트위터 사용자의 '방학'에 대한 토픽 분석

20190917 통계학과 신효민

트위터 API를 이용하여 곧 다가올 방학(vacation)에 대하여 트위터 사용자들이 어떤 토픽을 언급하는지 분석하고자 한다. 먼저 트위터에 존재하는 트윗 자료를 크롤링 하기 위해 tweepy 패키지를 이용하여 트위터 API 권한을 얻은 후 그 API Key, API Secret Key, Access Token, Access Token Secret을 이용한다. 다음으로 sklearn, numpy, matplotlib, re, nltk 등 필요한 패키지를 import 한다. 이를 이용하여 'vacation'이 언급 된 트윗 텍스트를 최신 순으로 100개 불러왔다.

```
['https://t.co/S8nGv1BxFn… #tiktok #hawaii #oahu #beach #mountain #travel #adventure #explore #photoshoot… https://t.co/PUvY2EJdWA',
 'I just moved our vacation so the kids can attend a #Pride Parade. Going to have many conversations during the parad… https://t.co/pErBJqjd1O',
 'The Better Business Bureau is warning consumers to look out for online scammers this summer, particularly in the fo… https://t.co/NYjUxS8cw9',
 '@shepski79 I brought some back from our vacation in Florida. Best smuggled souvenir in my bag. 😎',
 '"Cuba was the only vacation spot where Negroes could go without bias…." ￼Joe Louis￼#OTD June 6, 1960: African-Amer… https://t.co/ZtZIvaDoUS',
 'RT @ezhrjmn_: Selalu kita tgk kat tiktok, friends go for a vacation and malam malam dorg guna baju pyjamas yg sama? Haa boleh je beli kat @…',
 '@Digant5211 mera vacation bhaiya jaisa royal aur lamba nahi hota hai :P',
 'How To Take The Ultimate Vacation In Denver,￼Colorado https://t.co/iBLkIz2IbG',
 'RT @JiyaMahyavanshi: Instead of wondering when your next vacation is, you ought to set up a life you don' t need to escape from.￼￼RUBI VOCAL…',
 '@GOPChairwoman #CrookedDonald Trump spent 1/3 of his single term on vacation. Another 20% campaigning. Yes, Ronna, we remember 2 years ago.',
 '@LH_btc It was intense but amazing. Went to India, not a usual vacation type place. But saw so much and feels like… https://t.co/stLS8jQDnA',
 '@ms_maclean @nilikm Nobody wants a job, where they r the only 1 carrying all liability/no paid vacation/no mat leav… https://t.co/SdnYK7ZyFa',
 'I turned towards the sun on my vacation all day',
 'Bank holiday today so national museum of Korea !! This place was awesome !!now no more days off until week vacation… https://t.co/EVHwJd96b0',
```

<그림 1> 전처리 전 'vacation'이 언급된 트윗 텍스트

영문 텍스트 전처리를 위해 먼저 모든 단어를 소문자로 바꿔주고, 토큰화 해준 다음 표제어를 추출한다. 여기서 불용어를 제거할 때 링크와 관련 된 http가 포함 된 부분, 트위터에서 팔로워에게 트윗을 공유할 때 사용하는 RT(retweet)을 제거한다. 다음 숫자, 구두점, 특수문자와 같은 것들을 제거해주면 영문 텍스트 분석 전 전처리는 끝이 난다.

```
['tiktok hawaii oahu beach mountain travel adventure explore',
 'moved vacation kid attend pride parade going many conversation',
 'better business bureau warning consumer look online scammer summer particularly',
 'brought back vacation florida best smuggled souvenir bag',
 'cuba wa vacation spot negro could go without joe louis otd june',
 'ezhrjmn_ selalu kita tgk kat tiktok friend go vacation malam malam dorg guna baju pyjama yg sama haa boleh je beli kat',
 'mera vacation bhaiya jaisa royal aur lamba nahi hota hai p',
 'take ultimate vacation denver colorado',
 'jiyamahyavanshi instead wondering next vacation ought set life need escape rubi',
 'gopchairwoman crookeddonald trump spent single term vacation another campaigning yes ronna remember year ago',
 'lh_btc wa intense amazing went india usual vacation type place saw much feel',
 'ms_maclean nilikm nobody want job r carrying paid mat',
 'turned towards sun vacation day',
 'bank holiday today national museum korea place wa awesome day week',
```

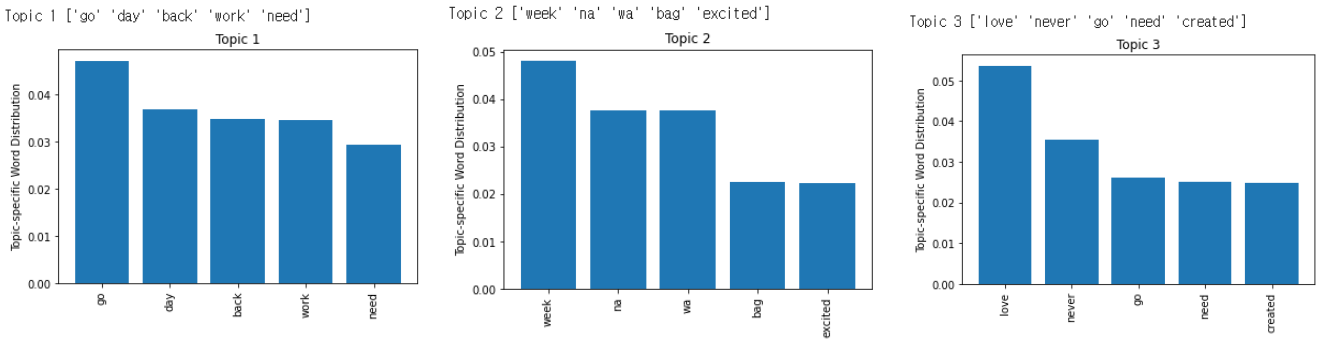
<그림 2> 전처리 된 트윗 텍스트

이제 전처리가 끝난 텍스트를 이용하여 해당 트윗 텍스트에 어떤 단어가 많이 언급 되었는지 알기 위해 문서 단어 행렬을 생성한다. 문서 단어 행렬이란 여러 개의 문서에 나타나는 단어들의 빈도를 행렬로 표현한 것을 의미한다. 이 텍스트의 문서 단어 행렬은 다음과 같았다.

```
단어 목록: ['addition' 'ako' 'amp' 'another' 'asked' 'attend' 'back' 'bag' 'beach'
 'best' 'better' 'boho' 'bureau' 'business' 'carryall' 'case' 'cause'
 'check' 'consumer' 'could' 'created' 'crocheted' 'day' 'denver'
 'designer' 'dubai' 'ebay' 'either' 'escape' 'etsy' 'event' 'everyone'
 'excited' 'finally' 'first' 'friend' 'fun' 'get' 'go' 'going' 'gon'
 'good' 'gun' 'guy' 'hand' 'handbag' 'happy' 'hippie' 'hold' 'holiday'
 'hope' 'iamnooter' 'ig' 'instead' 'ireisu_info' 'jiyamahyavanshi' 'job'
 'joy' 'june' 'kahit' 'kid' 'know' 'last' 'latest' 'lb' 'life' 'like'
 'lim' 'list' 'long' 'look' 'love' 'man' 'many' 'may' 'much' 'na' 'need'
 'never' 'next' 'nobody' 'normal' 'online' 'organic' 'ought' 'paid' 'park'
 'particularly' 'passport' 'people' 'permanent' 'phillip' 'place' 'plan'
 'playing' 'point' 'purse' 'radio' 'remember' 'rubi' 'sa' 'scammer' 'set'
 'share' 'shop' 'someone' 'something' 'spend' 'summer' 'sunrise' 'take'
 'talking' 'ten' 'think' 'three' 'tiktok' 'time' 'today' 'took' 'tourist'
 'travel' 'ultimate' 'wa' 'wait' 'wala' 'want' 'warning' 'wealth' 'week'
 'weekender' 'wondering' 'work' 'worst' 'woven' 'yan' 'year' 'yet']
문서단어행렬 모양 (문서의 수, 단어의 수): (100, 137)
```

<그림 3> 문서 단어 행렬

다음으로 LDA 모델을 이용하여 3가지 토픽으로 언급된 단어들을 나누고, 해당 토픽 안에서 어떤 단어가 많이 언급 되었는지 보고자 한다. 그 결과는 다음과 같다.



<그림 4> 토픽 모델링 결과

첫 번째 토픽에는 방학 때 어딘가 떠나는 것(go)에 관해 필요한 것(need), 일정(day, back), 스케줄(work)과 같은 계획에 관련된 단어들이 많이 보인 것으로 확인 된다. 두 번째 토픽에서는 na, wa와 같은 감탄사나 vacation week에 관련된 언급으로 인해 신나는 감정에 관련된 내용이 많이 보인다. 세 번째 토픽에서는 높은 수치로 love가 언급 되었다. 사랑하는 사람과 함께 방학을 보내는 내용이나, 사랑하는 것들에 대한 내용, 또는 방학 그 자체를 사랑한다는 내용이 주를 이뤘다.

한국어 텍스트를 이용하여 트위터 토픽 모델링을 할 때는, 한국인들의 실제 트윗이 많이 추출되었다. 하지만 알파벳의 경우 많은 나라에서 사용하고 있고, 특히 링크 또는 아이디에 사용되기 때문에 영문 텍스트 외에도 많은 예외적인 텍스트가 크롤링 할 때 걸려서 이를 들어내는 작업이 어려웠다. 또 적은 수의 트윗을 분석해서인지 다른 주제로 크롤링 할 때는 트위터 사용자 아이디가 분석 결과 중 상위권에 위치했다. 따라서 더 많은 양의 데이터의 필요성을 느꼈다. 또한 다음에는 실제 사람들이 방학이라는 주제에 대해 어떤 토픽을 언급하는지 분석한다는 취지에 맞게 트위터 사용자의 아이디 또한 함께 삭제하여 분석을 진행하면 더 좋은 결과를 낼 수 있을 것으로 보인다.