**분석과정**

1. **데이터 전처리(Data Preprocessing)**

* **텍스트 전체를 목차별로 정리함**

1. 알파벳, 숫자, 공백, 그리고 점 특수기호를 제외한 나머지 문자 삭제
2. 대분류, 중분류, 소분류에 대응하는 텍스트 추출하기
   1. 정규표현식을 이용하여 대분류,중분류,소분류를 추출
   2. 다중 딕셔너리를 생성하여 대분류, 중분류, 소분류에 대응하는 텍스트 추출하여 정리했다.
      1. 소분류에 대응하는 텍스트들을 정리. 소분류를 key, 텍스트를 value로 하는 딕셔너리를 생성했다.
      2. 그 다음 중분류를 key, 앞에서 생성한 딕셔너리를 value로 하는 이중 딕셔너리 생성했다
      3. 마찬가지로 대분류를 key, 앞에서 생성한 딕셔너리를 value로 하는 삼중 딕셔너리를 생성했다.
   3. 텍스트 추출시 딕셔너리의 key, value값을 이용하여 특정 목차의 텍스트를 추출할 수 있도록 만들었다.
   4. 목차별로 정리한 다중 딕셔너리를 pickle 확장자로 저장하여 언제든지 불러와서 분석에 활용할 수 있도록 했다.

* **전처리 과정**

1. 목차와 특수기호 삭제

* 텍스트 분석에 불필요한 ‘IV’, ‘Gong Son Chou II’ 등 로마숫자로 이루어진 목차와 ‘· ‘등 특수기호를 모두 삭제했다.

1. 개체명 인식으로 인물의 이름을 모두 삭제

* 인물의 성품은 특정 인물의 이름보다는 인물의 행위나 상태를 나타내는 어휘에서 직관적으로 파악할 수 있을 것이라고 생각했다.
* Humboldt University of Berlin에서 개발한 딥러닝 기반의 자연어처리 라이브러리인 ‘flair’을 활용했다. flair 라이브러리의 영어 텍스트 데이터용 개체명 인식 패키지를 활용하여 인물의 이름으로 인식되는 어휘를 찾은 뒤 해당 어휘들을 텍스트에서 모두 제거했다.

1. 전체 텍스트를 소문자로 변환

* 파이썬 프로그래밍 언어는 대소문자를 구분하므로 서로 스펠링이 같은 단어들이더라도 단어를 구성하는 알파벳의 대소문자가 다르면 각각 다른 단어로 인식한다. 이를 방지하기 위해 모든 텍스트를 소문자로 변환했다.

1. 동의어(Synonym) 통일

* 소리는 다르지만 문맥상 같은 의미로 판단되는 단어들을 하나의 표현으로 통일했다.
* 예를 들어 majesty, king, son of heaven 등의 단어는 문맥상 모두 ‘king’을 의미하는 단어이므로 모두 ‘king’으로 변환했다.

1. 명사구 추출

* 특정 형용사와 그 형용사의 직접적인 수식을 받는 명사로 구성된 명사구를 추출했다.
* 특정 명사에 어떤 형용사가 수식하는지에 따라 그 명사에 대한 반응이 달라질 수 있다. 예를 들어 king이라는 단어에 benevolent(자비로운), inferior(열등한)이 각각 수식하게 되면 왕에 대한 평가가 완전히 달라지게 된다. 따라서 명사구를 추출하는 작업은 매우 중요하다.
* spacy 패키지를 활용하여 형용사 뒤에 바로 명사가 따르는 형태의 표현을 명사구로 인식하도록 하여 명사구를 추출했다. 추출된 명사구의 형용사와 명사는 서로 공백으로 구분되어 있는데 형태소 분석과 표제어 추출 과정에서 명사구가 하나의 단어로 인식되도록 공백을 ‘-’으로 치환하였다.

1. 형태소 분석

* ‘NLTK’ 라이브러리에서 제공하는 품사 태깅 패키지를 활용하여 텍스트의 토큰별 형태소를 분석했다.
* 텍스트의 핵심어를 추출하기 위해 명사, 형용사, 동사로 태깅된 단어들만 추출했다. 핵심적인 의미를 내포할 수 있는 명사와 이러한 명사의 상태를 수식하는 형용사, 그리고 명사의 행위를 기술하는 동사 단어를 추출하고 각 품사의 단어들간의 관계를 유기적으로 분석하여 텍스트의 전체적인 주제를 파악하고자 했다.

1. 표제어 추출(Lemmatization)

* 자연어처리 라이브러리인 ‘spacy’의 표제어 추출 기능을 활용했다.
* 품사나 단수/복수에 따라 형태가 다르더라도 의미는 서로 같은 단어들을 모두 동일한 단어로 처리하기 위해 각 단어의 표제어를 추출했다.
* 한편 명사구의 표제어 추출 과정에서 명사구의 표제어를 추출한 뒤 형용사 단어와 명사 단어 사이에 표시되어 있는 ‘-’를 제거하여 두 단어가 접합된 형태로 명사구를 표현했다. 예를 들면 ‘benevolent-men’의 표제어 추출 결과인 ‘benevolent-man’을 ‘benevolentman’으로 바꾸어 표현했다. 텍스트 전처리 후 토픽모델링과정에서 ‘-’로 묶인 형태의 표현을 하나의 단어로 인식하지 않고 ‘-’으로 연결된 두 단어를 각각 따로 처리하기 때문에 토픽모델링과정에서 명사구가 하나의 단어로 인식될 수 있도록 불가피하게 ‘-’을 제거하였다.

1. 불용어(stop words) 제거

* he, it 등의 대명사처럼 중요한 의미를 지니지 않아 분석에 불필요한 단어들인 불용어들을 찾아 텍스트에서 삭제했다.
* NLTK 라이브러리에서 제공하는 불용어 사전을 참고하여 사전에 포함되어있는 불용어를 1차적으로 삭제한 뒤, 삭제된 결과를 확인하면서 추가로 삭제해야 할 불용어들을 찾아 사전에 추가했다. 이렇게 갱신된 사전을 다시 참고하여 더 많은 불용어를 삭제하고 분석 정확도를 높일 수 있었다.

**B. 빈도 분석(Frequency Analysis)**

* **텍스트에 등장하는 단어들의 빈도를 계산하여 고빈도 순으로 상위 15개씩 추출했다.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 삭제되지 않은 부분 | | |  | 삭제된 부분 | | |
| 순위 | 단어 | 빈도 | 순위 | 단어 | 빈도 |
| 1 | king | 223 | 1 | king | 443 |
| 2 | people | 208 | 2 | people | 220 |
| 3 | kingdom | 138 | 3 | kingdom | 161 |
| 4 | principle | 59 | 4 | officer | 63 |
| 5 | good | 55 | 5 | hear | 41 |
| 6 | parent | 55 | 6 | virtue | 38 |
| 7 | superior man | 54 | 7 | son | 37 |
| 8 | righteousness | 52 | 8 | benevolence | 35 |
| 9 | scholar | 47 | 9 | receive | 33 |
| 10 | water | 47 | 10 | word | 31 |
| 11 | great | 44 | 11 | righteousness | 31 |
| 12 | love | 43 | 12 | death | 30 |
| 13 | officer | 43 | 13 | office | 30 |
| 14 | benevolence | 42 | 14 | scholar | 30 |
| 15 | nature | 42 | 15 | send | 29 |

**<**삭제되지 않은 부분과 삭제된 부분 각 텍스트에 대한 빈도 분석 표**>**

* **먼저 각 텍스트별 고빈도 어휘에서 공통적으로 등장하는 어휘를 분석했다.**
* 두 텍스트의 고빈도 어휘 상위 15개를 분석한 결과 **people, king, kingdom, righteousness, benevolence, scholar, officer**이 공통적으로 등장했다. 특히 두 텍스트 모두 공통적으로 상위 3개의 고빈도 어휘로 **king, people, kingdom**이 분석되었는데 이를 통해 두 텍스트 모두 왕과 왕실 그리고 백성의 관계에 대해 주로 언급할 것으로 추측할 수 있다. 또한 다른 공통 어휘인 **righteousness, benevolence, scholar, officer**과 연계하여 내용을 분석하면 당시의 봉건적 시대상황과 맹자의 정치적 사상을 고려했을 때 두 텍스트 모두 왕과 왕실이 신하와 백성에게 베풀어야 하는 자비로움과 이러한 행위에 대한 학자들의 사상에 대해 언급할 것으로 추측할 수 있다.

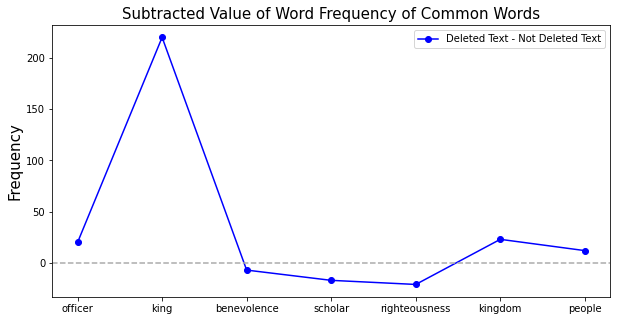


<삭제되지 않은 부분 빈도분석 시각화>



<삭제된 부분 빈도분석 시각화>

* 한편 7개의 공통 어휘에 대한 삭제된 부분과 삭제되지 않은 부분의 빈도 차이를 계산한 결과, 삭제된 부분의 **kingdom, king, people, officer**의 빈도가 삭제되지 않은 부분보다 더 높다는 것을 확인할 수 있었다. 특히 **king**의 빈도가 삭제된 부분에서 매우 큰 폭으로 더 높은 것을 확인할 수 있었는데 이를 통해 삭제된 부분에서 왕에 대한 언급이 많고 왕의 품행이나 역할에 대한 내용이 매우 빈번하게 다뤄졌을 것이라고 추측할 수 있다.



<공통 어휘에 대한 텍스트별 빈도수 비교 그래프>

* **다음으로 각 텍스트별 고빈도 어휘에서 공통 어휘를 제외한 어휘를 분석했다.**
* 우선 삭제되지 않은 부분의 고빈도 어휘중에서 공통 어휘를 제외한 어휘로는 **good, great, love, nature, parent, principle, superior man, water**이 있는데 이를 통해 삭제되지 않은 부분에서는 군자라면 갖춰야 할 선한 본성, 부모와 자식간의 사랑 등 기본적인 유학(유교)적 이론, 사상이 다뤄지고 있다는 것을 추측할 수 있다.
* 반면 삭제된 부분의 고빈도 어휘중에서 공통 어휘를 제외한 어휘로는 **death, hear, office, receive, send, son, virtue, word** 등이 있는데 이를 통해 삭제된 부분의 내용에는 국가의 조정을 살피는데 있어서 군주와 관료 등 국가의 지배층이 자식을 대하듯 백성의 이야기를 듣고 그들이 행복한 생활을 누릴 수 있도록 좋은 정책을 베푸는 등 좋은 정치 방식의 미덕에 대한 내용이 다뤄지고 있을 것으로 추측할 수 있다. 한편 ‘죽음’이라는 단어가 등장한 것으로 보아 삭제된 부분의 내용에는 죽음과 같은 부정적인 내용이 적지 않게 언급되고 있을 것이라고도 추측할 수 있다.
* 정리하자면 삭제되지 않은 부분에서는 선한 본성 등 일반적인 유교적 가치에 대한 내용 등 이론적인 부분이 주로 언급되는 반면 삭제된 부분에서는 이론적인 내용보다는 국가의 지배층이 권력자로서 국가와 백성에게 이로운 방향으로 나라를 이끌어가도록 하는 현실적인 조언이 언급된다고 추측할 수 있다. 즉, 삭제된 내용은 삭제되지 않은 내용에 비해 정치인 등 좋은 권력자가 갖춰야 할 자질에 대해 더욱 직접적으로 언급하고 있을 것으로 추측할 수 있다. 좋은 권력자의 자질에 대한 내용을 다룰 때 나쁜 권력자의 행위, 자질을 비교하기도 하는데 이러한 행위를 하는 권력자의 사례에 당시 왕이 본인을 투영함으로써 이러한 내용에 불쾌감을 가진 나머지 불쾌함을 주는 내용을 맹자 원본에서 삭제하도록 지시했을 것이라고 추론할 수 있다.

**C. 토픽 모델링(Topic Modeling)**

* **토픽모델링은 여러 문서에서 주제(topic)을 추출하는 기법이다. 토픽 모델링으로 추출한 주제들로부터 각 텍스트의 핵심 내용을 파악할 수 있다. 토픽 모델링의 최종 결과물로 텍스트 내용에 중요한 영향을 미칠 것으로 분석된 단어들의 집합이 도출되는데, 각 단어들을 조합하여 텍스트의 전반적인 내용을 함축하는 주제를 빠르게 추출할 수 있다. 이를 위해 ‘잠재 디리클레 할당(LDA)’ 토픽모델링 알고리즘과 ‘KeyBERT’ 키워드 추출 알고리즘을 활용했다.**

1. Latent Dirichlet Allocation(LDA) Topic Modeling

* 잠재 디리클레 할당(이하 LDA) 알고리즘은 대표적인 토픽 모델링 알고리즘으로, 분석하고자 하는 주제의 개수를 설정하면 분석 대상인 여러 개의 텍스트에서 주제를 구성할 만한 여러 개의 단어를 통계적으로 추출하는 주제 분석 기법이다.
* 파이썬 내장 라이브러리인 ‘사이킷 런(scikit-learn)’에서 제공하는 메소드를 활용했다.
* 분석하고자 하는 주제의 개수는 5개로 설정했고 각 주제를 구성하는 단어들 중 10개만 보이도록 설정했다.

1. 삭제되지 않은 부분의 LDA 토픽 모델링 결과

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **주제1** | **주제2** | **주제3** | **주제4** | **주제5** |
| parent | scholar | good | eat | superior man |
| superior man | office | mind | nourish | righteousness |
| son | officer | nature | hand | scholar |
| turn | parent | righteousness | grain | sage |
| father | desire | water | follow | seek |
| speak | field | benevolence | superior man | desire |
| water | accord | officer | pursue | good |
| disciple | order | brother | want | nature |
| great | receive | great | officer | eat |
| word | father | eat | mind | nourish |

주제1: 군자가 갖춰야 할 부모-자식 간의 관계(**parent, son, superior man, father, great**)

주제2: 부모-자식 관계처럼 정사를 다뤄야 할 정치인의 자세(**office, officer, parent, receive, father**)

주제3: 정치인이 좋은 정치를 하기 위해 가져야 하는 선한 본성(**good, mind, nature, righteousness, benevolence, officer, great**)

주제4: 백성의 생업이 풍요로워지기를 추구하는 군자의 마음을 따라야 하는 정치인(**eat, nourish, grain, follow, superior man, pursue, officer, mind**)

주제5: 백성의 생업이 풍요로워지는 것을 열망하는 군자와 성인의 본성(**superior man, righteousness, sage, seek, desire, good, nature, eat, nourish)**

1. 삭제된 부분의 LDA 토픽 모델링 결과

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **주제1** | **주제2** | **주제3** | **주제4** | **주제5** |
| speak | present | love | benevolent | brother |
| word | heaven | music | water | eat |
| mind | scholar | death | rule | send |
| court | office | pleasure | family | live |
| nature | virtue | hear | heart | superior man |
| serve | throne | kill | regard | support |
| carry | proper | wife | leave | carry |
| hear | govern | brother | wife | seek |
| seek | support | tell | hear | mean |
| leave | account | desire | mean | desire |

주제1: 정치인이 추구해야 할 본성(**mind, court, nature, seek**)

주제2: 왕이 왕위를 적절히 유지하기 위한 미덕있는 통치 방식(**office, virtue, throne, proper, govern, support**)

주제3: 음악 감상의 기쁨과 같이 기쁜 마음으로 베풀어야 할 가족에 대한 사랑(**love, music, pleasure, hear, wife, brother, desire**)

주제4: 가족을 대하는 마음과 같은 자비로운 통치(**benevolent, rule, family, heart, regard, wife**)

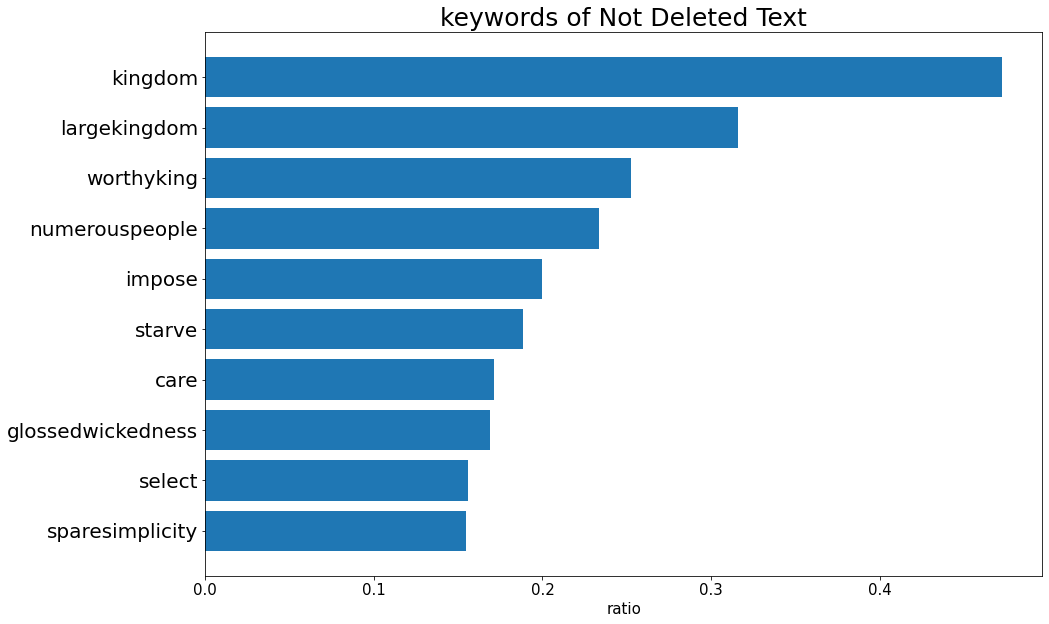
주제5: 군자의 삶을 살기를 추구하고자 하는 열망(**live, superior man, seek, desire**)

1. KeyBERT Topic Modeling

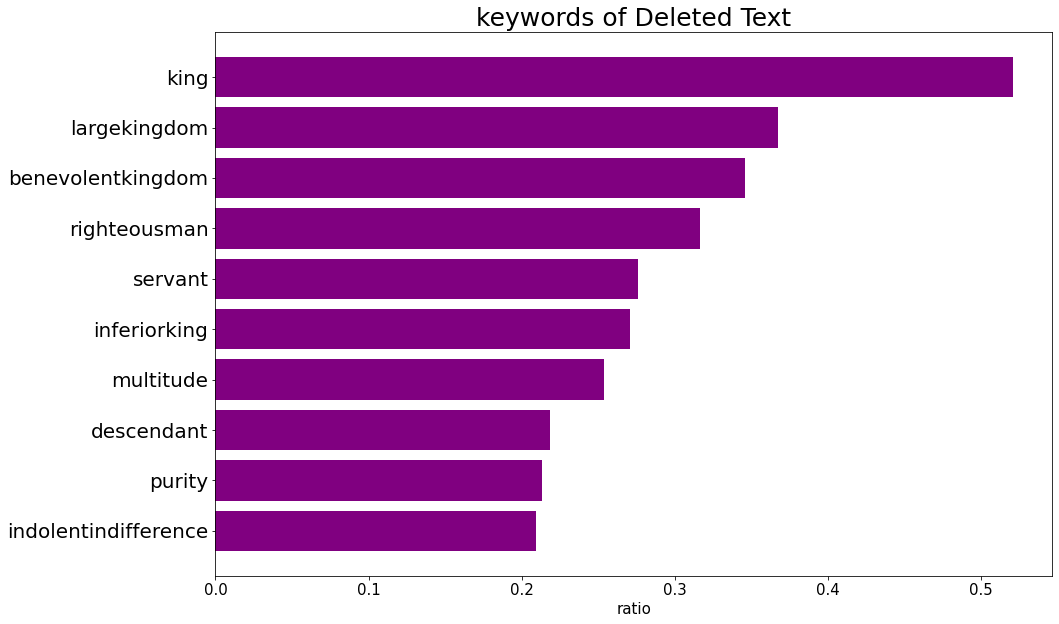
* KeyBERT는 대표적인 자연어처리 딥러닝 알고리즘인 BERT를 기반으로 제작된 키워드 추출 알고리즘이다. 텍스트의 문맥을 파악하여 텍스트의 내용을 구성하는 핵심 키워드를 추출한다.

1. 전체 텍스트 대상 키워드 추출

우선 삭제되지 않은 부분의 전체 텍스트에 대한 키워드 추출 결과, **kingdom, large kingdom, worthy king, numerous people, impose, starve, care, glossed wickedness, select, spare simplicity**가 키워드로 추출되었는데 ‘worthy king’처럼 왕에 대한 긍정적인 표현이 등장한다.



다음으로 삭제된 부분의 전체 텍스트에 대한 키워드 추출 결과, **king, large kingdom, benevolent kingdom, righteous man, servant, inferior king, multitude, descendant, purity, indolent indifference**가 키워드로 추출되었는데 ‘indolent indiffence’같은 부정적인 느낌의 단어와 ’inferior king’같은 왕에 대한 직접적인 부정적 표현이 등장한다.



1. 텍스트 대분류별 키워드 추출

우선 삭제되지 않은 부분의 대분류별 키워드 추출 결과는 다음과 같다.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Liang Hui Wang** | **Gong Sun Chou** | **Teng Wen Gong** | **Li Lou** | **Wan Zhang** | **Gao Zi** | **Jin Xin** |
| kingdom | king | king | righteousness | sorrow | righteousness | mental constitution |
| king | honourable dignity | worthy king | sincerity | desire | nature | spiritual nature |
| great kingdom | servant | successor | obey | murmur | philosopher | benevolent embrace |
| old people | common people | proper source | distress | noble | willow | perform |
| consume | surpass | inferior officer | tranquillity | beautiful people | propriety | purity |
| fowl | convey | mournfulness | seek | filial piety | admirable virtue | personal cultivation |
| virtuous king | taxis | ritual | mary | uncultivated person | benevolent man | righteous |
| grain | enforce | great urgency | provide | field | passage | sustain |
| burden | distress | accompany | narrow lane | doctrine | observance | inquire |
| select | select | exert | responsive attachment | conform | embarkment | mild harmony |

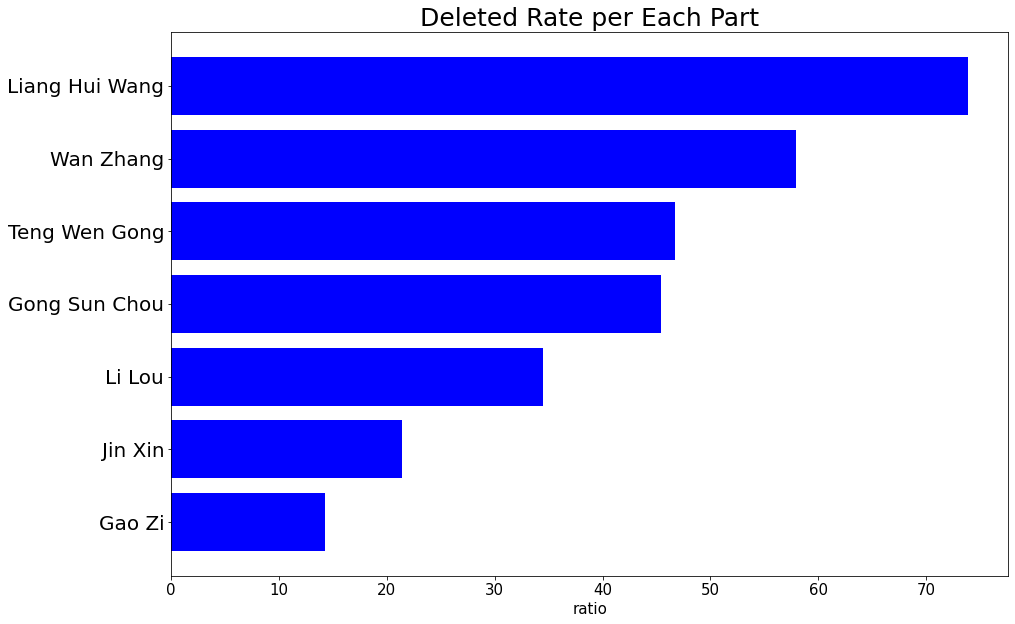
* **Consume, grain, field, uncultivated person, personal cultivation** 등 백성의 생업인 농업 활동과 관련된 내용이 포함되어 있다.
* **benevolent man, filial piety, conform, righteous, purity** 등 유교에서 중요하게 여기는 가치에 대한 내용이 포함되어 있다.
* **great kingdom, virtuous king, honourable dignity, worthy king** 등 왕과 왕실에 대한 긍정적인 내용이 언급된다.

다음으로 삭제된 부분의 대분류별 키워드 추출 결과는 다음과 같다.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Liang Hui Wang** | **Gong Sun Chou** | **Teng Wen Gong** | **Li Lou** | **Wan Zhang** | **Gao Zi** | **Jin Xin** |
| king | king | dynasty | kingdom | marry | king | antiquity |
| ancient kingdom | royal dignity | king | benevolent rule | servant | righteousness | nobility |
| righteousness | noble | ancient king | instrument | impose | wonder | virtuous king |
| servant | attainment | benevolent kingdom | superior | family | arise | king |
| inferior king | subordinate | inferior man | practise | honourable situation | intelligence | righteousness |
| chariot | tyrannical kingdom | hereditary | compass | middle kingdom | avert | follower |
| multitude | flourish | purity | flourish | inferior officer | expect | polite consideration |
| surpass | inquire | repay | example | destitute | persuade | chariot |
| fatherless | purity | distress | ancestral temple | refined idea | seek | sacrifice |
| common people | indolent indifference | necessary millet | violate benevolence | complete concert | right path | execute |

* **ancient kingdom, ancient king, antiquity** 등 과거의 왕과 왕조의 사례를 언급하고 있다.
* **Indolent indifference, violate benevolence, destitute, inferior man** 등 유교적 관점에서 부정적인 내용이 포함되어 있다.
* **Inferior king, tyrannical kingdom, inferior officer** 등 정치인에 대한 부정적인 평가가 드러나는데 특히 왕과 왕실에 대한 부정적인 내용이 직접적으로 드러난다.

특히 전체 대분류 중에서 ‘Liang Hui Wang’은 전체 소분류 개수 대비 삭제된 소분류의 개수의 비율이 73.91%로 가장 높아 가장 많은 내용이 삭제된 단락으로 분석되었다.



<단락별 전체 내용 대비 삭제된 내용 비율>

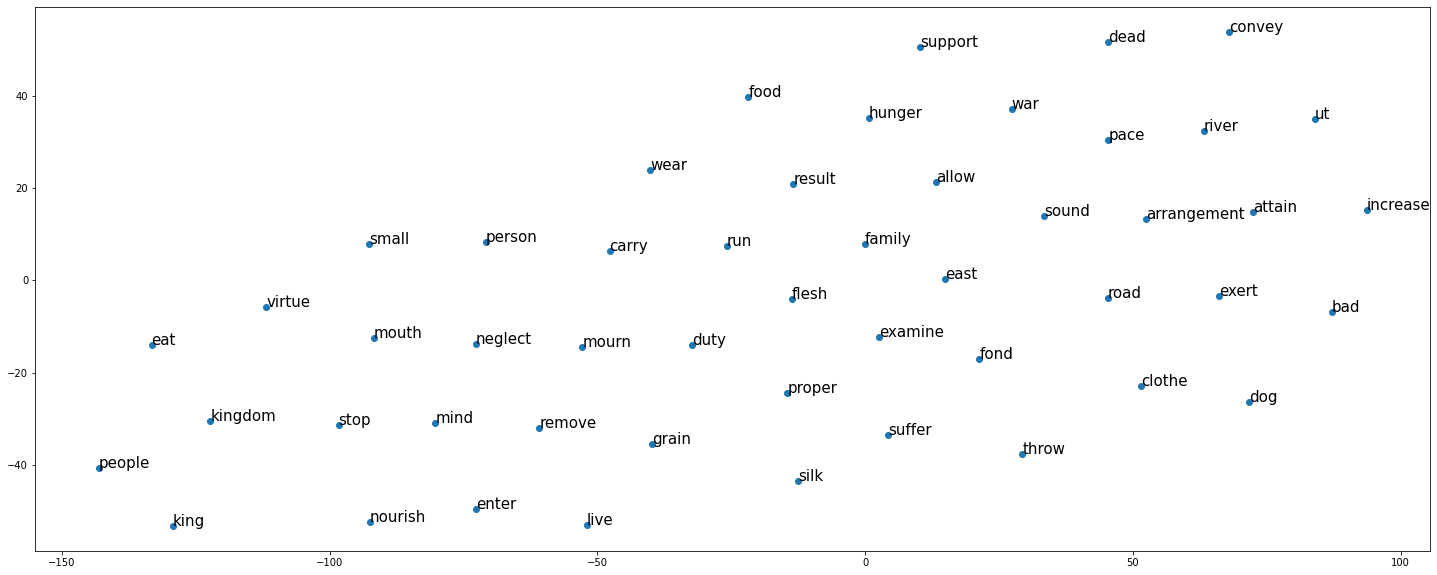
이 단락의 KeyBert 수행 결과에 **’inferior king’**같이 왕에 대해 직접적으로 부정적인 표현이 등장한 것을 고려하면 왕의 권위나 명예를 떨어뜨릴만한 내용이 포함되어 있어서 해당 단락의 내용이 가장 많이 삭제되었을 것이라고 추정할 수 있다.

**D. 연관어 분석(Word Association Analysis)**

* **연관어 분석은 특정 키워드를 중심으로 연관성이 높은 단어들을 추출하는 분석 기법이다. 인공신경망 기반의 단어 임베딩 기법인 Word2Vec을 활용하여 텍스트의 단어들을 같은 차원의 공간에 배치하고 단어들간의 코사인 유사도를 구하여 단어간의 유사 관계를 분석했다. window size를 5로 설정하여 모델을 학습시켰는데 텍스트에서 최소 5번 이상 등장한 단어들만 100차원의 공간에 임베딩했다.**

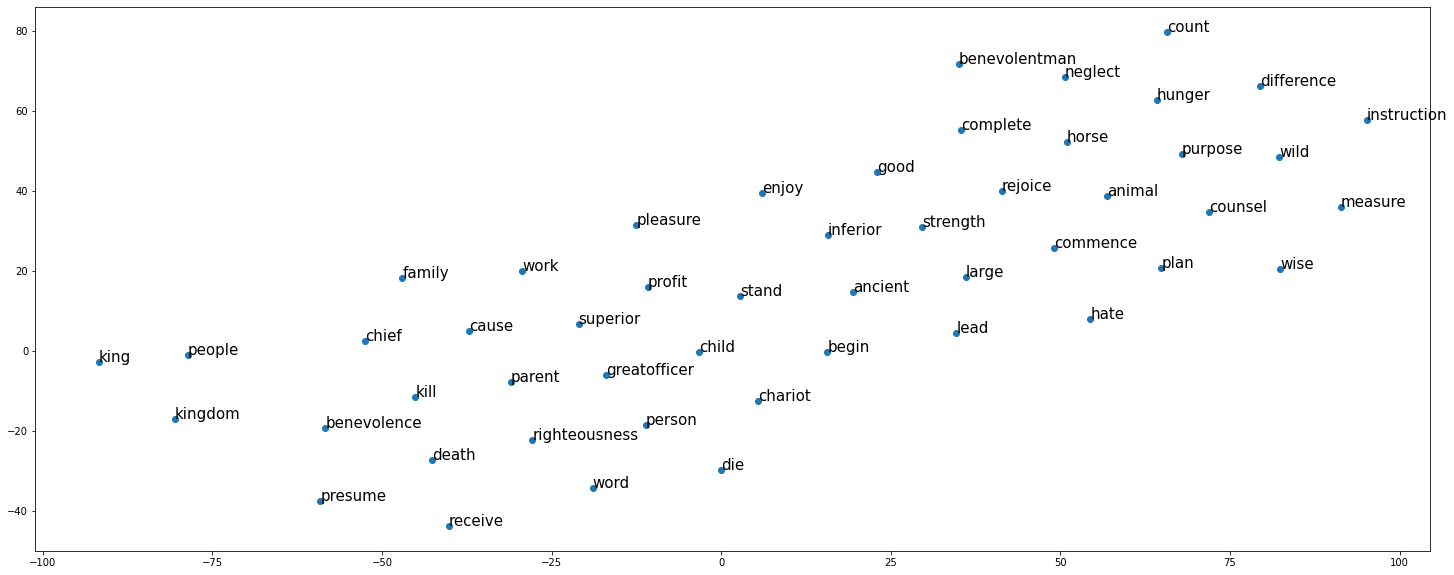
Word2vec으로 100차원 공간에 할당된 단어들의 위치를 시각적으로 확인하기 위해 T-SNE로 100차원을 2차원으로 차원축소하여 단어들의 위치를 시각화하였다. 전체 단어들 중 50개의 단어들만 선별하여 차원축소 기법으로 단어들의 위치를 2차원 공간에 시각화했다.

먼저 삭제되지 않은 부분의 단어들의 위치 시각화 결과는 다음과 같다.

****

king, kingdom 주위의 단어들을 살펴보면 **eat, nourish, grain, food** 등 백성의 생업과 관련된 어휘들이 등장하는 것을 확인할 수 있다.

다음으로 삭제된 부분의 단어들의 위치 시각화 결과는 다음과 같다.



king, kingdom 주위의 단어들을 살펴보면 **benevolence, benevolent man, superior, righteousness** 등 유교적 인간상과 관련된 단어들을 볼 수 있고 kill, death, die 등 죽음과 관련된 단어와 **great officer, profit** 등 정치, 행정 관련 단어들도 찾아볼 수 있다.

**결론**

삭제되지 않은 부분과 삭제된 부분 모두 공통적으로 유교적 인간상과 관련된 어휘들이 등장하는 것을 확인할 수 있었다. 한편 삭제되지 않은 부분은 주로 백성의 생업과 관련된 내용과 왕과 왕실에 대한 긍정적인 내용이 언급되는 반면 삭제된 부분에서는 정치, 행정적으로 부정적인 내용이 등장하는데 특히 왕과 왕실에 대한 직접적으로 부정적인 표현이 등장한다. 이러한 점에서 내용 삭제를 지시한 당시 왕은 이러한 내용이 왕의 권위를 실추시키는 것으로 받아들여 삭제를 지시했을 것이라고 추정할 수 있다.