NER Research for ML entities

- https://www.kaggle.com/datasets?search=Named+Entity+Recognition
  - NER datasets from Kaggle
- https://www.kaggle.com/Cornell-University/arxiv
  - lots of abstracts!!!
  - https://www.kaggle.com/artgor/arxiv-metadata-exploration
  - https://www.kaggle.com/Cornell-University/arxiv/discussion/173851 can filter arxiv by metadata (stat.ML)
  - https://www.kaggle.com/officialshivanandroy/transformers-generating-titles-from-abstracts Super interesting - uses BERT to generate titles from abstracts. But the titles generated often incorporated the methods!!
- https://www.kaggle.com/jl18pg052/word-embedding-word2vec-topic-modelling-lda
  - Extracts dominant topic
- https://www.kaggle.com/vetrirah/janatahack-independence-day-2020-ml-hackathon
  - Labeled abstracts with topic
- **https://www.kaggle.com/madjakul/iob-annotation-of-abstracts-in-computer-science**
  - **CS-related NER (has ALGO – algorithm) IOB tagging**
- **https://www.kaggle.com/madjakul/computer-science-gazetteers**
  - **Gazetteers (seeds) for ML algorithms**
- https://www.aclweb.org/anthology/W19-5807.pdf How to use Gazetteers for NER NN
- https://support.prodi.gy/t/ner-with-gazetteer/272/2 Prodigy patters (not quite the same as gazetteers) If you want a pure gazetteer entity recognition component, you can use spaCy's Matcher or PhraseMatcher classes: https://spacy.io/usage/linguistic-features#section-rule-based-matching The subsequent statistical NER is constrained by these existing entities: it can't propose any entities that overlap or overwrite the ones that are already set.
- https://support.prodi.gy/t/ner-or-phrasematcher/686/2 POS can be very useful. If needed, you can use the Prodigy recipes pos.teach and dep.teach to fine-tune the pre-trained model, in case your data includes constructions that weren't so frequent in the training data. You can then use spaCy to iterate around the tree and extract the information you need relative to the entities detected by your rules or entity recognizer.
- https://datascience.stackexchange.com/questions/9950/nlp-is-gazetteer-a-cheat
  - "However, when you train any kind of NLP model, which does rely on dictionary while training, you may get real world performance way lower than your initial testing would report, unless you can include all objects of interest into the gazetteer (and why then you need that model?) because your trained model will rely on the feature at some point and, in a case when other features will be too weak or not descriptive, new objects of interest would not be recognized. If you do use a Gazetteer in your models, you should make sure, that that feature has a counter feature to let model balance itself, so that simple dictionary match won't be the only feature of positive class (and more importantly, gazetteer should match not only positive examples, but also negative ones)."]

… gazetteers … they take into account features, which are most of the time:

- tokens
- part of speech
- capitalization
- gazetteers
- (and much more)

Actually gazetteers features provide the model with intermediary information that the training step will take into account, without explicitly being dependent on the list of NEs present in the training corpora. Let's say you have a gazetteer about sport teams, once the model is trained you can expand the list as much as you want without training the model again. The model will consider any listed sport team as... a sport team, whatever its name.

In practice:

1. Use any NER or ML-based framework
2. Decide what gazetteers are useful (this is maybe the most crucial part)
3. Affect to each gazetteer a relevant tag (e.g. sportteams, companies, cities, monuments, etc.)
4. Populate gazetteers with large lists of NEs
5. Make your model take into account those gazetteers as features
6. Train a model on a relevant corpus (it should containing many NEs from gazetteers)
7. Update your list as much as you want

Off-topic but interesting articles on the challenges and opportunities of ML is geosciences
https://arxiv.org/pdf/1711.04708.pdf

- Seems like it would be really useful to have links to other datasets or models outside of USGS (NASA, NOAA) for research purposes.
- To encourage use of ML in geoscience, encourage use of available big datasets
- Issues with spatiotemporal structure of data (to use multiple datasets)