Tóp. Especiais em Sistemas Inteligentes

Aluizio Lima
Christian Cardozo

Pré-processamento da base de dados

- Arquivos gerados
  - Nome do episódio
    - nome do arquivo
  - Número do episódio
    - Exp. Regular: \b[Ee]pisode [0-9][0-9]?
  - Mortes no episódio
    - 'Deaths', 'DeathsEdit' ou 'Deaths Edit'
  - Texto pré-processado(“limpo”)
    - Corte inicial: “Contents[show]”
    - Remoção de linhas inúteis: 'Contents[show]', 'PlotEdit', 'Plot Edit', 'Synopsis', etc.
    - Corte final: 'Recap', 'RecapEdit', 'Appearances', 'AppearancesEdit', 'Appearances Edit'
  - Ajustes finais
    - Remoção de ‘\n’, sequências de pontos finais e espaços, remoção do texto ‘Edit’

Identificação de Entidades Nomeadas

- Busca de entidades locais (por episódio)
- Mantém-se um dicionário global: atualizado a cada análise de episódio
- Para cada episódio:
  - Passo 1: Tokenizar texto por sentenças
  - Passo 2: Análise da sentença
    - Tokenizar sentença por palavras
    - Aplicar POS Tagger
    - Aplicar ne_chunk
  - Passo 3: Extração das entidades
    - Regra 1: Entidade Nomeada + (Entidade Nomeada | NNP)*
    - Regra 2: Entidade Nomeada + “‘s” + “Watch”
    - Exceções: '–', '–', '—', '[', '', 'imp', 'beyond', 'house', 'ser ser'

Identificação de Entidades Nomeadas

- Passo 4: Adicionar entidades locais ao dicionário global
  - Entidades Candidatas
    - Já possuem o termo na lista
    - Similaridade média >= 0.7
      - Remoção de “palavras de honra”: ser, lord, commander, king, etc
      - similaridade(str1, str2) = Pertinencia(str1, str2)*0.4 + JaroWinkler(str1, str2)*0.6
      - Índice de pertinencia
        - 0 se nenhuma string é substring da outra
        - 0.5 caso uma é substring da outra e a substring tem apenas um termo e é o último termo da string maior. Ex.: Ned Stark e Stark
        - 1 caso contrário. Ex.: Arya e Arya Stark
  - A entidade candidata com maior frequência é escolhida
  - Nova entrada no dicionário global caso não haja entidades candidatas

Identificação de Entidades Nomeadas

- Cada entidade possui um id próprio e aponta para uma entidade pai
- Texto Taggeado <entity></entity>
  - Id: identificador da entidade
  - class: HSE para as casas das famílias e NE para as outras entidades
- 5431 entidades encontradas (sem “mergear” no dicionário global)
- 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
    - Se houver sequência de verbos que tenha ‘,’ são removidos.
  - Entidade + ? (qualquer classificação) + Entidade
- 6595 relacionamentos encontrados
- 2806 relações diferentes

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)
- Para tokens que não são entidades nomeadas: Porter Stemmer
- 10424 tokens (756 entidades nomeadas)

TF-IDF

- Primeira versão 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

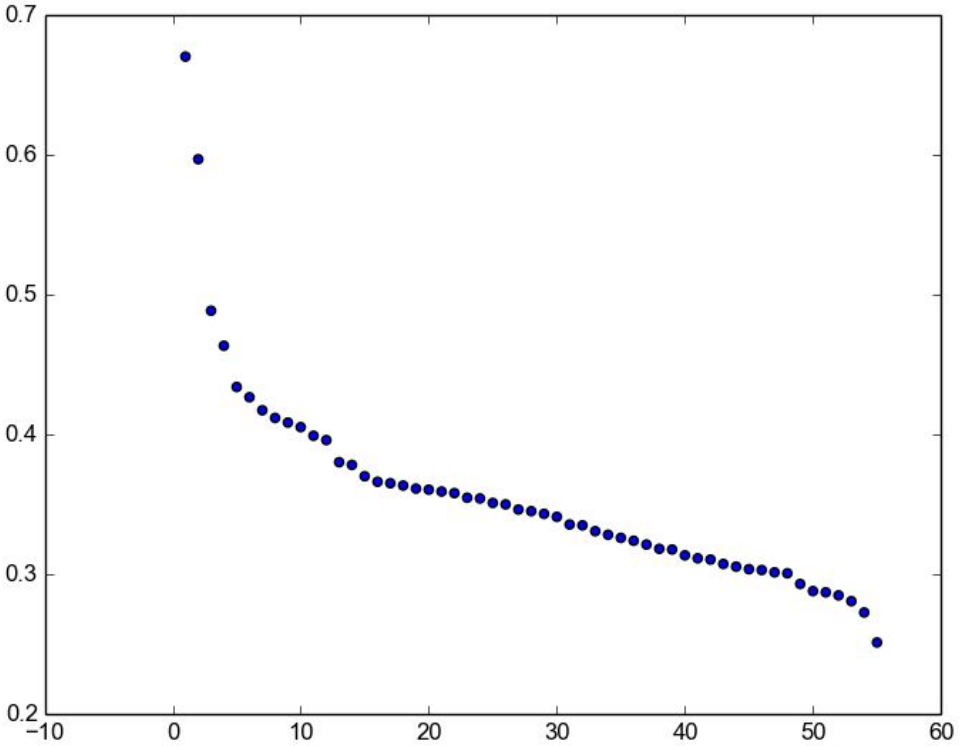
- Resultado de exemplo 1 - SE01EP01:
  - 1º Eddard Star = 0.16767325734416189
  - 2º Robert Baratheon = 0.10670116376446666
  - 3º Khal Drogo = 0.08383662867208094
  - 4º Viserys = 0.08383662867208094
  - 5º Daenerys Targaryen = 0.08383662867208094
- Resultado de exemplo 2 - SE04EP10:
  - 1º Jon Snow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

TF-IDF

- Segunda versão do TF-IDF
  - A matriz de termos vs. documentos é construída
  - Matriz 10424 x 55 = 573320 elementos (284825 células iguais à zero)
- Implementação do SVD para redução de dimensões da matriz
- K = 30

TF-IDF

- Gráfico K vs. Sigma



TF-IDF

- Busca de relevância para termos:
  - Identificar entidades nomeadas na query
    - Substituídas por __id__NUM
  - Tokens que não são entidades nomeadas
    - Remoção de pontuação e mantém-se apenas Alfanuméricos
    - Potter Stemmer
    - Minúscula
  - Similaridade Query x Documento: cosseno entre os vetores

TF-IDF

- Exemplo 1 de Query: “Battle of HARDHOME” (SE05 EP08)
  - TF-IDF sem SVD:
    - SE5 EP8 = 0.87579589937
    - SE3 EP1 = 0.932694443068
    - SE3 EP3 = 0.947773983468
    - SE5 EP10 = 0.958392269231
    - SE5 EP5 = 0.966069487072
  - TF-IDF com SVD:
    - SE5 EP8: 0.156036884294
    - SE5 EP5: 0.245279537001
    - SE3 EP1: 0.246801204662
    - SE5 EP10: 0.379971556554
    - SE3 EP3: 0.379994993036

TF-IDF

- Exemplo 2 de Query: “hold the door” (SE06 EP05)
  - TF-IDF sem SVD:
    - SE6 EP5: 0.888711651866
    - SE2 EP10: 0.930224497021
    - SE6 EP1: 0.932407403705
    - SE1 EP8: 0.935534782798
    - SE5 EP6: 0.945981698144
  - TF-IDF com SVD:
    - SE6 EP5: 0.888722716615
    - SE2 EP10: 0.930396452928
    - SE6 EP1: 0.932397856588
    - SE1 EP8: 0.93553658066
    - SE5 EP6: 0.945912312601

Repositório no GitHub

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