

개발일지0813



김카요

일지0813

TF-IDF를 통한 인물의 상세 설명에서 키워드 추출

2019.10.23 03:16   수정   공개   삭제



### TF-IDF

TF-IDF는 Term Frequency - Inverse Document Frequency 의 약자로서, 정보 검색과 텍스트 마이닝에서 이용하는 가중치이다. 여러 문서로 이루어진 문서들의 집합이 있을때, 어떤 단어가 특정 문서에 얼마나 중요한지 추출해 내는 통계적 수치를 이용하여 문서의 추출하는데 이용된다. 또한 문서들 간의 유사도를 판단하는 데 사용되기도 한다.

TF-IDF의 값은 TF와 IDF 값을 곱한 값이다.

단어 빈도, term frequency, 는 특정 단어가 문서내에 얼마나 자주 등장하는지 나타내는 값이다. 값이 증가함에 따라 해당 단어의 중요도 또한 증가 한다고 해석 할 수 있다.

로 표현하며 가장 단순한 방식은 다음과 같다.

$$tf(t, d) = f(t, d)$$

이때, 단어의 빈도수를 이용해서 계산 해주면 된다.

에는 빈도 계산 법이 크게 세가지가 있다.

**불린 빈도** : 단어 수가 그리 많지 않거나 등장 횟수가 적은 경우 자주 사용된다.

$$tf(t, d) = \begin{cases} 1 & (t \text{가 } d \text{에 한 번이라도 등장}) \\ 0 & (else) \end{cases}$$

수식 1.1 불린 빈도 함수

**로그 스케일 빈도**

$$df(t, d) = \log(f(t, d) + 1)$$

수식 1.2 log-scale 빈도

**역수 빈도**: 문서의 길이가 상대적으로 길 경우 단어의 빈도값을 조절할 경우 사용된다.

$$df(t, d) = 0.5 + \frac{0.5 \times f(t, d)}{\max \{f(w, d) : w \in d\}}$$

수식1.3 증가 빈도

개발일지0813

와 IDF



김카요

증가 빈도, document frequency, 는 단어 자체가 문서군 내에서 자주 사용 되는 경우 해당 단어가 얼마나 자주 등장 하는 것을 의미한다. 즉, 특정 단어  $t$ 가 등장한 문서의 수를 가리킨다.

때 우리가 필요한 것은 IDF 값이다. IDF는 DF값의 역수(반비례)로, Inverse document frequency의 약자다. IDF는 한 단어가 문서 집합 전체에서 얼마나 공통적으로 나타나는지 나타내는 척도이다.

DF는 단순 역수 관계는 아니지만 역수비스무리 한 관계성을 나타낸다. 결국엔 한쪽이 증가함에 따라 다른 한쪽은 감소한다.

$$idf(t, d) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

수식 1.4 IDF

$D$ : 문서 집합의 크기  
 $\{d \in D : t \in d\}$ :단어  $t$ 가포함된문서의수. 값이 0이 되는 것을 방지하기 위해 쓰임.

이므로는 TF-IDF 의 값을 다음과 같이 계산 할 수 있다.

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

수식 1.5 TF-IDF

문서 내에 단어 빈도가 높을 수록 && 전체 문서들 중 그 단어를 포함한 문서가 적을수록 TF-IDF 값이 높게 나온다. 즉, 해당 단어가 그 특정 문서에서 키워드로 작용을 한다는 의미이다.

이용하면 문서에서 공통적으로 나타나는 단어들은 미리 걸러낼 수 가 있다.

로그 함수값은 상 1 이상이므로, IDF값과 TF-IDF값은 항상 0 이상이 된다.

단어를 포함 하는 문서들이 많을 수록 로그 함수 안의 값이 1에 가까워지고, 해당 경우 IDF 와 TF-IDF의 값은 0에 가까워진다.

### 결론

≡ 적용하여 실험한 사항은 아니고 대략적인 구상이기 때문에 언제든지 변화 및 수정 사항이 존재 할 수 있음.

≡ 주어진 각 인물에 대한 상세설명은 몇 문장짜리( 대략 한~두문단) 길이를 지닌다.

≡ 상세 설명에서 TF-IDF를 응용하기 위해서는 '상세설명'을 두가지 형태로 처리 할 수 있다.

- 문장마다 나누어진 리스트 형태의 타입
- 한 인물의 상세설명 == 하나의 문서에 전부 대입.

kenize 이전에 명사들만 추출 해 내서 다시 합치는 재 처리 과정이 필요.  
과정에서 KoNLpy 파이썬 패키지 이용

kkma.nouns(string) 이용

# 개발일지0813



hingvectorization 등 다양한 경우 고려중.  
|는 이쪽으로 : <https://konlpy-ko.readthedocs.io/ko/v0.5.1/install/>

	설치하기 — KoNLPy 0.5.1 documentation konlpy-ko.readthedocs.io
--	---

)을 기반으로 Tfidfvectorizer 이용  
블로그 참고 하여 코드 테스트 가능

/blog.naver.com/PostView.nhn?  
d=vangarang&logNo=221072014624&categoryNo=35&parentCategoryNo=0

	[파이썬을 이용한 한글 NLP] 04. TF-IDF Scoring blog.naver.com
--	--

제 설명은 키워드를 추출하는 것이 목표기 때문에 A or B 형태 두 가지 모두에 대해 테스트가 필요하다.  
|면 B형태 일경우 단순 TF만 계산 한다거나 A 형태 일 경우 TF-IDF를 활용하는 방안 구상.  
|으로 각 경우에 대해 실험 필요

제로 위 블로그레 있는 코드를 참고하여 간단하게 실험을 해보았음. KoNLpy를 설치 못해서 명사 처리 하기 이전임. 해당 :  
제와 같음.

```
from sklearn.feature_extraction.text import TfidfVectorizer
from collections import defaultdict
import nltk

corpus = [ '생졸년 미상. 부여의 왕.', '서기전 1세기에 활동한 것으로 보이며, 그의 행적은 고구려 시조인 동명성왕(
'부여왕 해부루(解夫婁)는 늙도록 아들이 없어 산천에 후사를 구하러 다녔다.', '그러던 중 곤연(鯤淵)에서
'그뒤 태자로 책봉되어 해부루를 이어서 부여의 왕이 되었다. 그리고 태백산 남쪽의 우발수(優渤水)에서
'그런데 유화가 이상하게도 알을 낳자 이를 버리게 하였다. 그러나 곧 알의 신비함을 인정하고 유화에게
'그뒤 주몽이 달아나자 그를 추격하는 군대를 파견하였지만 잡지는 못하였다. 주몽이 고구려 건국을 위하여

vectorizer = TfidfVectorizer()
matrix = vectorizer.fit_transform(corpus)

word_idx = defaultdict(lambda : 0)
for idx, feature in enumerate(vectorizer.get_feature_names()):
    word2id[feature] = idx

for i, sent in enumerate(corpus):
    print('===== document[%d] =====' % i)
    print(' [ (token, sp_matrix[i, word2id[token]]) for token in sent.split() ] )

```

과

'|

개발일지0813

데이터에 대한 전처리를 하고 실험을 하면 좀 더 유의미한 결과가 나올 것 같음. 후에 인물 DB에 넣을 수 있는 형태로 키워드  
가 것이 최종 목표.



김카요

ERENCE>

tps://ko.wikipedia.org/wiki/Tf-idf

tps://konlpy-ko.readthedocs.io/ko/v0.4.3/#

tp://blog.naver.com/PostView.nhnblogId=vangarang&logNo=221072014624&categoryNo=35&parentCa  
=0

공감

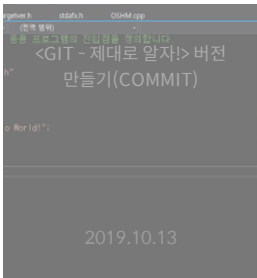
일지0813' 카테고리의 다른 글

DF를 통한 인물의 상세 설명에서 키워드 추출 (0)	2019.1
RCETREE 오류 - 원격저장소 로그인 정보 삭제 하기 (0)	2019.1
- 제대로 알자!> 버전 만들기(COMMIT) (0)	2019.1
-제대로 알자!> 저장소(REPOSITORY) 만들기 (0)	2019.1
- 제대로 알자!>GIT 과 SourceTree의 설치 (0)	2019.1
'0126 백준 1260 DFS BFS (0)	2017.0


!일지0813' Related Articles

SOURCETREE 오류 - 원  
격저장소 로그인 정보 삭제  
하기


2019.10.13



2019.10.13



2019.10.12



2019.10.12

omments

여러분의 소중한 댓글을 입력해주세요

Send