Lecture 10: Web Mining, Annotation

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

- Web and social media mining
  - Web pages: HTML basics
  - Twitter mining revisited
- Linguistic annotation
  - TimeML

Homework 2 pitfalls

	Total token #	Total sentence #	Avg sentence length
Low	300172	13349	22.48
Medium	2206853	94177	23.43
high	1678805	71067	23.62

1. Obtaining averages from flattened groups →
   - We did this back in LING 1330…
2. Producing per-essay measurements, but results are saved as a *separate series*, not inserted into essay_df.
   - Problem? The values become detached from their original essays: you can no longer connect them to additional attributes (L1, Prompt, token count…)
   - You should build out essay_df with per-essay measurements, and use it as your exploration base. Use .groupby() and filters to zone in on particular attributes and further narrow down… on the fly!
3. Only ever looking at three average numbers (for low, medium, high) in drawing conclusions. (And ANOVA.)
   - Remedy? Look at overall DISTRIBUTION.
   - Use .describe(), boxplots.
4. For-loops for processing each row, "+= 1".
   - This is NOT the pandas way!
5. Inefficiency. Don't tokenize 5 times!

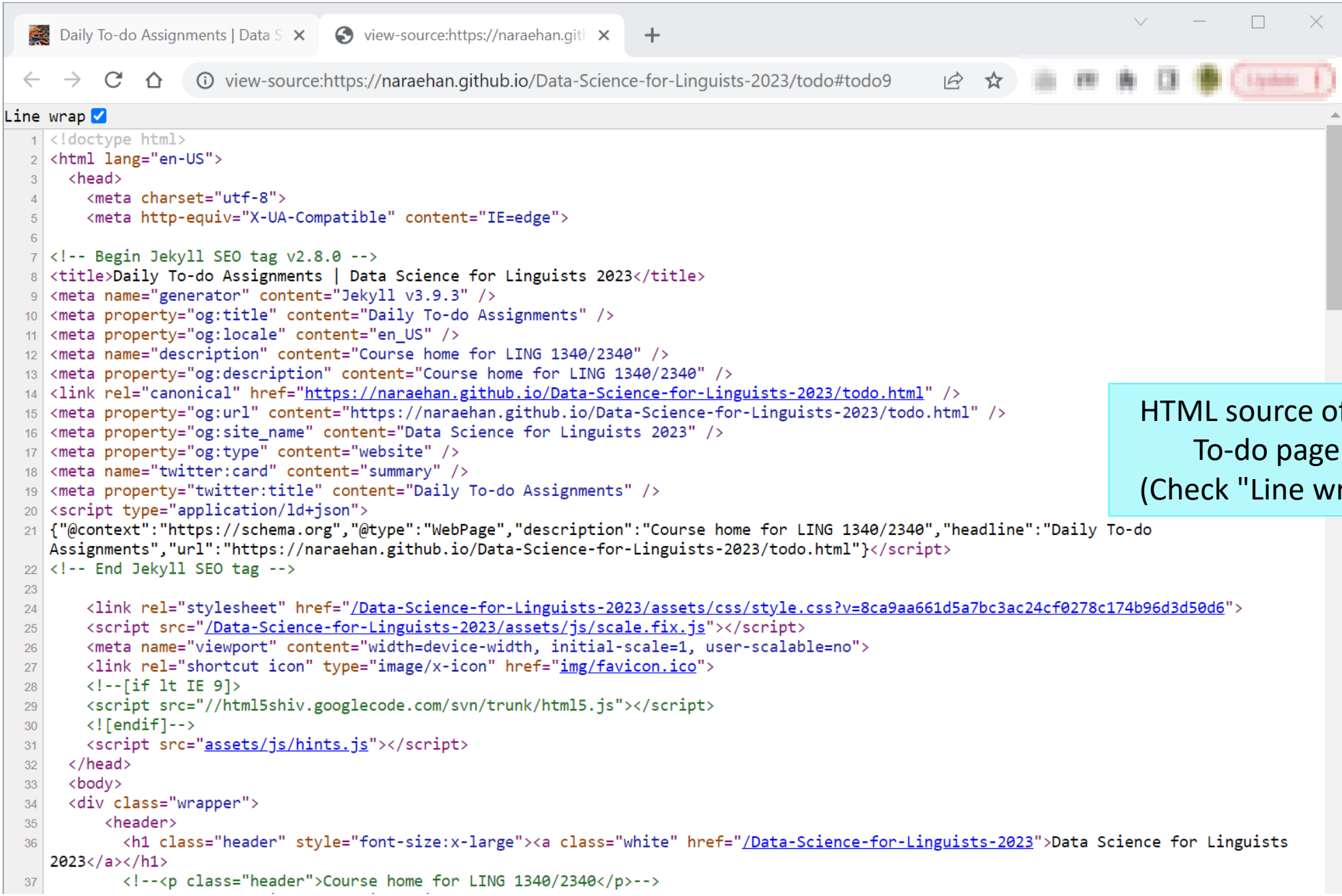
We must adapt to the new **pandas way**, which at long last allows us a proper statistics treatment.

Per-sample measurements are the ground truth! You then closely examine their **distribution**.

Web mining

- Involves "web crawling" "web spyder", ...
- `scrapy` is the most popular library.
  - https://scrapy.org/

  ← You will have to install it first.
- You have collected a set of web pages. Now what?
  - A web page typically has tons of non-text, extraneous data such as headers, scripts, etc.
    - Example: https://naraehan.github.io/Data-Science-for-Linguists-2024/todo
  - You will need to parse each page to extract textual data.
  - Beautiful Soup (bs4) is capable of parsing XML and HTML files.
- OK, so you've processed the web pages as data. Now what?
  - Linguistic analysis?



HTML source of our To-do page. (Check "Line wrap")

Processing a static Twitter corpus

- "Twitter Samples" corpus can be downloaded from http://www.nltk.org/nltk_data/

In [3]:
```
# One json object per line
jfile = 'D:/Corpora/twitter_samples/positive_tweets.json'
jlines = open(jfile).readlines()
jlines[0]
```

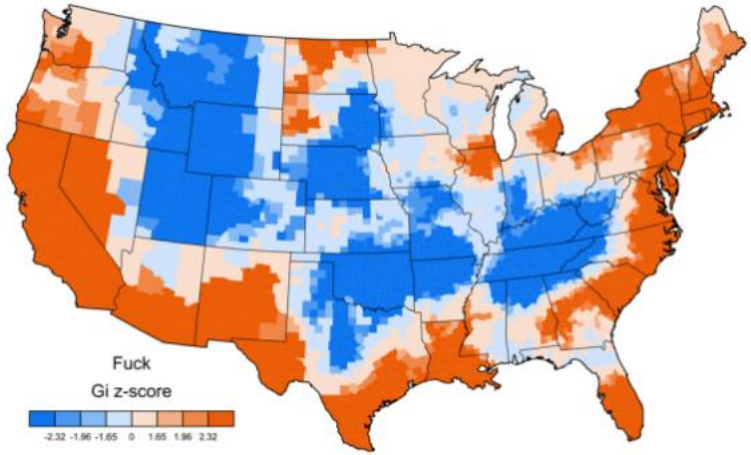
Out[3]:
```
'{"contributors": null, "coordinates": null, "text": "#FollowFriday @France_Int
e @PKuchly57 @Milipol_Paris for being top engaged members in my community this
 week :)", "user": {"time_zone": "Paris", "profile_background_image_url": "htt
```

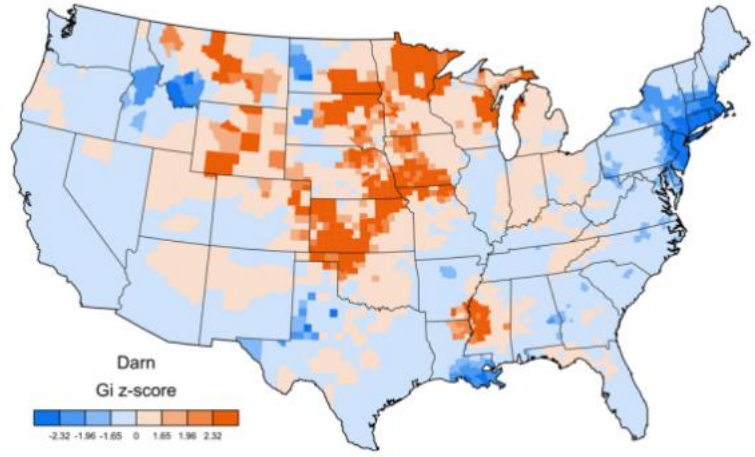
In [5]:
```
# using json library to read line.
import json
json.loads(jlines[0])
```

Out[5]:
```
{'contributors': None,
 'coordinates': None,
 'created_at': 'Fri Jul 24 08:23:36 +0000 2015',
 'entities': {'hashtags': [{'indices': [0, 13], 'text': 'FollowFriday'}],
  'symbols': [],
  'urls': [],
  'user_mentions': [{'id': 3222273608,
    'id_str': '3222273608',
    'indices': [14, 26],
    'name': 'France International',
```

Mining social media for swear words

- https://stronglang.wordpress.com/2015/07/28/mapping-the-united-swears-of-america/
  - Jack Grieve mined Twitter and mapped prominent swear words by geographic regions within the US





Linguistic annotation: representing meaning

- TimeML
- Abstract Meaning Representation (AMR)
  - https://amr.isi.edu/index.html
- What semantic theories and concepts does it use?

TimeML

- Markup Language for Temporal and Event Expressions
  - https://timeml.github.io/site/index.html
  - http://xml.coverpages.org/timeML.html
- Influenced by Reichenbach's theory of tense (1947)
  - Distinguishes: speech time, event time, and reference time
  - https://plato.stanford.edu/entries/tense-aspect/
- Published corpora ("Timebank"):
  - https://timeml.github.io/site/timebank/documentation-1.2.html
  - TimeBank 1.2 (released by Linguistic Data Consortium):
    - https://catalog.ldc.upenn.edu/LDC2006T08

TimeML exercise

- The following simple sentence, uttered on October 20, 2009, encodes **events** that occurred on a **time** axis.

> Mia visited Seoul to look me up yesterday.

- As a linguist, determine what pieces of semantic information are present, and think about how you will formally represent them.

Annotating event/time relation: TimeML

```
<maf xmlns:"http://www.iso.org/maf">
   <seg type="token" xml:id="token1">Mia</seg>
   <seg type="token" xml:id="token2">visited</seg>
   <seg type="token" xml:id="token3">Seoul</seg>
   <seg type="token" xml:id="token4">to</seg>
   <seg type="token" xml:id="token5">look</seg>
   <seg type="token" xml:id="token6">me</seg>
   <seg type="token" xml:id="token7">up</seg>
   <seg type="token" xml:id="token8">yesterday</seg>
   <pc>.</pc>
</maf>
```

Word tokens:
inline segmentation

```
<isoTimeML xmlns:"http://www.iso.org./isoTimeML">
   <TIMEX3 xml:id="t0" type="DATE" value="2009-10-20"
      functionInDocument="CREATION_TIME"/>
   <EVENT xml:id="e1" target="#token2" class="OCCURRENCE" tense="PAST"/>
   <EVENT xml:id="e2" target="#token5 #token7" class="OCCURRENCE"
      tense="NONE" vForm="INFINITIVE"/>
   <TIMEX3 xml:id="t1" type="DATE" value="2009-10-19"/>
   <TLINK eventID="#e1" relatedToTime="#t0" relType="BEFORE"/>
   <TLINK eventID="#e1" relatedToTime="#t1" relType="ON_OR_BEFORE"/>
   <TLINK eventID="#e2" relatedToTime="#t1" relType="IS_INCLUDED"/>
</isoTimeML>
```

Time Event Annotation:
stand-off annotation

Knowledge representation

Human-curated systems for meanings and concepts:

- Computerized & hierarchically organized lexicons
  - WordNet, Proposition Bank
- Ontology, taxonomy
  - Computerized conceptual hierarchies
  - Industry applications are often based on domain-specific ontologies/taxonomies

Wrapping up

- Next class
  - To-do #10: Try out AMR
  - More annotation
- Your project
  - Progress Report #1 specs published
  - Work on it! Focus on DATA.