Tóp. Especiais em Sistemas Inteligentes

Aluizio Lima
Christian Cardozo

Pré-processamento dos textos

- Nome e número do episódio
- Mortes
- Texto limpo:
  - Remover início: “Contents[show]”
  - Remover final: “Recap”, “RecapEdit”, etc
  - Remover linhas irrelevantes: “Plot”, “Summary”, etc
  - Replace de toda ocorrência de “Edit”
  - Replace em casos de dois espaços(“ “), dois pontos(“..”), etc

Extração das entidades nomeadas

- Tokenizar sentenças
- Para cada sentença:
  - Tokenizar palavras
  - Classificação gramatical(POS_Tagger)
  - Árvore sintática(ne_chunk)
  - Análise da árvore: “GPE”, “PERSON” e “ORGANIZATION”
- Resultado inicial: 3779 entidades

Identificação de entidades

- Dicionário(Bag Of Entities)
  - Chave: termo
  - Valor: lista de entidades similares
- Para cada entidade do processo de extração
  - Se já existe uma entrada no dicionário, ir para próxima iteração
  - Encontrar lista mais próxima usando similaridade de strings
    - jaro_winkler * 0.8 + SmithWaterman * 0.1 + nlevenshtein * 0.1
    - Média >= 0.79
  - Se encontrar, adiciona na lista
  - Se não, cria uma nova lista com o termo atual e inclui a entrada no dicionário
- 532 entidades

Identificação de entidades

- Resultados “bons”:
  - Bran,Bran Stark,Brandon
  - Jaqen,Jaqen H'ghar
  - Daenerys Targaryen,Daenerys,Dany,Daenerys Stormborn
  - Roose Bolton,Roose
  - Jon,Jon Snow,Jon Stark,Jon Sn
  - Samwell Tarly,Samwell,Samwall
- Resultados “ruins”:
  - Lord,Lord Commander,Lord Eddard Stark,Lord Commander Jeor Mormont
  - Greyjoy,Gregor,Greyjoys,Grey,Grey Worm,Greyscale,Greyguard
  - Queen Cersei,Queen Regent Cersei,Queen Regent,Queen Selyse
  - Littlefinger,Little Sam

Identificação de relações

- Separar texto em sentenças
- Identificar tokens que são entidades nomeadas
- Gerar duplas de entidades seguidas na sentença
- Procurar pelo verbo entre as duas entidades na sentença

Repositório no GitHub

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