텍스트 전처리 과정 공유

- 먼저 한가지 말씀드릴 것이 있는데, 저번 주에 확정했던 보조제 리스트에서 중복되는 Ashwagandha를 하나 제거하고, 대신 vitamin D를 추가해서 25개의 보조제 리스트를 다시 확정했습니다. (Vitamin D가 레딧에서 많이 언급되었는데 미처 모르고 있었습니다ㅜㅜ)

- **processing_submissions.py, processing_comments.py**가 전처리 하는 스크립트이고,

  결과를 저장한 파일이
  processed_insomnia_submissions_with_dosages,
  processed_insomnia_comments_with_dosages 입니다.

1. 기본적인 전처리

공백과 개행문자 정리, url 제거, 이메일 주소 제거, 소문자로 변환, 숫자와 단위 보존, 토큰화, 불용어 및 비문자 제거, 원형(lemma)로 복원

---> 이 과정이 완료된 것이 게시글 파일에는 **processed_text,** 댓글 파일에는 **processed_comment** 컬럼에 저장됩니다.

2. 복용량 정보 추출

저희가 사용하는 보조제 리스트에 있는 보조제와 복용량이 언급된 경우에는 정보를 추출하여 **dosages** 컬럼에 저장합니다.

복용량 정보는 mg, ml, pills, capsules, capsule, tablets, tablet oz, mcg 가 하나 이상의 숫자 뒤에 오는 경우를 추출했습니다.

-Dosages 컬럼 예시

1) 용량과 관련 없는 숫자들

: dosages 컬럼에서 빈 배열([ ])로 표시됨.

I've been sleeping a good 7-8 hours nearly

Benadryl will zonk me 50% of the time

it takes me between 3 and 6 hours to fall asleep

➔ [ ]

2) 용량과 관련이 있지만 보조제 관련이 아닌 내용들

: dosages 컬럼에서 용량을 볼 수 있지만 보조제 리스트에 없는 약물이나 음식등에 대한 용량이므로 'unknown'이라고 표시됨.

before I went to bed, I cut up a big (10oz) potato

➔ [('10oz', 'oz', 'unknown')]

bought some "sleep aids" with 50mg diphenhydramine HCL.

➔ [('50mg', 'mg', 'unknown')]

3) 보조제 리스트에 포함된 약물의 용량이 언급된 경우

: 용량과 이름이 함께 표시됨.

I started taking Metamucil in the morning and 5 mg Melatonin at night.

➔ [('5 mg', 'mg', 'melatonin')]

문제점: <여러 가지 보조제와 여러 개의 용량 정보가 포함된 텍스트>

currently take 200mg diphenhydramine 25mg melatonin cycle time prescribed ambien 10mg.....

와 같은 글에서, 3가지 용량 정보가 모두 입력 리스트에 있는 멜라토닌의 용량으로 인식됨

→ [('200mg', 'mg', 'melatonin'), ('25mg', 'mg', 'melatonin'), ('10mg', 'mg', 'melatonin')]

Chat gpt의 제안에 따라 텍스트에서 각 용량 정보의 위치와 보조제 이름의 위치를 찾아서 가장 가까운 것끼리 매핑하는 방법을 사용해봤는데 계속 오류가 나고 아직 해결은 못했습니다.

교수님께 조언을 구해보고 최악의 경우에는 여러 보조제의 용량들이 같이 언급된 글에서 직접 텍스트를 확인해서 용량 정보를 매핑해야 하는지....?도 여쭤보고 싶습니다.