**1.업무 분담 및 workflow**

1. 처리/탐색

- 전처리 부분이 작업량의 대부분을 차지하고 중요하기 때문에 4명이서 동시에 진행하는 것이 낫다는 의견을 반영해 데이터셋 확인해서 필터링 해야할 부분을 각자 찾아내고 공유할 예정(특수기호, 숫자, 이모티콘, slang, 영역을 나눠서 찾아봐야함)

<추가의견>

동언: 동사보다는 명사가 더 의미있는 경우가 많으니 명사만을 따로 추출하는

기능을 써보는 것도 좋을 것 같음(R에서 extractNoun 같은 기능)

지은: 특수기호, 숫자, 이모티콘 등은 정규표현식을 사용한 함수로 한꺼번에

날릴 수 있음, 관건은 오타나 잘못된 띄어쓰기, 은어 처리 부분이라고 생각함.

영문법 오타를 고쳐주는 교정시스템이 있는지 탐색해봤으면 함

1. embedding(vectorize)

- GloVe, TF-IDF 등 여러 word embedding 적용해보기

1. 필터링

- 일괄 필터링이 가능한 부분은 함수를 정의해서 사용

- 노가다가 필요한 부분은 혹시라도 다른 알고리즘을 적용 가능한지

알아봐야함

1. 모델 설계

- 이해도가 높은 조원 2명의 주도 하에 기본형태로 모델 설계할 예정

- 모델 설계 후 은닉층을 추가해보면서 정확도를 높여갈 2명, 노드의 갯수를

추가해보면서 확인할 2명으로 나눠서 진행하고, 각 팀은 어느 지점에서

오버피팅에 빠지는지를 확인해보면서 최적의 모델을 찾을 예정

-각 팀별로 진행한 경우의 수 history는 꼭 정리를 해놔야함!!(ex.은닉층 몇 개에

얼마의 정확도를 냈는지)

-최적의 은닉층과 노드를 찾았다면 두 팀의 결과를 공유하여 하이퍼

파라미터를 조정해 다시 한번 정확도를 체크할 예정(+ 튜닝을 거치면서

정확도가 올라가지 않는다면 다시 처리탐색 부분으로 돌아가 확인하는 작업을

반복할 수 있음)

p.s.

- 광현씨는 집안 사정으로 이번주 하루 빠질 수 있다는 점 양해 부탁드리고, 모델

설계할 때 따로 선생님께서 광현씨에게 설명해주시기로 했습니다

- BERT 라이브러리 다운받아서 모델이 잘 돌아가나 확인을 해봤는데 정~말 오래

걸리는 단점이 발견되었습니다..(무려 2시간째 돌아가는 중) - 어떻게 할지

생각해봐야함…..