Northwestern University

Named Entity Recognition on Business Insider Articles

Aarij Rehman

March 6th, 2020

*Introduction*

Business Insider is a portal for business news. Articles are posted daily about different financial, economic, and global topics. For that reason, Business Insider is an ideal candidate for natural language processing.

In this project, we’re analyzing an article from every day in 2013-2014. From these articles, we’re interested in extracting three particular pieces of information:

1. Names of CEOs
2. Names of Companies
3. Percentages

We use a combination of regular expressions and logistic regressions to build NERs, named entity recognizers. Ideally, we can give the NER features of any potential candidate, and the NER will be able to predict whether the candidate belongs to its class of item.

*Structure*

The corpus is the collection of two years of Business Insider articles. I’ve tokenized each article by sentence and combined all those sentences into a single list. This is acceptable because we’re concerning ourselves only with entity extraction, so all sentences can be viewed equally.

*CANDIDATE EXTRACTION TOO*

*Extracting CEO Names*

When extracting CEO names, a NER based on logistic regression was built. The most important aspect of the NER is the features which categorize a candidate. Features either consider the sentence with the candidate, the candidate, or a combination of both. The features implemented are listed and described below:

1. length\_sentence
2. ceo\_in\_sentence
3. length\_candidate
4. stop\_in\_name

*Extracting Company Names*

When extracting company names, a NER based on logistic regression was built. This is because in the context of a sentence, it’s easier to make predictions about

1. length\_sentence
2. ceo\_in\_sentence
3. length\_candidate
4. stop\_in\_name

*Extracting Percentages*

When extracting company names, a NER based on logistic regression was built. This is because in the context of a sentence, it’s easier to make predictions about

1. percentage after number
2. ‘percentage’ after number