

# Refined Name Entity Recognition (NER) by A Customized SpaCy Model and Pattern Rules of RegEx

Moha  
Yeghaneh

## Introduction

# Refined Name Entity Recognition (NER) by A Customized SpaCy Model and Pattern Rules of RegEx

Moha Yeghaneh

January 12, 2020



## 1 Introduction

## 2 Data Preparation

## 3 Data Annotation

## 4 Refined Named Entity Recognition Model

## 5 Explicit Observation Extraction

## 6 Data Publishing

Refined Name  
Entity  
Recognition  
(NER) by A  
Customized  
SpaCy Model  
and Pattern  
Rules of  
RegEx

Moha  
Yeghaneh

Introduction

Data  
Preparation

Data  
Annotation

Refined  
Named Entity  
Recognition  
Model

Explicit  
Observation  
Extraction

Data  
Prediction

The aim of this presentation is to show the initial result and direction of our research. Here is the main area that we have worked:

## ■ Data Preparation

Refined Name  
Entity  
Recognition  
(NER) by A  
Customized  
SpaCy Model  
and Pattern  
Rules of  
RegEx

Moha  
Yeghaneh

Introduction

Data  
Preparation

Data  
Annotation

Refined  
Named Entity  
Recognition  
Model

Explicit  
Observation  
Extraction

Data  
Prediction

The aim of this presentation is to show the initial result and direction of our research. Here is the main area that we have worked:

## ■ Data Preparation

## ■ Data Annotation

Refined Name  
Entity  
Recognition  
(NER) by A  
Customized  
SpaCy Model  
and Pattern  
Rules of  
RegEx

Moha  
Yeghaneh

Introduction

Data  
Preparation

Data  
Annotation

Refined  
Named Entity  
Recognition  
Model

Explicit  
Observation  
Extraction

Data  
Publication

The aim of this presentation is to show the initial result and direction of our research. Here is the main area that we have worked:

- **Data Preparation**
- **Data Annotation**
- **Data Investigation**

Refined Name  
Entity  
Recognition  
(NER) by A  
Customized  
SpaCy Model  
and Pattern  
Rules of  
RegEx

Moha  
Yeghaneh

Introduction

Data  
Preparation

Data  
Annotation

Refined  
Named Entity  
Recognition  
Model

Explicit  
Observation  
Extraction

Data  
Prediction

The aim of this presentation is to show the initial result and direction of our research. Here is the main area that we have worked:

- **Data Preparation**
- **Data Annotation**
- **Data Investigation**
- **Data Visualization**

The aim of this presentation is to show the initial result and direction of our research. Here is the main area that we have worked:

- **Data Preparation**
- **Data Annotation**
- **Data Investigation**
- **Data Visualization**
- **Refined Named Entity Recognition Model.**

The aim of this presentation is to show the initial result and direction of our research. Here is the main area that we have worked:

- **Data Preparation**
- **Data Annotation**
- **Data Investigation**
- **Data Visualization**
- **Refined Named Entity Recognition Model.**
- **Classification of Observational Sentences using NER**

The aim of this presentation is to show the initial result and direction of our research. Here is the main area that we have worked:

- **Data Preparation**
- **Data Annotation**
- **Data Investigation**
- **Data Visualization**
- **Refined Named Entity Recognition Model.**
- **Classification of Observational Sentences using NER**
- **Data Publishing.**

The aim of this presentation is to show the initial result and direction of our research. Here is the main area that we have worked:

- **Data Preparation**
- **Data Annotation**
- **Data Investigation**
- **Data Visualization**
- **Refined Named Entity Recognition Model.**
- **Classification of Observational Sentences using NER**
- **Data Publishing.**
- **Toward Relation Extraction!**

# Difficulty and Challenges

Refined Name Entity Recognition (NER) by A Customized SpaCy Model and Pattern Rules of RegEx

Moha Yeghaneh

Introduction

Data Preparation

Data Annotation

Refined Named Entity Recognition Model

Explicit Observation Extraction

Data Publishing

- Noisy and inconsistent text data.
- Time consuming and tedious manual modification of annotation.
- Unavailability of training data and research paper in the domain.



# Purpose of this Presentation

Refined Name  
Entity  
Recognition  
(NER) by A  
Customized  
SpaCy Model  
and Pattern  
Rules of  
RegEx

Moha  
Yeghaneh

Introduction

Data  
Preparation

Data  
Annotation

Refined  
Named Entity  
Recognition  
Model

Explicit  
Observation  
Extraction

Data  
Published

In this report we want to communicate what we have already done including:

- **Prepossessing** and preparation of text data for classification

# Purpose of this Presentation

Refined Name  
Entity  
Recognition  
(NER) by A  
Customized  
SpaCy Model  
and Pattern  
Rules of  
RegEx

Moha  
Yeghaneh

Introduction

Data  
Preparation

Data  
Annotation

Refined  
Named Entity  
Recognition  
Model

Explicit  
Observation  
Extraction

Data  
Published

In this report we want to communicate what we have already done including:

- **Prepossessing** and preparation of text data for classification
- Information extraction in an informative and **interactive** way

# Purpose of this Presentation

Refined Name  
Entity  
Recognition  
(NER) by A  
Customized  
SpaCy Model  
and Pattern  
Rules of  
RegEx

Moha  
Yeghaneh

Introduction

Data  
Preparation

Data  
Annotation

Refined  
Named Entity  
Recognition  
Model

Explicit  
Observation  
Extraction

In this report we want to communicate what we have already done including:

- **Prepossessing** and preparation of text data for classification
- Information extraction in an informative and **interactive** way
- Introducing a refined named entity recognition (NER) model using deep learning

# Purpose of this Presentation

Refined Name  
Entity  
Recognition  
(NER) by A  
Customized  
SpaCy Model  
and Pattern  
Rules of  
RegEx

Moha  
Yeghaneh

Introduction

Data  
Preparation

Data  
Annotation

Refined  
Named Entity  
Recognition  
Model

Explicit  
Observation  
Extraction

In this report we want to communicate what we have already done including:

- **Prepossessing** and preparation of text data for classification
- Information extraction in an informative and **interactive** way
- Introducing a refined named entity recognition (NER) model using deep learning
- evaluation of model using gold standard data

# Purpose of this Presentation

Refined Name Entity Recognition (NER) by A Customized SpaCy Model and Pattern Rules of RegEx

Moha Yeghaneh

Introduction

Data Preparation

Data Annotation

Refined Named Entity Recognition Model

Explicit Observation Extraction

In this report we want to communicate what we have already done including:

- **Prepossessing** and preparation of text data for classification
- Information extraction in an informative and **interactive** way
- Introducing a refined named entity recognition (NER) model using deep learning
- evaluation of model using gold standard data
- Using machine learning and deep learning methods for text classification

# Purpose of this Presentation

Refined Name  
Entity  
Recognition  
(NER) by A  
Customized  
SpaCy Model  
and Pattern  
Rules of  
RegEx

Moha  
Yeghaneh

Introduction

Data  
Preparation

Data  
Annotation

Refined  
Named Entity  
Recognition  
Model

Explicit  
Observation  
Extraction

In this report we want to communicate what we have already done including:

- **Prepossessing** and preparation of text data for classification
- Information extraction in an informative and **interactive** way
- Introducing a refined named entity recognition (NER) model using deep learning
- evaluation of model using gold standard data
- Using machine learning and deep learning methods for text classification
- Proposing some ideas for the future work toward relation extraction and causal inference

# Feeling the Data through Some Statistics!

Refined Name Entity Recognition (NER) by A Customized SpaCy Model and Pattern Rules of RegEx

Moha Yeghaneh

Introduction

Data Preparation

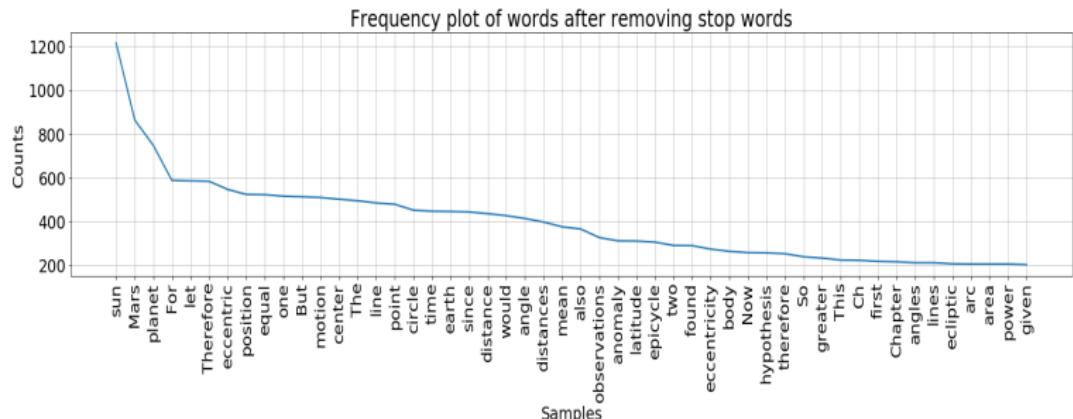
Data Annotation

Refined Named Entity Recognition Model

Explicit Observation Extraction

Data

- Corpus has 70 chapters including 1605 paragraph, 6699 sentences
- Corpus includes 169231 tokens (roughly speaking; words) and 9513 unique tokens
- lexical\_diversity which shows lexical richness is 1.2



Feeling the integrated data by NER and more  
attribute as dataframe

here you can see first 10 sentences with different attribute that has been add by our and can be used later for classification, relation extraction ...

.../.../.../Documents/Former\_projects

- You can find here the word cloud of the whole book and by chapter
  - It can give us some simple initial and simple intuition which can be used for the further text analysis



# NER workflow

Refined Name Entity Recognition (NER) by A Customized Spacy Model and Pattern Rules of RegEx

Moha Yeghaneh

Introduction

Data Preparation

Data Annotation

Refined Named Entity Recognition Model

Explicit Observation Extraction

Data Preprocess

- Data is annotated entity by entity using regex pattern.
- The result of each step is saved as jsonl fomat
- After troubleshooting (false tokenization,double punctuation...)
- Annotated data is merged and now the training data is ready!

Dragon was at 27° 20' 30'' above the horizon, which should have been 27° 19' 52'' above the horizon. We know that the sunset ends 2% minutes high. Therefore, the corrected distance of 3000 miles from the tail of **Dragon** was 27° 19' 52'' above the horizon. And since the **Latitude** of the **Sunset** was 27° 19' 52'', the remainder, by subtraction, is 2'. This means that **Dragon** was at precisely the same **Latitude**, **Altitude** as the **Sunset**. But because there was a difference of 3% degrees between them (as is clear from the following observation), a slight correction is required. For AB to be 27° 20' 30'' on a parallel close to the **Sunset**, **AB** = 1' + 27° 19' 52'' C the **Sunset** **AB** and BC = 27° 20' 30'' - 27° 19' 52''. Dividing the **arc** of BC by the **arc** of AB gives the **arc** of CA = 27° 20' 30'' - 27° 19' 52'', which, subtracted from 27° 20' 30'', the **Latitude** of **AB** and the **Latitude** of **BC** leaves 27° 19' 52'' - 27° 19' 52'' = 0'. However, **AB** is 27° 20' 30'' and **BC** is 27° 19' 52''. At the same time we found 27° 20' 30'' between **Dragon** and **Dragon** 18° 20% connected, and 27° 19' 52'' between **Dragon** and the bright star **Regulus** in the wing of **Vergilius**. This means that **Dragon** is correctly positioned. From these two distances (using the latitudes of the stars and **Dragon**) and the **Latitude** of **Dragon**, the **Altitude** of **Dragon** is found to be 27° 20' 30'' above the horizon. This is in accordance with the consensus of all measurements. Alternatively, the **Latitude** of **Dragon** is 27° 19' 52'' and the **Altitude** of **Dragon** is found to be 27° 20' 30''. Using tree quadrilaterals, to find the **Latitude** of **Dragon**, while the tail (end) was at 27° 19' 52''. Then, from the declinations and right ascensions of the fixed stars and our distances, the position of **Dragon** is determined as 27° 20' 30''.  
The **Latitude** of **Dragon** is 27° 19' 52''. The reason is that the declination and right ascension of the fixed stars and our distances are included in the **Latitude** of **Dragon**. The **Latitude** of **Dragon** is 27° 19' 52''. This is the **Latitude** procedure. It was included for the sake of showing a consensus, and also that it might fit the evidence that despite the lack of absolute perfection in the demonstration, short cuts either in computation or in our understanding can under certain circumstances be applied. For the previous procedure there is less in the actual work than in the reporting of it. At **Dragon**, **Latitude** was 27° 19' 52''. Therefore, the **Longitude** of **Dragon** was about 32° 5' from the zenith. And since **Dragon**'s distance from the earth was more than half the sun's distance, the resulting **Longitude** of about 32° 5' in our parallel table shows a latitudinal **Longitude** of 27° 19' 52''. Thus the **Longitude** is away from the middle of the earth would be 27° 19' 52'' **Longitude**. And because the **Longitude** of the **Dragon** was 32° 5', the **Longitude** of **Dragon** at the horizon was 27° 19' 52''. But since **Dragon** was 18° from the **Longitude** of **Dragon**, 27° 19' 52'' **Longitude**, corresponding to this position is 27° 19' 52'' **Longitude**, and if this is eliminated, **Dragon** could be placed at about 27° 19' 52'' **Longitude**. At that moment, the sun's position was 27° 19' 52'' **Longitude**. The distance between the bodies was 27° 19' 52'' **Longitude**, and that is 27° 19' 52'' **Longitude** (or it was 27° 19' 52'' **Longitude** + 27° 19' 52'' **Longitude**) in 1983, and 24 of 26 Vergilius in 1987. The sum of the **Longitude** motions was 27° 19' 52'' **Longitude**. The true **Longitude** of **Dragon** therefore followed 27° 19' 52'' **Longitude** in 1983, and on 21 February 1987 March before dawn at 27° 19' 52'' **Longitude**. Forty seconds must be subtracted to reduce the position to the **Longitude**, putting **Dragon** at 27° 19' 52'' **Longitude**.

# Training

Refined Name  
Entity  
Recognition  
(NER) by A  
Customized  
SpaCy Model  
and Pattern  
Rules of  
RegEx

Moha  
Yeghaneh

Introduction

Data  
Preparation

Data  
Annotation

Refined  
Named Entity  
Recognition  
Model

Explicit  
Observation  
Extraction

Data  
Prediction

- A Bert model with 100 iteration and batch size 16 has been used for NER classification
- The evaluation result per entity and overall is calculated by comparing with gold standard format as follows:

Metrics	ents_p'	ents_r	ents_f	
GEOG	100	99.85	99.92	
LONG	99.76	99.88	99.82	
PARA	98.51	99.76	99.13	
TIME	97.97	97.00	97.48	
STAR	84.61	74.15	79.04	

# CitableClass

Refined Name Entity Recognition (NER) by A Customized SpaCy Model and Pattern Rules of RegEx

Moha Yeghaneh

Introduction

Data Preparation

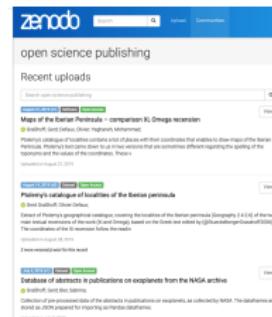
Data Annotation

Refined Named Entity Recognition Model

Explicit Observation Extraction

Data Publishing

- We have developed a useful framework CitableClass that we can use in order to publish and use the data.
- Any user can have access to data with notebooks using DOI number



*Thank you for your time and feedbacks :)  
Many thanks*

Refined Name  
Entity  
Recognition  
(NER) by A  
Customized  
SpaCy Model  
and Pattern  
Rules of  
RegEx

Moha  
Yeghaneh

Appendix

-  Ma, Y.; Zhou, G.; Wang, S.; Zhao, H.; Jung, W. SignFi: Sign Language Recognition Using WiFi. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2018, 2, 23.
-  Muller, Machine Learning and AI for the sciences – Towards Understanding