Extração de Entidades Nomeadas

Aluizio Lima
Christian Cardozo

Extração de Entidades Nomeadas

- Framework: NLTK
- Estratégia:
  - Passo 1: Tokenização das palavras com *nltk.word_tokenize*
  - Passo 2: Tagueamento gramatical com *nltk.pos_tag*
  - Passo 3: Construção da árvore sintática classificando entidades nomeadas com *nltk.ne_chunk* (ne = named entities)
  - Passo 4: analisar árvore sintática
    - 4.1: Buscar nós com label: “GPE” (geo-political entities), “PERSON” e “ORGANIZATION”
    - 4.2: Validar se a primeira letra é maiúscula e se não há ocorrências de minúsculas seguidas de maiúsculas (Exceto para strings com a substring “mc”)
    - 4.3: Se passou nos testes, adicionar à lista de entidades.

Resultado

- 3779 entidades nomeadas
  - Exemplos:
    - Missandei
    - Eurasia
    - Queen Rhaella Targaryen
    - Catelyn V
    - Sam IV
    - Ser Ilyn
    - Queen Margaery Tyrell
    - Master Mighdal

Repositório do código

- Disponível no GitHub:
  - https://github.com/NLP-TESI/NamedEntitiesTESI