PNL em Game of Thrones

TESI 2 - 2016.2

Limpando o texto

- A ideia é pegar somente a história do episódio em si.
  - Ou seja, pega tudo nas seções Plot e Summary

- Remove linhas em branco

- Remove caracteres não ASCII como “…”, “—” e “\u200B”

Encontrando entidades

- Entidades nomeadas geralmente começam com letra maiúscula

- São consideradas possíveis entidades sequencias de palavras com letras maiúsculas
  - Podem ter “of”, “the”, “‘s” e ‘’ entre os nomes

- Entre as possíveis, só são aceitas as que possuem pelo menos um token com as seguintes classes: “NN”, “NNP”, “NNS”, “NNPS”

Encontrando entidades

- Problema: Inicio de frases
  - Ex: “Enduring Lord Walder Frey 's insults directed at him(….)”
- Problema: Diálogos no meio do texto
  - Ex: “Daenerys turns (...) and bluntly says "A dragon is not a slave," (...)”

- Solução:
  - Dada uma entidade de começo de frase/diálogo
  - Caso sua primeira palavra seja “The” ou um titulo como “Lord”, aceita a entidade
  - Caso sua primeira palavra não apareça no texto com letras minúsculas, aceita a entidade
  - Fora esses casos, ignora a entidade

Agrupando e Classificando

- Usa conhecimento ‘hardcoded’ de titulos, casas e formas de se referir a lugares.
  - Titulos: “Princess”, “Lady”, “Commander”, “Lord Commander of”, ...
  - Casas: sempre possuem só um nome e são referenciadas por “House” + esse nome
  - Lugares: são geralmente precedidos de “return from”, “arrive at”, “back to”, “visit to”, ...

- Passos sequenciais:
  - Se uma entidade começa com um titulo, o nome canonico é sem o titulo e a entidade é uma pessoa. Ex:
    - “King Robert”, “Lord Snow”, “Lady Catelyn”

Agrupando e Classificando

- Passos sequenciais:
  - Procura por “House” + algo. Esses “algos” são consideradas casas.
    - Ex: “House Stark”, “House Targaryen”, “House Reed”
  - Se houver uma entidade que só tenha duas palavras e termina com uma casa, é uma pessoa.
    - Ex: “Eddard Stark”, “Daenerys Targaryen”, “Howland Reed”
  - Dada uma entidade anterior, se outra possui só um nome e é igual ao primeiro dela, elas são a mesma entidade.
    - Ex: “Eddard”, “Daenerys”, “Howland”

Agrupando e Classificando

- Passos sequenciais:
  - Procura se as palavras anteriores a cada entidade ainda não marcada possuem as expressões que indicam lugar
  - Se ela estiver depois de pelo menos 5 expressões de lugar, então é um lugar

- Não pensamos em nenhuma regra consistente para outras classes =(

Entidades

- Encontramos 13877 entidades, sendo 1273 unicas
- Analizando manualmente, em torno de 150 estão erradas
- Agrupando, temos 1073

Entidades

- Das 13877 entidades,
  - Entidades com “other”: 5277 (38%)
  - Entidades com “person”: 6977 (50%)
  - Entidades com “place”: 1529 (11%)
- Das 1073 entidades unicas,
  - Entidades com “other”: 891 (69%)
  - Entidades com “person”: 329 (25%)
  - Entidades com “place”: 31 (2%)

Encontrando relações

- Procurar por verbos entre entidades
  - Tokens com classes "VB", "VBG", "VBZ", "VBN"

- Se duas entidades possuem um token com classe de verbo, forma uma relação entre as entidades com o tipo token

Encontrando relações

- Resultados bons:
  - (Jon Stark, leave, Night 's Watch)
  - (Cersei Lannister, obey, Robert Baratheon)

- Resultados bens ruins...
  - (Sandor Clegane, is, Joffrey Baratheon)
  - (Gregor Clegane, burying, Gregor Clegane)

Encontrando relações

- 2322 triplas encontradas
- 998 relações
- Não tivemos boas ideias para relações =(

Marcando Entidades

- Substituir os nomes das entidades por placeholders
- Depois substituir os placeholders por:

  <entidade name=”nome canonico” type=”classificação”>

  nome original

  </entidade>

- Problema: Substituir “Tyrion” antes de “Tyrion Lanninster”
  - Solução: organiza entidades pelo tamanho do nome de forma crescente e substitui nessa ordem

TF-IDF

- Número de palavras na coleção: 8549
  - Removendo tokens como “ ‘m ”, “ ‘d “ e pontuações

- Quantidade de documentos: 56

- Tamanho da matriz do TF-IDF: 56 x 8549

TF-IDF

- Buscando por “Eddard Stark”, temos:
  - season_1_the_wolf_and_the_lion
  - season_1_winter_is_coming
  - season_1_you_win_or_you_die
  - season_1_cripples,_bastards_and_broken_things
  - season_1_lord_snow
- Buscando por “Battle of the Blackwater” temos:
  - season_3_valar_dohaeris
  - season_3_kissed_by_fire
  - season_3_walk_of_punishment
  - season_2_blackwater
  - season_3_and_now_his_watch_is_ended.txt

Fim