I tried two approaches (no dataset label string matching)

* Regular expression - 0.5/0.56
* BERT NER - 0.37/0.52

Ensemble, searching among known datasets and other approaches with BERT did not improve the results

**Regular expressions**

* Looking for uppercase letters
* Looking for nearby words with these letters
* Just looking for words beginning with uppercase letters
* Select good candidates
* I memorize all found and search among other documents

**BERT NER**

* I divide the document into sentences of 200-400 characters
* Select potential candidates
* I take 90% of sentences with tags and 10% without tags
* Use 3 classes - no class, first word and last word of dataset name
* Predict and select good candidates

**For train BERT**

* Download the data
* edit coleridge2.py
* run python coleridge2.py

**Link**

* [Data](https://www.kaggle.com/c/coleridgeinitiative-show-us-the-data/data)
* [BERT weights](https://www.kaggle.com/vlomme/07-model)
* [kaggle notebook](https://www.kaggle.com/vlomme/coleridge-initiative-5th-place?scriptVersionId=66412252)
* [Colab Regular expressions](https://colab.research.google.com/github/vlomme/Find-dataset-from-text/blob/main/re.ipynb)
* [Github](https://github.com/vlomme/Find-dataset-from-text)