PGCOMP-UFBA 2022
MATE34 - Atividade 01

No contexto de PLN - Processamento de Linguagem Natural, realizar um experimento de REN - Reconhecimento de Entidades Nomeadas.

O experimento consiste em reconhecer as entidades 'pessoa' e 'localidade' em textos de contexto geral escritos na língua portuguesa.

- tipo: core (vocabulary, syntax, entities, vectors)
- gênero: texto escrito (notícias, mídia)
- tamanho: LG (541 MB)
- componentes: tok2vec, morphologizer, parser, lemmatizer, senter, atribute_ruler, ner
- pipeline: tok2vec, morphologizer, parser, lemmatizer, attribute_ruler, ner
- vetores: 500 mil chaves, 500 mil vetores exclusivos (300 dimensões)
- fontes:
  - corpus Bosque (9.368 frases), formalmente UD-Portuguese-Bosque 2.8 (https://www.puc-rio.br/ensinopesq/ccpg/pibic/relatorio_resumo2018/relatorios_pdf/ctch/LET/LET-Luisa Rocha.pdf)
  - WikiNER
  - Explosion fastText Vector ...
  - autor: Explosion

O spaCy tem componentes que permitem o reconhecimento de entidades (NER)

O EntityRuler é um destes componentes, ele permite adicionar entidades nomeadas com base em dicionários de padrões, o que facilita a combinação de reconhecimento de entidades nomeadas baseado em regras e estatísticas para pipelines ainda mais poderosos.

Entity patterns são dicionários com duas chaves: "label", especificando o rótulo a ser atribuído à entidade se o padrão for correspondido, e "pattern", o padrão de correspondência. EntityRuler aceita dois tipos de padrões: Phrase patterns (para a string exata) e Token patterns (lista).