TF-IDF를 사용한 키워드 추출

[1] 0초
```
import requests
from bs4 import BeautifulSoup
url = "https://n.news.naver.com/mnews/article/018/0005596218?sid=101"
resp = requests.get(url)
soup = BeautifulSoup(resp.content, 'html.parser')
content = soup.find(class_='go_trans _article_content').text.replace('\n',' ')
```

[2] 0초
```
content
```

```
' 기술혁신대전·경영혁신대회 통합 행사17~18일 이틀간 부산 벡스코서 열어중소기업 혁신 이끈 유공자에 포상혁신기업 전시·홍보하고 세미나 진행[이데일리 김경은 기자] 중소기업의 기술·경영 혁신 성과를 다루는 국내 최대 행사인 '2023 중소기업 기술·경영 혁신대전'이 오는 17~18일 부산 벡스코에서 열린다.  (사진=2023 중소기업 기술·경영 혁신대전 공식 누리집)15일 중소벤처기업부에 따르면 중소기업 기술·경영 혁신대전은 '혁신형 중소기업, 세상을 바꾸다'라는 주제로 개최된다. △탄소중립·ESG △글로벌 △연구개발(R&D) 혁신 △디지털 △기술보호 등 5대 중점분야별 기술·경영 혁신성과를 공유하고 확산할 예정이다. 기술·경영 혁신대전은 급변하는 기술·경영환경에 대응해 중소기업의 혁신을 지원하기 위해 열리는 행사다. 2000년부터 시작한 '중소기업 기술혁신 대전'과 2018년부터 개최한 '중소기업 경영혁신대회'를 통합해 규모를 확대했다. 각 행사는 그동안 서울에서 열렸으나 올해는 2030 부산 엑스포 유치 홍보를 지원하기 위해 부산으로 옮겨 개최한다. 이틀간 열리는 행사에서는 중소기업 혁신 유공 포상을 비롯해 △기술·기능 인재 경진대회 △인수합병(M&A) 및 상생 투자를 위한 IR 피칭 포럼 △스케일업 팁스 컨퍼런스 등 중소기업 혁신성장을 위한 각종 세미나 및 컨퍼런스를 진행한다. 본 행사인 '중소기업 혁신 유공 시상식'은 이영 중기부 장관이 참석해 기술·경영혁신으로 우수한 성과를 달성하고 국가 경제 발전에 이바지한 중소기업 및 관계 유공자를 시상할 예정이다. 포상은 훈장(3점), 포장(4점), 대통령표창(20점), 국무총리표창(30점), 장관표창(171점) 등 총 228점이 수여된다.행사장에는 △탄소중립·ESG △글로벌 △R&D혁신 △디지털 △테마정책관 총 5개의 구역으로 나눠 주관기관별 전시·홍보부스를 운영하고 혁신 성과를 홍보한다.이번 행사는 온·오프라인에서 누구나 무료로 참석 가능하다. 온라인 참여는 공식 유튜브 채널과 누리집에서 생중계된다. '
```

0초
```
# !pip install konlpy
```

데이터 전처리

형태소분석

[4] 7초
```
# 코모란
from konlpy.tag import Komoran
tokenizer = Komoran()
preprocessed_docs = []

content_token = tokenizer.morphs(content)
print(content_token)
```

```
['기술', '혁신', '대전', '·', '경영', '혁신', '대회', '통합', '행사', '17', '~', '18', '일', '이틀', '간', '부산', '벡스코', '서', '열', '어', '중소기업', '혁신', '이끌', 'ㄴ
```

한글자 제거

[5] 0초
```
token_remove = []
for token in content_token:
  if len(token)>1:
    token_remove.append(token)
print(token_remove)
```

```
['기술', '혁신', '대전', '경영', '혁신', '대회', '통합', '행사', '17', '18', '이틀', '부산', '벡스코', '중소기업', '혁신', '이끌', '유공자', '포상', '혁신', '기업', '전시', '홍ㅂ
```

불용어처리

```
[6]  1 stop_word = ['17','18','2023','ㄴ다','에서']
     2
     3 token_list = []
     4 for token in token_remove:
     5   if token not in stop_word:
     6     token_list.append(token)
     7 print(token_list)
```

```
['기술', '혁신', '대전', '경영', '혁신', '대회', '통합', '행사', '이틀', '부산', '벡스코', '중소기업', '혁신', '이끌', '유공자', '포상', '혁신', '기업', '전시', '홍보', '세미나',
```

```
[7]  1 content_token = ' '.join(token_list)
     2 content_token
```

```
'기술 혁신 대전 경영 혁신 대회 통합 행사 이틀 부산 벡스코 중소기업 혁신 이끌 유공자 포상 혁신 기업 전시 홍보 세미나 진행 이데일리 김경 기자 중소기업 기술 경영 혁신 성과 다루 국내 최대 행사 중소기업 기술 경영 혁신 대전 부산 벡스코 열리 사진 중소기업 기술 경영 혁신 대전 공식 누리 15 중소 벤처기업 따르 중소기업 기술 경영 혁신 대전 혁신 중소기업 세상 바꾸 라는 주제 개최 탄소 중립 ESG 글로벌 연구 개발 혁신 디지털 기술 보호 중점 분야 기술 경영 혁신 성과 공유 확산 예정 기술 경영 혁신 대전 급변 기술 경영 환경 대응 중소기업 혁신 지원 위하 열리 행사 2000 부터 시작 중소기업 기술 혁신 대전 2018 부터 개최 중소기업 경영 혁신 대회 통합 규모 확대 행사 그동안 서울 열리 으나 올해 2030 부산 엑스포 유치 홍보 지원 위하 부산 으로 옮기 개최 이틀 열리 행사 중소기업 혁신 유공 포상 비롯 기술 기능 인재 경진 대회 인수 합병 상생 투자 위하 IR 피칭 포럼 스케일 팁스 컨퍼런스 중소기업 혁신 성장 위하 각종 세미나 컨퍼런스 진행 행사 중소기업 혁신 유공 시상식 이영 중기 장관 참석 기술 경영 혁신 으로 우수 성과 달성 국가 경제 발전 이바지 중소기업 관계 유공자 시상 예정 포상 훈장 포장 대통령 표창 20 국무총리 표창 30 장관 표창 171 228 수여 행사 탄소 중립 ESG 글로벌 혁신 디지털 테마 정책관 구역 으로 나누 주관 기관 전시 홍보 부스 운영 혁신 성과 홍보 이번 행사 오프라인 누구 무료 참석 가능 온라인 참여 공식 유튜브 채널 누리 중계'
```

```
[8]  1 from sklearn.feature_extraction.text import TfidfVectorizer
     2
     3 tfidf_vect = TfidfVectorizer()
     4 tfidf_v = tfidf_vect.fit_transform([content_token])
     5
     6 keyword = tfidf_v.tocoo()
     7 # coo : Coordinate 으로 0이 아닌 데이터만 별도의 배열에 저장하고, 그 데이터가 가리키는 행과 열의 위치를 별도의 배열에 저장하는 방식
```

참고 : https://bkshin.tistory.com/entry/NLP-7-%ED%9D%AC%EC%86%8C-%ED%96%89%EB%A0%AC-Sparse-Matrix-COO-%ED%98%95%EC%8B%9D-CSR-%ED%98%95%EC%8B%9D

희소 행렬 - COO 형식

COO(Coordinate: 좌표) 형식은 0이 아닌 데이터만 별도의 배열에 저장하고, 그 데이터가 가리키는 행과 열의 위치를 별도의 배열에 저장하는 방식입니다. 예를 들어 아래와 같은 2 x 3 행렬이 있다고 해봅시다.

3	0	1
0	2	0

0이 아닌 값은 [3, 1, 2]입니다. 3의 행과 열의 위치는 (0, 0)이고, 1의 행과 열의 위치는 (0, 2)이며, 2의 행과 열의 위치는 (1, 1)입니다. 행 위치 값만 모으면 [0, 0, 1], 열 위치 값만 모으면 [0, 2, 1]입니다.

COO 형식은 0이 아닌 값, 행 위치 값, 열 위치 값에 대한 배열로 표현하는 형식입니다. 0이 아닌 값 배열: [3, 1, 2], 0이 아닌 값의 행 위치 값 배열: [0, 0, 1], 0이 아닌 값의 열 위치 값 배열: [0, 2, 1]로 표현하는 것입니다. 이 세개의 배열만 저장해도 이를 통해 원본 행렬을 구할 수 있습니다. 따라서 원본 행렬을 다 저장하며 메모리를 낭비할 필요가 없습니다.

0초

```
1 sorted_words = sorted(zip(keyword.col, keyword.data), key=lambda x:(x[1], x[0]), reverse=True)
2 sorted_words
```

```
(74, 0.028227871846881837),
(73, 0.028227871846881837),
(72, 0.028227871846881837),
(69, 0.028227871846881837),
(68, 0.028227871846881837),
(67, 0.028227871846881837),
(66, 0.028227871846881837),
(65, 0.028227871846881837),
(64, 0.028227871846881837),
(63, 0.028227871846881837),
(62, 0.028227871846881837),
(60, 0.028227871846881837),
(58, 0.028227871846881837),
(57, 0.028227871846881837),
(56, 0.028227871846881837),
(55, 0.028227871846881837),
(54, 0.028227871846881837),
(52, 0.028227871846881837),
(50, 0.028227871846881837),
(49, 0.028227871846881837),
(47, 0.028227871846881837),
(46, 0.028227871846881837),
(45, 0.028227871846881837),
(44, 0.028227871846881837),
(43, 0.028227871846881837),
(40, 0.028227871846881837),
(38, 0.028227871846881837),
(37, 0.028227871846881837),
(36, 0.028227871846881837),
(34, 0.028227871846881837),
(33, 0.028227871846881837),
(32, 0.028227871846881837),
(31, 0.028227871846881837),
(30, 0.028227871846881837),
(28, 0.028227871846881837),
(27, 0.028227871846881837),
(26, 0.028227871846881837),
(24, 0.028227871846881837),
(23, 0.028227871846881837),
(22, 0.028227871846881837),
(21, 0.028227871846881837),
(20, 0.028227871846881837),
(19, 0.028227871846881837),
(18, 0.028227871846881837),
```

[10]
```
feature_name = tfidf_vect.get_feature_names_out()
feature_name
```

```
array(['15', '171', '20', '2000', '2018', '2030', '228', '30', 'esg',
       'ir', '가능', '각종', '개최', '경영', '경제', '경진', '공식', '공유', '관계', '구역',
       '국가', '국내', '국무총리', '규모', '그동안', '글로벌', '급변', '기관', '기능', '기술',
       '기업', '기자', '김경', '나누', '누구', '누리', '다루', '달성', '대응', '대전', '대통령',
       '대회', '디지털', '따르', '라는', '무료', '바꾸', '발전', '벡스코', '벤처기업', '보호',
       '부산', '부스', '부터', '분야', '비롯', '사진', '상생', '서울', '성과', '성장', '세미나',
       '세상', '수여', '스케일', '시상', '시상식', '시작', '엑스포', '연구개발', '열리', '예정',
       '오프라인', '온라인', '올해', '옮기', '우수', '운영', '위하', '유공', '유공자', '유치',
       '유튜브', '으나', '으로', '이끌', '이데일리', '이바지', '이번', '이영', '이틀', '인수',
       '인재', '장관', '전시', '정책관', '주관', '주제', '중계', '중기', '중립', '중소',
       '중소기업', '중점', '지원', '진행', '참석', '참여', '채널', '최대', '컨퍼런스', '탄소',
       '테마', '통합', '투자', '팁스', '포럼', '포상', '포장', '표창', '피칭', '합병', '행사',
       '혁신', '홍보', '확대', '확산', '환경', '훈장'], dtype=object)
```

[11]
```
# 핵심키워드 10개 추출
[(feature_name[i], score) for i, score in sorted_words[:10]]
```

```
[('혁신', 0.5927853087845185),
 ('중소기업', 0.3669623340094639),
 ('기술', 0.338734462162582),
 ('경영', 0.28227871846881836),
 ('행사', 0.2258229747750547),
 ('대전', 0.169367231081291),
 ('홍보', 0.11291148738752735),
 ('위하', 0.11291148738752735),
 ('열리', 0.11291148738752735),
 ('성과', 0.11291148738752735)]
```

[11]
```
1
```

[11]
```
1
```