

# AP



## Midterm Presentation

Karina Huang, Dianne Lee, Abhimanyu Vasishth, Phoebe Wong



# About the Partner

- AP is a cooperative, not-for-profit news agency that serves news in over 100 countries
- Cooperative: member news organizations grant permission for the AP to distribute their local news reports
- Non-member newspapers subscribe to AP and pay a fee to AP to get their material published
- Independent editors and publishers also pay for licensed AP content (such as images, videos and audio files)

**AP**



# Problem Statement

- One of the main sources of revenue for AP is the sale of image licenses to publishers who use these images in their articles
- 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 could be used on the AP web portals or in an API for external clients

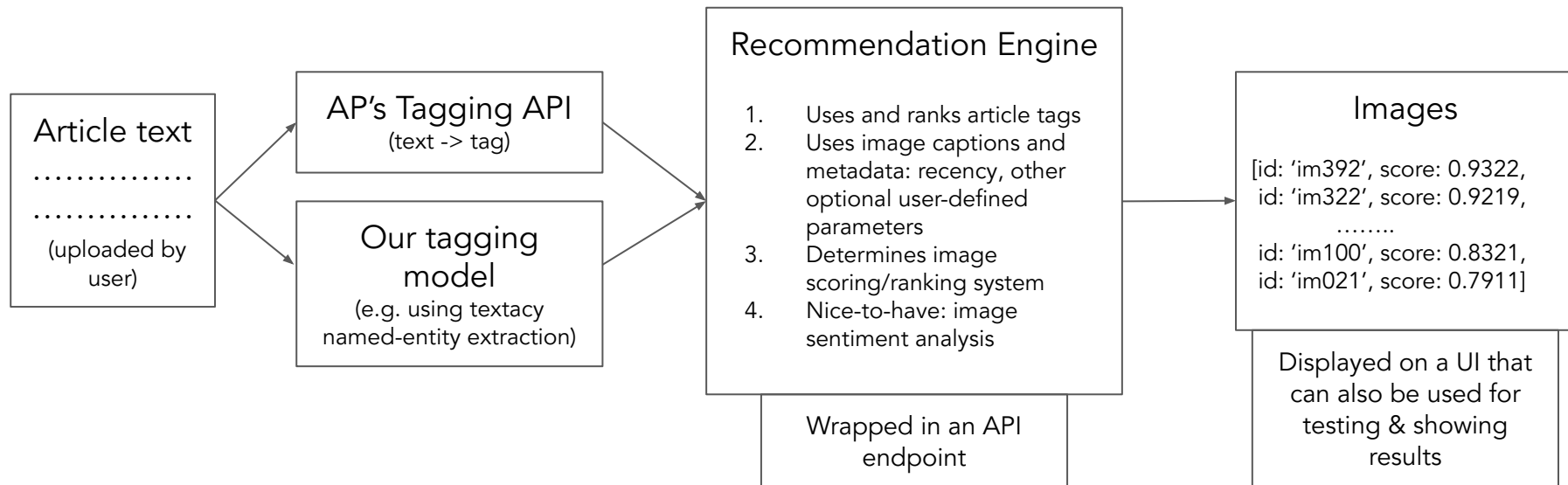


# Motivation

- AP Perspective: Efficient, automatic interface for internal and external clients to generate quicker & more fitting images
- NLP Perspective: Model the way a client would think about this problem through creative combination of NLP techniques, e.g. entity extraction to sentiment analysis, with a flavor of computer vision
- Challenges
  - What defines the “best” image recommendation?
- Potential Issues
  - Don't have access to entire image corpus, only those associated with existing articles
  - Exploring article-image relationships other than tag-to-tag matching, e.g. NN approaches



# Project Plan: Scope of Work



# 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

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

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

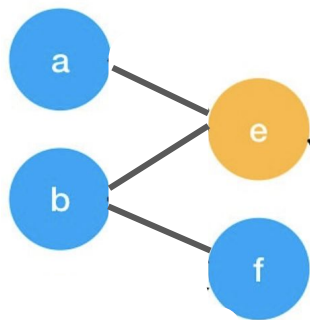
# Literature Review - TextRank

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
- 2. Ranking keywords
- Each keyword is a node and weight decided by the weights of its associated keywords
- Example: for an article,
  - Keyword list:  $[w_1, w_2, \dots, w_n]$
  - Select a  $k$  window of keywords:  $[w_1, w_2, \dots, w_k]$  ( $k: \{2-10\}$ )
    - Any 2-word co-occurrence pair = 1 undirected edge ( $w_1w_2, w_1w_3, \dots, w_1w_k, w_2w_3\dots$ )

# 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

## Parameter Suggestion from Literature

- 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
- Uses T keywords from final ranked keywords (5-20 or  $\frac{1}{3} * \text{size of text}$ )
- Can potentially combine with named entities recognition
  - e.g., only consider scores of named entities

# Literature Review - Next Steps

- Relevance Ranking
  - DeepRank<sup>1</sup>
  - Tag Importance Ranking + Multimodal Image Retrieval<sup>2</sup>
- Sentiment Analysis
  - Sentiment analysis for news<sup>3</sup>
  - Image sentiment analysis with deep networks<sup>4</sup>

## NOT REAL NEWS: A look at what didn't happen this week

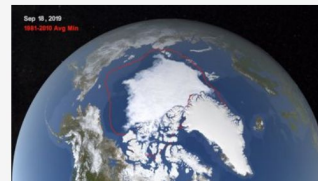
By BEATRICE DUPUY and ARIJETA LAJKA    October 11, 2019



A roundup of some of the most popular but completely untrue stories and visuals of the week. None of these is legit, even though they were shared widely on social media. The Associated Press...

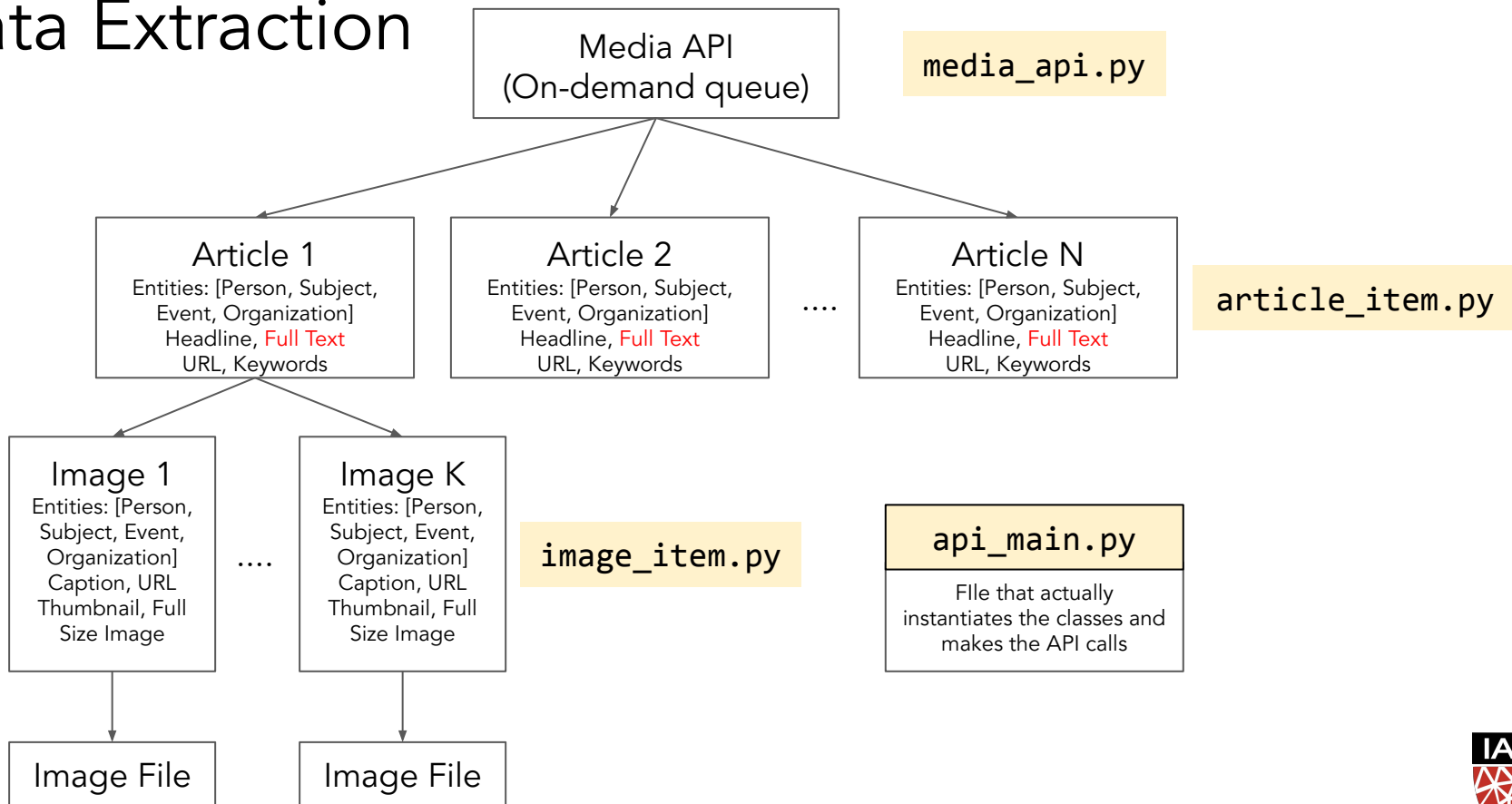
## NOT REAL NEWS: A look at what didn't happen this week

By AMANDA SEITZ, BEATRICE DUPUY and ARIJETA LAJKA    October 4, 2019



A roundup of some of the most popular but completely untrue stories and visuals of the week. None of these is legit, even though they were shared widely on social media. The Associated Press...

# Data Extraction



# EDA: Example

South Africa, Obama mark Mandela centennial with charity

By ANDREW MELDRUM July 18, 2018



Click to

## RELATED TOPICS

Nelson Mandela

AP Top News

Race and ethnicity

International News

JOHANNESBURG (AP) — 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

**Person**  
Barack Obama, Antonio  
Gutierrez, 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



Article

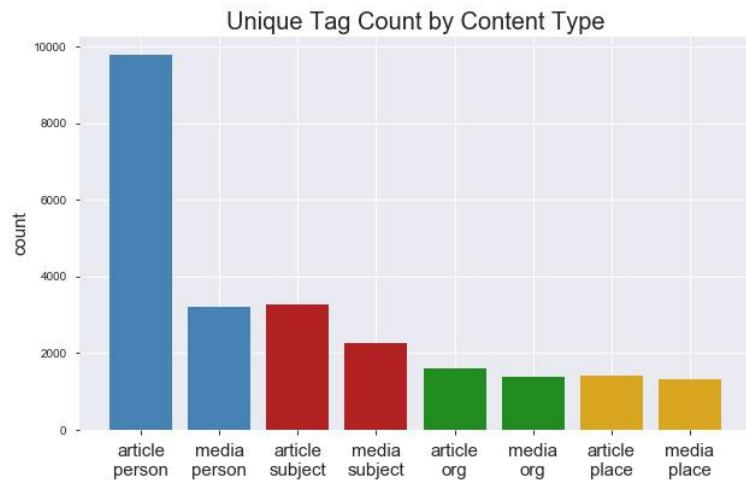
Tags

Associations

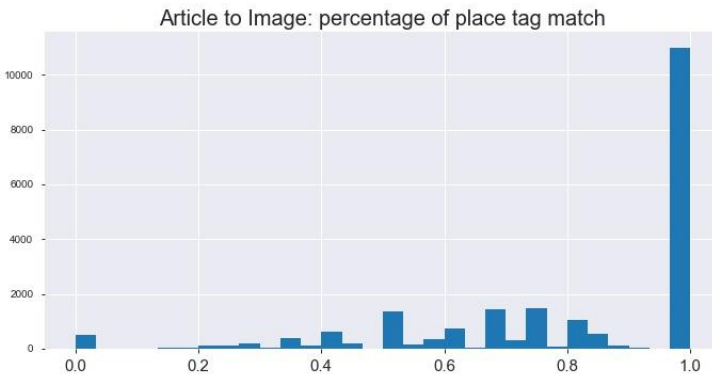
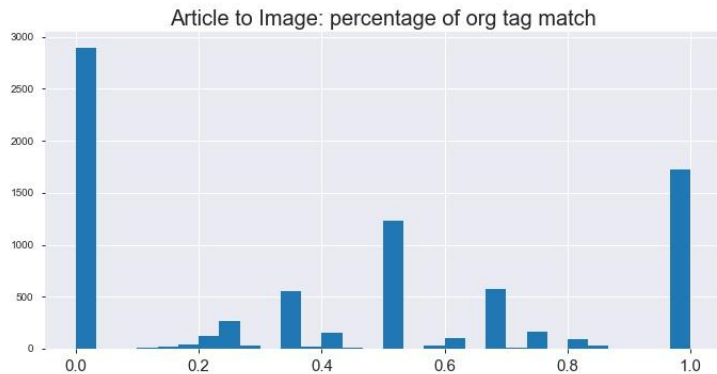
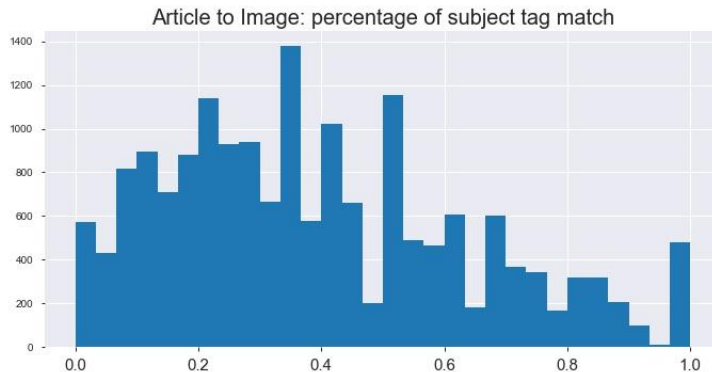
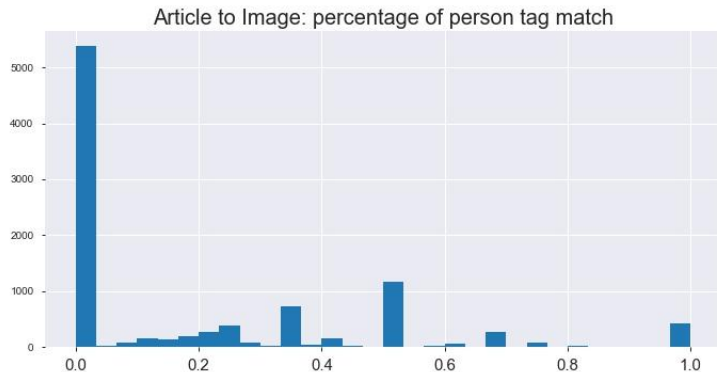


# EDA: Metadata Summary

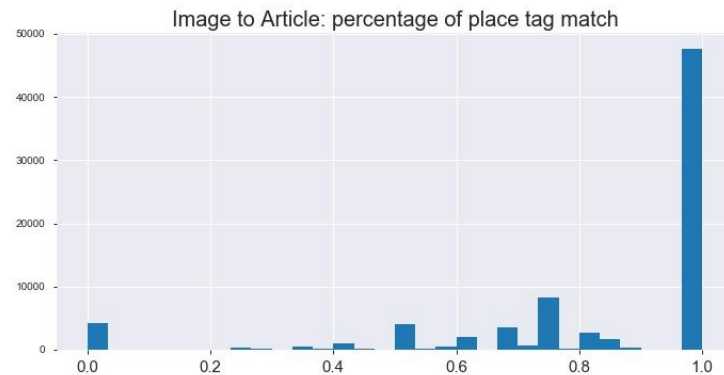
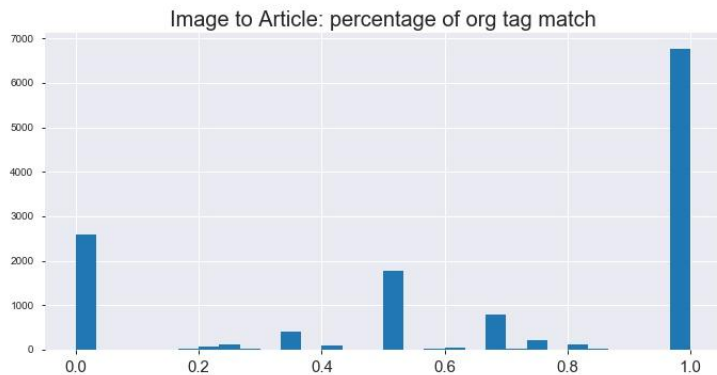
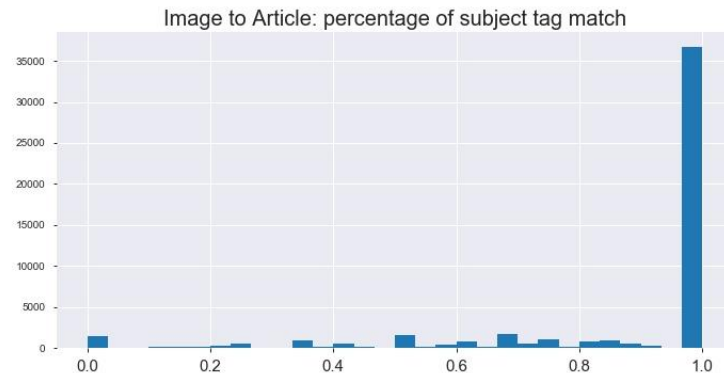
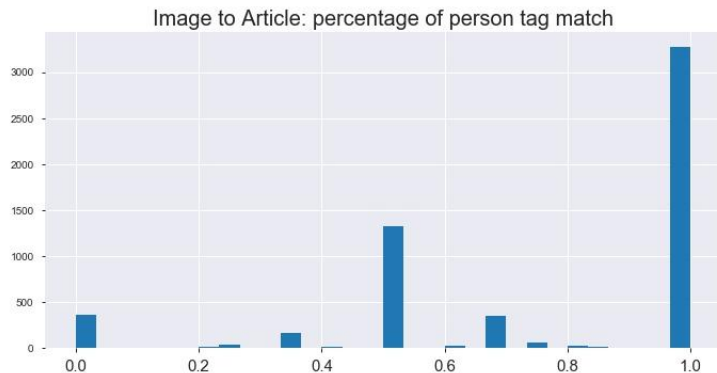
- Available Data
  - 30,098 articles
    - Created from Feb-Sep 2019
    - 99.9% in English
  - 81,126 media files
    - 96.8% images
    - Maximum number of media file per article: 67
  - 7,348 articles without media files



# EDA: Article to Image Tag Matching

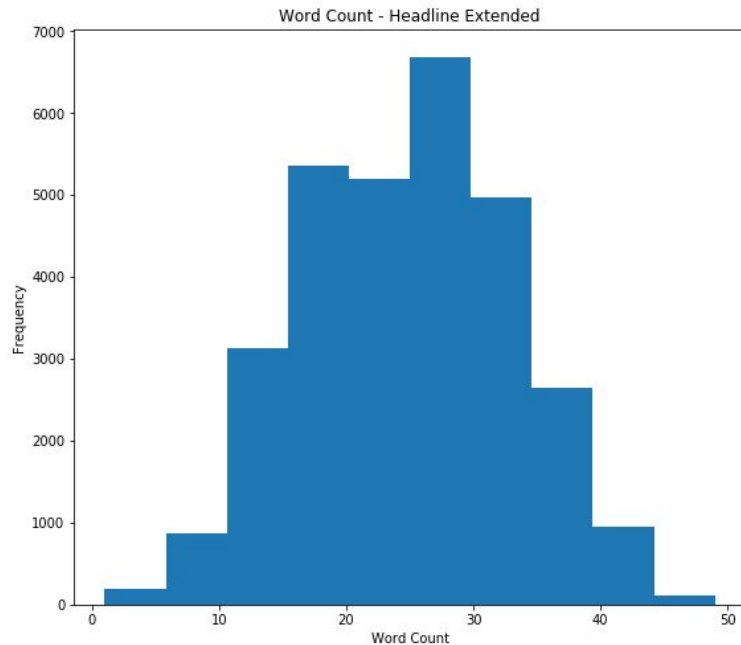
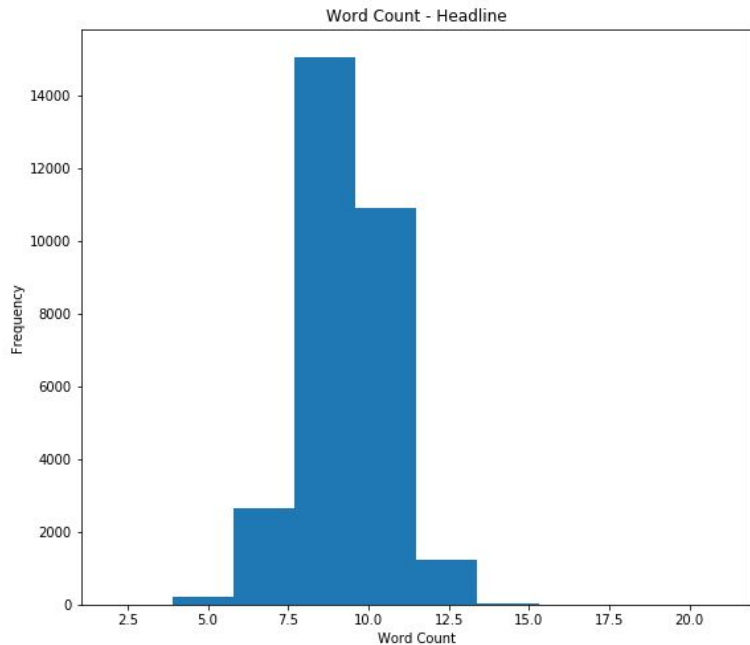


# EDA: Image to Article Tag Matching



# EDA: Headlines

- Word count by headlines and extended headlines
- Average: 9 words (headline) and 24 words (extended headline)





# 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
- We use spaCy to extract the named entities from the text
  - 18 predefined categories (geographic locations, person and etc)
- On the UI-side, we can 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

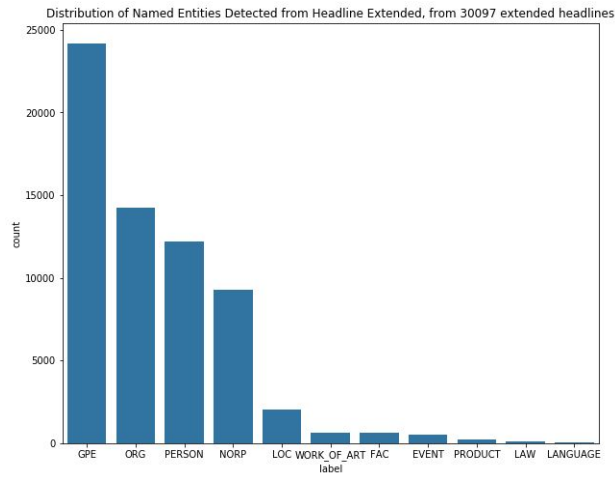
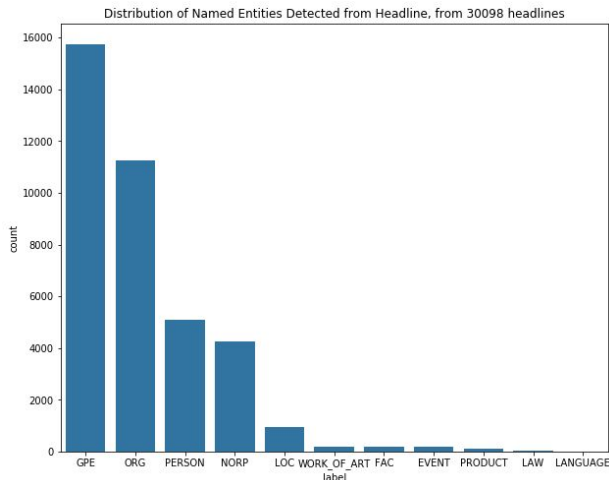
U.K.    **GPE**

startup for

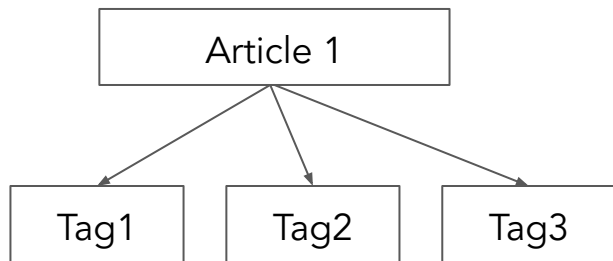
\$1 billion    **MONEY**

# Named Entity Extraction - EDA

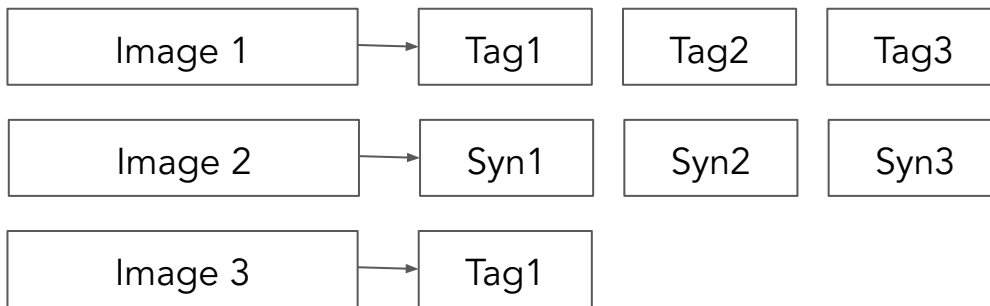
- Using article data provided by AP, we extracted named entities from "headline" and "headline\_extended"
- Length: Headline < headline\_extended < description summary < full text
- Most entities extracted are geographic locations
- Note that "Person" are mentioned more in headline extended



# Tag-To-Tag Matching



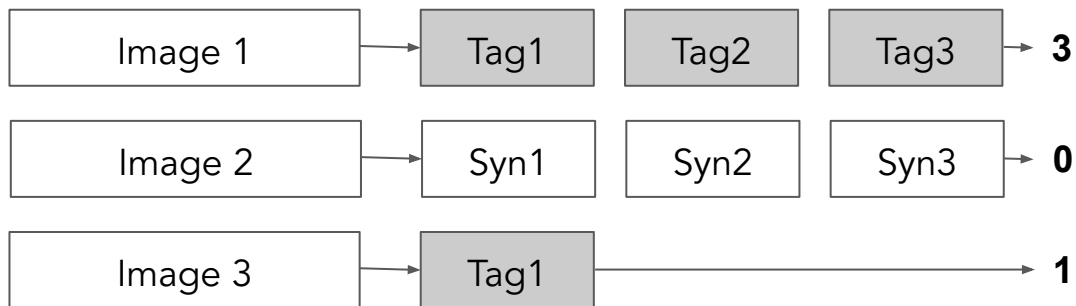
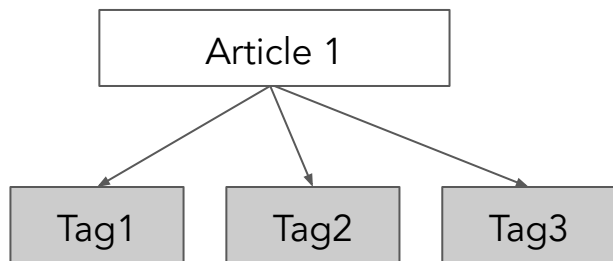
**INPUT**



**IMAGE CORPUS**

`models/baseline_model.py`

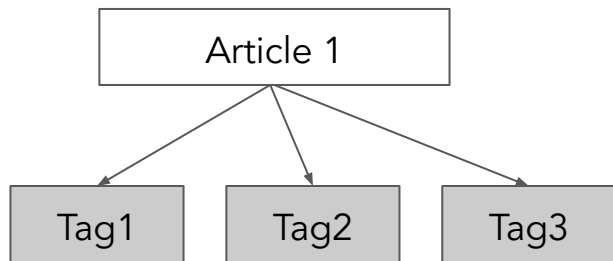
# Tag-To-Tag Matching



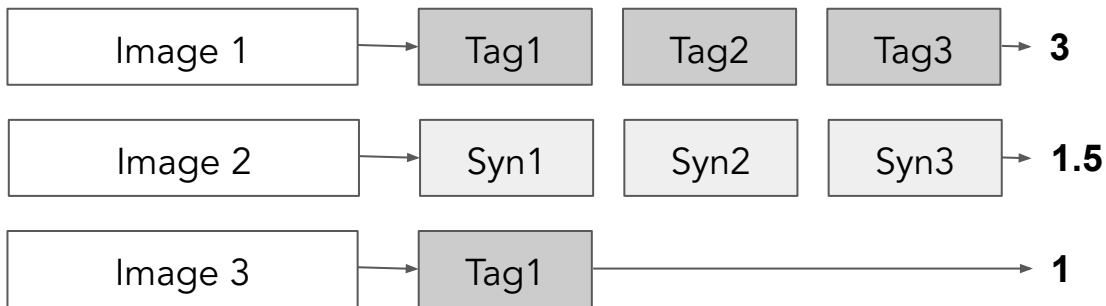
## 1. Exact Matching

`models/baseline_model.py`

# Tag-To-Tag Matching



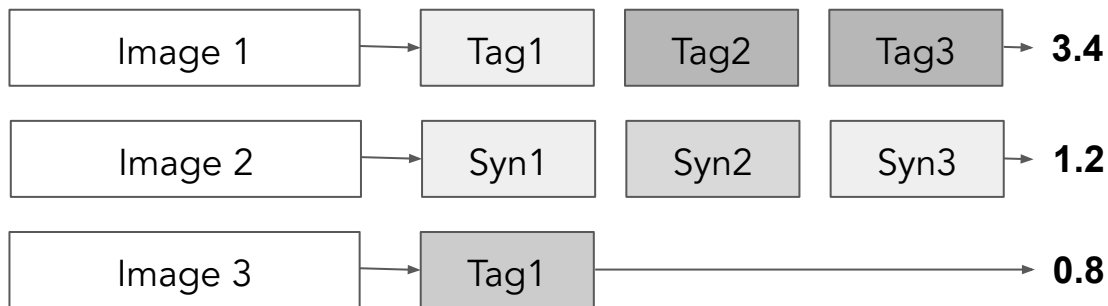
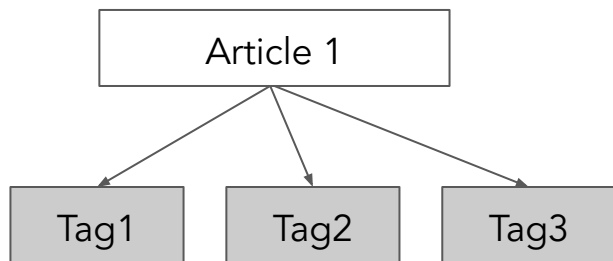
## 1. Exact Matching



## 2. Fuzzy Matching

`models/baseline_model.py`

# Tag-To-Tag Matching



1. Exact Matching




2. Fuzzy Matching

3. TF-IDF Weighted Match

`models/baseline_model.py`

# Tag-To-Tag Matching

Article Summary → Image Summary (Caption)

IMAGE	METRIC 1	METRIC 2	METRIC 3
	1	1	1
	3	3	3
	1	1	2

# Evaluation

- Challenges: Only have access to set article-image associations
- Rudimentary Numeric Method
  - Percent of images correctly predicted by top-ten ranking
  - Issues: Cannot rank predictions outside given article-image associations



- Need more qualitative method of evaluation



# UI First Version

## Input

Celtics convert rookie Tacko Fall to two-way contract

Search

The Boston Celtics have converted 7-foot-6 rookie Tacko Fall to a two-way contract, which will allow him to spend 45 days in the NBA this season. Fall has already gathered a significant fan following ε

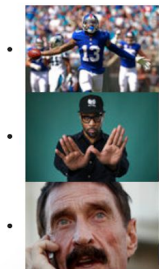
## Tags

- Sports: 0.99
- Basketball: 0.91
- Boston Celtics: 0.2

## Matches

- 0b5caa00d2a34db8a7d7c4bc30e6081b: 0.95
- 0b40eeb8cff64ac3a3fae568a748dd04: 0.81
- 0b328f5537d14be4bbe800ced89b5eec: 0.55

## Image Thumbnails



# Next Steps

Oct 15 - Oct 22:

- Model iteration
- UI improvement

Oct 22 - Oct 29:

- Model selection
- UI improvement
- Work with tagging API

Oct 29 - Nov 5:

- Model training
- Build an API wrapper for our models

Nov 5 - Nov 12:

- Model training
- Improve API wrapper

Nov 12 - Nov 19:

- Model training
- Iterate UI and API wrapper

Nov 19 - Nov 26:

- Training complete
- Work with AP on next steps

Nov 26 - Dec 3:

- Get code base ready for shipping

Dec 3 - Dec 12:

- Finalize deliverables



Thank You. Questions?