Tóp. Especiais em Sistemas Inteligentes

Aluizio Lima
Christian Cardozo

Identificação de Entidades Nomeadas

- Tokenizar sentenças
- Para cada sentença:
  - Tokenizar palavras
  - Classificação gramatical(POS_Tagger)
  - Árvore sintática(ne_chunk)
  - Identificação de Casas(“House <nome>”) como NE
  - Identificação de entidades pela sequência de “nltk.tree.Tree” e “NNP”
- Cada episódio terá um dicionário de entidades
- Mantém-se um dicionário global: atualizado a cada análise de episódio

Identificação de Entidades Nomeadas

- Identificação de entidades equivalentes:
  - Tentar encaixar cada entidade encontrada localmente(no episódio) em alguma entidade já salva no dicionário global
  - Remoção de “palavras de honra”: ser, lord, commander, king, queen, etc
  - Busca-se a melhor opção no dicionário global. São candidatos:
    - Entidades com “match” de termos
    - Entidades com similaridade média maior que 0.7

Identificação de Entidades Nomeadas

- Similaridade entre termos
- Comparar str1 e str2
  - Mantém apenas letras A-Z
  - “Índice de pertinência”:
    - Se “str1 in str2” ou “str2 in str1”
      - Se len(str1) == len(str2) == 1 , índice igual à 0.5
      - Se não, índice igual à 1
    - Se não, índice igual à 0
  - Similaridade: indice_pertinencia * 0.4 + jaro_winkler * 0.6

Identificação de Entidades Nomeadas

- Análise dos candidatos à entidade nomeada:
  - Se não há candidatos, cria-se uma nova entrada
  - Se há apenas um candidato, atribuímos a entidade ao candidato em questão
  - Em caso de mais de um candidato, escolhe-se aquele com maior frequência
- Cada entidade possui um inteiro como identificador único
  - Ex.:
    - id = 20
    - Valores = [‘Arya’, ‘Arya Stark’] (Referência de id = 20)
- As entidades são salvas em um arquivo CSV
- Salva-se também um texto “taggeado”
- 5431 entidades (756 entidades distintas no dicionário global)

Identificação de Entidades Nomeadas

- Exemplo de entidades encontradas:
  - Daenerys Targaryen;Daenerys;Queen Daenerys;Queen Daenerys Targaryen;Princess Daenerys Targaryen;Daenerys Stormborn
  - Lord Eddard;Eddard;Eddard Stark;Edd;Lord Eddard Stark
  - Prince Oberyn;Prince Oberyn Martell;Oberyn Martell;Oberyn
  - Theon Greyjoy;Theon
  - Cersei;Queen Cersei Lannister;Cersei Lannister;Queen Cersei;Queen Regent Cersei;Queen Regent Cersei Lannister
  - Jon Arryn;Lord Jon Arryn
  - Jon; Jon Snow;Jon Sn;Lord Commander Jon Snow

Identificação de Entidades Nomeadas

- Texto “taggeado”:

King **<entity class="NE" id=0>Robert Baratheon</entity>** majestically arrives in **<entity class="NE" id=1>Winterfell</entity>** , the home of his old and trusted friend , **<entity class="NE" id=2>Eddard Stark</entity>** , **<entity class="NE" id=3>Warden</entity>** of the **<entity class="NE" id=4>North</entity>** , with an important offer . On the eastern continent , the exiled **<entity class="NE" id=5>Princess Daenerys Targaryen</entity>** marries **<entity class="NE" id=6>Khal Drogo</entity>** , a warlord of the **<entity class="NE" id=7>Dothraki</entity>** with tens of thousands of warriors at his command.

Identificação de relações

- Regras para relações entre entidades:
  - Entidade + IN (prep.) + DT (deter.) + Entidade
  - Entidade + [sequência de verbos] + Entidade
  - Entidade + ? (qualquer classificação) + Entidade
- 6719 relacionamentos encontrados

Identificação de relações

- Exemplos (Entidade 1, relação, Entidade 2):
  - Robert Baratheon,arrives,Winterfell
  - Princess Daenerys Targaryen,marries,Khal Drogo
  - White Walker,arrives and kills,Ser Waymar
  - Will,return to warn,Castle Black
  - Ser Jaime Lannister,of the,Kingsguard
  - Jon,of the,Night's Watch
  - Dany,'s,Dothraki
  - Eddard,on the,Kingsroad

TF-IDF

- Tokeniza o texto “taggeado”
- Entidades nomeadas serão tokens no formado “__id__NUM”
- Normalização dos tokens:
  - Lower Case
  - Remoção de pontuação
  - Mantém-se apenas alfanuméricos
- Remoção de Stop Words (utilizando a lista do scikit)
- Cálculo do TF-IDF
  - Não é construído a matriz do espaço vetorial
  - É mantido um dicionário para cada documento(episódio) representando o vetor de tokens presentes naquele documento

TF-IDF

- Resultado de exemplo 1 - SE01EP01:
  - 1º Eddard Star = 0.16767325734416189
  - 2º Robert Baratheon = 0.10670116376446666
  - 3º Khal Drogo = 0.08383662867208094
  - 4º Viserys = 0.08383662867208094
  - 5º Daenerys Targaryen = 0.08383662867208094
- Resultado de exemplo 2 - SE04EP10:
  - 1º Jon Snow = 0.1585113589481058
  - 2º Tyrion Lannister = 0.14631817749055923
  - 3º Tywin Lannister = 0.10973863311791941
  - 4º Mance Rayder = 0.10567423929873722
  - 5º Bran Stark = 0.06096590728773301

Repositório no GitHub

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