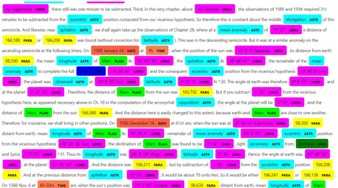
Refined Name Enitiy Recognition (NER) by A Customized SpaCy Model and Pattern Rules of RegEx

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

January 15, 2020



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**

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**

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**

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**
- **Customized 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
- **Customized 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**
- **Customized Named Entity Recognition Model.**
- **Classification of Observational Sentences using NER**
- **Data Publishing.**

Difficulty and Challenges

- Noisy and inconsistent text data.

Difficulty and Challenges

- Noisy and inconsistent text data.
- Time consuming and tedious annotation workflow.

Difficulty and Challenges

- Noisy and inconsistent text data.
- Time consuming and tedious annotation workflow.
- Unavailability of training data.

Difficulty and Challenges

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

Purpose of this presentation

communicate the workflow and ideas:

- **Preprocessing** and preparation of text data for classification.

Purpose of this presentation

communicate the workflow and ideas:

- **Preprocessing** and preparation of text data for classification.
- **Information Extraction** in an informative and interactive way.

Purpose of this presentation

communicate the workflow and ideas:

- **Preprocessing** and preparation of text data for classification.
- **Information Extraction** in an informative and interactive way.
- **Customized Named Entity Recognition** (CNER) model using sate of the art deep learning models.

Purpose of this presentation

communicate the workflow and ideas:

- **Preprocessing** and preparation of text data for classification.
- **Information Extraction** in an informative and interactive way.
- **Customized Named Entity Recognition** (CNER) model using sate of the art deep learning models.
- **Evaluation** of model using gold standard data.

Purpose of this presentation

communicate the workflow and ideas:

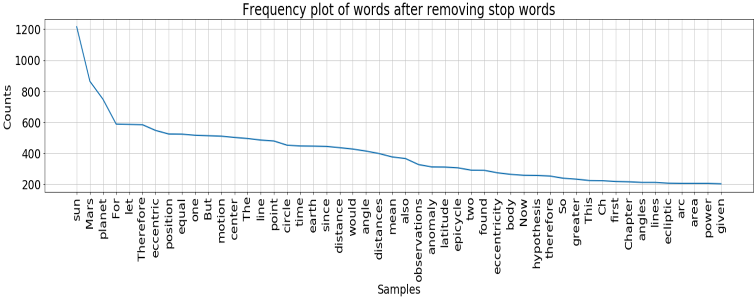
- **Preprocessing** and preparation of text data for classification.
- **Information Extraction** in an informative and interactive way.
- **Customized Named Entity Recognition** (CNER) model using sate of the art deep learning models.
- **Evaluation** of model using gold standard data.
- **Explicit Observation Extraction** from Text Corpus.

Purpose of this presentation

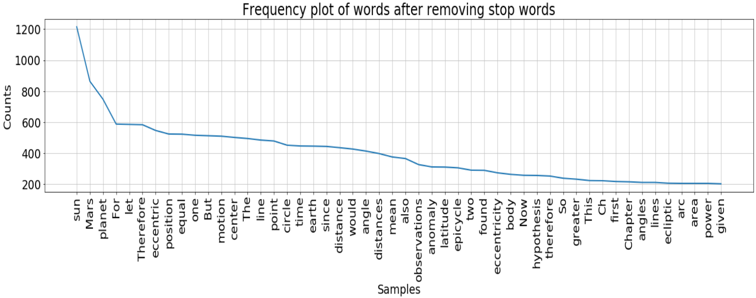
communicate the workflow and ideas:

- **Preprocessing** and preparation of text data for classification.
- **Information Extraction** in an informative and interactive way.
- **Customized Named Entity Recognition** (CNER) model using sate of the art deep learning models.
- **Evaluation** of model using gold standard data.
- **Explicit Observation Extraction** from Text Corpus.
- Toward **Relation Extraction** and Causal Inference.

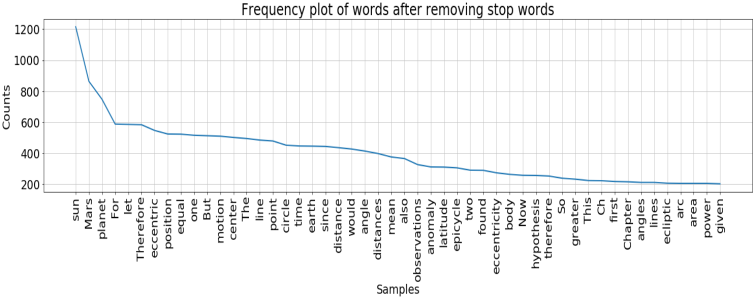
- Corpus has 70 chapters including 1605 paragraph, 6699 sentences



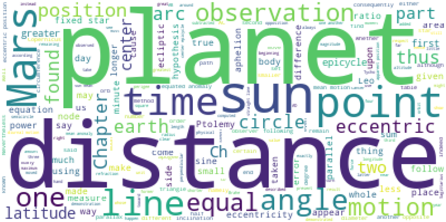
- Corpus has 70 chapters including 1605 paragraph, 6699 sentences
- Corpus includes 169231 tokens (roughly speaking; words) and 9513 unique tokens



- 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



- Word cloud of the whole book and by chapters as initial intution



- Couple of sentences with different attribute that has been add by our analysis and can be used later for classification,relation extraction and...

	Sentence	SentIndex	Chapter	Paragraph	ParaIndex	ASO	Entities	CNER	Label
158	On 1580 November 12 at 10h 50m,1 they set Mars down at 8° 36′ 50″ Gemini2 without mentioning the horizontal variations, by which term I wish the diurnal parallaxes and the refractions to be understood in what follows.	967	10	On 1580 November 12 at 10h 50m,1 they set Mars down at 8° 36′ 50″ Gemini2 without mentioning the horizontal variations, by which term I wish the diurnal parallaxes and the refractions to be understood in what follows. Now this observation is distant and isolated. It was reduced to the moment of opposition using the diurnal motion from the Prutenic Tables ^3.	218	{'act': 'set', 'subject': 'they', 'obj': 'Mars '}	[[1580 November 12, DATE], [10h 50m,1 they, TIME], [Mars, PLAN], [8° 36′ 50″, LONG], [diurnal, ASTR]]	[1, 1, 1]	1
168	On 1582 December 28 at 11h 30m, they set Mars down at 16° 47′ Cancer by observation ^6.	977	10	On 1582 December 28 at 11h 30m, they set Mars down at 16° 47′ Cancer by observation ^6. The moment of opposition assigned by Tycho comes 46 minutes later, during which the planet retrogressed less than one minute. Tycho therefore puts it at 16° 46′ 16″ Cancer ^7. On an inserted sheet here, an attempt was made to correct for a refraction of two minutes. This was, I think, first trial of the theory of refraction then being developed. Nevertheless, he followed the observed value unchanged, thus declining to consider the planet as something which could alter its position. Nor was there any need for correction, since it was in Cancer, beyond the reach of refraction, and was in mid-sky where, in Cancer, there is no longitudinal parallax.	221	{'act': 'set', 'subject': 'they', 'obj': 'Mars '}	[[1582 December 28, DATE], [11h 30m, TIME], [Mars, PLAN], [16° 47′ Cancer, LONG], [observation, ASTR]]	[1, 1, 1]	1
175	On 1585 January 31 at 12h 0m, Mars was placed at 21° 18′ 11″ Leo ^8.	984	10	On 1585 January 31 at 12h 0m, Mars was placed at 21° 18′ 11″ Leo ^8. The diurnal motion, by comparison of observations, was 24′ 15″. The moment of opposition followed at 19h 35m, 7 hours and 35 minutes later. To this period belongs 7′ 41″ of diurnal motion westward. Therefore, at the designated moment, it would have been at 21° 10′ 30″ Leo, which is what was accepted. There is no mention of parallax. Nothing had to be done about refraction, because Mars was high and at mid-sky. I therefore find the bit of advice in the table about refraction (properly) ignored.	222	{'act': 'placed'}	[[1585 January 31, DATE], [12h 0m, TIME], [Mars, PLAN], [21° 18′ 11″ Leo, LONG]]	[1, 1, 1]	1

- Data is annotated entity by entity using regex pattern.

.

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

- Data is annotated entity by entity using regex pattern.
- The result of each step is saved as jsonl fomat

.

- 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...)

.

- 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!

.

- A model word representation Bert with100 iteration and batch size 16 has been used for NER classifcation

Metrics	ents_p'	ents_r	ents_f	
GEOM	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	

- A model word representation Bert with100 iteration and batch size 16 has been used for NER classifcation
- The evaulation result per entity and overall is calculated by comparing with gold stadnard format as follows:

Metrics	ents_p'	ents_r	ents_f	
GEOM	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	

- you can see here an example of explicit observation extraction.

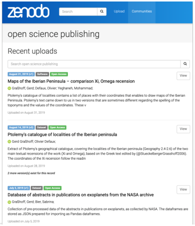
183	On 1587 March 7 at 19h 10m they deduced the position of Mars from the observations, which was 25° 10′ 20″ Virgo.	On 1587 March 7 at 19h 10m they deduced the position of Mars from the observations, which was 25° 10′ 20″ Virgo. This they kept in the table, but changed the time to 17h 22m. The difference of 1h 48m multiplied by a diurnal motion of 24′ gives the same number of minutes and seconds (that is, 1′ 48″), no more. It therefore should have been 25° 8′ 32″ Virgo, which also approaches nearer the point opposite the sun. The difference is of practically no importance ^9.

- you can see here an example of explicit observation extraction.
- moreover, you can find how the text has been structured.

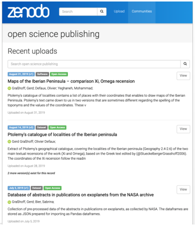
183	On 1587 March 7 at 19h 10m they deduced the position of Mars from the observations, which was 25° 10′ 20″ Virgo.	On 1587 March 7 at 19h 10m they deduced the position of Mars from the observations, which was 25° 10′ 20″ Virgo. This they kept in the table, but changed the time to 17h 22m. The difference of 1h 48m multiplied by a diurnal motion of 24′ gives the same number of minutes and seconds (that is, 1′ 48″), no more. It therefore should have been 25° 8′ 32″ Virgo, which also approaches nearer the point opposite the sun. The difference is of practically no importance ^9.

- We have a developed a usefull framework citableclass that we can used in order to publich and use the data.

- We have a developed a usefull framework citableclass that we can used in order to publich and use the data.
- Any user can have access to data with notebooks using DOI number

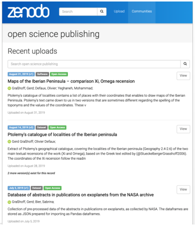


- We have a developed a usefull framework citableclass that we can used in order to publich and use the data.
- Any user can have access to data with notebooks using DOI number



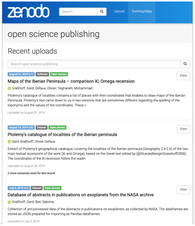
- how can we have reuslable notebooks flies(env, requ...), there is some answers like but it seems not perfect

- We have a developed a usefull framework citableclass that we can used in order to publich and use the data.
- Any user can have access to data with notebooks using DOI number



- how can we have reuslable notebooks flies(env, requ...), there is some answers like but it seems not perfect
- find the best envirment to cooperate with (colab,binder,

- We have a developed a usefull framework citableclass that we can used in order to publich and use the data.
- Any user can have access to data with notebooks using DOI number



- how can we have reuslable notebooks flies(env, requ...), there is some answers like but it seems not perfect
- find the best envirment to cooperate with (colab,binder,
- find the best way to presnt the result in way that non-technical readers could be able to go thorough the text and computational analysis (voila)

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