기초 텍스트마이닝 강의 교안 (5 주차 ~ 8 주차)

1 주차: 텍스트마이닝 개요와 텍스트 전처리

이론 내용

- 텍스트마이닝 정의와 활용 사례
- 텍스트 처리 흐름: 수집 → 정제 → 분석 → 시각화 → 예측
- 전처리 개념: 토큰화, 정제, 불용어 제거, 표제어 추출

실습 내용

- nltk, spaCy, KoNLPy 를 활용한 토큰화
- 웹크롤링 데이터 전처리 실습

예제 코드

```
# 예제: nltk 를 활용한 토큰화
import nltk
nltk.download('punkt')
from nltk.tokenize import word_tokenize
text = 'ChatGPT is an advanced language model.'
tokens = word_tokenize(text)
print(tokens)
```

2 주차: 텍스트 표현 기법과 임베딩

이론 내용

- BoW, TF-IDF, Word2Vec, 임베딩 개념
- 토큰의 의미와 Subword Tokenizer 개념

실습 내용

- scikit-learn 으로 TF-IDF 벡터화
- gensim Word2Vec 학습 및 유사도 분석

예제 코드

```
# 예제: TF-IDF 벡터화
from sklearn.feature_extraction.text import TfidfVectorizer
docs = ['I love data science', 'I love AI']
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(docs)
print(X.toarray())
print(vectorizer.get_feature_names_out())
```

3 주차: LLM 기반 텍스트마이닝

이론 내용

• LLM(GPT, BERT 등) 개요와 Transformer 구조
• Hugging Face 라이브러리와 프롬프트 기법

실습 내용

• Hugging Face pipeline 으로 감성 분석
• KoBERT 로 한국어 감성 분석 실습

예제 코드

```
# 예제: Hugging Face 감성 분석
from transformers import pipeline
classifier = pipeline('sentiment-analysis')
result = classifier('I love using transformers!')
print(result)
```

4 주차: 텍스트마이닝 실전 프로젝트

이론 내용

• 프로젝트 기획 및 분석 방법론 정리

실습 내용

• 네이버 리뷰 데이터 분석
• TF-IDF, 감성분석, 워드클라우드 등 활용

예제 코드

```
# 예제: 워드클라우드 생성
from wordcloud import WordCloud
import matplotlib.pyplot as plt
text = 'data science AI machine learning big data'
wc = WordCloud(font_path='malgun.ttf', background_color='white').generate(text)
plt.imshow(wc, interpolation='bilinear')
plt.axis('off')
plt.show()
```