



토픽 모델링 과제

주제: 유전적 심장질환, 고혈압

정보융합학부 2022204094 문창원

1. 서론

- 토픽 모델링 대상 주제 및 간단 소개

2. 데이터 수집 및 처리

- 데이터 수집 과정 및 결과, 전처리 과정 및 결과

3. 토픽 모델링 결과

- 토픽 모델링 적용 기법 및 도출 결과

4. ChatGpt 적용 및 비교 결과

- ChatGPT 적용 프롬프트 명시 및 결과 비교

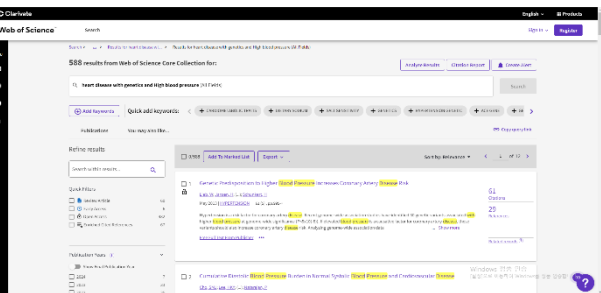
서론

토픽모델링 대상 주제: heart disease with genetics and High blood pressure

간단 소개: 심장 질환에 영향을 미칠 수 있는 유전적 정보와 고혈압의 관계를 포함하는 연구 문헌. 총 588개의 연구 문헌이 존재.

데이터 수집

A search interface with a text input field containing 'heart disease with genetics and High blood pressure'. Below the input field are filters for 'All' and 'Author'. There are buttons for '+행 추가' (Add row), '+기간 추가' (Add period), and '고급 검색' (Advanced search). At the bottom, there are buttons for 'x분명한' (Clear) and '찾다' (Find).



Web of Knowledge 플랫폼을 바탕으로 주제 수집. 총 588개의 연구 문헌이 존재.

데이터 전처리 과정 및 결과

```
In [22]: import pandas as pd

# 파일 경로 지정
file_path = 'C:/textwining.csv'

# 파일 불러오기
data = pd.read_csv(file_path, encoding='utf-8')
# 'Abstract' 컬럼만 포함하는 새로운 데이터프레임 생성
abstracts_only = data[['Abstract']]

abstracts_only
```

Out [22]:

| | Abstract |
|----------------------|---|
| 0 | Hypertension is a risk factor for coronary art... |
| 1 | BACKGROUND:The clinical significance of isolat... |
| 2 | Background Smoking is an important cardiovascu... |
| 3 | Background-Population studies have demonstrate... |
| 4 | Growing research suggests that posttraumatic s... |
| ... | ... |
| 583 | Although similar to 25% of colorectal cancer o... |
| 584 | The genetic determinants of fasting glucose (F... |
| 585 | Genetic influences on lipid traits have been s... |
| 586 | Central obesity, measured by waist circumference... |
| 587 | An increasing number of genome-wide associati... |
| 588 rows x 1 columns | |

```
data[['Abstract']]
```

```
import re
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

# 필요함: NLTK 리소스 다운로드
import nltk
nltk.download('punkt')
nltk.download('stopwords')

data = pd.read_csv(file_path, encoding='utf-8') # 파일 불러오기

# 불용어 목록 로드
stop_words = set(stopwords.words('english'))

# 텍스트 전처리 함수 정의
def preprocess_text(text):
    # 소문자 변환
    text = text.lower()
    # 숫자 및 특수 문자 제거
    text = re.sub(r'\d+', '', text) # 숫자 제거
    text = re.sub(r'[^\w\s]', '', text) # 특수 문자 제거
    # 토큰화
    tokens = word_tokenize(text)
    # 불용어 제거 및 길이가 1 이상인 단어만 포함
    tokens = [word for word in tokens if word not in stop_words and len(word) > 1]
    return ' '.join(tokens)

# 'Abstract' 컬럼 전처리
data['Abstract'] = data['Abstract'].apply(preprocess_text)
```

```
[nltk_data] Downloading package punkt to
[nltk_data] C:\Users\문정원\AppData\Roaming\nltk_data...
[nltk_data] Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\문정원\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
```

1. 위에서 export한 csv파일에서 연구 문헌의 abstract에 해당하는 컬럼만 추출.

2. NLTK를 사용하여 영어와 관련한 불용어 목록을 불러온 후, 소문자 변환, 숫자 및 특수 문자 제거, 토큰화, 불용어 제거를 통해 텍스트 전처리를 진행.

3. 왼쪽 그림은 전처리가 진행된 후 컬럼의 상태.

토픽 모델링 적용

토픽 모델링 적용 기법: LDA

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation

data = pd.read_csv('C:/textmining.csv') # 전처리된 파일 경로

# 문서-단어 행렬 생성
vectorizer = CountVectorizer(max_df=0.95, min_df=2, stop_words='english')
dtm = vectorizer.fit_transform(data['Abstract']) # 'Abstract' 컬럼에 대해 DTM 생성

# LDA 모델 설정 및 학습
lda = LatentDirichletAllocation(n_components=10, random_state=0) # 7개의 토픽 추출
lda.fit(dtm)

# 각 토픽의 상위 단어 출력
def print_top_words(model, feature_names, n_top_words):
    for topic_id, topic in enumerate(model.components_):
        print([Topic #{}: {}].format(topic_id + 1, ' '.join(feature_names[feature_names.index(word) < n_top_words] for word in topic.argsort()[::-1])))

n_top_words = 10
feature_names = vectorizer.get_feature_names_out()
print_top_words(lda, feature_names, n_top_words)
```

전처리된 데이터에서 문서-단어 행렬을 생성(DTM) 후, LDA 모델을 통해 10개의 토픽을 학습한 후, 각 토픽의 상위 10개 단어를 출력. n_components는 추출할 토픽의 수, max_df와 min_df는 단어가 문서에 등장하는 빈도의 상한과 하한을 결정.

토픽 모델링 결과

토픽 모델링 적용 결과

Topic #0: hypertension risk disease blood factors pressure high genetic cardiovascular cvd
Topic #1: disease genetic vascular sbp traits heritability pressure blood environmental factors
Topic #2: risk associated cardiovascular higher cholesterol ci disease pressure levels age
Topic #3: blood levels associated plasma pressure dna methylation concentrations heritability loci
Topic #4: genetic association genes loci identified genomewide studies variants linkage traits
Topic #5: risk factors disease cvd heart cardiovascular study obesity diabetes health
Topic #6: pressure blood associated genetic hf dbp systolic cardiovascular disease sbp
Topic #7: risk patients polymorphism allele subjects higher associated gene study genotype
Topic #8: bp variants rs hypertension study genetic associated rare eh using
Topic #9: associated disease bp risk genetic hypertension association blood variants chd

ChatGPT 적용 및 비교 결과

토픽 모델링 결과

Topic #0: 고혈압과 심혈관 위험 요인 관련 토픽
Topic #1: 유전적 요인과 혈압의 변이 관련 토픽
Topic #2: 심혈관 위험 요소와 콜레스테롤 관련 토픽
Topic #3: DNA 메틸화와 혈압 관련 토픽
Topic #4: 유전자 연관 연구 관련 토픽
Topic #5: 심혈관 질환과 생활 습관 관련 토픽
Topic #6: 심혈관 질환과 혈압 관리 관련 토픽
Topic #7: 유전적 다형성과 환자 연구 관련 토픽
Topic #8: 고혈압과 유전적 변이 관련 토픽
Topic #9: 고혈압, 유전, 심장 질환 관련 토픽

ChatGPT 적용 결과

ChatGPT 3.5

나
heart disease with genetics and High blood pressure 과 관련된 주요 토픽을 10개 뽑아줘!

ChatGPT
물론이죠! 여기 고혈압과 유전적인 심장 질환과 관련된 주요 토픽입니다:

1. 유전적 요인과 심혈관 질환의 연관성
2. 고혈압의 유전적 기원 및 유전자 연구
3. 유전자 검사를 통한 심장 질환 위험 평가
4. 가족력과 심혈관 질환 발병 위험
5. 고혈압의 유전적 영향과 생활습관의 상호작용
6. 혈압 관리를 위한 개인 맞춤형 치료법의 중요성
7. 유전적 특이점과 심혈관 질환 예방 전략
8. 유전자가 심장 질환 치료에 미치는 영향
9. 혈압 조절을 위한 유전자 치료 연구 동향
10. 심혈관 유전체학의 발전과 심혈관 질환 예방에 대한 잠재적 영향

공통점

고혈압, 심혈관 질환, 유전적 요인과 관련된 토픽은 겹치는 게 많았다.

차이점

연구 문헌에서의 토픽(예를 들어, DNA 메틸화와 혈압 관련 토픽)에서의 차이점이 나타났다.