Canoe Project

March 13

Aries, Sherry, Taimoor @ Columbia University

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Final Goal

1. Predict the document category (Account Statement, Call Notice, Distribution Notice, Other) given a document

2. Extract data from the document, if the document's category is Account Statement, or Call Notice or Distribution Notice

In this presentation

1. Show how we prepared the data set

2. Explain what kind of models/ techniques we are going to use (or have used) and why

Data Preparation

1. Used PyPDF2 to extract information from PDFs
2. Read all files and added the text to a 2d list containing the text of all files
3. Identified dates, entity and fund name and document type from PDF names
4. Cleaned data and removed whitespace and new lines, tabs etc

Tagging & Chunking

Tagging

1. Stanford CoreNLP tagger
2. NLTK part-of-speech tagger

- Automatic tagging (regular expression tagger, the look up tagger)
- N-gram tagging(2-gram or 3-gram)

Chunking: segments and labels multitoken sequences

1. Noun-Phrase chunking
2. Chunking with regular expressions

Named Entity Recognition

Identifying the boundaries of the NE and its type

1. NLTK
2. Stanford Named Entity Recognizer Model
3. Spacy NER model

NE type	Examples
ORGANIZATION	*Georgia-Pacific Corp., WHO*
PERSON	*Eddy Bonte, President Obama*
LOCATION	*Murray River, Mount Everest*
DATE	*June, 2008-06-29*
TIME	*two fifty a m, 1:30 p.m.*
MONEY	*175 million Canadian Dollars, GBP 10.40*
PERCENT	*twenty pct, 18.75 %*
FACILITY	*Washington Monument, Stonehenge*
GPE	*South East Asia, Midlothian*

Three standard approaches to do NER

1. Hand-written regular expressions
2. Classifiers
- Generative: Naive Bayes
- Discriminative: Maxent models

3. Sequence models
- HMMs(Hidden Markov Model)
- CMMs(Conditional Markov models) /MEMMs(Maximum Entropy Markov models)
- CRFs(conditional random fields)

The machine learning sequence model approach to NER

Training

1. Collect a set of representative training documents
2. Label each token for its entity class or other (O)
3. Design feature extractors appropriate to the text and classes
4. Train a sequence classifier to predict the labels from the data

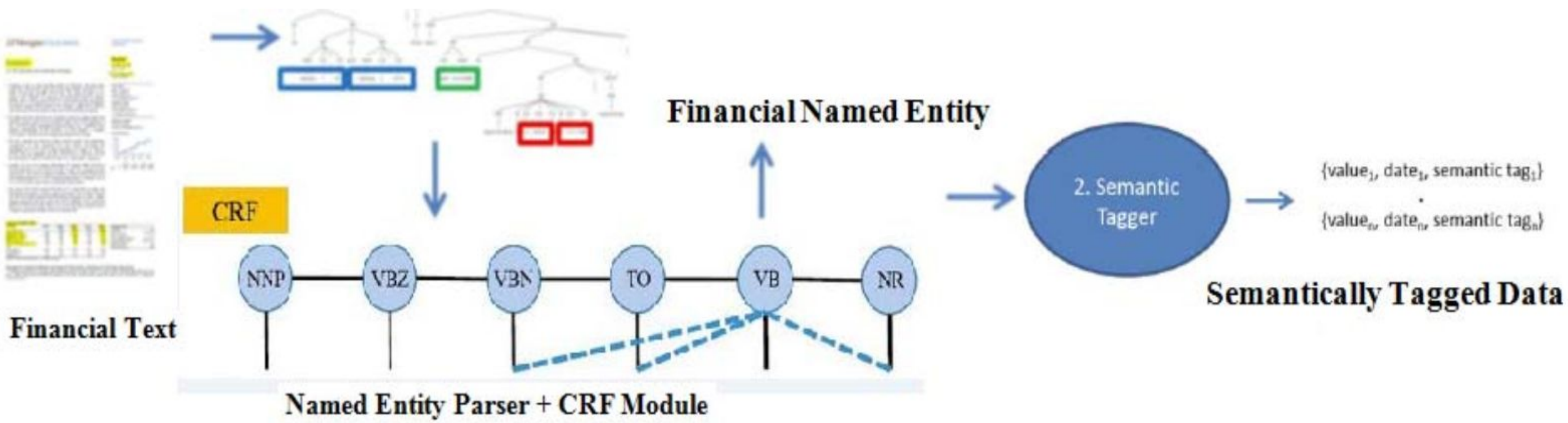
Testing

1. Receive a set of testing documents
2. Run sequence model inference to label each token
3. Appropriately output the recognized entities

Models for NER in financial documents

1. Skip-Gram Model
2. CRF(conditional random fields)

- Training is slower, but CRFs avoid causal-competition biases
- Takes into account the position of the current filed, so do not need to spending much time engineering structural features into our model



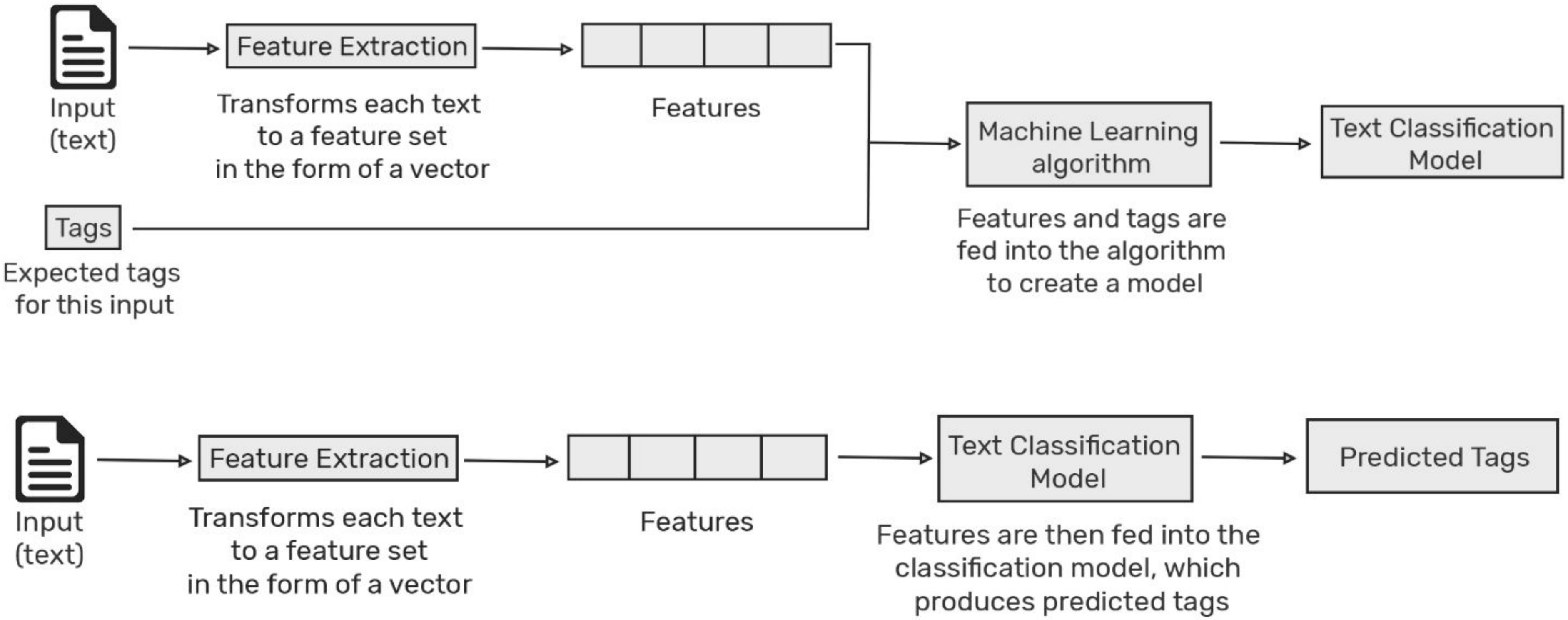
Problems - NER in financial documents:

1. Financial institutions have long complex names
2. Financial institution names appear in an individual line, be free of additional text, so it lacks context, natural language features and structure tags
3. Names can break across several lines
4. Names are often capitalized
5. An institution may be mentioned using different names

Data characteristics:

Financial institution names can typically be split into a root fragment and a suffix

Predict Category - Text Classification



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Reference: https://monkeylearn.com/text-classification/

Text Classification - Feature Extraction

sklearn.feature_extraction.text.CountVectorizer

Convert a collection of text documents to a matrix of token counts.

This implementation produces a sparse representation of the counts using scipy.sparse.csr_matrix.

Examples

```
>>> from sklearn.feature_extraction.text import CountVectorizer
>>> corpus = [
...     'This is the first document.',
...     'This document is the second document.',
...     'And this is the third one.',
...     'Is this the first document?',
... ]
>>> vectorizer = CountVectorizer()
>>> X = vectorizer.fit_transform(corpus)
>>> print(vectorizer.get_feature_names())
['and', 'document', 'first', 'is', 'one', 'second', 'the', 'third', 'this']
>>> print(X.toarray())
[[0 1 1 1 0 0 1 0 1]
 [0 2 0 1 0 1 1 0 1]
 [1 0 0 1 1 0 1 1 1]
 [0 1 1 1 0 0 1 0 1]]
```

Text Classification - Feature Extraction

sklearn.feature_extraction.text.TfidfTransformer

Transform a count matrix to a normalized tf or tf-idf representation (term-frequency times inverse document-frequency)

```
>>> transformer = TfidfTransformer()
>>> transformer.fit_transform(counts).toarray()
array([[0.85151335, 0.        , 0.52433293],
       [1.        , 0.        , 0.        ],
       [1.        , 0.        , 0.        ],
       [1.        , 0.        , 0.        ],
       [0.55422893, 0.83236428, 0.        ],
       [0.63035731, 0.        , 0.77630514]])
```

The weights of each feature computed by the `fit` method call are stored in a model attribute:

```
>>> transformer.idf_
array([1. ..., 2.25..., 1.84...])
```

Text Classification - Algorithms

- Naive Bayes
  - Multinomial Naive Bayes (MNB)
  - a couple of thousand tagged samples
- Support Vector Machines
  - more computational resources than Naive Bayes
- Deep Learning
  - Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN)
  - at least millions of tagged examples

Text Classification - Example

	index	Category Name	Sub Category	File Name	category	file	text
0	195	Capital Activity	Account Statement	Arsenal Capital Group - Wenger Partners - Acc...	1	Arsenal Capital Group - Wenger Partners - Acc...	Chelsea Partners ...
1	199	Capital Activity	Account Statement	Arsenal Capital Group - Wenger Partners - Acc...	1	Arsenal Capital Group - Wenger Partners - Acc...	Liverpool Investments 6/30/2017 To...
2	275	Capital Activity	Account Statement	Arsenal Capital Group - Wenger Partners - Acc...	1	Arsenal Capital Group - Wenger Partners - Acc...	Newcastle Ventu...
3	192	Capital Activity	Account Statement	Arsenal Capital Group - Wenger Partners - Acc...	1	Arsenal Capital Group - Wenger Partners - Acc...	4/14/2016 Courtois Investment Group 5 Fulham R...
4	196	Capital Activity	Account Statement	Arsenal Capital Group - Wenger Partners - Acc...	1	Arsenal Capital Group - Wenger Partners - Acc...	Liverpool Investments R 1/15/2...
5	200	Capital Activity	Account Statement	Arsenal Capital Group - Wenger Partners - Acc...	1	Arsenal Capital Group - Wenger Partners - Acc...	Tottenham Hotspur...
6	193	Capital Activity	Account Statement	Arsenal Capital Group - Wenger Partners - Acc...	1	Arsenal Capital Group - Wenger Partners - Acc...	...
7	197	Capital Activity	Account Statement	Arsenal Capital Group - Wenger Partners - Acc...	1	Arsenal Capital Group - Wenger Partners - Acc...	Arsenal Capital Group R 9/12/2...
8	201	Capital Activity	Account Statement	Arsenal Capital Group - Wenger Partners - Acc...	1	Arsenal Capital Group - Wenger Partners - Acc...	Capital Account Stateme...
9	194	Capital Activity	Account Statement	Arsenal Capital Group - Wenger Partners - Acc...	1	Arsenal Capital Group - Wenger Partners - Acc...	Chelsea Partners ...

Account Statement: 1, Distribution Notice: 2, Call Notice: 3, otherwise: -1

Text Classification - Example (CONT.)

	0	1	2	3	4	5	6	7	8	9	...	73	74	75	76	77	78	79	80	81	82
category	1	1	1	1	3	1	1	1	1	1	...	1	1	1	1	3	1	3	1	1	1
predict	3	1	1	2	3	3	3	2	3	1	...	1	1	3	2	1	3	3	1	1	3

	0	1	2	3	4	5	6	7	8	9	...	73	74	75	76	77	78	79	80	81	82
category	1	1	1	1	1	1	1	1	1	1	...	1	1	1	1	1	1	1	1	1	1
predict	1	1	1	3	1	1	1	1	1	1	...	3	3	1	2	1	3	1	1	-1	2