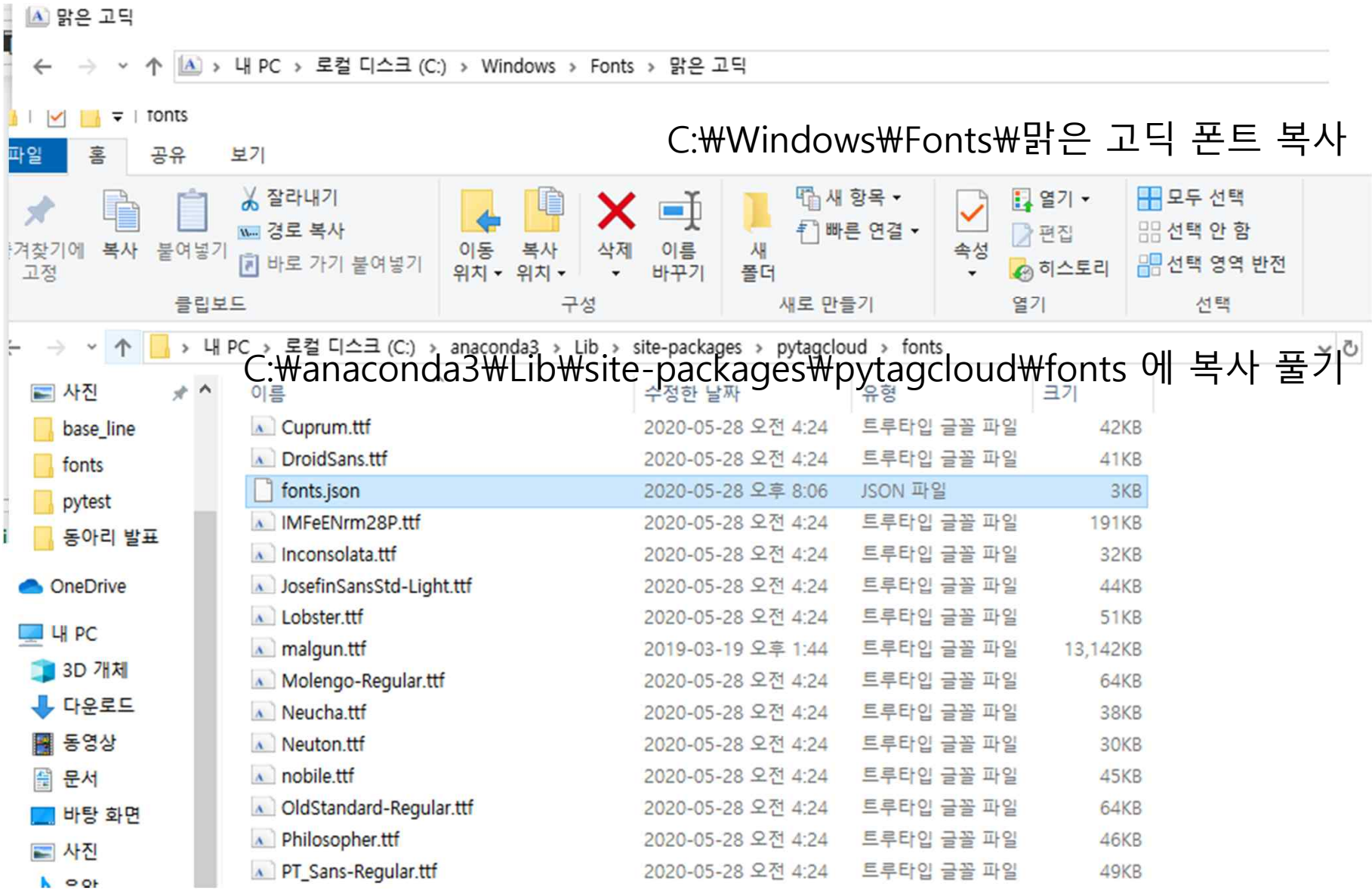
빈도,감성분석, 토픽모델링

2017010715 허지혜

목차


맑은 고딕
내 PC › 로컬 디스크 (C:) › Windows › Fonts › 맑은 고딕
fonts
파일
홈
공유
보기
잘라내기
경로 복사
바로 가기 붙여넣기
복사
붙여넣기
이동 위치
복사 위치
삭제
이름 바꾸기
새 폴더
새 항목
빠른 연결
속성
열기
편집
히스토리
모두 선택
선택 안 함
선택 영역 반전
클립보드
구성
새로 만들기
열기
선택
C:₩Windows₩Fonts₩맑은 고딕 폰트 복사
내 PC › 로컬 디스크 (C:) › anaconda3 › Lib › site-packages › pytagcloud › fonts
C:₩anaconda3₩Lib₩site-packages₩pytagcloud₩fonts 에 복사 풀기
사진
base_line
fonts
pytest
동아리 발표
OneDrive
내 PC
3D 개체
다운로드
동영상
문서
바탕 화면
사진
이름
수정한 날짜
유형
크기
Cuprum.ttf 2020-05-28 오전 4:24 트루타입 글꼴 파일 42KB
DroidSans.ttf 2020-05-28 오전 4:24 트루타입 글꼴 파일 41KB
fonts.json 2020-05-28 오후 8:06 JSON 파일 3KB
IMFeENrm28P.ttf 2020-05-28 오전 4:24 트루타입 글꼴 파일 191KB
Inconsolata.ttf 2020-05-28 오전 4:24 트루타입 글꼴 파일 32KB
JosefinSansStd-Light.ttf 2020-05-28 오전 4:24 트루타입 글꼴 파일 44KB
Lobster.ttf 2020-05-28 오전 4:24 트루타입 글꼴 파일 51KB
malgun.ttf 2019-03-19 오후 1:44 트루타입 글꼴 파일 13,142KB
Molengo-Regular.ttf 2020-05-28 오전 4:24 트루타입 글꼴 파일 64KB
Neucha.ttf 2020-05-28 오전 4:24 트루타입 글꼴 파일 38KB
Neuton.ttf 2020-05-28 오전 4:24 트루타입 글꼴 파일 30KB
nobile.ttf 2020-05-28 오전 4:24 트루타입 글꼴 파일 45KB
OldStandard-Regular.ttf 2020-05-28 오전 4:24 트루타입 글꼴 파일 64KB
Philosopher.ttf 2020-05-28 오전 4:24 트루타입 글꼴 파일 46KB
PT_Sans-Regular.ttf 2020-05-28 오전 4:24 트루타입 글꼴 파일 49KB

Fonts.json 파일 txt로 열기

2. 사전을 이용한 감성분석

- 감성 분석 방법

문서에서 긍정적 단어 : +1
부정적 단어 : -1으로 감성 점수 계산

감성 점수 > 0 : 긍정적 문서
감성 점수 = 0 : 종합적 문서
감성점수 < 0 : 부정적 문서

Sigmoid 함수를 사용하여 0~1 사이로 정규화 하는 방법도 있음

2. 사전을 이용한 감성분석

- 감성 사전 구축 => 쉽지 않음

문장 내 어휘의 도메인에 따라서 극성이 바뀜

예) 치솟다 -주가가 치솟다(긍정)
-부채가 치솟다(부정)
단순하다 -문제가 단순하다(긍정)
-그는 단순하다(부정)

2. 사전을 이용한 감성분석

- 감정어의 종류

1. 감정어 (Sentimental Words)

말하는 이의 감정을 주관적으로 표현하는 것으로 극성이 잘 바뀌지 않음.

2. 평가어(Opinion Words)

대상에 대한 감정을 사실적으로 평가하는 것으로

감정어를 기본으로 하면서
도메인 별로 평가어를 예외처리 하여 구축해야함
통상적으로는 감정어와 평가어를 모두 긍부정어로 통칭하여 언급

2. 사전을 이용한 감성분석

- 가중치 설정

실제 사용되는 감정어 및 긍부정어는 적음
‘아주,조금,매우,별로’ 같은 정도 부사에 의해 정도성이 설정됨

정도 부사가 감정 어휘의 정도성을 증가시킴

TF vs TF-IDF

지금까지 사용한 행렬의 값은 TF이다. 단순히 그 문서에서 출현하는 단어의 빈도이다.
그러나 TF-IDF 값을 이용할 수도 있다.
=(단어 빈도),(역문서 빈도)
IDF(t) = log(전체 문서 수/t가 나오는 문서 수)
IDF는 특정 문서에서만 나타나는 단어에 높은 값을 준다

따라서 영화,재미있다 등 문서에서 많이
나타나는 단어들의 IDF값은 작다
반면 '신기전, 라스베가스, 김지호' 등 문서에서만
나타나는 단어들의 IDF값은 크다

TF vs TF-IDF

여기서 일반적인 중요성을 나타내는 TF를 곱하게 되면 TF*IDF

그 문서에서 자주 나타난 단어(TF)가 몇몇 문서들에서만 자주 나타나는 단어(IDF)인 경우 높은 값을 가지게 된다

즉, TF-IDF는 불용어 가능성을 감안한 빈도이다

TF vs TF-IDF

- TF가 아닌 TF-IDF를 이용해 행렬을 구하고 모델을 만들면 결과가 더 좋을 때가 있다

N- 그램

앞에서 본 방식은 개별 단어만 보기 때문에 순서는 무시해도 된다.
위와 같은 방식을 **bow(bag of words)** 방식이라고 한다

그러나 모든 단어의 연결을 고려하는 것은 어렵지만 2개 혹은 3개 단어가 연결된 정도를 보는 것은 가능하다 일종의 문맥을 참고하는 방식

유니그램(unigram, 1개 단어), 바이그램(bigram, 2개 단어), 트라이그램(trigram, 3개단어)이라고 하며 통칭 n-그램 이라고 한다.

토픽 모델링

토픽 모델링은 문서의 주제를 찾는 것을 목적으로 한다.

연관된 단어들의 묶음을 하나의 토픽으로 간주한다.

자주 사용되는 단어와 확률 분포를 이용하여 어떠한 단어들이 많이 나타날지를 결정한다.

문서가 갖는 여러 토픽 중 어느 것이 토픽인지를 결정한다.

하나의 문서는 여러 개의 토픽(주제)를 가질 수 있고
각각의 토픽에는 자주 사용되는
단어가 있다고 가정한다.
토픽 별로 각 단어가 나타날 발생확률을 구한다

끝끝~