A pré-execução do Snorkel

Etapas

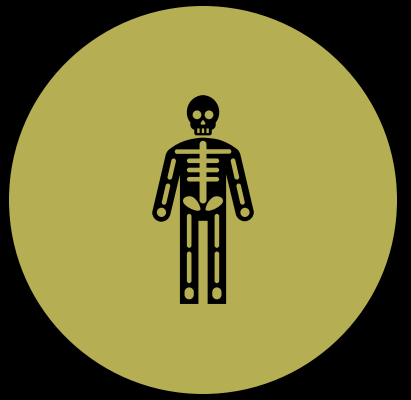
PRÉ-PROCESSAMENTO DE DADOS & GERAÇÃO DE CANDIDATOS

DEFINIÇÃO DAS GOLDEN_LABELS (MODELO)

DEFINIÇÃO DAS LABELING_FUNCTIONS (MINERAÇÃO(?))

Pré-Processamento de Dados

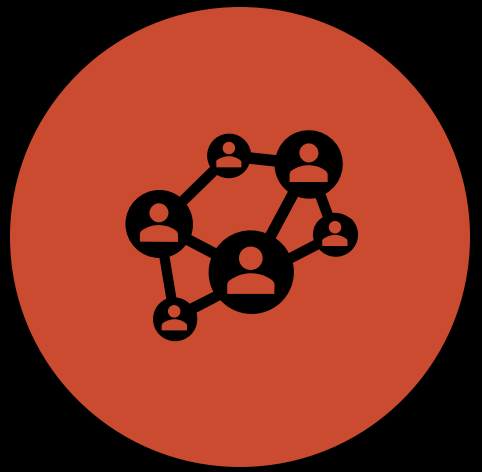
FORMATAÇÃO DO CORPUS EM .TSV (CONCLUÍDO)



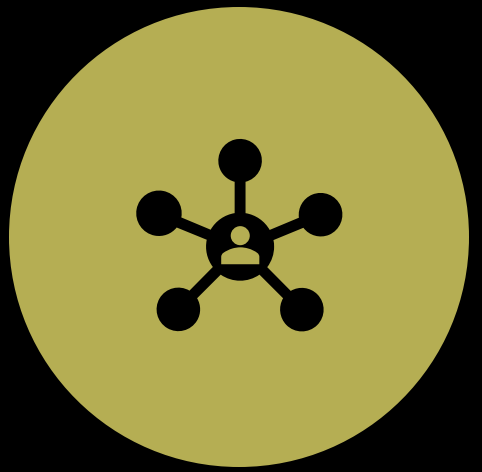
AGRUPAMENTO DOS DICIONÁRIOS TOPONÍMICOS RELEVANTES EM SUBSTÂNCIAS E EVENTOS ADVERSOS (CONCLUÍDO)

DEFINIR OS DICTIONARYMATCHERS COM OS DICIONÁRIOS (CONCLUÍDO)

Extração de Candidatos



DEFINIR A RELAÇÃO DE CANDIDATOS (APENAS A RELAÇÃO SUBSTÂNCIA->EVENTO FOI DEFINIDA)



DEFINIR O EXTRATOR DE CANDIDATO DA RELAÇÃO COM OS MATCHERS DEFINIDOS

Problemas com a Extração

- Dicionários Poluídos:
  - Palavras como "e", "um", "anti" e "sene" foram observadas como candidatos para Substâncias além de uma formatação estranha para espaços (\t).
    - Dúvidas:
      - Utilizar um stemmer? (tempo de execução, acurácia ou ambos?)
      - Utilizar stop words? (O que são e onde encontrá-las?)
      - Algum outro método para filtrar os dicionários?
- N-grama
  - O n-grama que foi usado na extração é o mesmo do tutorial, n_max = 7
    - Dúvidas:
      - Qual o efeito do n-grama na extração?
      - Vale a pena mudar ele?

Golden_Labels

AS GOLDEN_LABELS SERÃO OS EXEMPLOS QUE VÃO SER ALIMENTADOS AO MODELO PARA O TREINAMENTO

NO TUTORIAL, DE 27792 CANDIDATOS, TEMOS 5493 GOLDEN_LABELS, DE PROPORÇÃO 80/20 DE NEGATIVOS/POSITIVOS

Problemas com as Golden_Labels

- Volume:
  - Temos 6000 candidatos apenas na relação Substância->Evento
    - Para manter a proporção teríamos que definir "na mão" 1500 entradas pra Golden_Label (somente para uma relação)
  - Dúvidas:
    - Fabricar essas labels?
      - Analisar a estrutura de algumas frases e alternando as substâncias e eventos para gerar um número de labels
        - Problema:
          - Pode afetar a eficiência do modelo.

Labeling Functions

Funções que avaliam os candidatos extraídos e retornam 1, -1 ou 1, para confirmar a relação, negar ou abster.

Por estrutura:

- Avaliam a estrutura e/ou o conteúdo do candidato.

Por Distant Supervision:

- São funções que checam se a relação existe num grupo de relações já conhecidas.

Problemas com as Labeling Functions

- Estrutura:
  - Por onde começar?
    - Uma ideia é ter um dicionário de verbos que dão a entender a relação de x causou y. Algo como o Alexandre tentou fazer com o filtro semântico.
      - Como criar esse dicionário?
    - Analisar a estrutura dos candidatos e tentar reconhecer algum padrão.
- Distant Supervision:
  - "Rodar" cada item do dicionário de substâncias com cada item do dicionário de eventos.
  - Ou tentar algum scraping com o bulário / algum dataset já conhecido