Entities to focus on:

* companies
* definitions
* money and currency usage
* PII

Documents are in the test folder of the repository

Entity Name: Amounts: Definitions

Details of what the LexNLP.Extract package:
The `lexnlp.extract.en.definitions` module contains methods that allow for the extraction of definitional statements from text. Example statements that are covered by default in this module are:

- X shall [not] include …
- X shall have the meaning …
- X is hereby changed to …
- X shall be interpreted …
- X shall for purposes …
- X shall be deemed to …
- X shall refer to …
- X shall mean …
- X is defined …
- The word “X” includes every description of …
- The term “X” means …
- Description of term (the “x”)

Doc 1: amtd-digital

Manual Review Output:

Agreement
Investor
Company
Party
Parties
A Shares
Purchase Price
Purchased Shares
MAA
Closing
Group Companies
Material Adverse Effect
B Shares
Act
Confidential Information
Closing Date
Indemnifiable Loss
Indemnitee
Indemnifying Party
Agreement
Consultation Period
Dispute
HKIAC
Business Day

Program Output:

Agreement
Investor
Company
Party
Parties
A Shares
Purchase Price
Purchased Shares
MAA
Closing
Group Companies
Material Adverse Effect

B Shares
Act
Confidential Information
Closing Date
Indemnifiable Loss
Indemnitee
Indemnifying Party
Agreement
Consultation Period
Dispute
HKIAC

Summary of differences between manual review and predicted output:
The Module classified 23/24 entities successfully.
The Output has returns a generator object . (https://wiki.python.org/moin/Generators)
To get output in readable format :

```
print(list(lexnlp.extract.en.definitions.get_definitions(text)))
```

Enclosing the function in list() will return a python list (string type) containing definition names
Possible reason why “Business Day” entity was not classified is that it was not preceded by the word “the”, eg “the Act” was successfully classified.

Doc 2: cboe global

Manual Review Output:

- Example 1
- Example 2
- Award Date
- Corporation
- Participant
- Agreement
- Restricted Stock Units
- Retirement Vesting
- disability
- Good Reason
- Committee
- ECN
- FX
- ATS
- secret or confidential information
- Addendum
- Target

- Threshold
- Maximum
- Total Shareholder Return
- TSR Percentile
- Agreement
- Employer
- Plan
- Performance Period
- Cboe
- separation from service
- specified employee

Program Output:

- Award Date
- Corporation
- Participant
- Agreement
- Restricted Stock Units
- Retirement Vesting
- disability
- Good Reason
- Committee
- ECN
- FX
- ATS
- secret or confidential information
- Addendum
- Target
- Threshold
- Maximum
- Total Shareholder Return
- TSR Percentile
- Agreement
- Employer

Summary of differences between manual review and predicted output:

Model classifies 21/26 entities present

Doc 3: ContractKen_Word_Addin_Testing_ContractDoc_1.doc

Company

Atrium Nominee
Dowling Shareholders
Trident Shareholders
Initial Shareholders
Agreement
Effective Date
Recapitalization Agreement
North Bay
Affiliate
Applicable Law
Arden Re
Atrium
Atrium Group
Beneficial Owner
Business Day
Bye-laws
Change of Control
Committee
Common Shares
Exchange Act
Excluded Securities
Government Approval
Governmental Authority
Initial Public Offering
Joinder Agreement
Leaver Sale Provisions
Lien
Material Subsidiary
Memorandum of Association
Nominee
Nominee Agreements
Organizational Documents
Permitted Transferee
Person
Plans
Related Party Agreement
Representative
Securities Act
Subsidiary

Third Party Purchaser
Transfer
Director
Board
Enstar Director
Trident Directors
Trident Designees
Trident Director
Directors
Offered Shares
Offering Shareholder
Offering Shareholder Notice
ROFO Notice Period
ROFO Notice
Trident Shareholders
Waived ROFO Transfer Period
Drag-along Sale
Drag-along Shareholder
Drag-along Notice
Selling Shareholder
Proposed Transferee
Tag-along Sale
Tag-along Shareholder
Sale Notice
Tag-along Notice
Tag-along Period
Atrium Nominee Tag Notice
Pre-emptive Shareholder
New Securities
Issuance Notice
Pre-emptive Pro Rata Portion
Exercise Period
Over-allotment Notice
Overallotment New Securities
Non-Exercising Shareholder
Over-allotment Exercise Period
Exercising Shareholder
Information
Leaver

Leaver Sale
include
including
Exhibits
The Exhibits
Company

Program Output :

- Company
- Atrium Nominee
- Dowling Shareholders
- Trident Shareholders
- Initial Shareholders
- Agreement
- Effective Date
- Recapitalization Agreement
- North Bay
- Affiliate
- Applicable Law
- Arden Re
- Atrium
- Atrium Group
- Beneficial Owner
- Business Day
- Bye-laws
- Change of Control
- Committee
- Common Shares
- Exchange Act
- Excluded Securities
- Government Approval
- Governmental Authority
- Initial Public Offering
- Joinder Agreement
- Leaver Sale Provisions
- Lien
- Material Subsidiary
- Memorandum of Association
- Nominee
- Nominee Agreements
- Organizational Documents
- Permitted Transferee
- Person
- Plans

- Related Party Agreement
- Representative
- Securities Act
- Subsidiary
- Third Party Purchaser
- Transfer
- Director
- Board
- Enstar Director
- Trident Directors
- Trident Designees
- Trident Director
- Directors
- Offered Shares
- Offering Shareholder
- Offering Shareholder Notice
- ROFO Notice Period
- ROFO Notice
- Trident Shareholders
- Waived ROFO Transfer Period
- Drag-along Sale
- Drag-along Shareholder
- Drag-along Notice
- Selling Shareholder
- Proposed Transferee
- Tag-along Sale
- Tag-along Shareholder
- Sale Notice
- Tag-along Notice
- Tag-along Period
- Atrium Nominee Tag Notice
- Pre-emptive Shareholder
- New Securities
- Issuance Notice
- Pre-emptive Pro Rata Portion
- Exercise Period
- Over-allotment Notice
- Overallotment New Securities
- Non-Exercising Shareholder
- Over-allotment Exercise Period
- Exercising Shareholder
- Information
- Leaver
- Leaver Sale
- include
- including
- Exhibits

- The Exhibits
- Company

Summary of differences between manual review and predicted output:
Model classifies 85/85 entities

The package has a high level of accuracy for definitions
After checking the code, i have observed that an actual NER algorithm has been implemented.

Entity Name: Money and Currency usages
The `lexnlp.extract.en.money` module contains methods that allow for the extraction of money and currency amounts from text. Example references that are covered by default in this module include:

- five dollars
- 5 dollars
- 5 USD
- $5

Comprehensive ISO 4217 codes can be captured, but frequently result in false positive matches in real documents. By default, only the following ISO 4217 codes and currency symbols are detected:

- USD/$: US Dollars
- EUR/€: Euros
- GBP/£: Great British pounds
- JPY/¥: Japanese Yen
- CNY/RMB/元/¥: Chinese Yuan/Renminbi
- INR/Rs/₹: Indian Rupee

-

amtd digital

- amtd-digital
- Entities
- US$0.0001,
- US$3,000,000,
- US$368,000,000,
- US$1,000,000,
- US$100,000,

Output

- US0.0001
- US$3
- US$368000000
- US$1000000
- US$1.2
- US$100000
-

Detects 4 /5 of the entities present. Overpredicts and gives $3 and $1.2 as output also. They are not monetary amounts.

cboe-global

Entities

no monetary amounts

Output

Detects US$20 as a monetary amount. It is not a monetary amount. Model has overpredicted.

ContractKen_Word_Addin_Testing_ContractDoc_1

Entities

US$1

Output

US$1

US$10

Model overpredicts again including 10 as an entity

On looking at code:

Regex is used for extraction

`lexnlp.extract.en.pii`: Extracting personally-identifiable information (PII)

The `lexnlp.extract.en.pii` module contains methods that allow for the extraction of personally identifying information from text. Examples include:

- phone numbers
- US social security numbers
- Names

Entities

1st floor, Building 3, Zhongguancun Yonghe Hangxing Science Park, No. 11 Hepingli East Street, Dongcheng District, Beijing, PRC

Output

Does not detect any entity. returns empty list.

cboe global

Entities

no entites

Output

returns empty list

Returns empty list for all documents except

ISDA-Example.txt

Entities

(312) 234-273

(866) 255-1444

(212) 548-8622

(312) 992-6351

(312) 621-1950

(312) 621-1969

(312) 280-3708

(312) 280-3730

Output

(312) 234-273

(866) 255-1444

(212) 548-8622

(312) 992-6351

(312) 621-1950

(312) 621-1969

(312) 280-3708

(312) 280-3730

Detects 8/8 entities

ContractKen_Word_Addin_Testing_ContractDoc_1

Entities

Atrium Underwriting Group Limited Lloyd’s Building, Room 790 1 Lime Street London, England EC3M 7DQ Email: peter.hargrave@atrium-uw.com Attention: Peter Hargrave, Human Resources Director

c/o Dowling Capital Partners

190 Farmington Avenue

Farmington, CT 06032

Hogan Lovells US LLP

1735 Market Street, Suite 2300

Philadelphia, Pennsylvania 19103

c/o Enstar Group Limited

Windsor Place, 3rd Floor

22 Queen Street

Hamilton HM 11

Bermuda

Willkie Farr & Gallagher LLP

787 Seventh Avenue

New York, New York 10019

Output:

Detects none returns empty list

For all other contracts it also returns an empty list(except ISDA-Example.txt)

After taking a look at the code I realised :

1. The code relied completely on regex and did not use any natural language processing algorithm to extract the entity
2. There were only 2 methods to extract said entity, one for US phone numbers and the other for US Social Security numbers

`lexnlp.extract.en.entities.nltk_re`: Extracting companies

The `lexnlp.extract.en.entities.nltk_re` module contains methods that allow for the extraction of company names from text. Example statements that are covered by default in this module include:

- Deutsche Bank Securities Inc.
- ACME, INC.
- Wells Fargo Bank Minnesota, National Association
- Lexpredict LLC

When function

```
print(list(lexnlp.extract.en.entities.nltk_re.get_entities.nltk_re(text)))
```

Is used an error is thrown :

```
AttributeError: module 'lexnlp.extract.en.entities.nltk_re' has no attribute 'get_entities' occurs
```

```
This issue has been raised on their github
https://github.com/LexPredict/lexpredict-lexnlp/issues/55
```