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

Resource-specific (ad-hoc) formats

- Brown corpus

```
The/at Fulton/np-tl County/nn-tl Grand/jj-tl Jury/nn-tl said/vbd
Friday/nr an/at investigation/nn of/in Atlanta's/np$ recent/jj
primary/nn election/nn produced/vbd ``/`` no/at evidence/nn ''/''
that/cs any/dti irregularities/nns took/vbd place/nn ./.
```

- Korean Treebank corpus:

```
;;05:127: 저는 그 일을 할 수 있는 한 빨리 하겠습니다 .

(S  (NP-SBJ 저/NPN+는/PAU)
   (VP (NP-OBJ-LV 그/DAN
                  일/NNC+을/PCA)
       (VP (NP-ADV (S (NP-SBJ (S (NP-SBJ *pro*)
                                  (VP 하/VV+ㄹ/EAN))
                               (NP 수/NNX))
                      (ADJP 있/VJ+는/EAN))
                   (NP 한/NNX))
           (ADVP 빨리/ADV)
           (VP (LV 하/VV+겠/EPF+습니다/EFN))))
   ./SFN)
```

NOT standard
(cf. XML, JSON).
Project-dependent.

It is up to end users to understand the data format, then write code to parse data files.

Refer to documentation!

Dependency annotation: format

- https://raw.githubusercontent.com/UniversalDependencies/UD_English-EWT/dev/en_ewt-ud-dev.conllu

```
# sent_id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-0002
# newpar id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-p0002
# text = President Bush on Tuesday nominated two individuals to replace retiring jurists on federal courts in the Washington area.
1       President       President       PROPN   NNP     Number=Sing     5       nsubj   5:nsubj _
2       Bush    Bush    PROPN   NNP     Number=Sing     1       flat    1:flat  _
3       on      on      ADP     IN      _       4       case    4:case  _
4       Tuesday Tuesday PROPN   NNP     Number=Sing     5       obl     5:obl:on        _
5       nominated       nominate        VERB    VBD     Mood=Ind|Tense=Past|VerbForm=Fin        0       root    0:root  _
6       two     two     NUM     CD      NumType=Card    7       nummod  7:nummod        _
7       individuals     individual      NOUN    NNS     Number=Plur     5       obj     5:obj   _
8       to      to      PART    TO      _       9       mark    9:mark  _
9       replace replace VERB    VB      VerbForm=Inf    5       advcl   5:advcl:to      _
10      retiring        retire  VERB    VBG     VerbForm=Ger    11      amod    11:amod _
11      jurists jurist  NOUN    NNS     Number=Plur     9       obj     9:obj   _
12      on      on      ADP     IN      _       14      case    14:case _
13      federal federal ADJ     JJ      Degree=Pos      14      amod    14:amod _
14      courts  court   NOUN    NNS     Number=Plur     11      nmod    11:nmod:on      _
15      in      in      ADP     IN      _       18      case    18:case _
16      the     the     DET     DT      Definite=Def|PronType=Art       18      det     18:det  _
17      Washington      Washington      PROPN   NNP     Number=Sing     18      compound        18:compound     _
18      area    area    NOUN    NN      Number=Sing     14      nmod    14:nmod:in      SpaceAfter=No
19      .       .       PUNCT   .       _       5       punct   5:punct _
```

Known as the
CoNLL-U format
https://universaldependencies.org/format.html

Do not re-invent the wheel.

- If you can, avoid parsing them manually!
- There are Python libraries. Import and use them.
  - CSV & TSV: `pandas`
  - HTML & XML: Beautiful Soup (`bs4`)
  - JSON:
    - `json` library
    - `pandas.read_json`
- NLP-specific formats (Treebank, Universal Dependency, CoNLL):
  - Look at NLTK, see if it has reader
  - If not, chances are there is parser library written by someone somewhere (likely on GitHub)

Data-mining web & social media

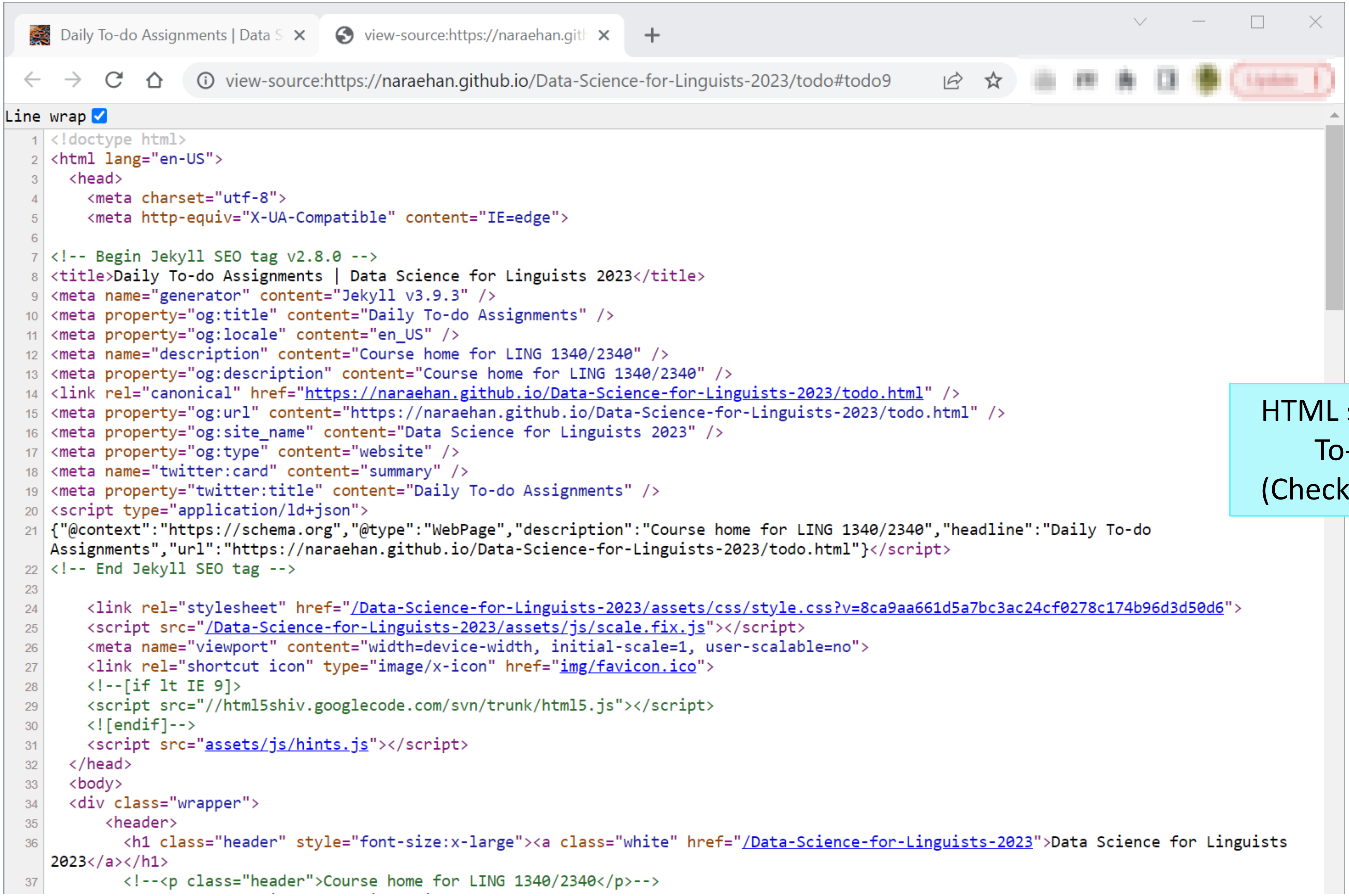
- Twitter sample corpus
  - Static corpus: download from the NLTK data page
- How does one data-mine Twitter?
  - Answer: through **API** (Application Program Interface)
  - Getting acquainted with JSON format
  - Tutorials on on the Learning Resource page
  - I also gave a brief tour last week.
- Libraries used: `tweepy`, `json`

If you can pay for it…
(as of last week)

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-2023/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?


Daily To-do Assignments | Data S
view-source:https://naraehan.git
view-source:https://naraehan.github.io/Data-Science-for-Linguists-2023/todo#todo9

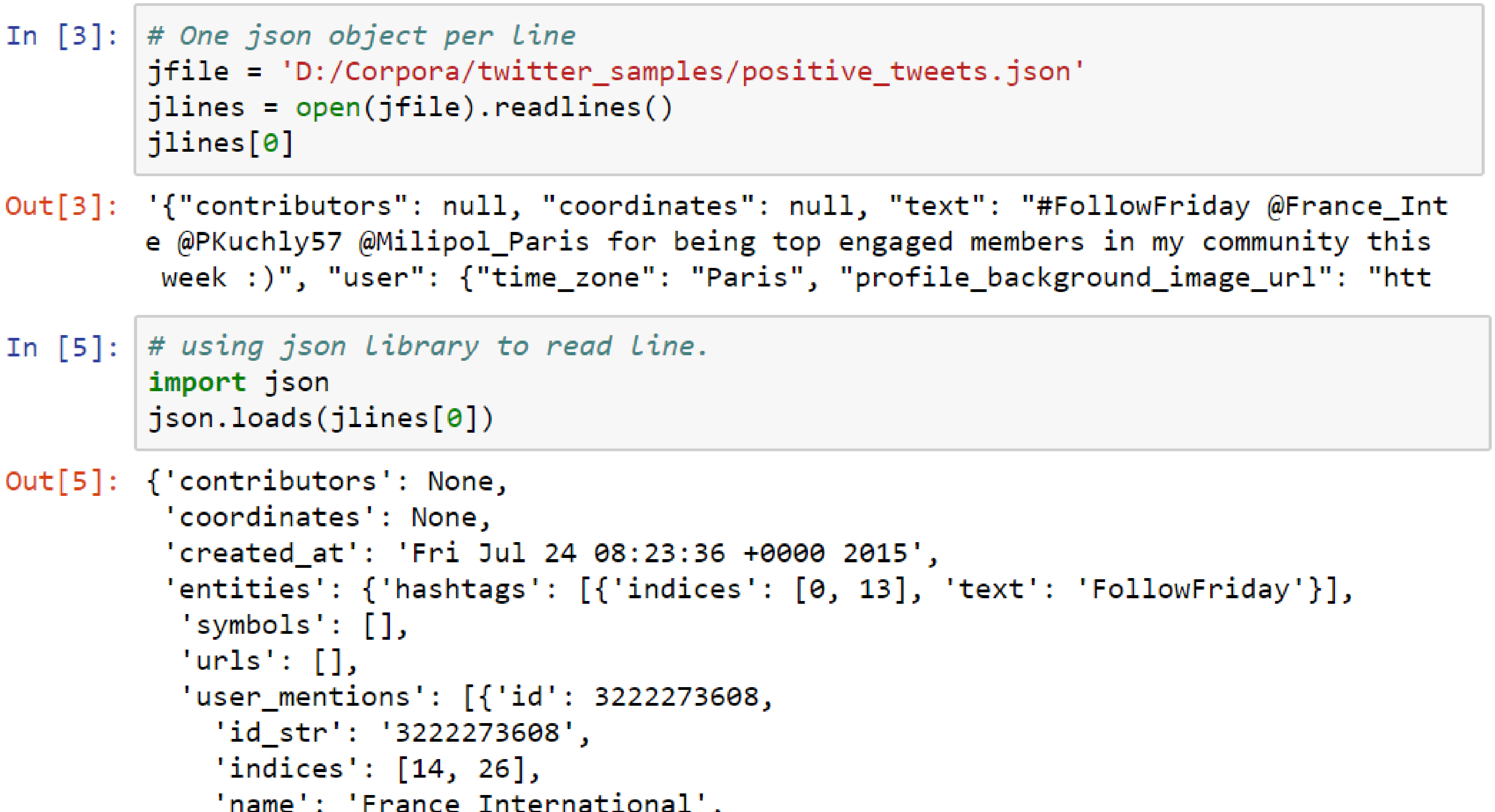
Line wrap ☑

```
<!doctype html>
<html lang="en-US">
  <head>
    <meta charset="utf-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">

<!-- Begin Jekyll SEO tag v2.8.0 -->
<title>Daily To-do Assignments | Data Science for Linguists 2023</title>
<meta name="generator" content="Jekyll v3.9.3" />
<meta property="og:title" content="Daily To-do Assignments" />
<meta property="og:locale" content="en_US" />
<meta name="description" content="Course home for LING 1340/2340" />
<meta property="og:description" content="Course home for LING 1340/2340" />
<link rel="canonical" href="https://naraehan.github.io/Data-Science-for-Linguists-2023/todo.html" />
<meta property="og:url" content="https://naraehan.github.io/Data-Science-for-Linguists-2023/todo.html" />
<meta property="og:site_name" content="Data Science for Linguists 2023" />
<meta property="og:type" content="website" />
<meta name="twitter:card" content="summary" />
<meta property="twitter:title" content="Daily To-do Assignments" />
<script type="application/ld+json">
{"@context":"https://schema.org","@type":"WebPage","description":"Course home for LING 1340/2340","headline":"Daily To-do
Assignments","url":"https://naraehan.github.io/Data-Science-for-Linguists-2023/todo.html"}</script>
<!-- End Jekyll SEO tag -->

    <link rel="stylesheet" href="/Data-Science-for-Linguists-2023/assets/css/style.css?v=8ca9aa661d5a7bc3ac24cf0278c174b96d3d50d6">
    <script src="/Data-Science-for-Linguists-2023/assets/js/scale.fix.js"></script>
    <meta name="viewport" content="width=device-width, initial-scale=1, user-scalable=no">
    <link rel="shortcut icon" type="image/x-icon" href="img/favicon.ico">
    <!--[if lt IE 9]>
    <script src="//html5shiv.googlecode.com/svn/trunk/html5.js"></script>
    <![endif]-->
    <script src="assets/js/hints.js"></script>
  </head>
  <body>
  <div class="wrapper">
      <header>
        <h1 class="header" style="font-size:x-large"><a class="white" href="/Data-Science-for-Linguists-2023">Data Science for Linguists
2023</a></h1>
        <!--<p class="header">Course home for LING 1340/2340</p>-->
```

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

Processing a static Twitter corpus

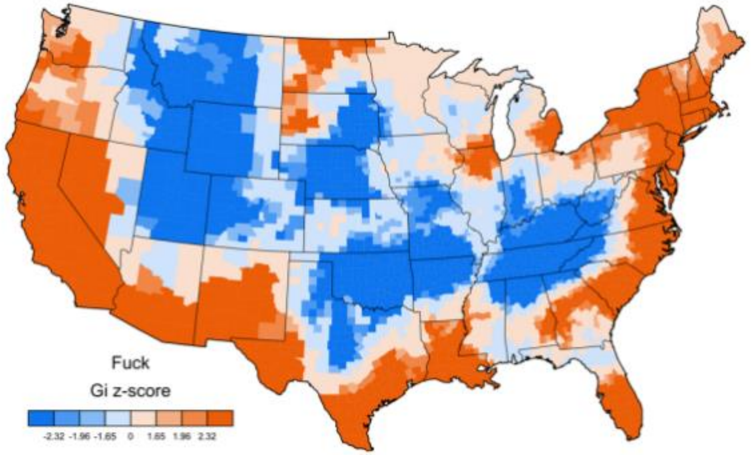
- "Twitter Samples" corpus can be downloaded from http://www.nltk.org/nltk_data/

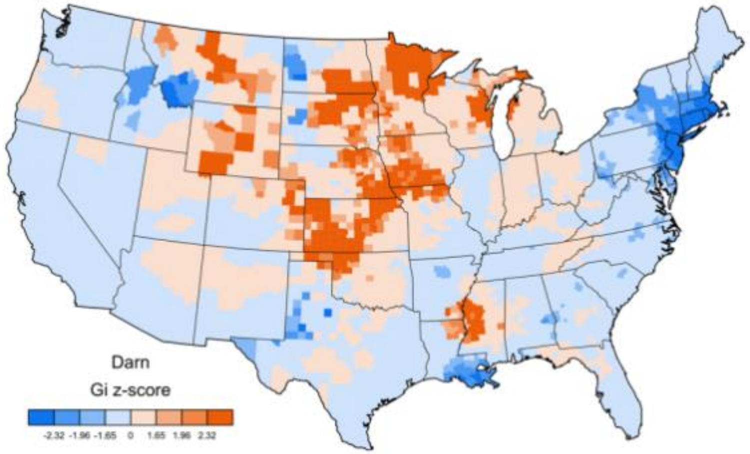


```
In [3]:  # One json object per line
         jfile = 'D:/Corpora/twitter_samples/positive_tweets.json'
         jlines = open(jfile).readlines()
         jlines[0]

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
  - http://timeml.org/site/publications/timeMLdocs/timeml_1.2.1.html
  - http://xml.coverpages.org/timeML.html
- Published corpora ("Timebank"):
  - http://www.timeml.org/timebank/timebank.html (currently down)
  - 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.