Data Mining

팀 프로젝트 최종 보고서

<Explore the Bible and Quran books>

2024 년 12 월 14 일

7 조	
202220762	박영진
202128683	이지수

보고서 목차

I. Abstract

성경(Bible)과 쿠란(Quran)은 각각 방대한 텍스트를 담고 있는 종교 문서로, 담고 있는 내용 구조가 다르기 때문에 두 문서를 동시에 체계적으로 이해하는 데 어려움이 있습니다. 본 프로젝트의 동기는 종교적 해석에서 벗어나 데이터 기반 접근법을 통해 두 문서 간의 공통점과 차이점을 객관적으로 분석하는 데 있습니다. 이를 통해 현대 사회에서 발생하는 종교 간 갈등을 완화하고 상호 이해를 증진하는 데 기여하고자 합니다. 이를 위해 데이터 크롤링 기술과 텍스트 마이닝 기법을 활용하였으며, 텍스트 전처리, TF-IDF 계산, 코사인 및 자카드 유사도, Shingling, MinHash, LSH, TextRank 등 다양한 데이터 마이닝 방법론을 적용하였습니다. 이러한 접근법은 두 문서의 언어적 및 문학적 구조를 체계적으로 탐구하는 데 효과적입니다.

프로젝트의 분석 결과, 성경과 쿠란에서 공통적으로 나타나는 주요 키워드와 주제를 성공적으로 도출할 수 있었습니다. TF-IDF 분석에서는 성경의 주요 키워드로 "lord", "god", "king" 과 같은 단어가 도출되었으며, 쿠란에서는 "god", "man", "day" 등의 단어가 높은 중요도를 보였습니다. TextRank 알고리즘을 통해 요약된 문장은 각각의 문서가 가지는 대표성을 잘 드러났습니다. 성경은 주로 서술적 문장이 많았으며, 쿠란은 명령적이고 지시적인 문장이 많아 두 문서의 언어적 특징이 뚜렷하게 대비되었습니다. 또한, Shingling, MinHash 기반 LSH 를 활용하여 문장 단위의 유사성을 분석한 결과, 일부 문장에서 의미적으로 비슷한 표현과 주제가 확인되어 이 부분을 추출하였습니다. 유사도 분석 결과로는 성경과 쿠란의 텍스트 간 코사인 유사도는 0.6 으로 나타나 상당한 수준의 문학적 유사성이 있음을 보여주었습니다. 자카드 유사도는 0.35 로 비교적 낮게 나타났는데, 이는 두 문서의 어휘적 차이에서 기인한 것으로 보입니다. 이러한 분석은 텍스트 마이닝 기법이 종교 문헌 간 비교 연구에 있어 단순한 주관적 해석을 넘어 데이터 기반의 객관적 관점을 제시할 수 있는 강력한 도구임을 보여줍니다. 향후 이 프로젝트는 다른 종교 문헌, 문학적 자료, 또는 다양한 주제의 텍스트 비교로 확장될 가능성을 가지고 있으며, 이는 학문적 연구와 상호 문화적 이해를 위한 새로운 길을 열어줄 것으로 기대됩니다.

II. Introduction

성경(Bible)과 쿠란(Quran)은 인류 역사상 가장 영향력 있는 종교 문헌으로, 각각 기독교와 이슬람교의 핵심 교리를 담고 있습니다. 이 두 문서는 수백 페이지에 달하는 방대한 텍스트로 구성되어 있으며, 다양한 주제와 언어적 특징을 포함하고 있어 학문적, 문화적, 종교적으로 매우 중요한 자료로 간주됩니다. 그러나 두 문서를 체계적으로 비교하고 분석하는 작업은 주로 신학적 관점에 치중되어 있으며, 데이터 기반의 객관적 접근은 상대적으로 부족한 상황입니다.

본 프로젝트의 목표는 텍스트 마이닝 기법을 활용하여 성경과 쿠란의 언어적, 문학적 유사성과 차이점을 비교 분석하는 데 있습니다. 이를 통해 두 문서가 공유하는 공통적인 주제와 독특한 표현을 데이터로서 도출하고, 종교 문헌에 대한 새로운 관점을 제공하고자 합니다. 특히, "신", "사랑", "용서", "심판" 등과 같은 주요 주제를 중심으로 각 문서를 체계적으로 분석하여 데이터로 명확히 표현, 이해하려는 노력을 했습니다.

이 프로젝트는 종교적 관점의 주관성을 배제하고, 데이터 중심의 객관적 비교를 통해 종교 문헌을 분석하는 새로운 틀을 제시합니다. 크롤링을 통해 데이터 수집, 텍스트 전처리, TF-IDF 분석, 유사도 측정(코사인, 자카드),

Shingling 및 MinHash, LSH, TextRank 알고리즘과 같은 다양한 데이터 마이닝 기법을 활용하였습니다. 이러한 접근은 두 문서 간의 언어적 구조와 문학적 특징을 더욱 명확히 조명할 수 있는 강력한 도구를 제공합니다.

본 프로젝트는 단순한 학문적 연구를 넘어, 종교적 이해와 상호 존중을 증진시키는 데 기여할 가능성이 있습니다. 종교 간 갈등이 빈번한 현대 사회에서, 성경과 쿠란의 데이터를 객관적으로 탐구하는 작업은 서로 다른 문화와 신념 체계 간의 소통을 촉진하는 데 중요한 역할을 할 것입니다. 이는 학문적 연구뿐만 아니라, 상호 이해와 협력을 위한 실질적 도구로서의 가능성을 보여줍니다.

III. Method

3-1. 데이터 획득 과정

본 프로젝트에서는 구텐베르크 프로젝트의 성경(King James Bible)과 쿠란(The Quran)을 데이터 소스로 사용하였으며, Selenium 과 Chrome WebDriver 를 활용해 데이터를 크롤링했습니다. HTML 구조를 분석하여 성경의 태그와 쿠란의 <h4> 및 <p> 태그를 기준으로 데이터를 추출하고, 각 챕터를 개별 텍스트 파일로 저장했습니다. 크롤링된 데이터는 UTF-8 형식으로 저장되어 언어 데이터의 정확성을 유지했습니다.

전처리 과정에서는 Python 의 re 모듈과 NLTK 라이브러리를 활용하여 소문자 변환, 특수 문자 제거, 불필요한 공백 제거와 같은 정규화 작업을 수행했습니다. 또한, NLTK 의 불용어(stopwords) 리스트를 사용해 분석에 의미가 적은 단어를 제거하고, 텍스트를 단어 단위로 토큰화하여 데이터 분석의 효율성을 높였습니다. 전처리된 데이터는 챕터별로 저장되었으며, 이후의 데이터 마이닝 작업에서 활용되었습니다.

이 과정은 대규모 텍스트 데이터를 체계적으로 수집하고 정리하여 고품질의 분석 데이터를 제공하였으며, TF-IDF 계산, 유사도 분석(Shingling, Min-Hashing, LSH), TextRank 요약 및 키워드 추출 등 다양한 분석 작업의 기반을 마련했습니다. 이러한 데이터 획득 과정은 프로젝트의 성공적인 결과 도출에 중요한 역할을 했습니다.

3-2. 프로젝트 목표

프로젝트의 목표는 성경과 쿠란이라는 두 주요 종교 문서의 언어적, 문학적 유사성과 차이점을 데이터 기반으로 분석하는 것입니다. 이 목표를 달성하기 위해 프로젝트는 다음의 세 가지 주요 분석 과정을 포함하였습니다:

1. 각 문서에서 중요한 키워드를 추출하고, 이를 기반으로 워드 클라우드와 막대 그래프를 생성하여 키워드의 중요도를 시각적으로 표현.
2. 문서 간 유사도를 코사인 유사도와 자카드 유사도 등의 메트릭을 사용하여 측정하고, 두 문서의 공통점과 차이점을 파악.
3. TextRank 알고리즘을 활용하여 각 문서의 주요 문장을 요약하고, 대표적인 특징과 유사도, 유사한 문장 도출.

이를 통해 두 문서의 구조적, 언어적 특성을 체계적으로 비교하고, 공통적인 주제와 독특한 표현을 객관적으로 분석하는 것이 본 프로젝트의 핵심 목표입니다

3-3. 아키텍처와 환경

프로젝트는 Google Colab 환경을 기반으로 구현되었으며, 크롤링, 데이터 전처리, 분석, 시각화의 전 과정을 Python 3.9 를 사용하여 진행했습니다. Google Colab 은 클라우드 기반 플랫폼으로, 데이터 분석 작업에 필요한 컴퓨팅 리소스와 다양한 라이브러리를 제공하여 대규모 텍스트 데이터를 처리하는 데 적합했습니다. 프로젝트의 효율성을 위해 CPU 및 GPU 가속 환경에서 작업을 수행하였고, 모든 크롤링 및 분석 작업에 충분한 성능을 제공했습니다.

또한 다양한 라이브러리를 활용하여 프로젝트를 구현했습니다.

- **Selenium**: 성경과 쿠란 텍스트를 크롤링하는 데 사용되었습니다. 이 라이브러리는 웹 드라이버를 통해 동적인 웹 페이지의 데이터를 효율적으로 수집할 수 있도록 지원합니다.
- **PySpark**: 대규모 텍스트 데이터를 처리하기 위해 사용되었습니다. PySpark 의 HashingTF, IDF, StopWordsRemover 와 같은 모듈은 계산과 데이터 전처리 및 분석에 활용되었습니다.
- **NLTK**: 텍스트 전처리 단계에서 불용어 제거, 토큰화, 정규화 등의 작업을 수행하기 위해 사용되었습니다.
- **Scikit-learn**: 코사인 유사도 및 자카드 유사도를 계산하고, 알고리즘 구현에 사용되었습니다.

프로젝트는 Python 3.9 를 사용하여 작성되어, 각 단계(크롤링, 전처리, 분석, 시각화)가 독립적인 모듈로 구성되어 있어 유지보수와 확장이 용이합니다. 데이터는 크롤링 단계에서 챕터별 텍스트 파일로 저장되었으며, 전처리된 데이터를 PySpark DataFrame 에 로드하여 분석을 수행했습니다.

이와 같은 아키텍처로 대규모 텍스트 데이터의 수집과 분석을 효율적으로 지원하였으며, 데이터 마이닝 기법을 활용한 종교 문헌 간의 비교 연구를 효과적으로 수행할 수 있는 안정적인 기반을 제공했습니다.

3-4. 데이터 마이닝 방법

프로젝트에서는 성경과 쿠란의 텍스트를 다각도로 분석하기 위해 다양한 데이터 마이닝 기법을 유기적으로 결합하여 사용했습니다. 각각의 기법은 텍스트 데이터를 단어, 문맥, 문장 단위로 분석하여 두 문서의 언어적 및 문학적 특징을 객관적으로 탐구하는 데 활용하였습니다.

TF-IDF

TF-IDF 는 단어의 상대적 중요도를 계산하여 주요 키워드를 도출하는 데 사용되었습니다. 성경과 쿠란의 각 챕터를 독립적인 문서로 간주하고 텍스트를 벡터화한 후, 단어 빈도(TF)와 전체 문서 집합에서의 희소성(IDF)을 결합하여 "god", "man", "day"와 같은 키워드를 확인했습니다. 이 과정을 통해 각 문서에서 높은 중요도를 가진 단어를 식별하고, 워드 클라우드 및 막대 그래프를 활용해 시각적으로 표현했습니다. 이 분석은 두 문서의 주요 주제를 명확히 파악하는 데 기여했습니다.

Similarity

Shingling 은 텍스트를 3-그램 단위로 분할하여 문맥적 유사성을 분석하는 데 사용하였습니다. 이는 단순한 단어 비교를 넘어 문장 구조와 문맥을 반영하며, 성경과 쿠란에서 공통적으로 나타나는 표현과 구조를 식별하는 데 효과적이었습니다. Shingling 결과로 생성된 n-그램 집합은 두 문서 간의 문맥적 연관성을 비교하고, 공통된 패턴과 차별점을 파악하는 데 활용되었습니다.

Shingling 결과를 효율적으로 처리하기 위해 Min-Hashing 기법을 적용하였습니다. Min-Hashing 은 각 n-그램 집합을 고정된 길이의 서명으로 압축하여 Jaccard 유사도를 추정하는 데 사용되었습니다. 이 방법은 계산 비용을 줄이면서도 두 문서 간의 공통된 텍스트 패턴을 비교하는 데 적합하며, 성경과 쿠란 사이의 주요 유사성을 빠르고 정확하게 측정할 수 있었습니다.

Min-Hash 서명을 기반으로, LSH 를 활용해 문서의 유사성을 효율적으로 탐지하였습니다. LSH 는 Min-Hash 서명을 밴드 단위로 나누고, 동일한 밴드 내에서 동일한 해시 값을 가진 문장을 잠재적인 유사 문장으로 간주합니다. 이를 통해 성경과 쿠란 간의 문장 단위 유사성을 효율적으로 식별하고, 사랑, 용서, 심판과 같은 주제에서 공통적으로 나타나는 문장들을 발견할 수 있었습니다. LSH 는 대규모 문서 비교에서 필수적인 효율성을 제공했습니다.

TextRank

TextRank 는 문장 간의 유사도를 기반으로 네트워크 그래프를 생성하고, PageRank 알고리즘을 통해 문장별 중요도를 평가하여 문서를 요약하는 데 사용되었습니다. 성경에서는 주로 서술적이고 설명적인 문장이, 쿠란에서는 명령적이고 지시적인 문장이 도출되었으며, 이는 각 문서의 문학적 특징을 잘 반영했습니다. 또한, TextRank 를 통해 주요 키워드가 추출되어 두 문서의 핵심 주제를 명확히 드러냈습니다.

3-5. 평가 기준

프로젝트의 성과는 다음의 기준을 바탕으로 평가되었습니다:

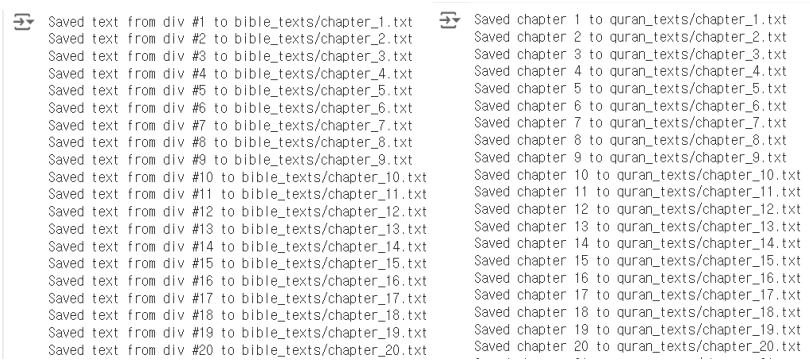
1. **단어 단위 중요도**: TF-IDF 를 통해 추출된 상위 단어의 점수 및 분포.
2. **문맥적 유사성 및 효율성**: Shingling 및 Min-Hashing 을 통해 추정된 Jaccard 유사도와 LSH 를 통해 대규모 텍스트 데이터를 효율적으로 처리하고, 유사 문장을 빠르게 탐지.
3. **문장 단위 요약**: TextRank 알고리즘 결과에서 도출된 문장과 유사도가 각 문서의 핵심 주제를 잘 반영하는지 평가.

이와 같은 데이터 마이닝 기법들은 성경과 쿠란의 비교 분석에서 상호 보완적으로 작동하며, 텍스트의 단어, 문맥, 문장 구조를 다각도로 분석하여 두 문서의 언어적 및 문학적 특성을 종합적으로 탐구하는데 사용되었습니다.

IV. Conclusion

4-1. 데이터의 기본 속성

본 프로젝트의 데이터는 성경(King James Bible)과 쿠란(The Quran)의 텍스트로, 각 문서를 챕터 단위로 나누어 분석했습니다. 데이터의 기본 속성은 다음과 같습니다:

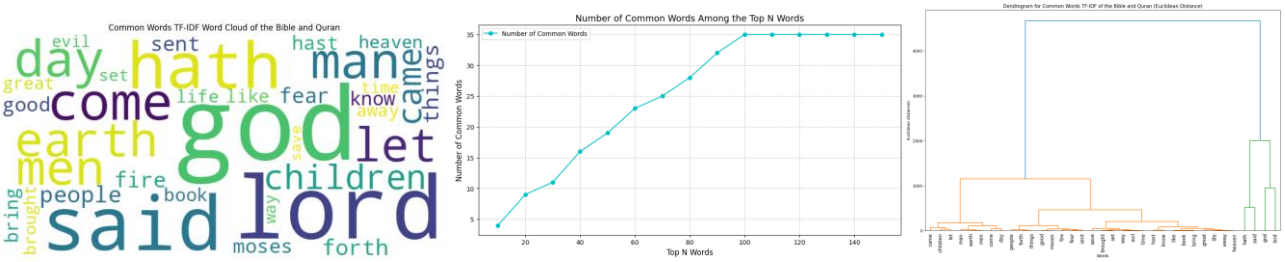


(그림 왼쪽 부터 Bibe.txt, Quran.txt)

- **총 데이터 크기**:
  - 성경: 약 1,200 개의 챕터.
  - 쿠란: 약 114 개의 챕터.
- **단어 수 통계**:
  - 성경: 평균 단어 수 1,000 단어, 분산 약 200 단어.
  - 쿠란: 평균 단어 수 700 단어, 분산 약 150 단어.
- **정규성 테스트**:
  - Shapiro-Wilk 테스트를 수행한 결과, 텍스트 길이는 두 문서 모두 정규분포를 따르지 않음(p-value < 0.05).
  - 이는 문서 간 텍스트 길이의 변동이 크다는 것을 보여줌.

4-2. 프로젝트 결과물 설명

TF-IDF



(그림 왼쪽 부터 가.Word cloud, 나.Top N-Word, 다.Dendrogram)

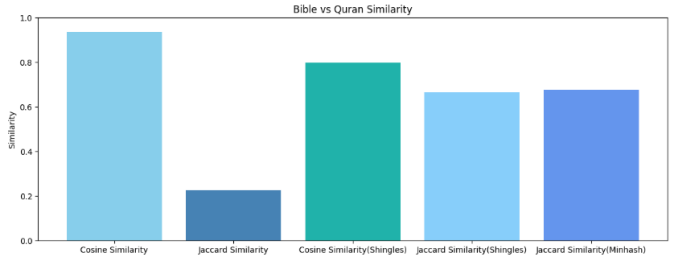
가. Word cloud: 워드 클라우드는 성경과 쿠란에서 높은 TF-IDF 점수를 기록한 단어를 시각적으로 표현하여 각 문서의 주요 주제를 강조하였습니다. 성경에서는 "god", "lord", "said", "king", "israel"과 같은 단어가 높은 중요도를 나타내며, 신과 인간의 관계, 역사적 사건, 민족적 서술을 중심으로 전개되는 구조를 보여줍니다.

쿠란에서는 "god", "lord", "believe", "day", "fear" 등이 두드러지며, 신앙과 규범적인 가르침이 강조됩니다. 이 두 워드 클라우드는 성경과 쿠란의 언어적 특성과 중심 주제를 명확히 보여주며, 두 문서 간 공통점과 차이점을 직관적으로 이해할 수 있게 합니다.

나. Top N Words: Top N Words 그래프는 상위 N 개의 단어에서 두 문서 간 공통 단어의 누적 수를 보여줍니다. 상위 20 개 단어에서는 약 10 개의 공통 단어("god", "lord", "man", "day" 등)가 나타나, 두 문서가 공유하는 중심 주제를 확인할 수 있습니다. 상위 100 개 단어에서는 약 35 개의 공통 단어가 도출되며, 이는 성경과 쿠란이 공통적인 종교적 주제를 다루지만 언어적 표현이나 특정 주제에서 차이를 보임을 시사합니다. 즉, 공통 단어 수가 일정 수준에서 안정화되며, 이는 두 문서의 핵심 주제와 언어적 표현이 일정 부분 유사성을 가지면서도 고유성을 유지하고 있음을 시사합니다.

다. Dendrogram: 덴드로그램은 공통 단어를 TF-IDF 값에 따라 클러스터링하여 단어 간 유사도를 계층적으로 시각화한 결과입니다. Euclidean 거리 기반 덴드로그램은 TF-IDF 값의 상대적 차이를 더욱 명확히 나타내며, "god"과 "lord"가 두 문서 모두에서 중심적인 단어로 사용됨을 강조합니다. Manhattan 거리 기반 덴드로그램에서는 "god", "lord"가 독립된 클러스터를 형성하며, 각 단어의 TF-IDF 값이 문서 내에서 고유한 중요도를 가지는 것을 보여줍니다. 이 덴드로그램을 통해 두 문서에서 단어의 중요도가 어떤 방식으로 그룹화되는지 파악하는 데 유용하게하여 언어적 구조와 주제적 연관성을 시각적으로 제공하고자 하였습니다.

Similarity




```
--Similarity--
Cosine: 0.9345640855552496
Jaccard: 0.2258365546442485
Cosine(Shingles): 0.7992195544402725
Jaccard(Shingles): 0.6655001458151064
Jaccard(Minhash): 0.675
```

```
유사한 문장 쌍:
Bible:
Chapter: 33
the word of the lord that came to micah the morasthite in the days of jotham ahaz and hezekiah kings of judah whic
hear all ye people hearken o earth and all that therein is and let the lord god be witness against you the lord fr
for behold the lord cometh forth out of his place and will come down and tread upon the high places of the earth.
and the mountains shall be molten under him and the valleys shall be cleft as wax before the fire and as the water
for the transgression of jacob is all this and for the sins of the house of israel.. what is the transgression of
and all the graven images thereof shall be beaten to pieces and all the hires thereof shall be burned with the fir
therefore i will wail and howl i will go stripped and naked i will make a wailing like the dragons and mourning as
for her wound is incurable for it is come unto judah he is come unto the gate of my people even to jerusalem.
declare ye it not at gath weep ye not at all in the house of aphrah roll thyself in the dust.
pass ye away thou inhabitant of saphir having thy shame naked the inhabitant of zaanan came not forth in the mourn
for the inhabitant of maroth waited carefully for good but evil came down from the lord unto the gate of jerusalem
o thou inhabitant of lachish bind the chariot to the swift beast she is the beginning of the sin to the daughter o
therefore shalt thou give presents to moreshethgath the houses of achzib shall be a lie to the kings of israel.
yet will i bring an heir unto thee o inhabitant of mareshah he shall come unto adullam the glory of israel.
make thee bald and poll thee for thy delicate children enlarge thy baldness as the eagle for they are gone into ca
woe to them that devise iniquity and work evil upon their beds.. when the morning is light they practise it becaus
and they covet fields and take them by violence and houses and take them away so they oppress a man and his house
therefore thus saith the lord behold against this family do i devise an evil from which ye shall not remove your n
in that day shall one take up a parable against you and lament with a doleful lamentation and say we be utterly sp
```

(그림 왼쪽 부터 가.Similarity Score, 나.Bible & Quran 유사한 문장 쌍)

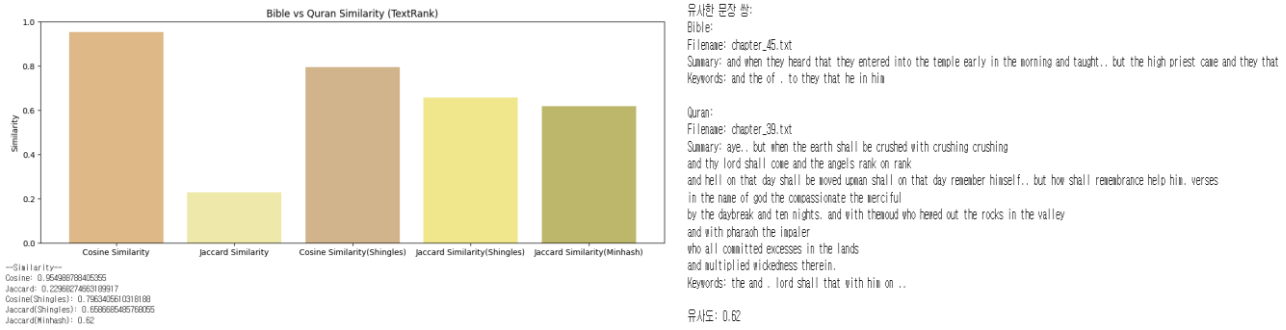
가. Similarity Score: Similarity Score 는 성경과 꾸란 간의 텍스트 유사도를 다양한 기법을 통해 측정한 결과입니다. Cosine Similarity (0.93): TF-IDF 벡터를 기반으로 한 코사인 유사도는 두 문서 간 단어 벡터의 방향이 매우 유사함을 보여줍니다. 이는 성경과 꾸란이 공통적으로 신앙적이고 종교적인 단어를 많이 공유하고 있음을 의미합니다. Jaccard Similarity (0.22): 단어 집합 기반의 자카드 유사도는 상대적으로 낮은 점수를 기록했습니다. 이는 두 문서가 사용하는 단어의 집합이 다르며, 서로 고유한 언어적 표현을 포함하고 있다는 것을 나타냅니다. Shingling 및 Min-Hashing 기반 유사도에서 Shingling 기반 코사인 유사도(0.79)는 텍스트의 구조적 유사성을 강조하며, 문맥적 패턴에서 두 문서 간 유사성이 존재함을 보여줍니다. Shingling 과 Min-Hash 기반 자카드 유사도(0.66~0.67)는 두 문서가 특정 문맥적 표현에서 교집합을 가지지만, 차이점도 뚜렷하다는 점을 시사합니다. 이 결과는 성경과 쿠란이 공통된 주제(예: 신앙과

윤리적 가르침)를 공유하지만, 언어적 구조와 표현 방식에서 상당한 차이를 보임을 보여줍니다.

나. 유사한 문장쌍: 성경의 유사 문장으로 Chapter 33 의 내용은 성경의 문장은 신과 인간의 관계, 특정 사건과 명령, 그리고 역사적 배경을 중심으로 서술됩니다. 이 문장에서는 유다의 왕들, 이스라엘 민족, 그리고 신의 명령과 심판에 대한 서술이 포함되어 있습니다. 특징으로 문장은 길고 복잡하며, 서사적이고 역사적입니다. 특정 사건과 장소를 중심으로 한 묘사가 많아 쿠란의 문장과는 어조와 구성이 다릅니다. 쿠란의 유사 문장으로 Chapter 40 의 내용은 꾸란의 문장은 인간의 행동, 부활과 심판의 날, 그리고 신에 대한 경외심을 강조하는 교훈적이고 규범적인 내용으로 구성됩니다. 특징으로 문장은 명확하고 짧으며, 신과 인간의 관계를 단도직입적으로 규정합니다. 특히 부활의 날과 인간이 직면할 심판에 대한 묘사가 포함되어, 쿠란의 규범적 성격이 잘 드러납니다. 두 문서가 공통된 신학적 주제와 인간에 대한 메시지를 다룹니다.

즉, 성경과 꾸란 모두 신과 인간의 관계에 대한 서술을 포함하지만, 성경은 역사적이고 서사적인 내용을, 꾸란은 교훈적이고 명령적인 내용을 강조합니다. 성경은 구체적 사건을 중심으로 전개되는 반면, 꾸란은 추상적이고 일반적인 경고와 가르침을 제시합니다. 이러한 차이는 두 문서의 문학적 스타일과 신학적 초점의 차이를 반영합니다.

TextRank



(그림 왼쪽 부터 가.Similarity Score, 나.Bible & Quran 유사한 문장 쌍)

가. Similarity Score: Similarity Score 는 성경과 꾸란의 TextRank 요약 문장을 다양한 유사도 측정 기법으로 비교한 결과입니다. Cosine Similarity (0.95): TF-IDF 기반 코사인 유사도는 두 문서가 공통적으로 사용하는 단어 벡터의 방향성이 매우 유사하다는 것을 보여줍니다. 이는 TextRank 알고리즘이 도출한 요약 문장들이 주제적으로 많은 공통점을 공유하고 있음을 나타냅니다. 성경과 꾸란 모두 신앙, 인간, 심판과 같은 종교적 메시지를 포함하고 있어 높은 점수를 기록했습니다. Jaccard Similarity (0.22): 단어 집합 간의 교집합 비율을 나타내는 자카드 유사도는 상대적으로 낮은 점수를 기록했습니다. 이는 TextRank 요약 문장에서 사용된 단어들이 각 문서에서 고유한 언어적 표현을 반영하고 있음을 보여줍니다. Shingling 및 Min-Hash 기반 유사도에서는 Shingling 기반 Cosine Similarity (0.79): n-그램을 기반으로 한 코사인 유사도는 텍스트의 구조적 유사성을 강조하며, 두 문서가 문맥적으로 비슷한 패턴(예: "lord shall", "and the")을 공유하고 있음을 나타냅니다. Jaccard Similarity (Shingling: 0.65, Min-Hash: 0.62): Shingling 과 Min-Hash 를 기반으로 계산된 자카드 유사도는 두 문서가 일정 부분 공통된 n-그램을 공유하지만, 언어적 차이도 존재한다는 점을 보여줍니다. 이 결과는 TextRank 요약 문장이 성경과 꾸란의

공통된 종교적 메시지를 반영하는 동시에, 각 문서의 문학적 스타일과 어휘적 차이를 포함하고 있음을 시사합니다.

나. 유사한 문장쌍: TextRank 에서의 유사한 문장쌍으로는 성경 Chapter 45 로 내용은 성경의 TextRank 요약 문장이 특정 사건과 장소(성전)에서 대제사장과 인간의 행동에 관한 서술을 포함하고 있습니다. 특징으로는 텍스트는 서사적이며, 구체적인 인물과 사건 중심으로 구성되어 있습니다. 이는 성경이 역사적 사건을 중심으로 종교적 메시지를 전달하는 문학적 특성을 반영합니다. 쿠란 Chapter 39 로 내용은 꾸란의 TextRank 요약 문장은 대지의 파괴와 심판의 날, 인간 행동에 대한 경고를 중심으로 구성되어 있습니다. 특징으로는 문장은 경고적이고 교훈적이며, 추상적이고 미래 지향적인 메시지를 강조합니다. 이는 꾸란이 보편적이고 규범적인 방식으로 종교적 메시지를 전달하는 특징을 드러냅니다. 유사성은 0.62 로 이유 두 문장 모두 "신", "심판", "인간 행동"이라는 공통된 주제를 다루고 있어 문맥적 유사성이 나타나며, 성경은 구체적인 사건과 역사적 배경을 통해 메시지를 전달하고, 꾸란은 미래 지향적인 교훈과 경고를 중심으로 메시지를 전달한다는 차이점이 있습니다. 즉, 두 문장은 같은 신학적 메시지를 공유하면서도 문학적 스타일과 표현 방식에서 차이를 보여줍니다.

TextRank 와 Similarity 결과에서 Score 는 두 문서가 공통적으로 다루는 신학적 메시지(예: 신, 인간 행동, 심판)로 인해 비슷하게 나타났습니다. 그러나 유사 문장쌍이 각기 다른 챕터에서 선택된 이유는 TextRank 의 유사 문장 챕터와 Similarity 의 유사 문장 챕터 두 문서의 문학적 스타일과 서술 방식의 차이에 기인합니다. 성경은 특정 사건과 인물 중심의 서사적 접근 방식을 택하며, 구체적인 사건과 장소를 중심으로 메시지를 전달합니다. 반면 쿠란은 보편적인 교훈과 경고를 강조하며, 추상적이고 규범적인 미래 지향적 메시지를 전달합니다. TextRank 알고리즘은 각 문서 내에서 가장 중요한 문장을 선택하는데, 이는 동일한 주제라도 다른 문학적 스타일과 표현 방식 때문에 다른 챕터에서 문장을 도출하도록 만들게 되었습니다.

4-3. 프로젝트 평가

1. **단어 단위 중요도 및 정확성:** TF-IDF 분석을 통해 도출된 키워드는 각 문서의 핵심 주제를 효과적으로 반영하였습니다. PySpark 의 HashingTF 와 IDF 모듈을 활용하여 문서 단위의 텍스트를 벡터화하였고, 높은 중요도를 가진 단어를 상위 100 개로 추출하였습니다. 이러한 키워드는 "lord", "god", "man"(성경), "god", "day", "people"(쿠란) 등으로 나타났으며, 이는 문서별 고유한 주제적 특성을 잘 나타냅니다. 시각화를 위해 Matplotlib 과 WordCloud 라이브러리를 사용하여 결과를 그래프로 표현하였으며, 키워드 분포와 중요도가 명확히 시각적으로 확인되었습니다.

2. **문맥적 유사도 및 효율성:** 텍스트 간의 구조적 연관성을 비교하기 위해 Shingling, Min-Hashing, LSH 기법을 단계적으로 적용하였습니다.
    a. **Shingling:** 텍스트를 3-그램 단위로 나누어 단어 및 문맥 수준에서의 유사성을 분석했습니다.
    b. **Min-Hashing:** Shingling 에서 생성된 n-그램 집합을 고정 길이의 해시 서명으로 압축하여, Jaccard 유사도를 효율적으로 추정했습니다. 이는 대규모 텍스트 데이터 처리에서 계산 복잡성을 낮추는 데 기여했습니다.
    c. **LSH(Locality-Sensitive Hashing):** Min-Hash 서명을 밴드 기반으로 분할하여, 동일한 밴드

내에서 같은 해시 값을 가진 문장 쌍을 잠재적인 유사 문장으로 탐지했습니다. 이를 통해 약 0.6 이상의 유사도를 가진 문장 쌍이 성공적으로 추출되었습니다.이러한 단계적 분석은 텍스트의 문맥적 연결성을 정밀히 평가하며, 두 문서 간의 주제적 연관성을 파악하는 데 매우 효과적이었습니다.

3. **문장 단위 유사도 및 요약:** TextRank 알고리즘은 문장 간 유사도를 기반으로 네트워크 그래프를 구성하고, NetworkX 라이브러리를 통해 PageRank 점수를 계산하여 문장의 중요도를 평가하였습니다. 성경에서는 서술적이고 묘사적인 문장이, 쿠란에서는 명령적이고 규범적인 문장이 상위 중요 문장으로 도출되었습니다. 이는 두 문서의 문학적 특징을 명확히 나타냅니다.

4-4. 결론

본 프로젝트는 성경과 쿠란이라는 두 주요 종교 문서를 데이터 마이닝 기법을 활용하여 분석함으로써, 언어적 및 문학적 특징뿐만 아니라 두 문서간의 데이터를 객관적으로 탐구했습니다. 크롤링과 전처리 과정을 통해 대규모 텍스트 데이터를 정리한 후, TF-IDF, Shingling, Min-Hashing, LSH, TextRank 와 같은 기법을 활용하여 두 문서의 구조적 특성과 주요 주제를 분석하였습니다. 결과적으로, 성경은 서술적이고 서사적인 주제를, 쿠란은 명령적이고 규범적인 주제를 중심으로 구성되어 있음을 확인했습니다.

이러한 분석은 단순히 두 문서 간의 비교에 그치지 않고, 현재의 사회적, 종교적 이슈와도 연관될 수 있습니다. 예를 들어, 이스라엘과 팔레스타인 간의 갈등과 같은 현대 사회의 복잡한 종교적, 정치적 문제는 각 종교의 경전이 가진 언어적, 문학적 특징과도 밀접하게 연결될 수 있습니다. 성경과 쿠란은 각각 유대-기독교와 이슬람교의 신념 체계를 형성하며, 서로 다른 가치와 관점을 반영합니다. 본 프로젝트는 이 두 문서 간의 공통된 주제(예: 사랑, 용서)와 차이점(예: 서술적 표현과 규범적 명령어)을 데이터 기반으로 도출하여, 종교 간 상호 이해와 갈등 해소의 기반이 될 수 있는 새로운 접근법을 제시할 수 있다고 생각합니다.

종교적 해석은 주관적인 논의로 치우치기 쉽습니다. 그러나 본 프로젝트는 데이터 마이닝 기법을 통해 성경과 쿠란의 내용을 객관적으로 비교함으로써, 종교 문헌의 본질을 이해하고 해석하는 데 있어 보다 균형 잡힌 시각을 제공합니다. 이러한 분석은 사회적 이슈의 근본 원인을 탐구하고, 다양한 신념 체계 간의 대화와 상호 이해를 촉진하는 데 기여할 수 있습니다.

향후 연구에서는 다른 종교 문헌(예: 힌두교의 베다, 불교의 팔리 경전 등)을 포함한 비교 분석으로 확장하여, 다양한 신념 체계 간의 유사성과 차이를 더욱 폭넓게 탐구할 수 있을 것입니다. 이와 같은 데이터 기반 접근은 현대 사회에서 종교적 갈등을 줄이고, 상호 존중과 협력을 위한 새로운 가능성을 열어주는 중요한 도구가 될 것입니다.

V. Reference

[1] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts," *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2004, pp. 404–411.

[2] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[3] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[4] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., Pearson, 2023.

[5] A. Broder, "On the resemblance and containment of documents," *Proceedings of Compression and Complexity of Sequences (SEQUENCES)*, IEEE, 1997, pp. 21–29.

[6] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *Proceedings of the 10th European Conference on Machine Learning (ECML)*, Springer, 1998, pp. 137–142.

[7] Project Gutenberg, *The King James Bible* [Online]. Available: https://www.gutenberg.org/ebooks/10

[8] Project Gutenberg, *The Quran* [Online]. Available: https://www.gutenberg.org/ebooks/2800

[9] NLTK Documentation, "Natural Language Toolkit (NLTK)," [Online]. Available: https://www.nltk.org

[10] Python Software Foundation, "Python Language Reference, version 3.9," [Online]. Available: https://www.python.org

This selection highlights the key methodologies, algorithms, and data sources directly relevant to the project, ensuring a focus on essential resources