2021-2 코퍼스언어학 기말과제

20191245 노유정

최근 한국에서 뿐만이 아니라 전세계에서 인정받은 자랑스러운 한국인들이 많이 생겼습니다. 작년 한 해 ‘기생충 열풍’을 몰고 오며 아카데미 4관왕을 하게 된 봉준호 감독, 한국인 최초로 동시에 아시아계 2번째로 오스카 여우주연상을 수상한 배우 윤여정, 넷플릭스에서 제공중인 모든 컨텐츠 중 모든 국가에서 종합 순위 1위라는 역대급 기록을 세우며 수많은 패러디물을 남긴 드라마 오징어게임(의 제작진 및 배우들)이 있습니다.  
하지만 이들에 앞서 해외에 한국과 자신들을 알려온 사람들을 묻는다면 대한민국인 모두가 답할 것입니다. 바로 BTS, 방탄소년단입니다. 그들은 데뷔 이후로 외국 팬들에게 많은 인기를 얻었고 특히 2017년부터 빌보드를 시작으로 해외에서 더욱 활발하게 활동하게 되었습니다. 이에 방탄소년단의 RM은 리더로서 책임감을 갖고 영어를 공부하여, 외국인들의 감탄을 자아낼 정도로 능숙하게 영어로 인터뷰, 수상소감, 연설 등을 하여 주목받은 바가 여럿 있습니다. 여기서 저는 ‘방탄소년단의 영어’를 분석해보는 것에 관심을 두게 되었습니다.  
그러던 중 2018년 당시 유명했던 그들의 UN 연설 ‘Speak Yourself’를 떠올리게 되었고 알아본 결과 때마침 작년에 다시 UN 연설 ‘Life goes on, Let’s live on’을 하였다는 것을 알게 되었습니다. 따라서 저는 방탄소년단의 두 UN 연설을 분석 데이터로 선정하였고 텍스트에 문장부호들이 포함되어 있어 이를 삭제하였습니다. 분석으로 알고자 하는 것은 크게 방탄소년단 영어 화법의 특징, 코로나를 기점으로 유의미한 변화가 있는지 등이 있습니다.

1. 빈도분석(for 제어문, if 제어문, paste, substring, gsub 사용)

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

우선 분석을 위해 두 연설을 어휘단위 벡터로 불러와 2018년의 연설을 UN18, 2020년의 연설을 UN20에 저장하였습니다. 분석하고자 하는 파일이 2개뿐이므로 굳이 제어문을 사용하기보다는 scan을 2번 사용하는 것이 더욱 효율적이겠으나, 만약 방탄소년단이 앞으로 더 많은 연설을 하게 된다면 약간의 수정으로 모든 연설을 같이 분석할 수 있게 되므로 for과 if의 반복 제어문을 사용하였습니다. 정확한 분석을 위해 gsub함수를 이용하여 어휘 양옆 문장부호 연쇄를 제거하였고 빈도분석을 위해 내림차순으로 table을 형성하고 데이터프레임을 만들었습니다. 또한 for 제어문 속 scan 과정에서 paste함수를, 파일 이름에서 숫자를 추출해내기 위해 substring을 사용하였습니다.

BNC 문어 15개 상위 빈도어는 the, of, and, a, in ,to, is, to, was, it, for, that, with, he, be  
구어 15개 상위 빈도어는 the, I, you, and, it, a, ‘s, to, of, that, n’t, in, we, is, do입니다. 이중 가장 눈에 띄게 나타나는 차이는 문어에서는 for로 목적 및 방향을 나타내는 표현을 사용하고 구어에서는 I, you, we등의 인칭대명사를 많이 사용한다는 것입니다. 연설은 기본적으로 구어에 속하나 문어의 성격을 많이 띈다고 알려져 있습니다. 방탄소년단의 연설에서는 i, my, we, our등의 표현이 많이 쓰이므로 보통의 연설과는 다르게 구어의 성격을 강하게 띈다고 해석할 수 있으며 따라서 이 연설은 정보 전달이나 교훈을 주는 것이 목적인 보통의 연설보다는 다르게, 사람들과 대화하듯이 정서적인 교감을 목적으로 하였다고 해석할 수 있습니다.

1. N-gram

텍스트이(가) 표시된 사진

자동 생성된 설명

자주 사용하는 숙어를 살펴보기 위해 앞서 만든 UNtotal을 이용하여 위와 같은 바이그램을 만들었습니다. UNtotal이 두 어휘벡터를 c로 묶어 놓았기 때문에 UN2018의 마지막 단어와 2020의 첫번째 단어도 바이그램에 포함되었겠지만 확인 결과 위의 head 결과에 영향을 미칠 조합이 아니었으므로 그대로 사용하였습니다.

우선 눈에 띄는 것은 보통의 바이그램은 맥락을 잃어버리는데 그렇지 않은 love myself, speak yourself입니다. 이는 연설의 제목이기도 한데, 매우 길지 않은 연설임에도 불구하고 연설에서 말하고자 하는 바를 반복하여 언급하였음을 알 수 있습니다. 그리고 또 눈에 띄는 것은 i가 가장 많이 쓰인 단어인 만큼 i를 포함한 바이그램이 상위에 위치해있다는 것입니다. 이에 대해 더 자세히 알 수 있을지를 기대하며 trigram을 만들어보았습니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

원래 trigram이 그 빈도수가 적어 분석하기가 어렵다는 단점이 있는데 가뜩이나 어휘수가 그리 많지 않은 텍스트에 적용하였더니 더욱 작은 값이 나왔습니다. 그래도 상위의 trigram들을 살펴보자면 앞서 bigram에서 알고자 하였던 i의 경우 for who i, i used to, i’d like to 등의 자신의 대한 설명을 하며 본인의 경우로 말하고자 하는 바의 예시를 들고자 하였음을 예측해볼 수 있습니다.

1. 키워드분석(집합론 함수 union사용)

텍스트이(가) 표시된 사진

자동 생성된 설명

키워드 분석을 위한 TDM을 만들기 위해 merge를 이용하여 UN18의 텍스트 빈도 내림차순을 기준으로 합집합(all = T)을 만들었습니다. union 사용 결과 합집합을 하여도 전체 크기가 크지 않다는 것을 알았고, wordcloud에서 max.word 제한을 주면 되기에 합집합으로 처리하였습니다. 또한 merge로 만든 matrix를 데이터프레임으로 만들고 빈도이기 때문에 na값을 0으로 처리해도 의미상 문제가 없어 0의 값으로 치환하였습니다.

이렇게 만든 TDM으로 unigram comparison cloud를 만들어 비교해보았습니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

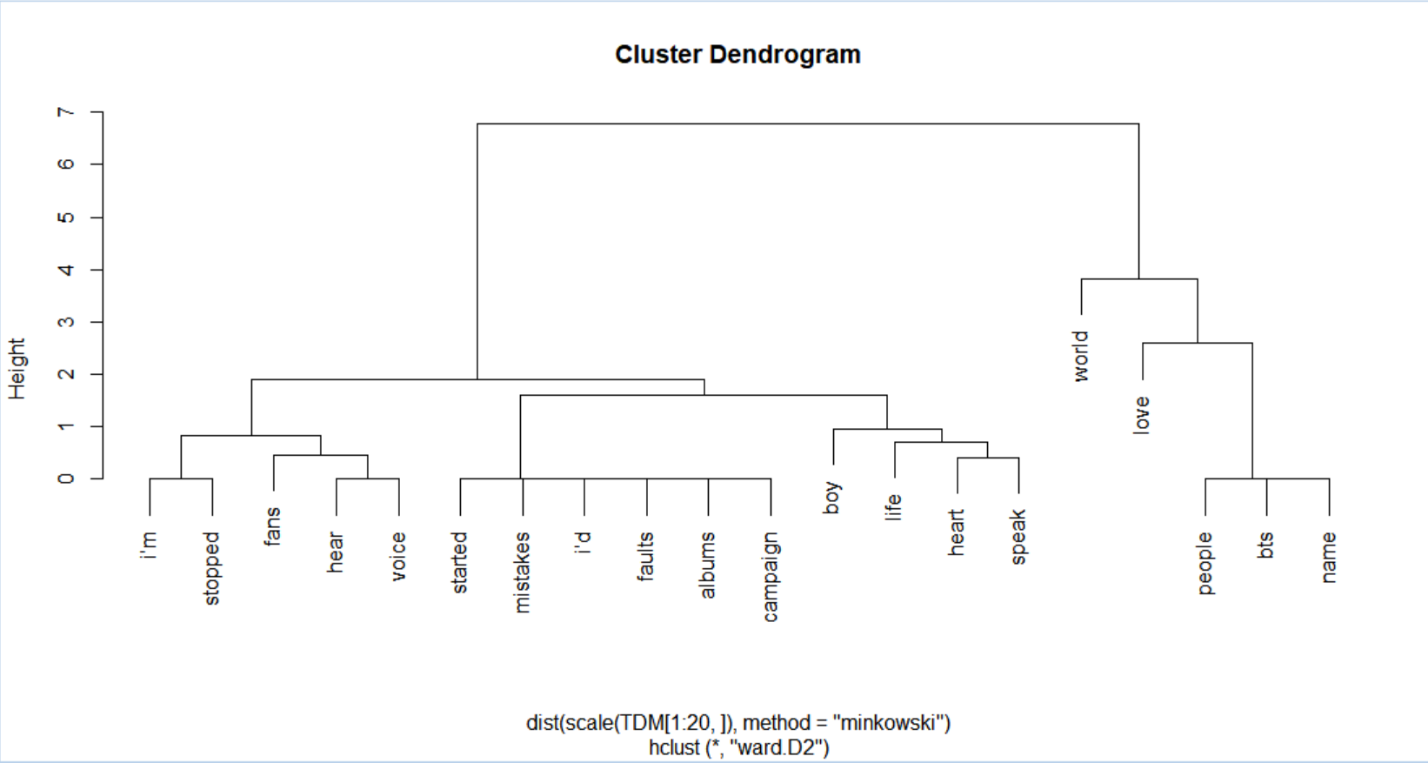
여기서 Freq.x는 2018년의, y는 2020년의 연설을 뜻합니다. 눈에 띄는 차이는 2018년에는 비교적으로 bts, i’m, me 등의 자신의 이야기를 했다면 2020년에는 we, our, world, together, friends 등의 표현을 사용하였습니다. 이는 2020년의 주제인 Life goes on, Let’s live on에서도 알 수 있듯이 코로나 팬데믹으로 지친 전세계인들에게 전하는 말인 만큼 팬들에게 힘을 주고 이전보다 더욱 소속감을 느낄 수 있는 표현을 사용하려고 하였다고 해석할 수 있습니다.

1. 탐색적분석( %in%사용)

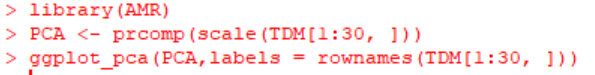
텍스트이(가) 표시된 사진

자동 생성된 설명

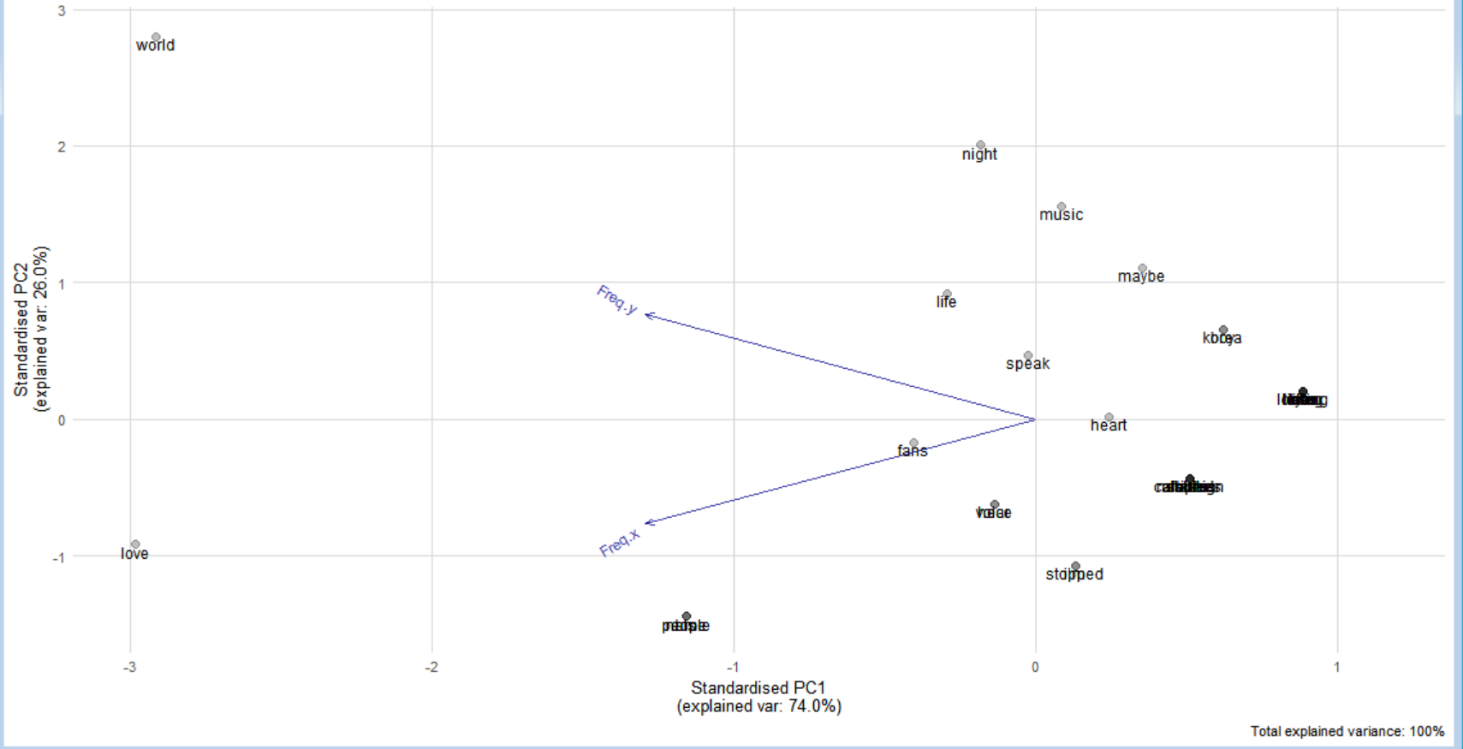
마지막으로 탐색적 분석에서는 수업 중 배운 EnglishStopwords를 이용하여 !와 %in%을 사용해 큰 의미 없는 단어를 제외, 즉 이 연설에서 특별히 쓰인 단어를 추출하였습니다. 그 결과를 이용하여 위의 method를 사용하여 아래의 Cluster Dendrogram을 만들었습니다.



군집분석 결과 크게 2개의 군집이 나타났음을 볼 수 있었습니다. 하지만 마찬가지로 텍스트의 크기가 크지 않아 더 정확한 분석을 위하여 같은 탐색적 분석에 속하며 단어간 연관관계를 알 수 있는 주성분 분석을 진행하였습니다.



AMR패키지를 설치하고 PCA를 진행하여 아래와 같이 2차원 그래프로 나타내었습니다.



하지만 이 역시 마찬가지로 큰 연관성을 알기가 어려웠습니다. 거리상 근접하고 각도로 양의 상관성인지 음의 상관성인지를 파악해야 하는데 육안상 거리와 각도가 거의 같고 일부 어휘들이 겹쳐 있어서 적은 어휘의 한계를 더욱 알게 되었습니다. 우선 군집분석 결과를 살펴보자면 world, love, people, bts 등이 하나의 군집으로 나타나고 fan, voice, mistake, fault, album, campaign 그리고 i’m, i’d 등이 나머지 하나의 군집으로 나타났습니다. 따라서 love 등의 긍정적인 단어와 mistake, fault, album 등의 부정적인(힘듦을 표현한) 단어로 군집이 나뉘었다고 해석해볼 수 있습니다. 이때 fan과 bts, love는 하나의 군집으로 나타날 것이라고 생각했는데 그렇지 않은 것을 보아 계층적 군집분석의 단점인 개별적인 유사도가 고려되기 어렵다는 것이 나타났다고 해석할 수 있습니다.

방탄소년단의 UN 연설을 분석해본 결과 우선 빈도분석에서는 그들의 연설이 일반적으로 정보전달의 목적을 가지는 연설과는 다르게 통상적인 구어처럼 인칭대명사를 많이 사용하였다는 것을 알았고 따라서 정서적인 발언을 많이 하였음을 알 수 있었습니다. 그리고 연어분석에서는 연설의 제목처럼 자신들이 말하고자 하는 바를 반복하여 언급했다는 것을 알 수 있었으며 특히 키워드 분석에서는 코로나 팬데믹 이전이었던 2018년과 그 이후인 2020년의 두 연설 사이에서 유의미한 변화가 있었음을 알 수 있었습니다. 2018년에는 본인들의 이야기를 더 많이 했다면 2020년에는 코로나로 지친 세계인들에게 희망을 주는 메시지를 담으려고 노력했습니다. 마지막으로 탐색적 분석에서는 긍정적인 단어와 부정적인 단어로 군집이 나뉘었으나 텍스트의 길이가 길지 않은 만큼 만족할 만한 결과를 얻지는 못하였습니다.  
 이번 방탄소년단 UN연설 텍스트 분석 및 해석을 진행하면서 가장 아쉬웠던 점은 아무래도 텍스트의 길이가 부족했다는 것입니다. 모든 분석 방법이 그렇지만 특히 탐색적 분석에서 어휘가 더욱 많았다면 훨씬 정확한 결과를 얻을 수 있었을 것이라는 아쉬움이 남습니다. 처음에는 UN연설 뿐만 아니라 빌보드 및 AMA의 수상소감 및 여러 영어 인터뷰도 포함시켜서 분석을 진행하려고 하였으나, 텍스트를 찾기가 어려웠던 것도 있고 연설과 수상소감, 인터뷰는 그 특성이 모두 다르기에 각 텍스트의 길이가 가장 길고 또한 의미 전달이 확실한 연설문만 분석하게 되었습니다. 하지만 분석해본 결과, 이들의 연설은 구어체의 성격을 강하게 띄었기 때문에 차라리 다 가져와서 연설, 수상소감, 인터뷰 3개로 나누어 분석을 진행했다면 지금보다 더 좋은 결과를 내었을 수도 있겠다는 생각을 하였습니다. 그래도 분석의 목적이었던 방탄소년단의 영어 화법과 코로나를 기점으로 한 유의미한 변화에 대해서는 알 수 있었으므로 다행이라고 생각합니다.  
또한 분석을 하면서 역시 프로그램은 숫자로 분석을 해주고, 이를 보기 편하게까지는 만들어주지만 역시 이에 대한 해석은 오롯이 실험자 본인의 몫이라는 것을 깨닫게 되었습니다. 같은 결과를 보고도 어떤 방향으로 해석해야 맞을지가 계속 고민이 되었습니다. 따라서 제 자신의 분석적인 식견을 넓히는 동시에 코딩 역량을 키우기 위해 노력 해야한다는 생각을 하였습니다. 앞으로 제가 진행할 텍스트 분석에서는 결과에 대해 더 정답에 가까운 해석을 하여 결론을 내릴 수 있도록 경험을 쌓아야겠다고 생각했습니다.