News Diggers

Sunday 02/12/2021 The NYU Times Edition: 002

Members:
Nick Wu, Amy Zhu, Noah Kurrack, Zixiao Chen, Hui Wen Goh

Our Team

Sunday 02/12/2021 The NYU Times Edition: 002

Nick Wu

Amy Zhu

Hui Wen Goh

Noah Kurrack

Zixiao Chen

Table of Contents

Sunday 02/12/2021 The NYU Times Edition: 002

01 Introduction

02 Implementation

03 Visualization

04 Analysis

05 Future Research

06 Wrap Up

Introduction: A Scenario

Sunday 02/12/2021 The NYU Times Edition: 002

- Samantha is an NYU Professor of Anthropology. Recently, she has been trying to investigate how advertisements trends change over time for her next research paper.
- She has access to a large database of newspaper articles, including advertisements, but how can she process all of this data automatically and accurately to show meaningful trends?

FINANCIAL
Media futures
Tension entre Damas et Ankara
Le Marais poitevin vidé de se
ZEIT
ZEITUNG FÜR POLITIK · WIRTSCHAFT · HANDEL UND KULTUR
Frankfur
Tribune
YORK TIMES AND THE WASHINGTON POST
Friday, June 21, 1996
n Dismisses Hard-I
Russie : un
La T
QUOTIDIEN ■ MARDI 18 JUIN 1996

Introduction: The Goal

Sunday 02/12/2021 The NYU Times Edition: 002

- **Assisting in handling first hand historical information**
  - Our project aims to assist historians in handling first hand historical resources more efficiently
  - We want to process raw text or image data into more comprehensive datasets (e.g. From XML to readable CSV files, removing errors generated from OCR)

- **Extracting entities through entity recognition**
  - Extracting key advertisements information such as product name, company name and company location

- **Model building and generalization**
  - Building a base model to extract valuable information out of training set.
  - Generalizing the model and applying it to other text datasets.

Introduction: The Data and Variables

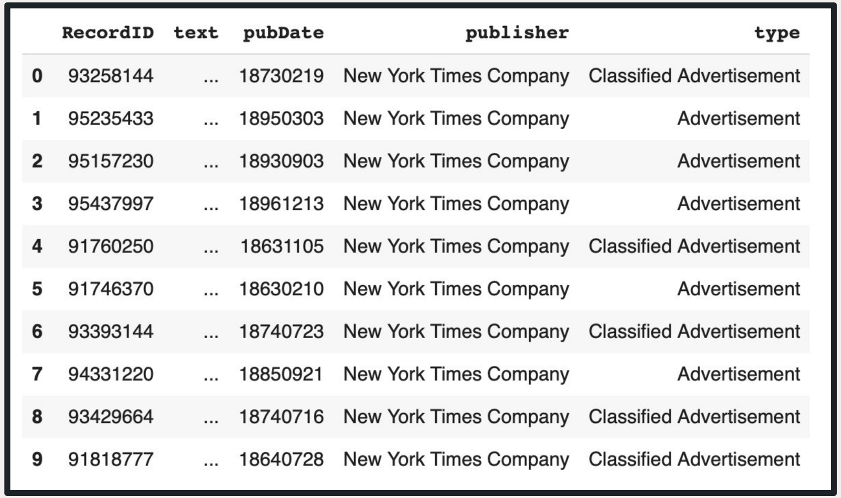
Sunday 02/12/2021 The NYU Times Edition: 002

The Data Given:

- 1.8+ million New York Times newspaper pages in the form of ZIPs to XML files
- 200,000+ Atlanta Daily World newspaper pages in the form of ZIPs to XML files

Important Variables in Each File:

- RecordID
- NumericPubDate
- FullText
- ObjectType



	RecordID	text	pubDate	publisher	type
0	93258144	...	18730219	New York Times Company	Classified Advertisement
1	95235433	...	18950303	New York Times Company	Advertisement
2	95157230	...	18930903	New York Times Company	Advertisement
3	95437997	...	18961213	New York Times Company	Advertisement
4	91760250	...	18631105	New York Times Company	Classified Advertisement
5	91746370	...	18630210	New York Times Company	Advertisement
6	93393144	...	18740723	New York Times Company	Classified Advertisement
7	94331220	...	18850921	New York Times Company	Advertisement
8	93429664	...	18740716	New York Times Company	Classified Advertisement
9	91818777	...	18640728	New York Times Company	Classified Advertisement

Implementation: Parsing & Exporting

Sunday 02/12/2021 The NYU Times Edition: 002

Unzipping

First, we unzipped the data by iterating over the nested zip files in the dataset folders to reach individual XML files.

- Ensured that only files that contained an advertisement would be opened

Parsing

Then, we used BeautifulSoup to parse through the XML advertisement files to find key variables:

- FullText
- RecordID
- NumericPubDate
- ObjectType

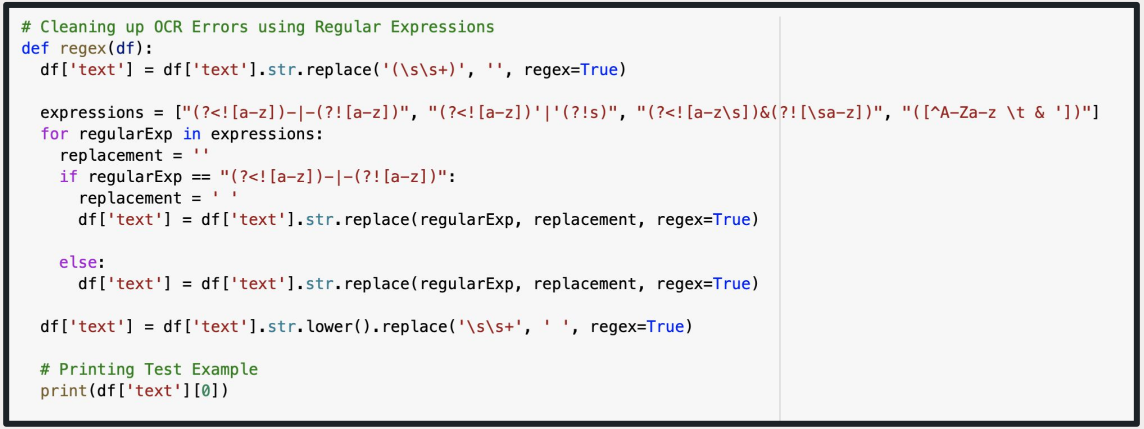
XML -> CSV

We exported the variables we found to a .csv file named AdData

Implementation: Cleaning

To clean the data and simplify the text, we used regex to:

1. Remove all punctuation except for hyphens between two words and apostrophes that signal possession
2. Make all text lowercase
3. Remove all numbers

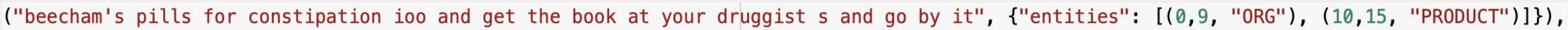


```
# Cleaning up OCR Errors using Regular Expressions
def regex(df):
  df['text'] = df['text'].str.replace('(\s\s+)', '', regex=True)

  expressions = ["(?<![a-z])-|-(?![a-z])", "(?<![a-z])'|'(?!s)", "(?<![a-z\s])&(?![\sa-z])", "([^A-Za-z \t & '])"]
  for regularExp in expressions:
    replacement = ''
    if regularExp == "(?<![a-z])-|-(?![a-z])":
      replacement = ' '
      df['text'] = df['text'].str.replace(regularExp, replacement, regex=True)

    else:
      df['text'] = df['text'].str.replace(regularExp, replacement, regex=True)

  df['text'] = df['text'].str.lower().replace('\s\s+', ' ', regex=True)

  # Printing Test Example
  print(df['text'][0])
```

Implementation: Training & NER

Creating Training Dataset

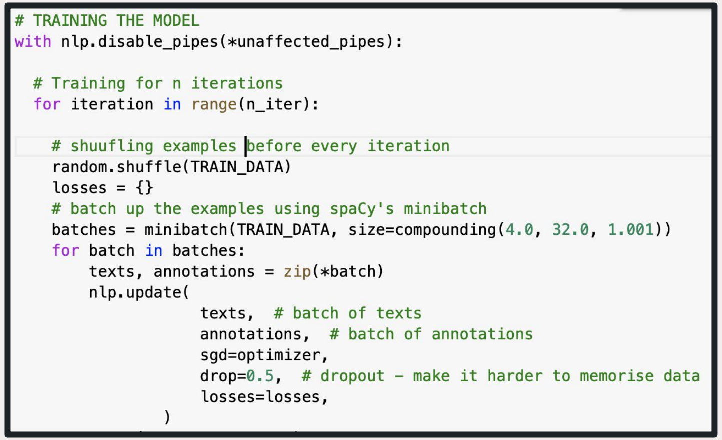
- Since our target variable, entity (such as product, company, location), is not directly accessible in the cleaned CSV file, we manually extracted entities and their position in a sentence to create a training dataset.
- E.g:



```
("beecham's pills for constipation ioo and get the book at your druggist s and go by it", {"entities": [(0,9, "ORG"), (10,15, "PRODUCT")]}),
```

Training spaCy model to identify the following key entities:

1. "ORG" - the organization name
2. "PRODUCT" - the product being advertised
3. "GPE" - the location/address being provided

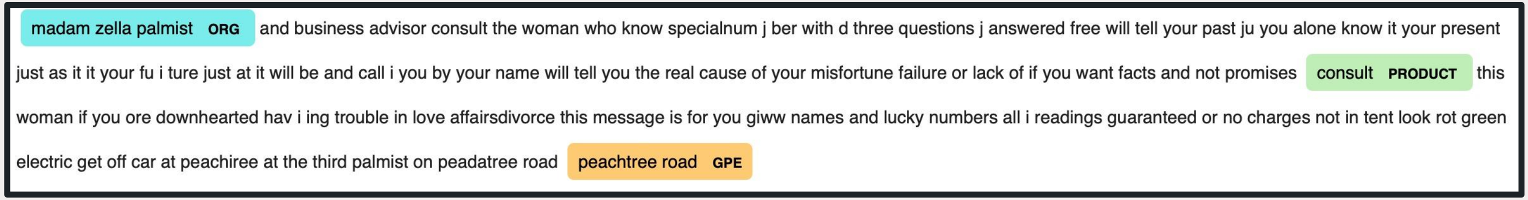


```
# TRAINING THE MODEL
with nlp.disable_pipes(*unaffected_pipes):

  # Training for n iterations
  for iteration in range(n_iter):

    # shuufling examples before every iteration
    random.shuffle(TRAIN_DATA)
    losses = {}
    # batch up the examples using spaCy's minibatch
    batches = minibatch(TRAIN_DATA, size=compounding(4.0, 32.0, 1.001))
    for batch in batches:
        texts, annotations = zip(*batch)
        nlp.update(
                    texts,  # batch of texts
                    annotations,  # batch of annotations
                    sgd=optimizer,
                    drop=0.5,  # dropout - make it harder to memorise data
                    losses=losses,
                )
```

Visualization

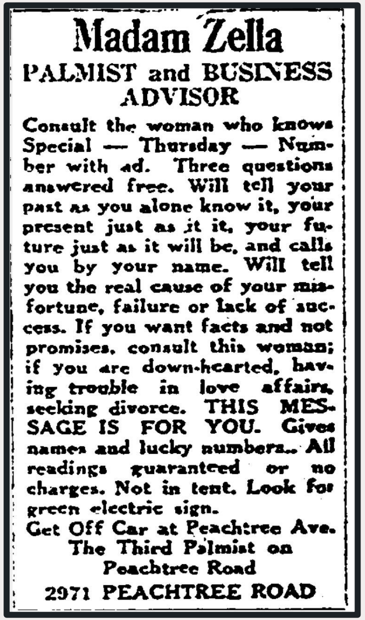
Sunday 02/12/2021 The NYU Times Edition: 002

spaCy Visualizations:



madam zella palmist ORG and business advisor consult the woman who know specialnum j ber with d three questions j answered free will tell your past ju you alone know it your present just as it it your fu i ture just at it will be and call i you by your name will tell you the real cause of your misfortune failure or lack of if you want facts and not promises consult PRODUCT this woman if you ore downhearted hav i ing trouble in love affairsdivorce this message is for you giww names and lucky numbers all i readings guaranteed or no charges not in tent look rot green electric get off car at peachiree at the third palmist on peadatree road peachtree road GPE

*Entity recognition based on example ad --->



Madam Zella

PALMIST and BUSINESS ADVISOR

Consult the woman who knows Special — Thursday — Number with ad. Three questions answered free. Will tell your past as you alone know it, your present just as it it, your future just as it will be, and calls you by your name. Will tell you the real cause of your misfortune, failure or lack of success. If you want facts and not promises, consult this woman; if you are down-hearted, having trouble in love affairs, seeking divorce. THIS MESSAGE IS FOR YOU. Gives names and lucky numbers.. All readings guaranteed or no charges. Not in tent. Look for green electric sign.

Get Off Car at Peachtree Ave.
The Third Palmist on
Peachtree Road
2971 PEACHTREE ROAD

Madam Zella Palmist - ORG
Consult- PRODUCT
Peachtree Road- GPE/Location

Implementation: Accuracy Scoring

Precision: 0.364

- Ratio of correctly predicted positive observations to the total predicted positive observations

Recall: 0.097

- Ratio of correctly predicted positive observations to all observations in the actual class (sensitivity)

F1-score: 0.154

- Weighted average of Precision and Recall

Analysis

Project Considerations

- Organizations were the hardest to recognize, as most of the companies advertised were local businesses with people’s names as the name of the company
- Project may be more accurate for newspaper articles since OCR quality is better

Limitations in Predictions

- Quality of OCR extraction or clean-up can greatly improve entity recognition
- Having more training data with labeled outputs would greatly increase the predictive ability of the custom model

Future Research

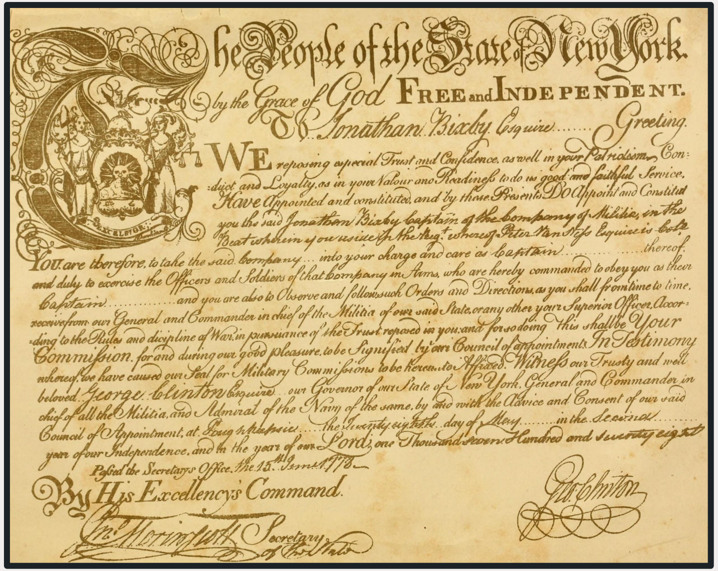
Sunday 02/12/2021 The NYU Times Edition: 002

- Apply the OCR and entity recognition technique to other historical first-hand resources
- Assist historian and museums to automatically process large amount of text data in a more efficient way
- Digitizing historical data to make more accessible and easier to be searched

To expand this project in the future, it would be helpful to compile more training data to make the model more accurate in detecting named entities.

Wrap-up

- Using our model, Samantha and other historians/anthropologists now have a way to receive quick insights derived from millions of historical data, including intuitive visualizations for easy analysis



The People of the State of New York.
by the Grace of God FREE and INDEPENDENT.
To Jonathan Bixby Esquire .... Greeting.
WE reposing especial Trust and Confidence, as well in your Patriotism, Con:
:duct and Loyalty, as in your Valour and Readineſs to do us good and faithful Service,
Have Appointed and constituted, and by these Presents DO Appoint and Constitut
you the said Jonathan Bixby Captain of the Company of Militia, in the
Beat wherein you reside in the Regt. whereof Peter Van Neſs Esquire is Colo.
YOU are therefore, to take the said Company ... into your charge and care as Captain .... thereof.
and duly to exercise the Officers and Soldiers of that Company in Arms, who are hereby commanded to obey you as their
Captain .... and you are also to Observe and follow such Orders and Directions, as you shall from time to time,
receive from our General and Commander in chief of the Militia of our said State, or any other your Superior Officer, Accor:
-ding to the Rules and dicipline of War, in pursuance of the Trust reposed in you: and for so doing this shall be Your
Commiſsion, for and during our good pleasure, to be Signified by our Council of appointment. In Testimony
whereof, we have caused our Seal for Military Commiſsions to be hereunto Affixed. Witneſs our Trusty and well
beloved George Clinton Esquire ... our Governor of our State of New York, General and Commander in
chief of all the Militia, and Admiral of the Navy of the same, by and with the Advice and Consent of our said
Council of Appointment, at Poughkeepsie .... the Twenty eighth day of May .... in the Second ....
year of our Independence, and in the year of our Lord, one Thousand seven Hundred and seventy eight
Paſsed the Secretarys Office, the 15th June 1778
By His Excellency's Command.
Jno. Morin Scott Secretary of the State
Geo. Clinton

Thank You!