**Information Extraction** – là tiến trình rút trích ra các thông tin có cấu trúc một cách tự động từ các nguồn dữ liệu không cấu trúc hay bán cấu trúc (unstructured/semi-structure) ví dụ như các tài liệu văn bản hay các trang web.

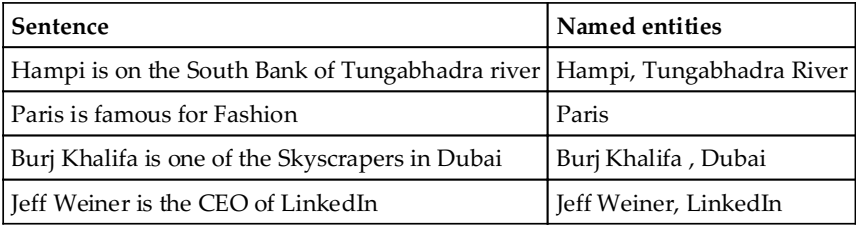


Hiểu các thực thể được đặt tên

Cho đến nay, chúng ta đã thấy làm thế nào để phân tích văn bản, xác định các phần của lời nói và trích xuất các đoạn từ các văn bản. Điều tiếp theo mà chúng ta cần xem xét là danh từ propernoun, cũng là gọi là thực thể có tên.

Các thực thể được đặt tên giúp chúng tôi hiểu thêm về những gì đang được đề cập trong một văn bản nhất định để rằng chúng ta có thể phân loại thêm dữ liệu. Vì các thực thể được đặt tên bao gồm nhiều hơn một từ, nó đôi khi rất khó để tìm thấy những điều này từ văn bản.

Hãy lấy các ví dụ sau để hiểu thực thể được đặt tên là gì:



Hãy xem xét kỹ hơn những điều này và cố gắng hiểu:

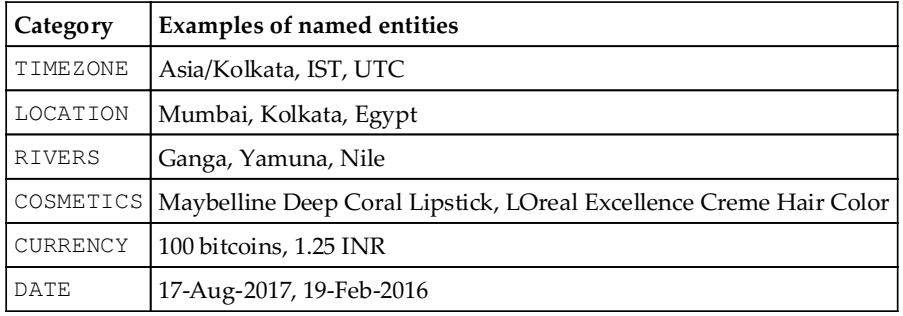
1.Mặc dù Ngân hàng Nam đề cập đến một hướng, nhưng nó không đủ điều kiện là một thực thể được đặt tên bởi vì chúng ta không thể xác định duy nhất đối tượng từ đó.

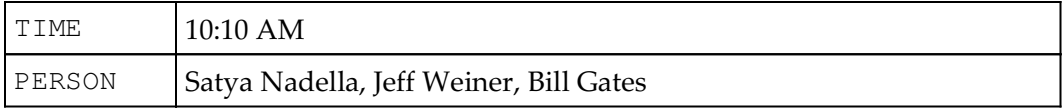
2.Mặc dù Thời trang là một danh từ, chúng ta không thể hoàn toàn đủ điều kiện nó là thực thể được đặt tên.

3.Nhà chọc trời là một danh từ, nhưng có thể có nhiều khả năng cho Nhà chọc trời.

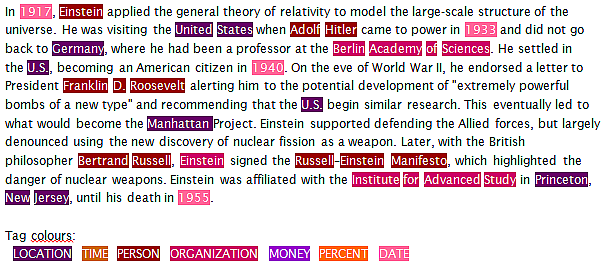
4.CEO là một vai trò ở đây; có rất nhiều người có thể giữ danh hiệu này.

Vì thế,đây cũng không thể là một thực thể được đặt tên.





**Named Entity Recognition (NER)** – là tiến trình xác định và phân loại các phần tử trong văn bản vào các danh mục được định nghĩa trước như tên người, tên tổ chức, địa điểm, giá trị tiền tệ, tỷ lệ phần trăm,…



Sử dụng NERs sẵn có

Python NLTK đã tích hợp hỗ trợ cho Nhận dạng thực thể có tên (NER). Để sử dụng tính năng này, đầu tiên chúng tôi

 cần phải nhớ lại những gì chúng ta đã làm cho đến nay:

1.Break một tài liệu lớn thành câu.

2.Break câu thành từ (hoặc mã thông báo)

3.Xác định các phần của lời nói trong câu.

4. Trích xuất các đoạn của các từ liên tiếp (không chồng chéo) khỏi câu.

5.Đăng ký các thẻ IOB cho các từ này dựa trên các mẫu phân đoạn.

Bước hợp lý tiếp theo sẽ là mở rộng hơn nữa các thuật toán để tìm ra các thực thể được đặt tên là bước thứ sáu. Vì vậy, chúng tôi sẽ cơ bản được sử dụng dữ liệu được xử lý trước khi bước 5 làmột phần của ví dụ này.

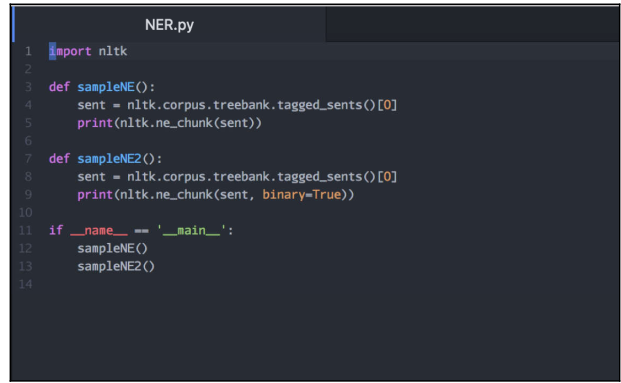
Chúng tôi sẽ sử dụng dữ liệu treebank để hiểu quy trình NER. Hãy nhớ rằng, dữ liệu đã được gắn thẻ trước ở định dạng IOB. Không có quá trình đào tạo, không có thuật toán nào được thấy ở đây đang đi làm (Vì vậy, không có phép thuật!)

Để hiểu được tầm quan trọng của quá trình đào tạo, chúng ta hãy lấy một ví dụ. Nói, cần có bộ phận khảo cổ để tìm ra địa điểm nổi tiếng nào ở Ấn Độ đang được tweet và đề cập trong các trang web mạng xã hội ở Kannada

Ngôn ngữ.

Giả sử rằng họ đã có dữ liệu ở đâu đó và nó được tính bằng terabyte hoặc thậm chí là petabyte, làm thế nào để họ tìm ra tất cả những tên này? Đây là nơi chúng ta cần lấy một tập dữ liệu mẫu từ đầu vào ban đầu và thực hiện quy trình đào tạo để tiếp tục sử dụng bộ dữ liệu được đào tạo này để trích xuất các thực thể được đặt tên trong Kannada.

Tạo một tệp mới gọi là NER.py



You will see the following output:

