2019/7/25

FVID Glassdoor

Intern Task 05

프로젝트 방향 업데이트 (중요! 꼭 자세히 읽어주세요)

1. **Employee Satisfaction Level & Stock Returns**: 평점과 주식값을 일대일로 비교하기는 쉽지 않습니다. 따라서 평점에 따라 회사를 5개로 분류(최상, 상, 중, 하, 최하)하여 각 집단의 회사로만 이루어진 주식 포트폴리오(equally weighted)를 작성하고, 각 집단의 return이 차이를 보이는지 확인할 것입니다.

   A. 이규진 인턴이 찾은 2019년도 논문이 이미 이와 비슷한 작업을 수행한 바 있습니다. 저희는 차별화를 위해 추후에 더 높은 신뢰도를 보이는 평점(긴 리뷰, 근속연도가 긴 사람의 리뷰 등)에 가중치를 더하는 방법 등을 써볼까 합니다.

2. **Company Categorization Based on Employee Text Review with Doc2Vec**: Word2Vec은 NLP에 통용되는 기법입니다. 이를 발전시킨 Doc2Vec의 아이디어를 적용해, 하나의 회사를 하나의 document로 보고, 각 회사의 리뷰데이터를 학습시킬 것입니다. 이후 생성되는 어떤 벡터공간에서 상대적 거리가 가까운 회사들끼리 정말로 비슷한 회사들인지 살펴보고자 합니다.

3. **Database & Github**: 오늘 설명드린대로, 저희가 앞으로 데이터를 더 쉽게 공유하기 위해 적절한 데이터베이스를 사용하고자 합니다. 또한 코드를 공유하기 위해 Github를 사용하기 시작하면 좋을 듯합니다.

4. **Related Literature Research:** Glassdoor과 잡플래닛 데이터를 이용한 논문이 더 많이 발견된 만큼, 꾸준히 관련 논문을 조사해 저희 프로젝트의 차별성을 강구하고자 합니다.

2019/7/25

FVID Glassdoor

Intern Task 05

다음 단계 및 역할

1과 2의 아이디어를 시험해보기 위해서 우선 데이터의 범위를 정해야 합니다. 우선 2018년도 S&P 500 회사 (504개) Glassdoor 리뷰(50만+개)를 proof of concept 개념으로 사용하면 좋을 것 같습니다. 차차 회사의 수와 범위를 확대할 것이고, 이후 KOSPI 200 회사 (200여개) 잡플래닛 리뷰(?개)를 활용해보고자 합니다.

<table>
<tr><th>프로젝트 방향</th><th>8/1</th><th>8/15</th></tr>
<tr><td>Employee Satisfaction Level & Stock Returns</td><td rowspan="2">데이터 수집
- 2018년도 S&P 500 회사 Glassdoor 데이터 크롤링 (박서영, 이규진)
- 동일회사 주식데이터 크롤링 (이명원)
- 잡플래닛 크롤러 발전 (이명원, 정지수) (완료)

분석
- 세부항목별 missing data를 imputation 기법으로 채우거나 제외 (박서영)
- 504개 회사의 리뷰데이터를 Doc2Vec에 태우기 (박서영)
- Word2Vec/Doc2Vec 공부하기 (이규진)</td><td>포트폴리오의 stock return과의 관계성 분석</td></tr>
<tr><td>Company Categorization Based on Employee Text Review with Doc2Vec</td><td>Doc2Vec이 분류한 회사 cluster의 특징 분석</td></tr>
<tr><td>Database & Github</td><td>적합한 데이터베이스 2개씩 조사해오기 (이규진, 이명원, 정지수)</td><td>데이터베이스로 모은 데이터 옮기기 (모두)</td></tr>
<tr><td>Related Literature Research</td><td colspan="2">관련 논문 조사하고 모으기 (박서영, 이규진)</td></tr>
</table>