Stanford Named Entity Recognizer (NER) & its usage in Law

CS-410 Technology Review – By Sunitha Vijayanarayan (sunitha3)

Abstract:

Named Entity Recognition (NER) labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names. This review is about the NER implementation developed by Stanford and explore some real-world use-cases.

Full Text:

Introduction:

In most documents available online , there will be some known entities most commonly people , organizations or locations. Some medical journals may have medical codes, judicial documents may reference statues , codes or other cases. A core subtask of information extraction seeks to locate and classify named identifiers and entities mentioned in unstructured text into predefined categories and this process is generically termed as Named-entity recognition (NER) (also known as (named) entity identification, entity chunking, and entity extraction) . Stanford NER is one of the popular implementations of Named entity recognition available for use. It comes with well-engineered feature extractors for Named Entity Recognition, and many options for defining feature extractors. Included with the download are good named entity recognizers for English, particularly for the 3 classes (PERSON, ORGANIZATION, LOCATION), and we also make available on this page various other models for different languages and circumstances, which we will explore in some detail in the following paragraphs

Details :

Stanford NER is a Java implementation of a Named Entity Recognizer. Stanford NER is also known as CRFClassifier. The software provides a general implementation of (arbitrary order) linear chain Conditional Random Field (CRF) sequence models. Conditional random fields (CRFs) are a class of statistical modeling method often applied in pattern recognition and machine learning and used for structured prediction. Whereas a classifier predicts a label for a single sample without considering "neighboring" samples, a CRF can take context into account. To do so, the prediction is modeled as a graphical model, which implements dependencies between the predictions. Here, by training your own models on labeled data, you can use this code to build sequence models for NER or any other task. Stanford provides two online web-based interfaces for users to try their NER capabilities. They are available at Stanford NER CRF classifiers and Stanford Core NLP . It is also available under a GNU General Public License for download and use.

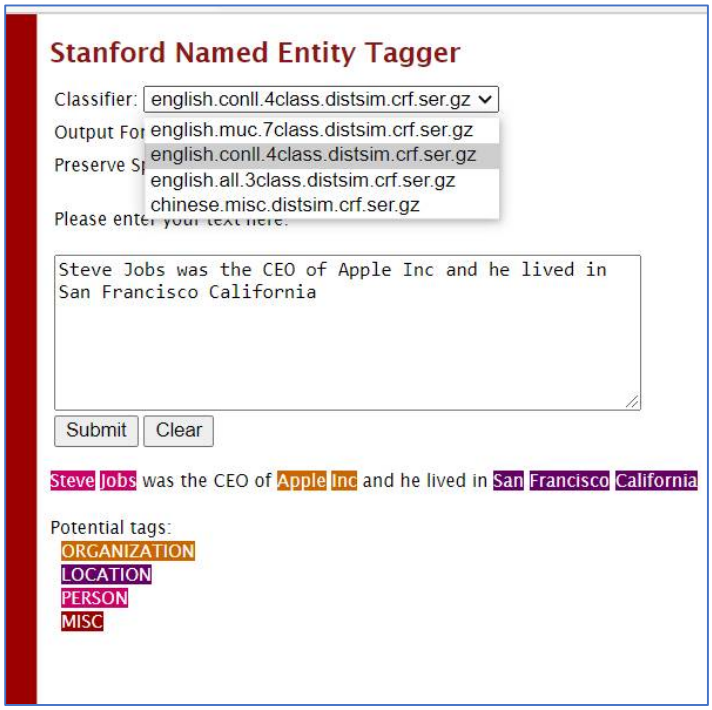
One of the biggest advantages of Stanford NER is that we can download the code and use the models to train for more specific use cases. For example , we could use it to annotate all the novels by a certain author or to cross link knowledge sharing articles authored by all employees in a company, identify citations within government documents and so on.

Experiments& Examples:

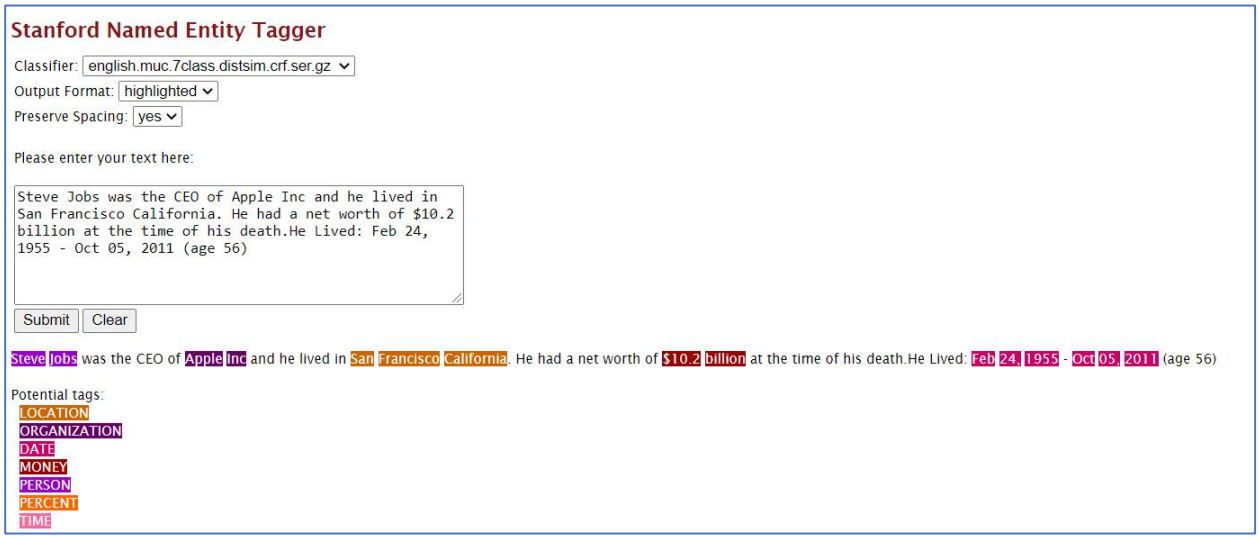
After reading through the documentation and looking at some details on CRF classifiers, I decided to try out the web interfaces to see how Stanford NER CRF classifiers worked.

Experiments on the Stanford Named Entity Tagger :

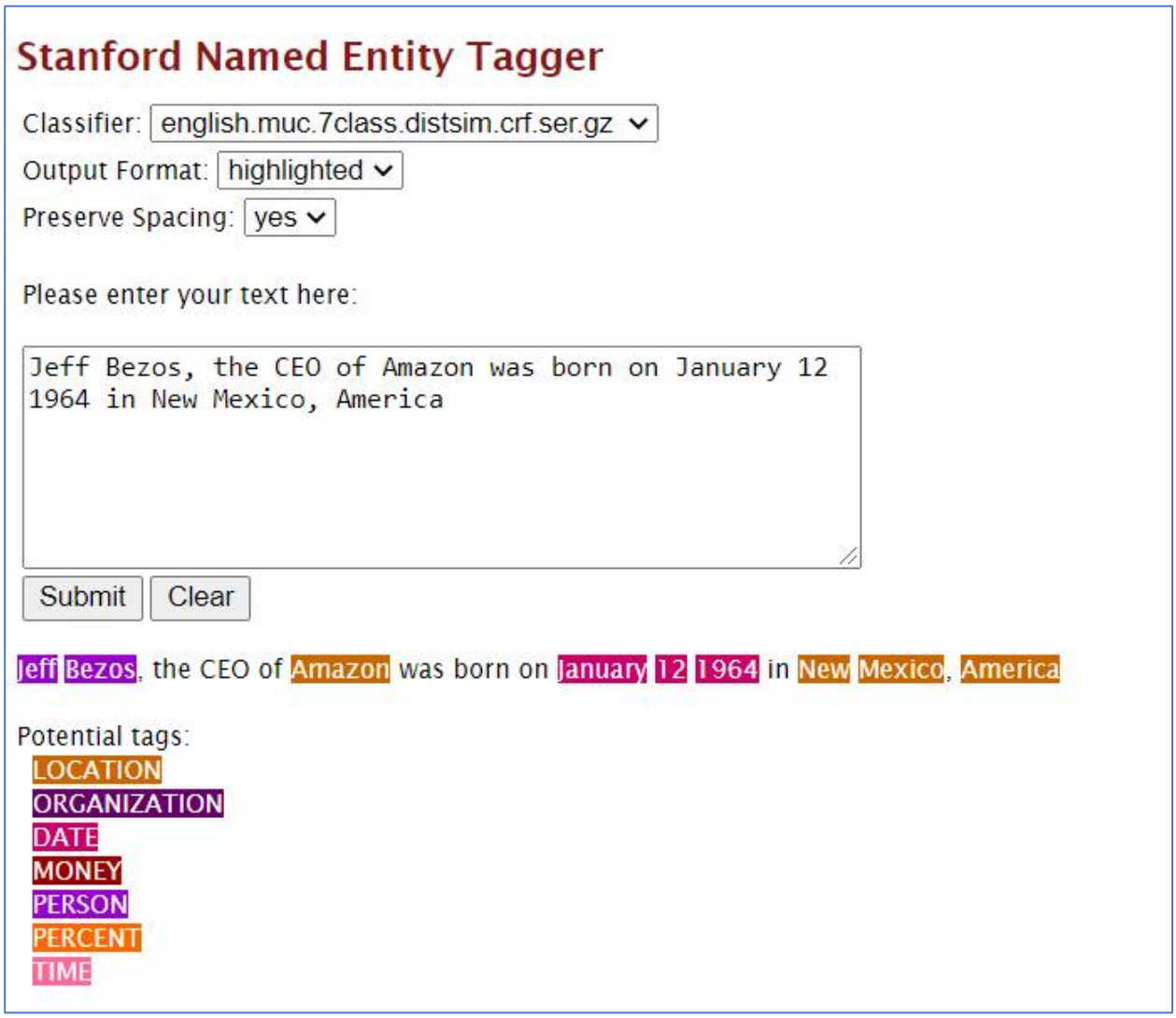
There are 4 classifiers available. The English classifiers are differentiated by the number of entity types they are trained to recognize. Here I tried the 4-class classifier which successfully identified a person , organization , and two locations.



The 7-class classifier was able to identify additional types of entities like Dates, money, percentage and time.



The 3-class option did not have the MISC classification which in my experiments I did not hit anyway, and I did not try the Chinese classifier to check accuracy. In my various experiments , here is a quirk I saw. Sometimes the NER wrongly tagged some organization names as places or people like the below. All the classifiers made the same mistake of marking Amazon as a location rather than an organization even though obvious clues were present.



Customizations:

Stanford NER allows us to customize and add our own list of entities. This is really helpful if you have a better list of organizations or people or locations that you are looking for in the documents and want to classify as much of them as possible. In the above example where NER mis-classified Amazon as a location , we can rectify it by modifying the NLP code and adding our own "Regexner" with the missed location. This will ensure that Amazon will be classified as a company instead of a location.

It is also possible to define custom types or custom entities by training the model on a different set of entities. Given enough training ,Stanford NER can identify brands, car models , Case citations , US Code and many more.

Comparisons with other NER :

Some comparisons have been performed by T D Perry [i]on the different NER tools available like Core NLP, Spacey and Flair. From the experiments conducted, for Legal data which my company is interested in , CoreNLP far outperforms the other offerings. Incidentally , another comparison study between Spacey and CoreNLP (Stanford NER) by Ori Cohen [ii], Stanford provided better results , though Spacey is orders of magnitude faster.

Applications in Legal entity identification :

Stanford NLP is a viable option to identify and tag US Code Citations on legal documents. A typical US Code citation looks like this. **42 U.S.C. § 1983 (2006).** Stanford NER model can be trained to recognize text in this pattern as a code citation and this can be used by applications to link the citations to an actual page that details the law. Such an application would be very helpful for Lawyers and law students to refer laws in real time while reading a case.

Conclusion:

Stanford NER is a very robust and useful tool for Named Entity Recognition. Though it has some quirks , it is easily customizable to be used in a variety of use-cases.

References

- https://www.lighttag.io/blog/spacy-vs-stanford/
- https://towardsdatascience.com/a-comparison-between-spacy-ner-stanford-ner-using-all-us-city-names-c4b6a547290
- https://nlp.stanford.edu/software/CRF-NER.shtml
- https://nlp.stanford.edu/projects/project-ner.shtml

[i] https://www.lighttag.io/blog/spacy-vs-stanford/

[ii] https://towardsdatascience.com/a-comparison-between-spacy-ner-stanford-ner-using-all-us-city-names-c4b6a547290