2019/7/19

FVID Glassdoor

Intern Task 04

전체 과제

1. 잡플래닛 아이디와 비번을 만들었습니다. 이를 바탕으로 잡플래닛 웹사이트에서 어떻게 데이터를 크롤링해올 수 있을지 생각해주세요. 직접 크롤러를 만드셔도 좋고, 어떤 정보를 크롤링하면 좋을지 고민하셔도 좋습니다.

   아이디: elainespak@gmail.com 비번: glassdoorteam! (느낌표 포함입니다)

2. 저희가 모은 데이터의 리뷰 텍스트를 기반으로 텍스트 마이닝 기법을 적용할 것입니다. 다양한 텍스트 마이닝 기법에 무엇이 있는지 간단하게 공부해 오시면 좋겠습니다. 특히, NLP의 기본 아이디어 중 하나인 Word2vec과 Doc2vec에 대해 공부하시면 좋을 듯합니다. 조금 난이도 있는 개념인데, 전 온라인의 다양한 정보 외에도 데이터마이닝 연구실의 선배님께서 쓰신 이 링크가 특히 유용했습니다.

   https://lovit.github.io/nlp/representation/2018/03/26/word_doc_embedding/

정지수 인턴 과제

1. 상장된 기업이 많을 것 같고 개인적으로 관심 있는 인더스트리를 고르고, 관련 기업 리스트를 10개 정도 작성해 주세요.

이명원 인턴 과제

1. Glassdoor 리뷰 중, 5가지 세부 항목은 평점을 부여하지 않은 리뷰들이 존재합니다. 추후 데이터분석에 있어 이 리뷰들을 어떻게 해야 할지 방법론을 조사해보세요. (예: 아예 분석에서 제거한다. Imputation 방법을 써 proxy 값을 넣는다 등)