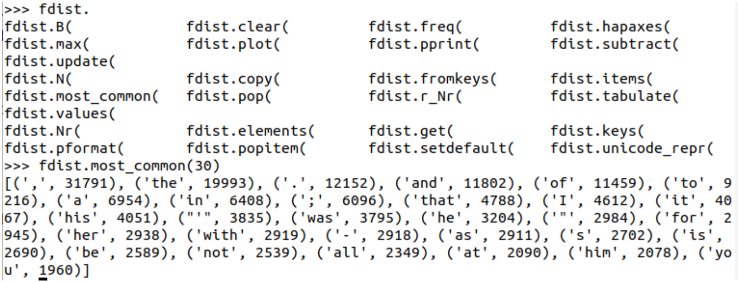


```
>>> largercorpus.
largercorpus.collocation_list(  largercorpus.concordance(       largercorpus.dispersion_plot(
  largercorpus.index(             largercorpus.readability(       largercorpus.unicode_repr(
largercorpus.collocations(      largercorpus.concordance_list(  largercorpus.findall(
  largercorpus.name               largercorpus.similar(           largercorpus.vocab(
largercorpus.common_contexts(   largercorpus.count(             largercorpus.generate(
  largercorpus.plot(              largercorpus.tokens
```

Frequency Distributions

```
fdist = FreqDist(largercorpus)
```

What we can do with this fdist object:

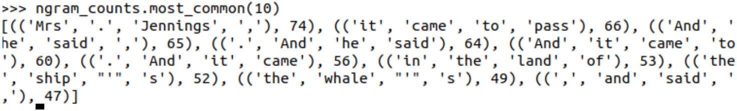


```
>>> fdist.
fdist.B(             fdist.clear(         fdist.freq(          fdist.hapaxes(
fdist.max(           fdist.plot(          fdist.pprint(        fdist.subtract(
fdist.update(
fdist.N(             fdist.copy(          fdist.fromkeys(      fdist.items(
fdist.most_common(   fdist.pop(           fdist.r_Nr(          fdist.tabulate(
fdist.values(
fdist.Nr(            fdist.elements(      fdist.get(           fdist.keys(
fdist.pformat(       fdist.popitem(       fdist.setdefault(    fdist.unicode_repr(
>>> fdist.most_common(30)
[(',', 31791), ('the', 19993), ('.', 12152), ('and', 11802), ('of', 11459), ('to', 9
216), ('a', 6954), ('in', 6408), (';', 6096), ('that', 4788), ('I', 4612), ('it', 40
67), ('his', 4051), ("'", 3835), ('was', 3795), ('he', 3204), ('"', 2984), ('for', 2
945), ('her', 2938), ('with', 2919), ('-', 2918), ('as', 2911), ('s', 2702), ('is',
2690), ('be', 2589), ('not', 2539), ('all', 2349), ('at', 2090), ('him', 2078), ('yo
u', 1960)]
```

How? - Frequency Distributions of ngrams

```
from collections import Counter
from nltk import ngrams
ngram_counts = Counter(ngrams(largercorpus, 4)) #4 for bigrams, 3 for trigrams ...
ngram_counts.most_common(10)
```

What we can do with this fdist object:

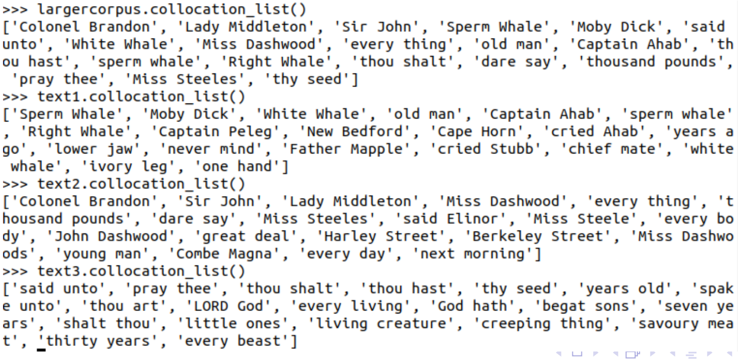


```
>>> ngram_counts.most_common(10)
[(('Mrs', '.', 'Jennings', ','), 74), (('it', 'came', 'to', 'pass'), 66), (('And', '
he', 'said', ','), 65), (('.', 'And', 'he', 'said'), 64), (('And', 'it', 'came', 'to
'), 60), (('.', 'And', 'it', 'came'), 56), (('in', 'the', 'land', 'of'), 53), (('the
', 'ship', "'", 's'), 52), (('the', 'whale', "'", 's'), 49), ((',', 'and', 'said', '
,'), 47)]
```

How? - Collocations

"A collocation is a sequence of words that occur together unusually often. Thus red wine is a collocation, whereas the wine is not. A characteristic of collocations is that they are resistant to substitution with words that have similar senses; for example, maroon wine sounds definitely odd."

```
largercorpus.collocation_list()
```



```
>>> largercorpus.collocation_list()
['Colonel Brandon', 'Lady Middleton', 'Sir John', 'Sperm Whale', 'Moby Dick', 'said
unto', 'White Whale', 'Miss Dashwood', 'every thing', 'old man', 'Captain Ahab', 'th
ou hast', 'sperm whale', 'Right Whale', 'thou shalt', 'dare say', 'thousand pounds',
 'pray thee', 'Miss Steeles', 'thy seed']
>>> text1.collocation_list()
['Sperm Whale', 'Moby Dick', 'White Whale', 'old man', 'Captain Ahab', 'sperm whale'
, 'Right Whale', 'Captain Peleg', 'New Bedford', 'Cape Horn', 'cried Ahab', 'years a
go', 'lower jaw', 'never mind', 'Father Mapple', 'cried Stubb', 'chief mate', 'white
 whale', 'ivory leg', 'one hand']
>>> text2.collocation_list()
['Colonel Brandon', 'Sir John', 'Lady Middleton', 'Miss Dashwood', 'every thing', 't
housand pounds', 'dare say', 'Miss Steeles', 'said Elinor', 'Miss Steele', 'every bo
dy', 'John Dashwood', 'great deal', 'Harley Street', 'Berkeley Street', 'Miss Dashwo
ods', 'young man', 'Combe Magna', 'every day', 'next morning']
>>> text3.collocation_list()
['said unto', 'pray thee', 'thou shalt', 'thou hast', 'thy seed', 'years old', 'spak
e unto', 'thou art', 'LORD God', 'every living', 'God hath', 'begat sons', 'seven ye
ars', 'shalt thou', 'little ones', 'living creature', 'creeping thing', 'savoury mea
t', 'thirty years', 'every beast']
```

Know your pre-processing

- If I am using a bag of words approach for feature engineering, what exactly are the features that are extracted? (cv.get_feature_names())
- What are the stopwords I am eliminating when I import a stopword list from nltk/sklearn etc.?
- How does my text look like after doing these pre-processing steps?

etc...

useful link: https://kavita-ganesan.com/how-to-use-countvectorizer/

Concluding Remarks on Corpus Analysis

- How to make these more useful/meaningful?
  - Do some corpus pre-processing (e.g., lowecaser, remove stop words, punctuation etc)
  - Explore other plotting functions like wordcloud, disperson plots etc to understand data.
- useful tutorials: on Kaggle and neptune.ai

Note: Go through the first 2 chapters of the NLTK book. It has good examples of doing such exploratory analysis of a corpus.

Another Exercise

Time 20 minutes, in breakout rooms.
Go to Linguistic Data Consortium's Top-10 most distributed NLP corpora (https://catalog.ldc.upenn.edu/topten). Pick any corpus you want. See if you can build a datasheet for this dataset (Look at the full list of questions at the end of Gebru et.al. 2018 paper). One person per team will summarize their observations at the end. Don't do exhaustively, do what you can easily understand from available material.

Next Class

- Topic: Data Labeling, using Snorkel
- To Do for you:
  1. Decide on a team for group discussion (13th Jan 2021 - today!)
  2. Decide on a paper for group discussion (15th Jan 2021)
- Remember: You can email me for a meeting or post your questions in the forum for today's session "Data Collection".