Practice 5

Perform following text mining tasks on the NIV Bible text files.

NIV bible 성경 text files들을 사용해서 아래 text mining task를 수행하라

Task 1 Preprocessing and Data Exploration

1-1 Perform tokenization on the Bible text dataset and provide descriptive statistics.

Tokenization을 수행하고, 데이터에 대한 전반적인 통계를 제시하시오. (단어 빈도, 글자 수 등)

1-2 For improved Result of text mining, what can be applied during the tokenization and preprocessing phase? Explain further process you might attempt.

보다 나은 분석 결과를 위해서, Tokenization과 전처리 단계에서 추가적으로 반영할 수 있는 작업이 있다면 적용하고 설명해보시오.

1-3 Using the preprocessed the dataset, Perform EDA process to know more insight about Bible. For example, which book is the longest and the shortest? Or How many words are in Bible? How many names are in the Bible? What are the most frequent verbs used in the Bible?

전처리한 데이터를 활용해서 성경에 대해서 인사이트를 얻을 수 있는 데이터 탐색을 수행하여라. 예를 들어 가장 긴 성경과 짧은 성경은 무엇인가? 전체 몇 개의 단어가 성경에 사용되었는가? 성경에 나오는 사람이름의 수는 몇 개나 되는가? 가장 많이 사용되는 동사는 무엇인가?

Task 2 TF/IDF

Bible can be divided into 8 groups. (4 groups for old and new testament).

Old Testament

The Old Testament is made up of 39 books. These 39 books can be divided into four main groups.

Law

The law includes Genesis through Deuteronomy. This section discusses how God brought Israel into existence and the law He gave them.

History

The history section begins with Joshua and ends with Esther. These books cover hundreds of years, from when Israel entered Canaan until they were carried into captivity.

Poetry

The books of poetry begin with Job and end with the Song of Solomon. These books give a detailed glimpse into what servants of God felt as they went through various situations in life.

Prophecy

The books of prophecy are usually divided into the Major Prophets and Minor Prophets. God sent the prophets with a specific message — usually urging Israel to change their sinful ways and turn back to God before it was too late. They also include many prophecies about Jesus Christ.

New Testament

The New Testament is made up of 27 books. These 27 books are usually divided into four main groups as well.

Gospels

The Gospel accounts, or the Life of Christ, include Matthew through John. These four books tell about the life of Jesus while emphasizing certain aspects of His nature. For example, Matthew emphasizes that Jesus is King, and He has a kingdom.

History

The book of Acts is the history of the beginning and growth of Jesus' church. It also tells a great deal about the work of Peter and Paul — two apostles of Jesus Christ. Several sermons are recorded in this book, and we also read how people responded to the preaching of Christ.

Letters

The letters — or epistles — include Romans through Jude. These letters are all about how to live and grow as a Christian. Some confront false teachings in the church; some encourage suffering Christians; some are teaching new Christians to grow in their faith.

Prophecy

The book of Revelation is a book of prophecy of "things which must shortly take place" (Revelation 1:1) written to the seven churches of Asia near the end of the first century. The emphasis of the book is God wins. It encourages Christians to remain faithful through difficult times (Revelation 2:10).

2-1 Calculate TF/IDF for each group of Bible. Show top 20 words of high TF/IDF for each group of Bible.

8개의 서로 다른 성경 그룹의 단어들에 대해서 TF/IDF를 계산하고, 높은 TF/IDF를 가지는 top 20 개의 단어들을 제시하라.

2.2 What can be inferred from the result of 2-1? Which groups are similar to each other? How different are those groups of books? Explain the insight you obtained from the result?

2-1의 결과로부터 무엇을 추론할 수 있는가? 비슷한 성경의 그룹은 무엇이 있으며, 또 각 성경 그룹은 각각 어떻게 다른가? 어떤 인사이트를 얻을 수 있는지 설명해보라.

3. Topic Modeling

3-1 Perform topic modeling for both the New and Old Testaments separately. How the topics are different compared to when you perform topic modeling to entire Bible?

신약과 구약성경에 대해서 각각 토픽 모델링을 적용하라. 전체 성경에 대해서 토픽 모델링을 했을 때와 비교해서 어떻게 다르게 나오는가?

3-2 Is your result of topic modeling easily interpretable? If not, how can you adjust your work to get more interpretable result? Try and learn from errors.

3-1에서 얻은 결과는 쉽게 해석이 되는가? 해석이 어렵다고 한다면 토픽 모델링을 어떻게 다르게 해서 해석 가능한 결과를 얻을 수 있을까? 다양하게 시도해보고 경험을 통해서 배워보자.

3-3 Choose books of bible of your interest and perform topic modeling on the specific parts. Share your result and insight you obtained about the Bible.

여러분이 관심있는 성경의 특정한 부분(예, 복음서, 시편, 모세 오경 등)에 대해서 topic modeling 을 적용하고, 그 결과와 인사이트를 공유해보자? 성경에 대해서 새롭게 알게 된 사실이 있는가?