Preprocessing



Andrei Pruteanu
PHD

@andrei_pruteanu

https://sites.google.com/site/andreipruteanu

Overview



Dataset Finding & Encoding

Analyzing Dataset

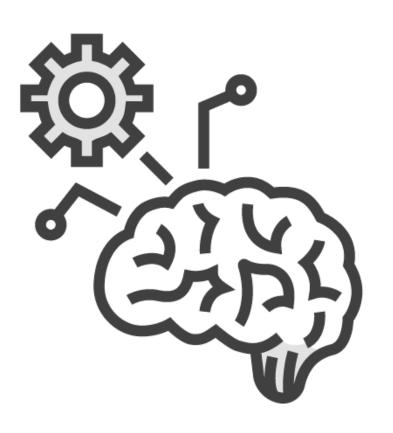
Preparing Dataset



Dataset Finding & Encoding



Dataset Requirements



IOB representation

Extensive and well maintained

Freely available

IOB representation



Standardized and common



IOB2 uses B-tag to mark start of a chunk



Popular in Computational Linguistics



Extensions for single-word chunks



Does not encode sentence nesting

Example - Raw Text

"Alex is going to Los Angeles"



Example - IOB Formatting

"Alex(B-PER) is(O) going(O) to(O) Los(B-LOC) Angeles(B-LOC)"



Kaggle Dataset



https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus

Annotated Corpus for NER

GMB(Groningen Meaning Bank)

Entities Information

Supported List

```
geo = Geographical Entity
```

org = Organization

per = Person

gpe = Geopolitical Entity

tim = Time indicator

art = Artifact

eve = Event

nat = Natural Phenomenon



Content Analysis



Dataset Preparation



Scikit-learn DictVectorizer

Scikit-learn

Feature extraction functionality

Transforms

Lists of feature-value mappings to vectors

Feature values

Numpy arrays



Summary



Criteria for finding a dataset

How to analyze it

Transforming to numerical format

