Milestone I

Karina Huang, Dianne Lee, Abhimanyu Vasishth, Phoebe Wong

Problem Statement

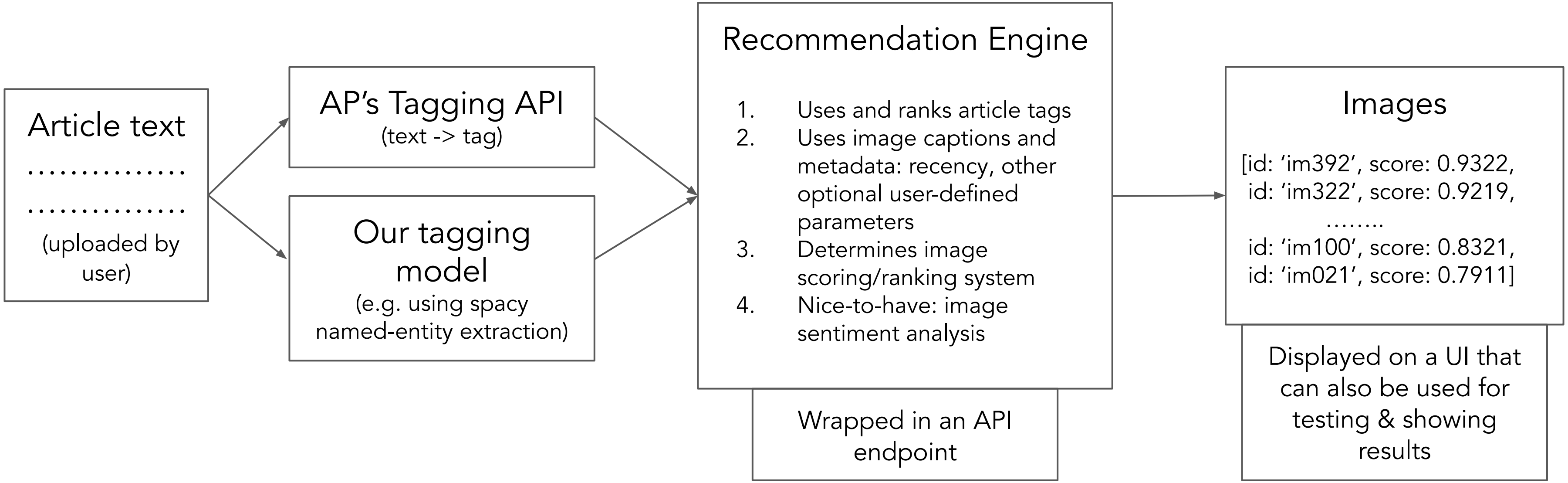
- One of the main sources of revenue for AP is the selling of image licenses

- Publishers currently search text associated with the images they are looking for through manually generated keywords, which is time consuming and may not make full use of the wide range of images at AP's disposal.

- Instead, we build an image recommendation system that takes in article text and outputs a set of images.

- This service could be used on the AP web portals or in an API for external clients

Project Plan: Scope of Work



Article text

....

....

(uploaded by user)

AP's Tagging API

(text -> tag)

Our tagging model

(e.g. using spacy named-entity extraction)

Recommendation Engine

1. Uses and ranks article tags
2. Uses image captions and metadata: recency, other optional user-defined parameters
3. Determines image scoring/ranking system
4. Nice-to-have: image sentiment analysis

Wrapped in an API endpoint

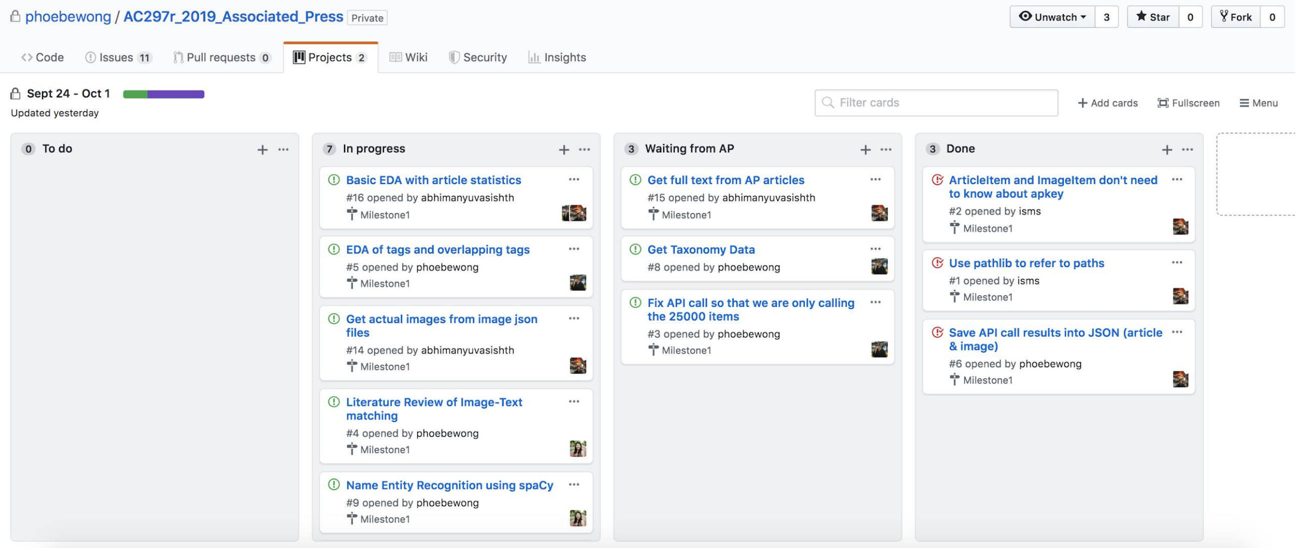
Images

[id: 'im392', score: 0.9322,
id: 'im322', score: 0.9219,
....
id: 'im100', score: 0.8321,
id: 'im021', score: 0.7911]

Displayed on a UI that can also be used for testing & showing results

Team And Collaboration Infrastructure

- Weekly Tue & Fri work sessions
- Github project with weekly issues assigned to each of us
- Google drive for presentations, documents and storing data



phoebewong / AC297r_2019_Associated_Press Private

Unwatch 3 Star 0 Fork 0

Code Issues 11 Pull requests 0 Projects 2 Wiki Security Insights

Sept 24 - Oct 1

Updated yesterday

Filter cards

Add cards Fullscreen Menu

0 To do

7 In progress

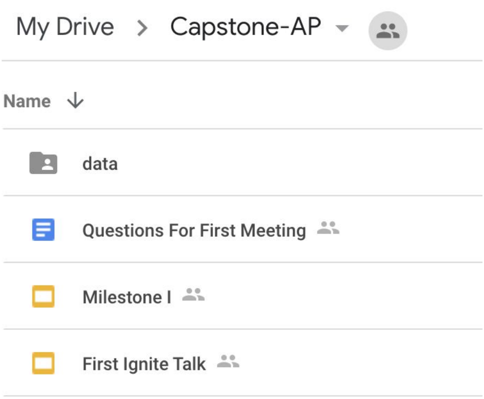
- Basic EDA with article statistics
  #16 opened by abhimanyuvasishth
  Milestone1
- EDA of tags and overlapping tags
  #5 opened by phoebewong
  Milestone1
- Get actual images from image json files
  #14 opened by abhimanyuvasishth
  Milestone1
- Literature Review of Image-Text matching
  #4 opened by phoebewong
  Milestone1
- Name Entity Recognition using spaCy
  #9 opened by phoebewong
  Milestone1

3 Waiting from AP

- Get full text from AP articles
  #15 opened by abhimanyuvasishth
  Milestone1
- Get Taxonomy Data
  #8 opened by phoebewong
- Fix API call so that we are only calling the 25000 items
  #3 opened by phoebewong
  Milestone1

3 Done

- ArticleItem and ImageItem don't need to know about apkey
  #2 opened by isms
  Milestone1
- Use pathlib to refer to paths
  #1 opened by isms
  Milestone1
- Save API call results into JSON (article & image)
  #6 opened by phoebewong
  Milestone1



My Drive > Capstone-AP

Name

- data
- Questions For First Meeting
- Milestone I
- First Ignite Talk

Literature Review

Keyword ranking

- TextRank, PageRank
- Theme-weighted Ranking of Keywords from Text Documents using Phrase Embeddings

Image-Sentence Retrieval

- Bidirectional Image-Sentence Mapping
- Learning Two-Branch NN for Image-Text Matching Tasks
- Overview of Text Similarity Metrics

Literature Review

WordNet: a lexical database for English (Miller 1995)

-

Literature Review

Using TF-IDF to Determine Word Relevance in Document Queries (Ramos 2003)

- Results of applying Term Frequency Inverse Document Frequency (TF-IDF) to determine favorable words in a corpus of documents to use in a query
- (1) Term Frequency

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

- (2) Inverse Document Frequency

$$idf(w) = log(\frac{N}{df_t})$$

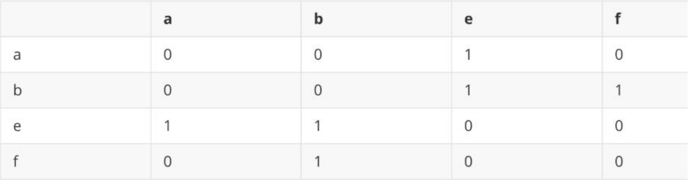
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Literature Review - TextRank

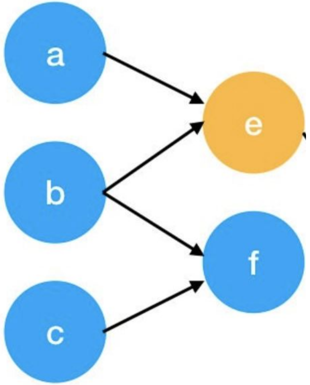
TextRank (Mihalcea, R., & Tarau, P. (2004))

- A graph-based ranking model, rank extracted keywords with “importance”
    - Inspired by Google’s PageRank to rank web pages
- 1) Extract keywords from full text (NER of full text)
- 2) Ranking keywords
    - Each keyword is a node
    - Weight of the keyword is decided by the weights of its associated keywords
    - Example: for an article,
        - Keyword list: [w1, w2, …, w_n]
        - Select a k window of keywords: [w1, w2, …, w_k] (k: {2-10})
            - Any 2-word pair = 1 undirected edge (w1w2, w1w3, …, w1w_k, w2w3...)

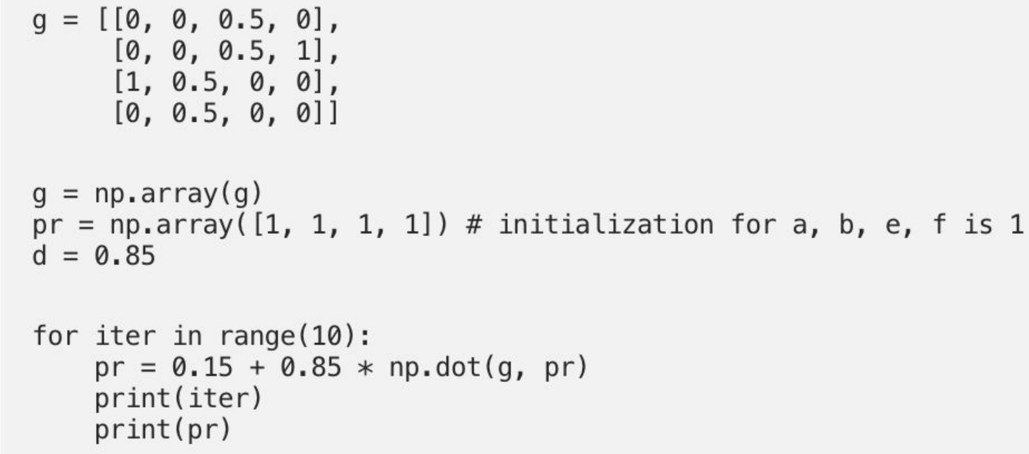
Literature Review - TextRank



	a	b	e	f
a	0	0	1	0
b	0	0	1	1
e	1	1	0	0
f	0	1	0	0



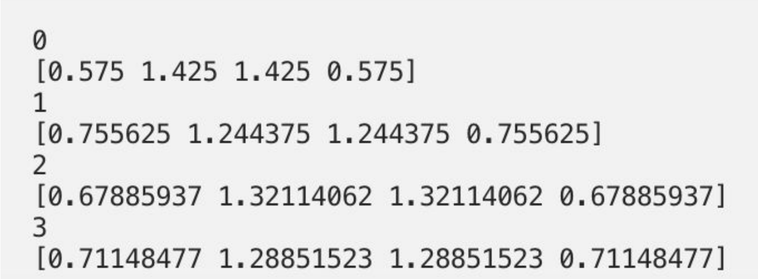
$$S(V_i) = (1-d) + d * \sum_j \frac{1}{|V_j|} S(V_j)$$



```
g = [[0, 0, 0.5, 0],
     [0, 0, 0.5, 1],
     [1, 0.5, 0, 0],
     [0, 0.5, 0, 0]]


g = np.array(g)
pr = np.array([1, 1, 1, 1]) # initialization for a, b, e, f is 1
d = 0.85


for iter in range(10):
    pr = 0.15 + 0.85 * np.dot(g, pr)
    print(iter)
    print(pr)
```

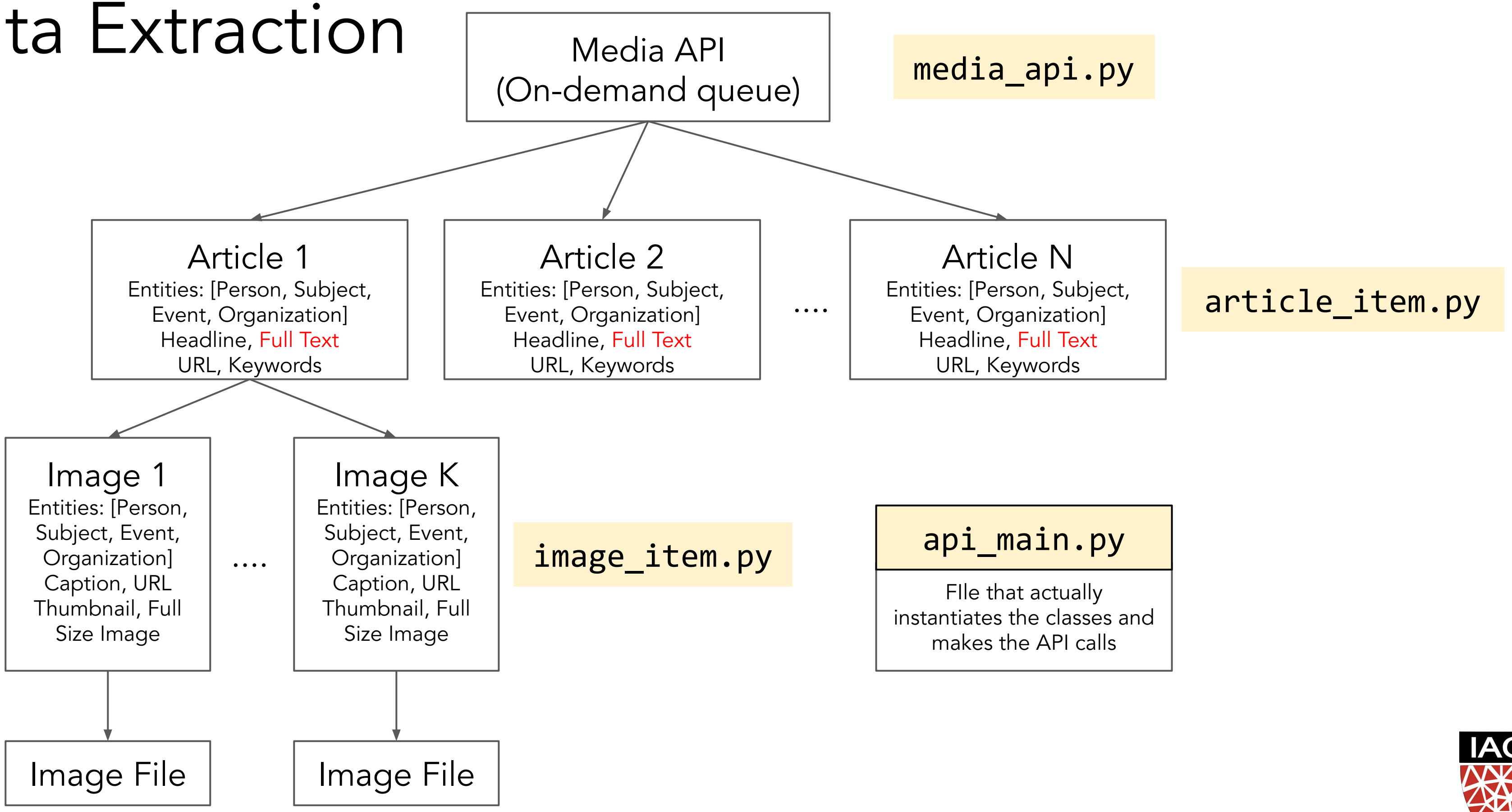


```
0
[0.575 1.425 1.425 0.575]
1
[0.755625 1.244375 1.244375 0.755625]
2
[0.67885937 1.32114062 1.32114062 0.67885937]
3
[0.71148477 1.28851523 1.28851523 0.71148477]
```

Literature Review - TextRank

- Converges usually for 20-30 iterations, at a threshold of .0001
- N-gram (for example, combine neighboring keywords to be n-gram keywords)
  - “Matlab code for plotting ambiguity functions” with keywords {‘Matlab’, ‘code’}
  - Then combine ‘Matlab code’ to be one keyword
- Extract T keywords from final ranked keywords (5-20 or ⅓ * size of text)

Data Extraction



Media API
(On-demand queue)

media_api.py

Article 1
Entities: [Person, Subject, Event, Organization]
Headline, Full Text
URL, Keywords

Article 2
Entities: [Person, Subject, Event, Organization]
Headline, Full Text
URL, Keywords

....

Article N
Entities: [Person, Subject, Event, Organization]
Headline, Full Text
URL, Keywords

article_item.py

Image 1
Entities: [Person, Subject, Event, Organization]
Caption, URL
Thumbnail, Full Size Image

....

Image K
Entities: [Person, Subject, Event, Organization]
Caption, URL
Thumbnail, Full Size Image

image_item.py

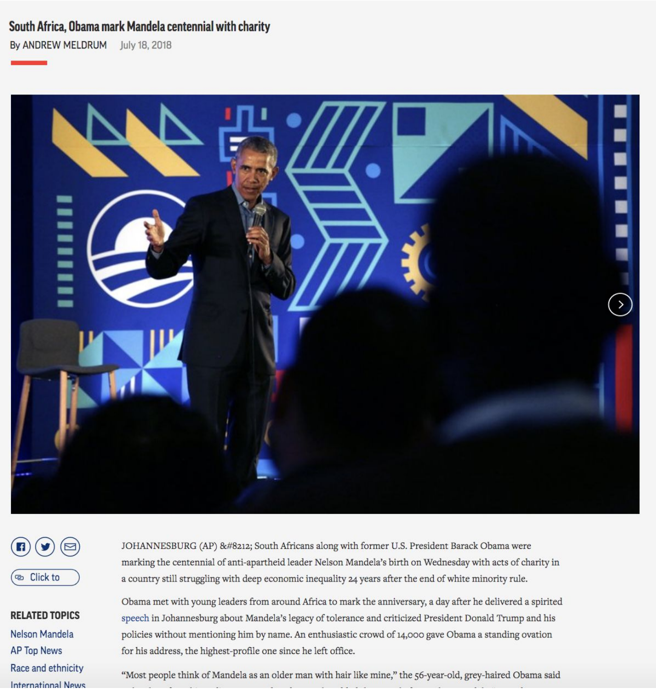
Image File

Image File

api_main.py

FIle that actually instantiates the classes and makes the API calls

EDA: Example



South Africa, Obama mark Mandela centennial with charity

By ANDREW MELDRUM July 18, 2018

JOHANNESBURG (AP) &#8212; South Africans along with former U.S. President Barack Obama were marking the centennial of anti-apartheid leader Nelson Mandela's birth on Wednesday with acts of charity in a country still struggling with deep economic inequality 24 years after the end of white minority rule.

Obama met with young leaders from around Africa to mark the anniversary, a day after he delivered a spirited speech in Johannesburg about Mandela's legacy of tolerance and criticized President Donald Trump and his policies without mentioning him by name. An enthusiastic crowd of 14,000 gave Obama a standing ovation for his address, the highest-profile one since he left office.

"Most people think of Mandela as an older man with hair like mine," the 56-year-old, grey-haired Obama said

Article

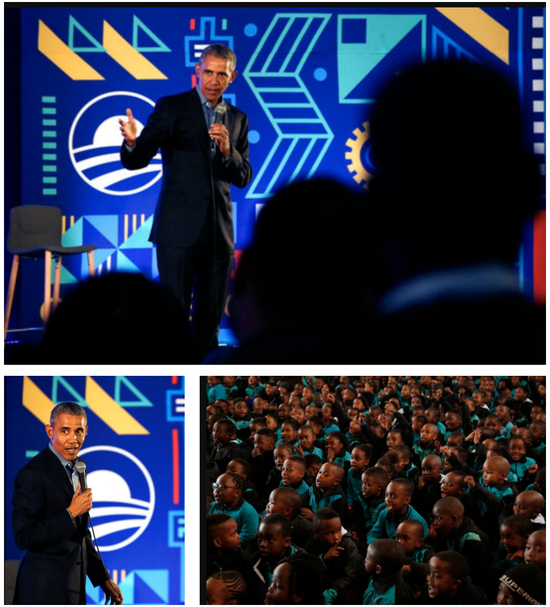
Person
Barack Obama, Antonio Gutierres, Beyonce Knowles, Nelson Mandela

Subject
General News, Government and Politics, Race and Ethnicity, Social Issues, African Americans

Organisation
South Africa Government

Place
United States, South Africa, Southern Africa

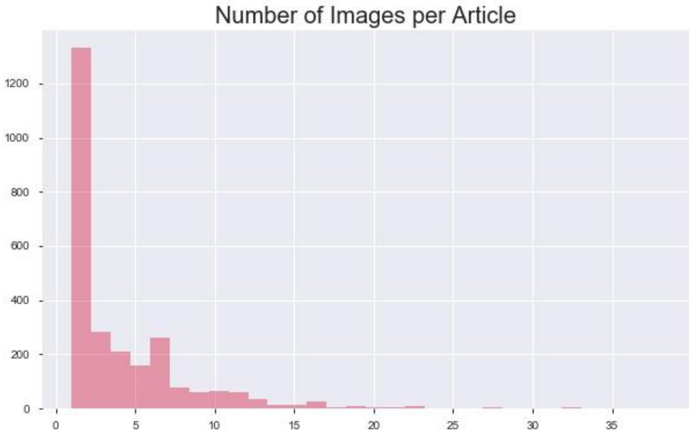
Tags



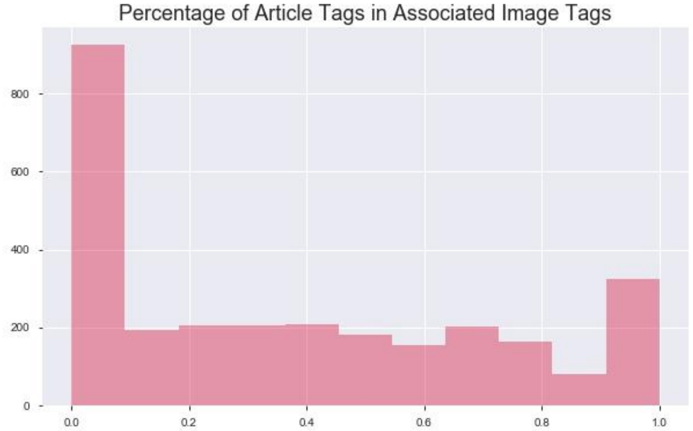
Associations

EDA with Metadata

- Current Available Data
  - 2,841 articles
  - 10,819 images
  - 199 articles without images
- Maximum # images per article = 38
- 1,793 unique subject tags
  - Editorial AP Category Tags
    - * remove for analysis
  - Tagging Service API Tags
- Next Steps:
  - EDA process set-up complete
  - 30, 000+ incoming data points



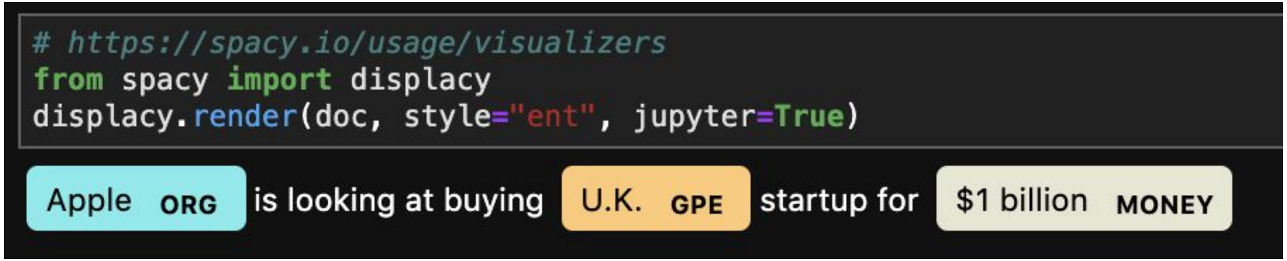
Number of Images per Article



Percentage of Article Tags in Associated Image Tags

Named Entity Extraction

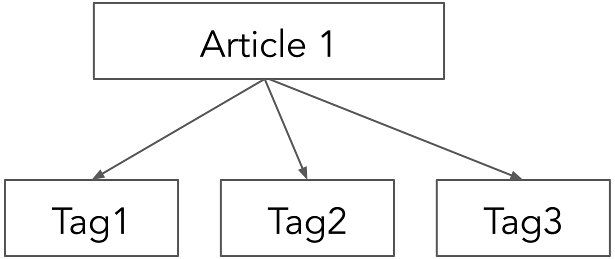
- AP has an existing rule-based named entity extraction of text
  - Might have limitations on how many categories/entities to be identified
- Once we have the full article text data, we plan to use spaCy to extract the named entities from the text
  - 18 predefined categories
- On the UI-side, we can potentially add a nice visualization with the highlights on named entities
- For example, “Apple is looking at buying U.K. startup for $1 billion”



```
# https://spacy.io/usage/visualizers
from spacy import displacy
displacy.render(doc, style="ent", jupyter=True)
```

```
Apple ORG is looking at buying U.K. GPE startup for $1 billion MONEY
```

Baseline Model Progress



Article 1

Tag1 Tag2 Tag3

INPUT

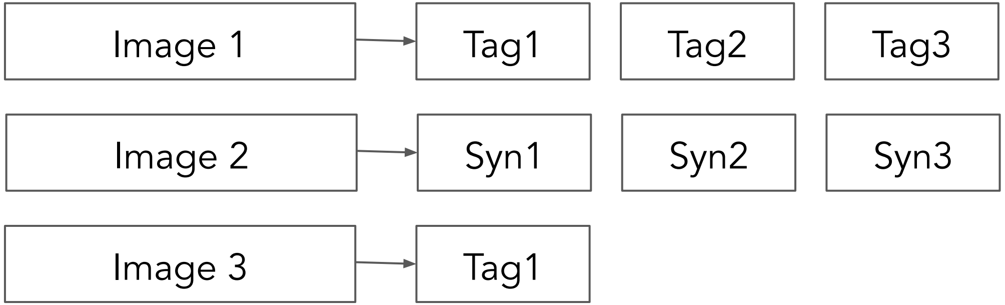


Image 1 → Tag1 Tag2 Tag3

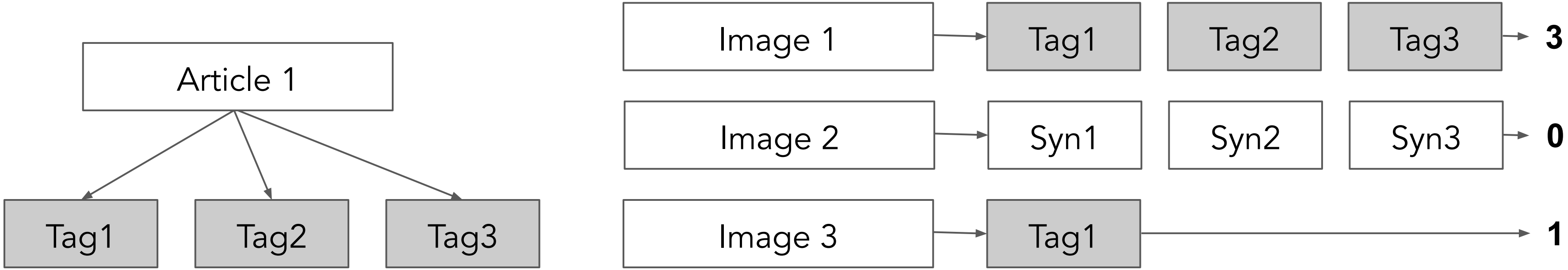
Image 2 → Syn1 Syn2 Syn3

Image 3 → Tag1

IMAGE CORPUS

```
models/baseline_model.py
```

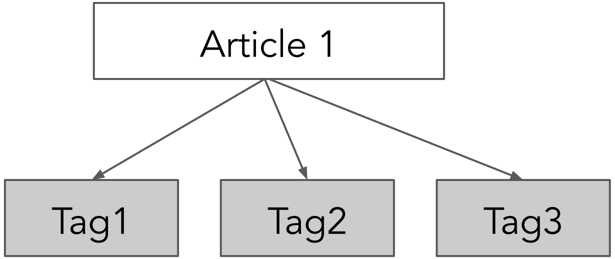
Baseline Model Progress



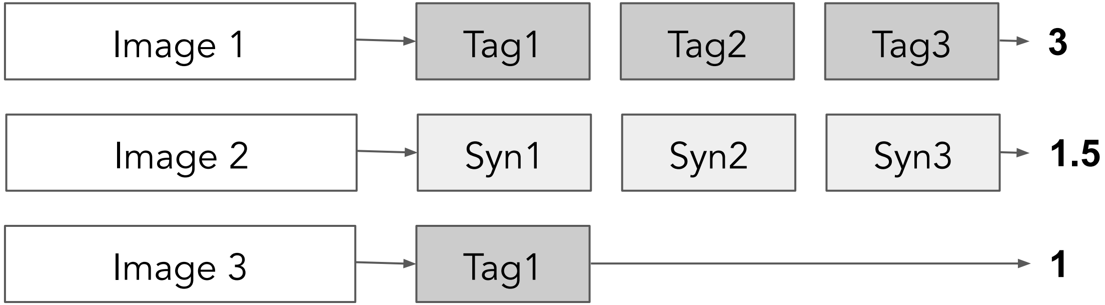
Similarity Metric 1: Exact Match

```
models/baseline_model.py
```

Baseline Model Progress



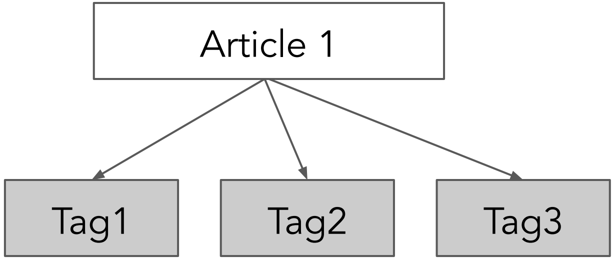
Similarity Metric 1: Exact Match

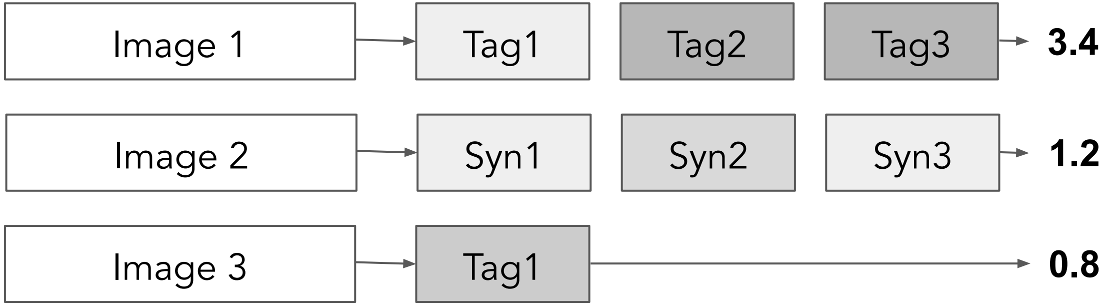


Similarity Metric 2: Epsilon Synonym Match

```
models/baseline_model.py
```

Baseline Model Progress





Similarity Metric 1: Exact Match

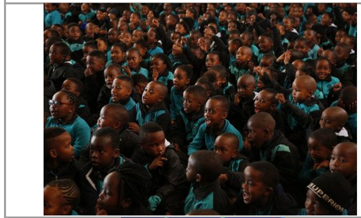
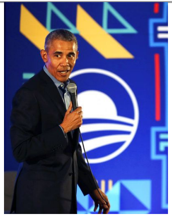
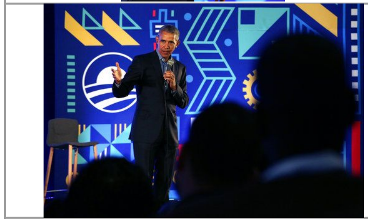
Similarity Metric 2: Epsilon Synonym Match

Similarity Metric 3: TF-IDF Weighted Match

```
models/baseline_model.py
```

Baseline Model Progress

Article Tags: ['General news', 'Government and politics', 'African-Americans', 'Race and ethnicity', 'Social issues', 'Social affairs', 'Racial and ethnic discrimination', 'Discrimination', 'Human rights and civil liberties']

IMAGE	METRIC 1	METRIC 2	METRIC 3
	42	167.5	5.656
	18	32.5	2.923
	42	167.5	5.598

Next Steps

Oct 1 - Oct 8

- Feature exploration: similarity comparison between available text features
- Set evaluation criteria: Human (MTurk) vs image similarity
- Baseline model: tag-to-tag matching

Oct 8 - Oct 15:

- Model iteration
- UI first version

Oct 15 - Oct 22:

- Model iteration
- UI improvement
- Review progress and plan next steps with AP