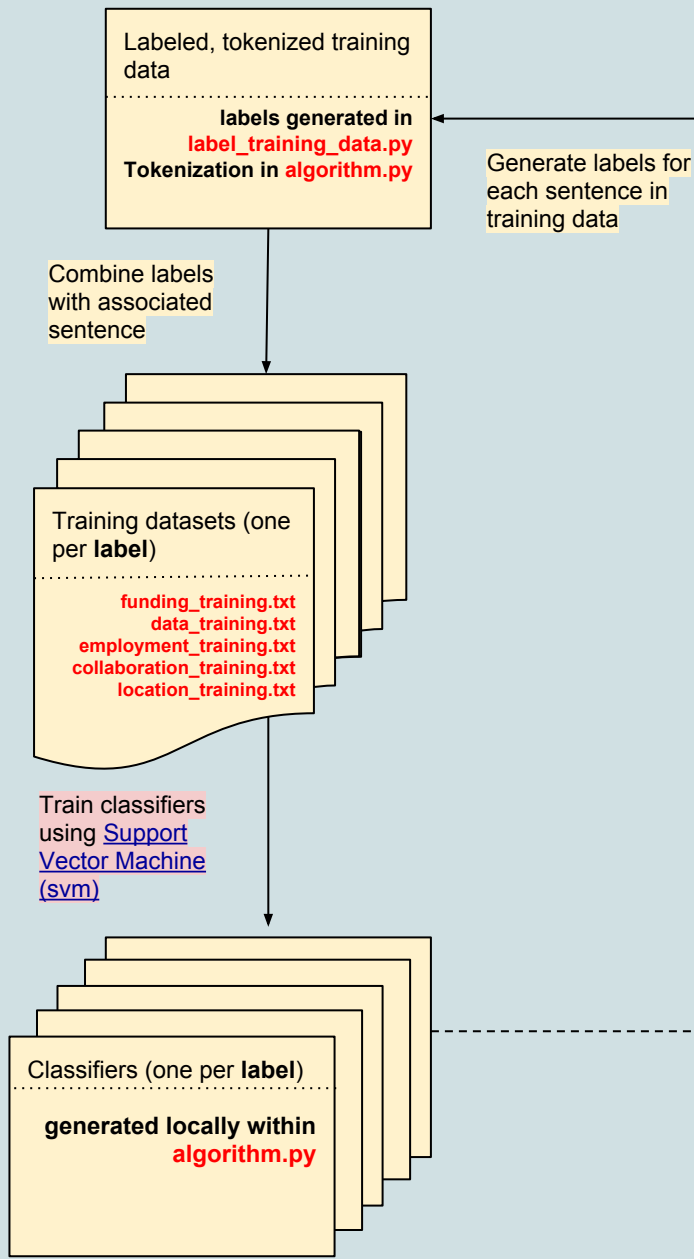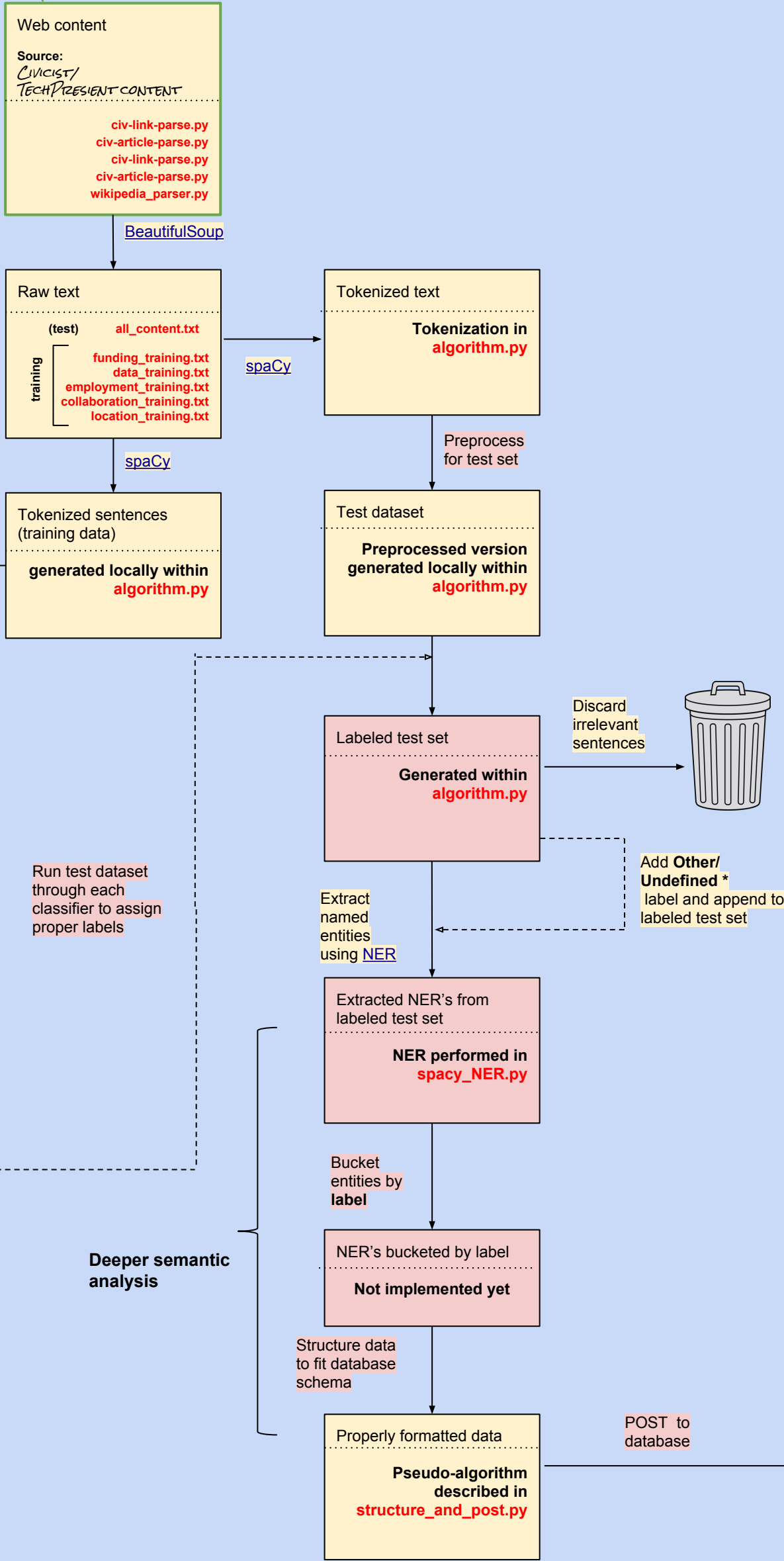# Classifiers

Labels:
(one for each connection type in Civic Graph)

- Funding
- Data
- Employment
- Collaboration
- Location
- Other/Undefined *
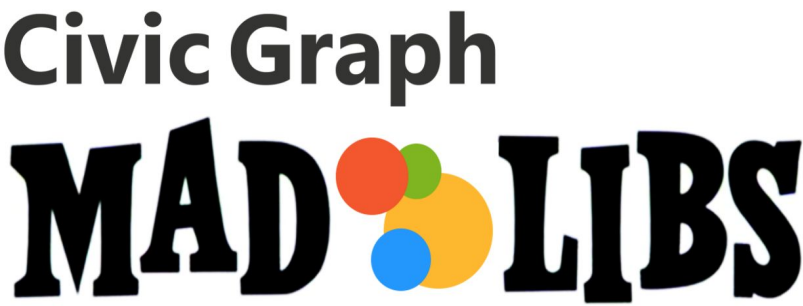
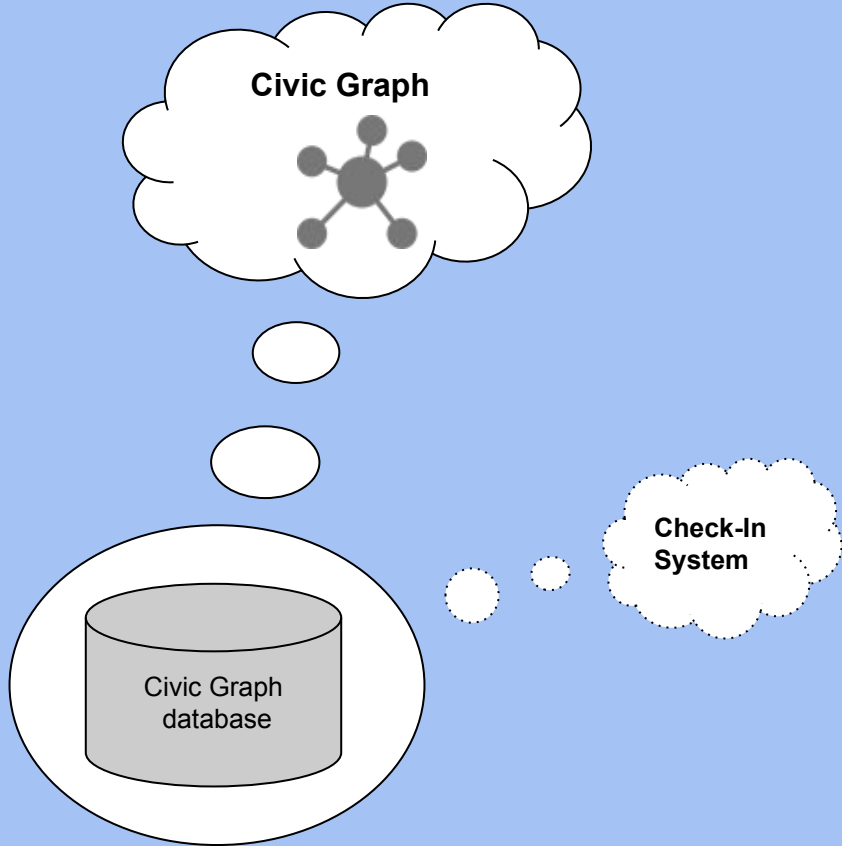# Scrapers/Tokenizers

Web content

Source:
*Civicist/
TechPresident content*

civ-link-parse.py
civ-article-parse.py
civ-link-parse.py
civ-article-parse.py
wikipedia_parser.py

BeautifulSoup

Raw text

(test)   all_content.txt

training
funding_training.txt
data_training.txt
employment_training.txt
collaboration_training.txt
location_training.txt

spaCy

Tokenized text

**Tokenization in
algorithm.py**

Preprocess
for test set

spaCy

Tokenized sentences
(training data)

**generated locally within
algorithm.py**

Test dataset

**Preprocessed version
generated locally within
algorithm.py**

Labeled, tokenized training
data

**labels generated in
label_training_data.py
Tokenization in algorithm.py**

Generate labels for
each sentence in
training data

Combine labels
with associated
sentence

Training datasets (one
per **label**)

funding_training.txt
data_training.txt
employment_training.txt
collaboration_training.txt
location_training.txt

Train classifiers
using Support
Vector Machine
(svm)

Classifiers (one per **label**)

**generated locally within
algorithm.py**

Run test dataset
through each
classifier to assign
proper labels

Labeled test set

**Generated within
algorithm.py**

Discard
irrelevant
sentences

Add **Other/
Undefined ***
label and append to
labeled test set

Extract
named
entities
using NER

Extracted NER's from
labeled test set

**NER performed in
spacy_NER.py**

**Deeper semantic
analysis**

Bucket
entities by
label

NER's bucketed by label

Not implemented yet

Structure data
to fit database
schema

Properly formatted data

**Pseudo-algorithm
described in
structure_and_post.py**

POST to
database

Integration with Civic Graph

Key:
- Code written
- Not done
- Source subject
  to change
- **corresponding file**

Civic Graph

Check-In
System

Civic Graph
database

Civic Graph

MAD LIBS

Automating Data Collection for
Microsoft Civic Graph

(Pipeline)