캡스톤 디자인(2) 계획서

1. 프로젝트 제목

- 자연어 처리를 활용한 한국어 유튜브 댓글 종류 분류

2. 프로젝트 구성원

3. 프로젝트 개요

SNS, 개인 블로그, 동영상 채널 등 다양한 방식의 컨텐츠 생산이 보편화된 오늘날, 방대한 양의 데이터가 인터넷에 저장되고 있습니다. 이를 마케팅 등 원하는 분야에서 적절히 사용하기 위해서는 필요한 데이터를 수집하고 분석하여 원하는 정보를 얻어내는 능력이 필요합니다.

본 프로젝트에서는 Youtube 사이트에 작성된 한국어 댓글들을 수집하고 목적에 맞는 다양한 기준으로 분류해 보는 것이 목표입니다.

4. 진행 계획

1) 적용 기술

- Python 사용 예정, github로 소스코드 관리
- 웹 크롤링 : Selenium, Google Youtube api
- 전처리 : KoNLPy, kss(Korean Sentence Splitter)
- 모델 학습 : scikit-learn, PyTorch

2) 진행 방법 & 일정 계획

9월 : 데이터 추가 수집 및 테스트용 데이터 만들기, 분류 모델 구상

10월 : 분류 모델 구현, 시각화 작업

11월 : 모델 테스트 결과 비교, 개선

3) 1학기 완료사항

- 유튜브 댓글 데이터 수집
- 자연어 데이터 전처리
- word2vec 오픈소스 활용 문장 embedding
- 문장 간 유사도 계산

```
print('문서 1과 문서2의 유사도 :',cos_sim(sentence2vec[0], sentence2vec[1]))
print('문서 1과 문서3의 유사도 :',cos_sim(sentence2vec[0], sentence2vec[2]))
print('문서 1와 문서4의 유사도 :',cos_sim(sentence2vec[0], sentence2vec[3]))

similarity = []

input_sentence_vec = sentence2vec[0]
temp = 0
for sentence in sentence2vec:
    similarity.append(cos_sim(input_sentence_vec, sentence))
```

```
문서 1과 문서2의 유사도 : 0.29738772
문서 1과 문서3의 유사도 : 0.4482292
문서 2와 문서3의 유사도 : 0.5610777
```

4) 2학기 계획

1. 댓글 데이터 추가 수집
2. 기존 모델과 다른 방법을 사용한 분류 모델 설계
3. 모델 개선 및 분류 결과 비교(긍정, 부정 감성분석 테스트)
4. 분류 문장 관계 시각화(embedding vector 시각화 툴 사용 예정)