**SNU 빅데이터 아카데미**

**-데이터 기반 통계분석 시스템구축-**

학생명: 허현

프로젝트명: 네이버 블로그 API를 활용한 블랙핑크 텍스트 분석 및 시각화

목차

1. 분석 계기 및 개요
2. API 활용 크롤링 과정
3. 데이터 핸들링
4. 워드클라우드2를 활용한 시각화
5. 네트워크 분석을 위한 핸들링 및 시각화
6. 한계점 및 개선사항

**1. 분석 계기 및 개요**

최근 여자친구와 대화 중 블랙핑크의 “지수”라는 멤버가 춤을 못 추지 않냐는 얘기를 나눴습니다. 여자친구가 춤을 좋아하다 보니 춤을 못 추는 아이돌 멤버에 거부감을 느끼는 것 같았습니다. “블랙핑크를 검색해서 나온 글 중에 이런 내용을 다룬 블로그가 많이 있을까?”, “특정 단어가 언급될 때 다른 어떤 단어가 많이 언급될까?” 등이 궁금해졌고, 주제를 블랙핑크로 정했습니다. 사용한 코드는 일부 수업에서 사용한 코드를 차용하였습니다.

**2. API 활용 및 크롤링 과정**

트위터, 네이버, 유튜브 등 다양한 곳에서 개발자를 위해 API를 제공하고 있습니다. 정리가 잘 된 형태로 제공받기 때문에 일부 정해진 코드만 잘 활용하면 쉽게 데이터를 받아올 수 있습니다. 저의 경우 네이버 API를 활용하여 1000개의 데이터를 모았습니다. 검색시 정렬 방법은 유사도로 설정하였고, 블로그 내용으로 분석을 진행하였습니다.

**3. 데이터 핸들링**

받아온 데이터는 분석이 불가능한 수준의 형태였기 때문에, gsub 기능을 활용해 특수 문자를 모두 제거해주었습니다. 그 후 한글 자연어 분석 패키지인 KoNLP로 각 문서별 명사를 추출했습니다. 블랙핑크라는 아이돌 그룹에 대한 글이 사전에 등록된 단어로만 구성되어 있지 않았기 때문에, 처음 몇 개의 글의 명사 추출 형태를 보고 필요하다고 생각되는 멤버 이름, “블핑”과 같은 주요 약자, “아이돌”이나 “블로그” 등의 단어를 추가해주었습니다.

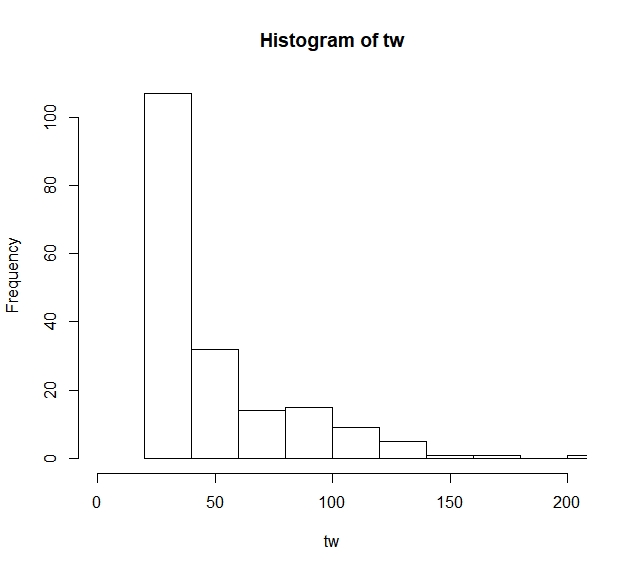
이렇게 단어를 추출하고 나서 어떻게 분석할까에 대한 고민을 시작했는데, 결과적으로 빈도분석과 연관도 분석을 하기로 했습니다. 빈도분석은 4에서, 연관도 분석은 5에서 진행하겠습니다.

**4. 워드클라우드2를 활용한 시각화**

워드클라우드 형태로 분석하기 위해 명사 추출을 통해 리스트 타입이었던 데이터를 데이터프레임 형태로 바꿔주어야 했습니다. 이 때 리스트 안에 있던 모든 단어를 하나의 벡터에 넣고 그 벡터를 테이블형태로 바꿔주어 각 단어가 몇 번씩 나왔는지 알 수 있었습니다. 5회 이상 나온 데이터만 가지고 히스토그램을 그린 결과는 다음과 같았습니다.

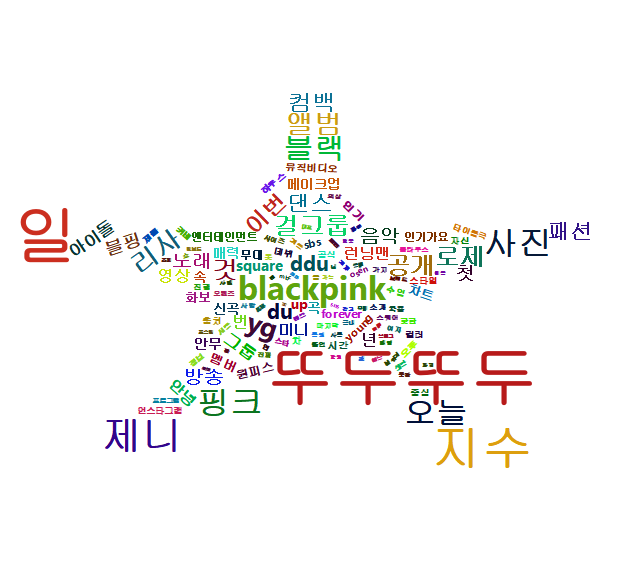


그래프를 통해 5회를 기준으로 사용하기에는 사소하고 정제가 잘 되지 않은 단어들도 많이 나올 것이라고 생각하여 20회를 기준으로 새로 히스토그램을 그렸고, 결과가 예측했던 정도의 멱분포를 따른다고 판단하여 이 데이터를 활용해 다음 단계를 진행했습니다.



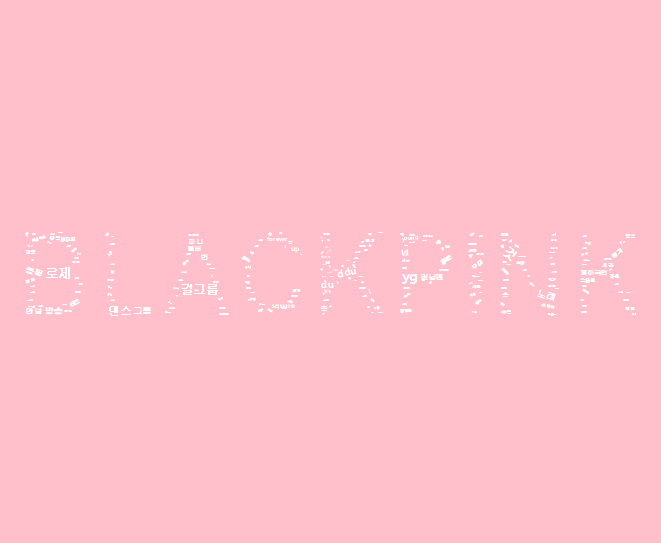
20회 이상의 빈도를 가진 단어로 다음으로 한 것은 또 한 번의 정제였습니다. 데이터프레임에서 words 컬럼에는 각 단어가 들어있었고 Freq 컬럼에는 빈도 수가 들어있었는데, 이 데이터프레임을 모두 살펴보면서 분석으로 사용하기 어려운 단어들의 목록을 만들었습니다. 그 중에는 한 글자짜리 의미를 알 수 없는 단어가 많았고, ‘www’, ‘com’과 같은 왜 나왔는지는 알 수 있지만 분석하고자 하는 바와 연관이 없다고 판단되는 단어도 있었습니다.

Wordcloud2는 내장된 모양이 있을 경우 그 모양으로 형태를 맞춰 단어를 시각화합니다. 블랙핑크는 연예인이고 연예인은 Star라는 단어가 쉽게 연상되기 때문에 ‘shape = “star”’를 입력해서 시각화했고, 결과는 다음과 같습니다.



위의 그래프에서 큰 글자일수록 많이 언급되었음을 알 수 있습니다. 그 중에는 “블랙”과 “핑크”처럼 본래 하나로 다뤄져야 할 단어들이 두 개 혹은 세 개로 나눠졌음을 알 수 있습니다. 최신곡 “뚜두뚜뚜”도 “ddu”와 “du”로 나눠졌는데 단어장에 삽입을 해도 해결이 되지 않아 차후 개선해야 할 부분입니다. 또한 총 1000개의 블로그 글을 순차적으로 받아서 비교적 최신의 내용이 부각되었습니다. “뚜두뚜두”는 최신 타이틀곡으로 굉장히 많이 언급된 반면, 기존 타이틀곡 중에는 위 그림에 포함된 곡이 없었습니다. “런닝맨” 키워드 역시 블랙핑크가 7월 중순에 해당 프로그램에 출연한 것의 영향력으로 많이 언급되었다고 추측할 수 있습니다.

워드클라우드2는 내장된 형태 외에도 사용자가 원하는 그림에 맞게 혹은 사용자가 지정한 글에 맞게 시각화해주기도 합니다.



위의 형태로 시각화를 하는 것은 많은 컴퓨팅 파워와 일정량 이상의 단어가 필요합니다. 위의 별 모양 워드클라우드는 20회 이상 언급된 단어로만 진행된 반면, 직접 입력한 단어를 통한 시각화는 단어가 9글자나 되다보니 같은 단어로 다 채울 수가 없었습니다. 그래서 위의 그래프는 10회 이상 언급된 단어로 만들었습니다. 이 그래프는 강한 의미를 보이기보다는 이 방법도 가능하다는 점을 보이기 위해 만들었습니다.

**5. 네트워크 분석을 위한 핸들링 및 시각화**

네트워크 분석은 한 블로그에서 어떤 단어를 사용했는지의 정보를 살리기 위해 다시 명사를 추출했던 리스트를 사용했습니다. 모든 문서 있는 단어들로 컬럼명을 만들고 각 문서마다 해당 단어의 수를 표기하게끔 매트릭스를 만들었습니다. 다음은 이렇게 만든 매트릭스의 일부입니다.

블랙핑크 제 몸매 다이어트 자극 사진 안녕

[1,] TRUE TRUE TRUE TRUE TRUE TRUE TRUE

[2,] TRUE FALSE FALSE FALSE FALSE FALSE FALSE

[3,] TRUE FALSE FALSE FALSE FALSE FALSE FALSE

[4,] TRUE FALSE FALSE FALSE FALSE FALSE FALSE

[5,] TRUE FALSE FALSE FALSE FALSE FALSE FALSE

[6,] TRUE FALSE FALSE FALSE FALSE FALSE FALSE

[7,] TRUE TRUE FALSE FALSE FALSE FALSE FALSE

궁극적으로 관심있었던 특정 단어가 나왔을 때 다른 단어가 얼마나 언급되었는지를 알기 위해, 위의 매트릭스를 활용해 transpose(Y) X Y 형태로 행렬곱을 해주었습니다. 이렇게 하면 대각선의 값은 각 단어가 총 얼마나 언급되었는지를 알려주고, 대각 외의 값은 i번째 줄의 단어와 j번째 열의 단어가 하나의 문서에서 동시에 언급된 횟수를 알려줍니다. 결과는 다음과 같았습니다.

블랙핑크 제 몸매 다이어트 자극 사진 안녕

블랙핑크 924 180 19 8 5 90 66

제 180 184 6 7 5 25 15

몸매 19 6 19 2 2 4 2

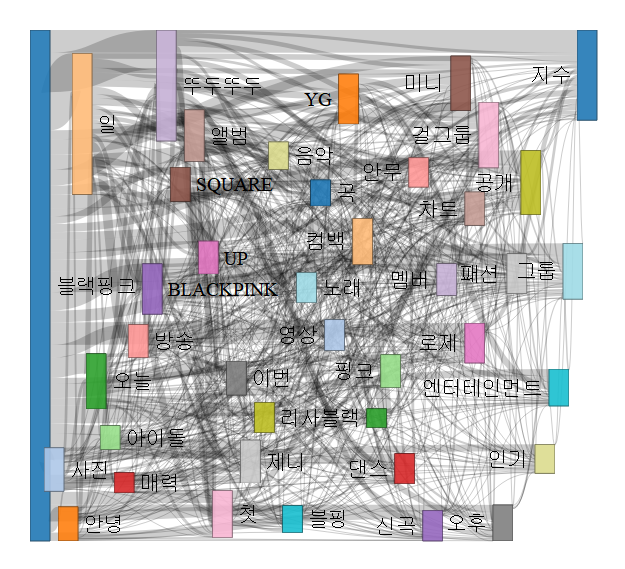
다이어트 8 7 2 8 5 6 2

자극 5 5 2 5 5 5 2

사진 90 25 4 6 5 95 9

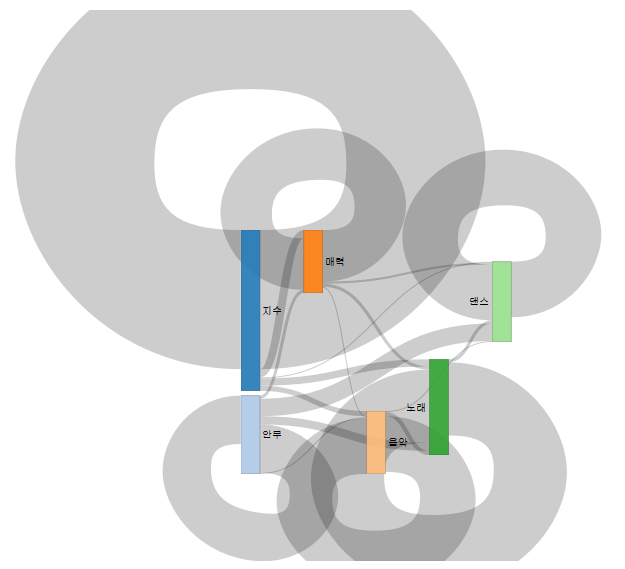
안녕 66 15 2 2 2 9 69

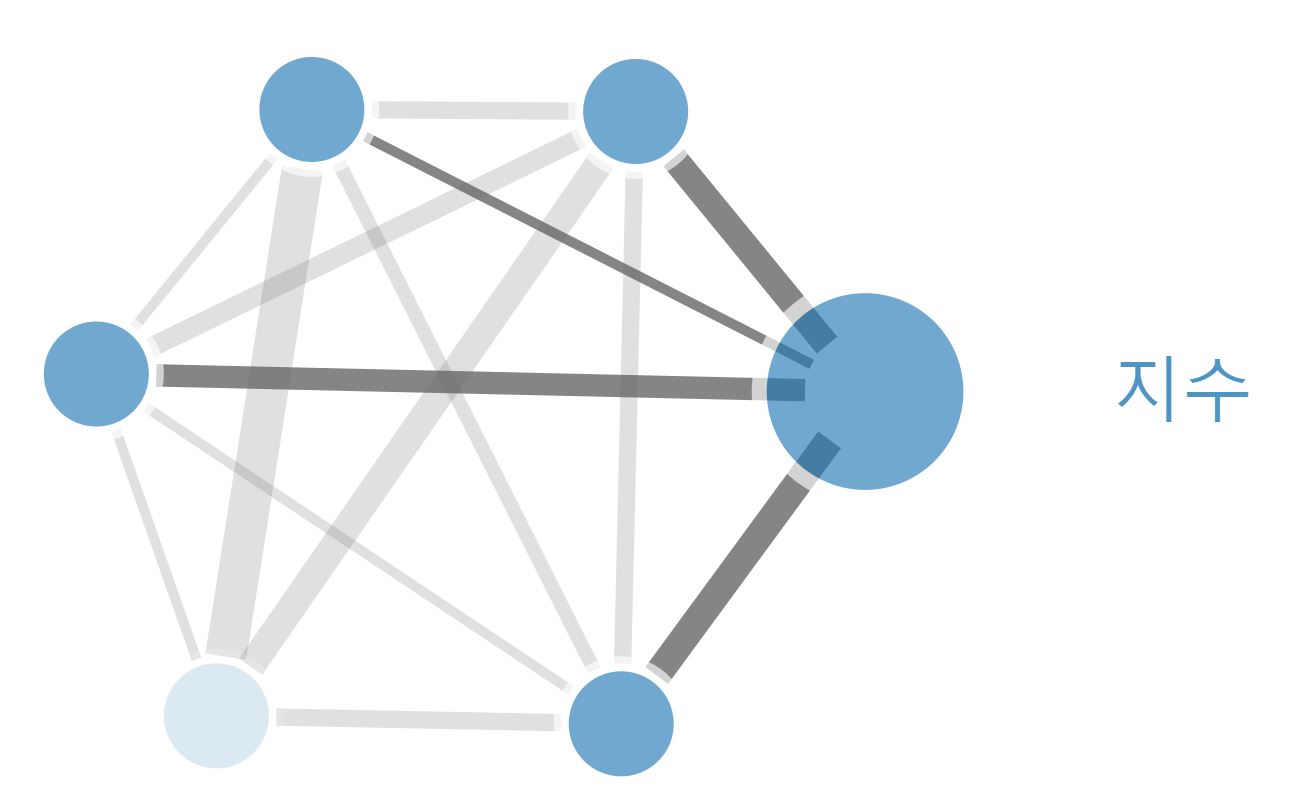
이렇게 다시 처음부터 시작한 것의 문제점은 “제”와 같이 위에서 제거했던 불필요한 단어 모음이 다시 생겼다는 점이었지만, 그 묶음을 재활용해서 해당 단어가 있는 인덱스를 모두 제거해주는 방법으로 문제를 해결했습니다. 또한 이번에는 네트워크 시각화를 할 것이기 때문에 더욱 더 큰 컴퓨팅 파워를 요구했고 이런 한계 때문에 총 언급 횟수가 40회 이상인 단어만 활용하였습니다. networkD3 패키지와 igraph 패키지를 활용하여 SankeyNetwork와 ForceNetwork 그래프를 그렸습니다.



워드클라우드처럼 막대기의 길이가 긴 것이 가장 많이 언급된 단어라고 볼 수 있습니다. 추가적으로 각 막대기 주위에 있는 선은 단어간의 연관성의 정도를 알려주는 것이고 선이 굵을수록 연관도가 높다고 볼 수 있습니다. SankeyNetwork의 경우 난잡해 보이는 효과가 커서 이번 분석에 아주 적절한 시각화는 아니지만 레포트 제출이라는 형식상 ForceNetwork가 동적인 상황에서 더 좋다는 점을 고려하여 레포트에는 정적으로 충분히 표현 가능한 SankeyNetwork를 사용했습니다.

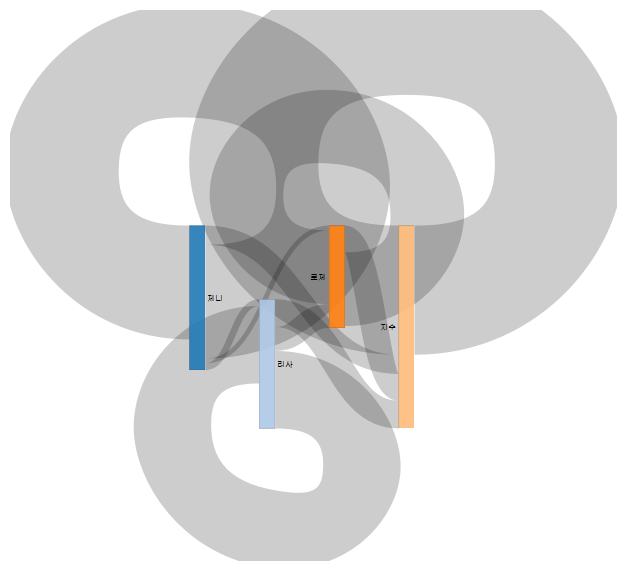
다음으로는 처음 문제를 제기한 “지수”라는 멤버와 춤의 관계를 알아보기 위해, 춤을 포함한 몇 가지 이 멤버에 대해 궁금한 특성만을 가지고 같은 시각화를 해보았습니다. 이번에는 네트워크 형태의 시각화도 시도해보았습니다. 아래의 시각화를 해본 결과, 소수의 특성만을 이용할 때에는 SankeyNetwork가 정보 전달에 더 좋다고 여겨집니다.





분석에 사용한 특성은 "지수", "안무", "댄스", "매력", "노래", "음악"이었고, 아래 네트워크 그림에서 지수라는 노드가 유일하게 연결되지 않은 노드가 “안무”였습니다. “댄스”와의 연결선도 얇은 것으로 보아 춤에 대한 연상력이 떨어지는 멤버인 것은 맞는 것 같습니다. 그래도 “매력”, “노래”, “음악”의 키워드와는 괜찮은 연결성을 보여주어 매력있고 노래를 잘하는 멤버로 보여지는 것 같습니다. 여기까지 시각화를 하고 나서 “댄스”와 “안무”라는 키워드가 “지수”와만 연관성이 없는 것인지 다른 멤버들도 없는 것인지를 보이지 않으면 소위 말하는 “악마의 편집”이 될 수 있다는 생각이 들었습니다. 그래서 “지수”로 되어있던 키워드를 각각 “제니”, “로제”, “리사”로 바꾸어 진행해봤습니다. 그 결과 멤버 “로제”는 “댄스”, “안무” 키워드와의 연결선이 없었고, “제니”는 모두와 연결선이 있었으며, 그룹의 메인 댄서인 “리사”는 “댄스”, “안무” 키워드와 연결선이 없었습니다. “리사”라는 멤버가 가장 인기가 없는 것인지 절대적 언급 양 자체가 부족해서 이런 결과가 나온 것으로 예상되기는 하지만, 적어도 제가 가진 데이터로 나온 결과로는 “지수”라는 멤버가 춤을 못 춰서 연결선이 적다고 할 수는 없겠습니다.

마지막으로 멤버들간 연상 정도를 보기위해 4명의 멤버 이름을 넣고 Sankey Network 시각화를 했습니다.



일단 멤버의 언급 순위를 보면, “지수” 멤버가 가장 인기가 많고, “제니”, “리사”, “로제” 순으로 나왔습니다. 그리고 4명 중 2명을 연결시키는 관계는 6개가 나오는데, “지수” 멤버는 다른 모든 멤버와의 연결선이 굵은 것으로 보아 다른 모든 멤버와의 연상성이 높은 것으로 추정됩니다. “리사” 멤버와 “로제” 멤버는 둘 사이의 연결선이 굵게 나타나서 역시 연상성이 높은 것으로 추정됩니다. “제니” 멤버는 “지수” 멤버 외에는 연결선이 비교적 굵지 않아 다른 멤버들과 연상성이 떨어지는 것으로 추정됩니다.

**6. 한계점 및 개선사항**

개인적으로 재미있는 분석이었지만, 앞서 언급했듯 여러 문제점과 한계점을 가지고 있는 분석이었습니다. 그 중에는 이 분석 자체의 유의성을 해치는 것도 있을 수 있습니다. 무엇보다도 최신 페이지 위주로 자료를 수집하여 최신성은 강조되었지만 “블랙핑크”라는 그룹에 대한 총체적 의미는 잘못 파악되었을 수 있습니다. 이 문제의 경우 네이버에서 허용하는 가져올 수 있는 글의 수에 한계가 있고, API를 활용한 자료 수집이 에러가 많이 일어났기 때문에 발생한 것으로 추후 더 좋은 분석을 위해서는 API 활용법에 대한 공부와 여러 일에 거친 데이터 수집이 요구됩니다.

위의 문제와 연결되어 1000개라는 데이터 수가 데이터 분석을 하기에 적다는 것이 또 하나의 한계점입니다. 분석 과정에서 적게 나온 단어들은 제거했는데, 이런 단어 중에는 제거하지 말았어야 할 단어들도 있었을 수 있습니다. 이는 샘플 데이터가 1000개 밖에 되지 않았다는 점에 기인합니다. 이 역시 추후에 더 많은 데이터 수집으로 해결해야 할 것입니다.

마지막으로 블로그 글의 내용을 명사로 추출하는 과정에서 사전에 등록되지 않았지만 인터넷 상에서 많이 쓰는 표현이나 한글이 가지고 있는 텍스트 분석 시의 어려움 때문에, 추출했어야 할 단어가 추출되지 않았을 수 있고 의미상 한 글자인 것이 여러 글자에 나뉘어 추출되었을 수 있습니다. 전자의 경우에는 원본 텍스트를 보면서 충분히 필요한 단어라면 사전에 등록하는 방식으로 해결할 수 있으며, 후자의 경우에는 좀 더 고급 텍스트 분석 기법을 배워서 활용해야 할 것으로 보입니다. 물론 한글이 가지고 있는 고유의 문제점 때문에 완벽한 해결은 불가능할 것으로 보입니다.