기말 팀 프로젝트 브리핑

주제 : 토픽모델링을 활용한 최근 5년 간 국내 문헌정보학의 연구동향 분석

➔ 분석 데이터를 학술데이터로 잡은 이유는, 직관적으로 판단해보았을 때 학술데이터는 비교적 시간에 대해서 강건(Robust)하기 때문임. 만약 학술데이터가 시간에 따라 영향을 많이 받는다면, 학술데이터가 비정상 시계열 데이터가 될 것이고, 이를 정상 시계열 데이터로 변환하는 것 자체로 많은 시간을 잡아먹을 것임.

분석 도구 : R, Google Colab

➔ 별다른 특이사항이 없으면 Google Colab으로 진행하고자 하였지만, 한글 텍스트를 다루기 때문에 R의 KoNLP를 사용하여 불용어 처리를 진행해야 하고, 따라서 일단은 Rstudio를 활용하여 분석을 진행하고자 함. 만약 KoNLP를 Colab에 설치할 수 있다면, 발표 내용대로 Google Colab을 통해서 전체적인 분석을 진행할 예정

분석 흐름 :

1. 국가 학술정보 Open API의 통합검색 API를 통하여 문헌정보학의 학회지 별 논문을 수집하고, 논문의 고유한 번호인 lodID를 추출함.
2. 추출한 lodID를 바탕으로 상세보기 API를 통해 논문에 대한 세부정보(주제, 초록, 제목)등을 수집하여 데이터프레임화
3. 데이터프레임화 한 논문 초록에서 영문 설명을 제거하고, KoNLP를 활용하여 한글 설명에서 불용어를 제거함과 동시에 토큰화 진행.
4. 토큰화한 단어들을 바탕으로 TDM 행렬 생성, 이를 바탕으로 LSA, LDA를 진행
5. LSA, LDA를 통해 추출된 토픽의 주제어를 확인하고, 주제어를 바탕으로 토픽이 가리키는 연구동향 정의 (이 과정에서 주관적으로 주제를 정의해야함)
6. 이를 시각화 하여 확인하고, 필요하다면 TF-IDF등으로 모델의 성능을 확인해보기

일단은 해결된 과제 :

1. R과 API의 연동. 시범적으로 API 연동을 진행하여서 가져온 데이터를 데이터프레임화 하는데 성공함
2. LSA와 LDA를 R에서 적용시키는 방법. LSA는 다른 package는 찾지 못하였지만, 기본적으로 행렬분해를 바탕으로 진행되기 때문에, R에서 기본적으로 제공하는 svd()를 통하여 행렬분해를 진행하고, 이 결과 값을 확인하는 것으로 보임.

앞으로 해결해야할 과제 :

1. 상세검색을 통해 얻은 논문 초록의 영문 설명 제거. 아마 정규표현식을 통해서 제거할 수 있을 것으로 생각하지만, 이를 어떻게 구현해야 할 지는 생각을 해보아야 함.
2. 추가 텍스트 마이닝 기법 도입. 사실 텍스트 마이닝 기법에는 LSA와 LDA 이외에도 꽤 많은 방법들이 있는 것으로 알고 있고, LSA를 사용할 때의 단점을 보완한 것이 LDA이기 때문에, 어느정도 LSA와 LDA가 내용이 겹칠 수 있다는 생각이 들었음. 이는 나중에 각자 추가로 조사하여 R에서 적용할 수 있는지 확인해야 함.

대략적인 기말 프로젝트 진행 안내:

제가 자기소개에서 말씀드렸다시피 제가 현재 통계분석학회에서 활동을 하고 있습니다. 그래서 제가 중간고사가 끝나면 아마 바로 3주동안 학회에서 진행하는 프로젝트에 참가하게 되어 개인적으로 매우 바쁜 일정을 소화해야 할 것으로 보입니다. 따라서 이에 맞춰서 기말 프로젝트 일정이 조정될 것 같아서 미리 양해의 말씀드리겠습니다. 이에 따라서, 전체적인 기말프로젝트는 다음과 같이 진행될 예정입니다.

프로젝트 1~3주차 : 프로젝트에서 사용하게 될 API를 통한 데이터 수집, LSA, LDA에 대한 소개, 프로젝트 페이퍼 서론과 선행연구 부분 작성

앞에서 말씀드린 것처럼, 제가 다른 프로젝트를 진행하는 3주동안 실제적으로 데이터를 분석하는 것은 무리라고 판단되고, 여러분들도 프로젝트에서 다루어야 할 개념들과 코드에 대해서 익숙해져야 할 시간이 필요하다고 생각합니다. 따라서, 이 기간동안에는 매주 팀 프로젝트 시간에 API 활용법, LSA, LDA에 대한 개념과 R을 통해 진행하는 법에 대해서 소개하는 시간을 가지고자 합니다. 또한, 이 기간동안 각자 이 주제에 대한 선행연구가 있는지 살펴보시고, 이를 바탕으로 미리 프로젝트 페이퍼의 서론 부분과 선행연구 부분을 작성하고자 합니다. 이 과정에서, 아마 대략적인 분석 흐름에 대해서 이해

하실 것이라고 생각하고, 저도 관련해서 학습할 수 있는 자료들을 준비하여 만들고, 같이 작성할 예정입니다.

프로젝트 4~5주차 : 실제 학술 데이터 API에서 데이터 추출 후 데이터 분석 진행

프로젝트 1~3주차에 진행한 내용을 바탕으로 실제 데이터 분석을 진행하고자 합니다. 아마 데이터량 자체는 작아보이기는 하지만, 이게 텍스트 데이터다보니 이 과정에서 데이터가 너무 커질수도 있어서 이때 아마 주제를 조금 축소해서 진행할 수도 있고, 분석도 LSA와 LDA를 바탕으로 진행을 하되 결과에 따라서 다른 분석 방법을 도입할 수도 있습니다. 그래도 제 생각에는 전처리만 잘 끝내면 분석 자체는 금방 진행할 수 있을 것으로 예상하기 때문에, 2주동안 분석을 진행하는 것으로 일정을 잡았습니다.

아마 이 기간중에는 수업시간 외에 따로 오프라인 혹은 구글미트를 통해 화상으로 팀 프로젝트에 대해서 논의해야 할 것 같다고 생각합니다. 따라서, 이 기간에는 여러분들의 시간표를 바탕으로, 서로 가능한 시간에 모여서 팀 프로젝트를 진행해야 할 것 같습니다. 이 점 미리 유의해주시길 바랍니다.

프로젝트 6~7주차 : 데이터 분석 내용을 바탕으로 프로젝트 페이퍼 및 PPT자료 작성

아마 데이터 분석을 성공적으로 마쳤다면, 프로젝트 페이퍼를 작성하는 것은 금방 끝날 것이라고 생각합니다. 프로젝트 페이퍼를 작성할 때, 각자 범위를 잘 나누어서 페이퍼를 작성하면 괜찮을 것이라고 생각합니다.

또한 마지막 주차에 프로젝트 분석 결과에 대해서 발표하는 시간이 있는 것으로 알고 있는데, 이 때 사용할 PPT 자료를 만들어야 할 것으로 보입니다. 이건 생각보다 좀 시간이 걸릴 것 같아서, 미리 만들 수 있으면 만들어 놓는 것으로 하고, 전체적으로 템플릿이나 이런 것에 대해서 조금 고민을 해봐야 할 것 같습니다.

이렇게 해서 프로젝트 주제, 분석방법, 분석흐름, 전체적인 진행 일정에 대해서 소개해드렸습니다. 혹시 보시고 질문 있으시면 편하게 질문 남겨주시면 되겠습니다!