

NIV bible 성경 text files들을 사용해서 아래 text mining task를 수행하라

1-1 전체 데이터를 읽어와서 아래와 같은 형식으로 정리하여라

Book	Chapter	Verse	Script
Genesis	1	1	In the beginning God created the heavens and the earth.
Genesis	1	2	Now the earth was formless and empty, darkness was over the surface of the deep, and the Spirit of God was hovering over the waters.
...

1-2 Tokenization을 수행하고, 빈도분석 및 frequent word에 대한 시각화를 수행하여라. (+해석하기)

2-1 Genesis부터 Malachi까지는 구약성서(Old testament)이고 Matthew부터는 신약성서(New Testament)이다. 신약과 구약에 대해 따로 빈도 분석을 수행하고, frequent words가 어떻게 다른지 비교하여라

2-2 stop word를 비롯하여 단어 빈도 분석에서 배제되어야 할 단어들을 선정하여 1-2, 2-1에 반영한 후 달라진 결과를 해석하여라. (1-2, 2-1에서 이미 적용하였다면 어떤 단어를 배제하였는지 설명만 하면 됨)

2-3 신약과 구약 모두에서 10회 이상 나타나는 단어들을 추출하여라. 이 단어들의 출현 비율 (해당 단어의 빈도 / 데이터 내 모든 단어의 수)을 계산하라. (신약/구약 대해서 각각 따로)

해당 단어의 신약성경에서 출현비율 (prop A)와 구약성경에서의 출현비율 (prop B)를 각각 계산한 후 $\log(\text{propA} / \text{propB})$ - **log ratio**를 계산하여 **log ratio** top 20개 단어와 bottom 20개 단어를 추출하여보라. 이를 통해 알게된 사실을 설명하여라.

3. 감성분석

3-1 신약과 구약성경의 감성분석을 각각 수행하여라

3-2 신약과 구약에서 나타나는 감성 단어의 빈도를 비교

3-3 각 성경(Book) 별로 감성을 분석하여 비교하여라. 긍정적인 감성의 성경과 부정적인 감성의 성경은 어떤 것들이 있는가?

3-4 각 chapter들을 감성분석하여보라. 창세기, 역대서, 열왕기, 시편 등 길이가 상대적으로 긴 성경을 일부 선정하여, 각 성경이 chapter의 흐름에 따라서 감성이 어떻게 변화하는지 분석하고, 본인들이 알고 있는 성경 지식과 대조하며 설명하여라.