Chapter. 03

자연어 분석 기초

자연어 모델링 II

FASTCAMPUS
ONLINE
금융공학/퀀트 I

강사. 장순용

I 키포인트

- BOW 모형.
- 단어 빈도 (TF).
- 역문서 빈도 (IDF).
- TF IDF 모형.
- Cosine 유사도.
- 지도학습에 의한 문서의 분류 예측.

FAST CAMPUS
ONLINE
장순용 강사.

I BOW (Bag of Words) 모형

- 개개 단어는 문서의 의미를 나타내는 가장 기본적인 단위이다.

  ⇒ 개개 단어는 feature라고도 불리운다.

- BOW는 단어간의 연관성을 고려하지 않은 추상화 모형이다.

- 행렬의 셀은 0과 1의 binary 값을 갖는데 불포함 or 포함을 의미한다.

I BOW (Bag of Words) 모형

- 다음과 같은 세개의 문서 (문장)을 가정한다.

문서 #1 : “삼성 호재, 주가 삼성 매수세력”

문서 #2 : “주가 매도세력 삼성 경제”

문서 #3 : “주가 매도세력, 주가 삼성, 주가 경제”

BOW =

단어	문서 #1	문서 #2	문서 #3
삼성	1	1	1
호재	1	0	0
주가	1	1	1
매수세력	1	0	0
매도세력	0	1	1
경제	0	1	1

I BOW (Bag of Words) 모형

- Bag of Words의 문제점:

⇒ 모든 단어는 동일하게 중요하다는 전제.

⇒ Semantic information이 포함되지 않는다.

예). “She is beautiful”에서 “she” , “is” 보다는 “beautiful”이 더 중요하다.

I 단어 빈도 (TF)

- 단어 빈도 (Term Frequency, TF):

⇒ 단어 빈도 (TF)는 특정 단어가 문서 내에서 얼마나 자주 등장하는가를 나타냄.

⇒ TF가 높을수록 문서 내에서 단어가 중요한 역할을 함.

$$TF(word) = \frac{\text{문서 내 } word\text{의 수}}{\text{문서 내 모든 단어의 수}}$$

⇒ TF는 문서 하나씩 별도로 계산한다.

I 단어 빈도 (TF)

- 다음과 같은 세개의 문서 (문장)을 가정한다.

문서 #1 : “삼성 호재, 주가 삼성 매수세력” ⇒ 길이 = 5

문서 #2 : “주가 매도세력 삼성 경제” ⇒ 길이 = 4

문서 #3 : “주가 매도세력, 주가 삼성, 주가 경제” ⇒ 길이 = 6

TF =

단어	문서 #1	문서 #2	문서 #3
삼성	2/5 = 0.4	1/4 = 0.25	1/6 = 0.17
호재	1/5 = 0.2	0	0
주가	1/5 = 0.2	1/4 = 0.25	3/6 = 0.5
매수세력	1/5 = 0.2	0	0
매도세력	0	1/4 = 0.25	1/6 = 0.17
경제	0	1/4 = 0.25	1/6 = 0.17

I 역문서 빈도 (IDF)

- 역문서 빈도 (Inverse Document Frequency, IDF):

⇒ 문서 빈도 (DF)는 특정 단어를 포함하는 문서의 빈도를 나타낸다.

⇒ 역문서 빈도 (IDF)는 특정 단어의 희소성과 연관된다.

⇒ IDF는 DF의 역수에 로그를 적용하여 계산한다. $Log() = Log_{10}()$ 사용.

$$IDF(word) = Log\left(\frac{\text{말뭉치 내 모든 문서의 수}}{word\text{를 포함한 문서의 수}}\right)$$

⇒ IDF는 말뭉치 전체적 특성이므로 문서 하나 하나 별도로 계산할 필요가 없다.

I 역문서 빈도 (IDF)

- 다음과 같은 세개의 문서 (문장)을 가정한다.

문서 #1 : “삼성 호재, 주가 삼성 매수세력” ⇒ 길이 = 5

문서 #2 : “주가 매도세력 삼성 경제” ⇒ 길이 = 4

문서 #3 : “주가 매도세력, 주가 삼성, 주가 경제” ⇒ 길이 = 6

IDF =

단어	DF	IDF
삼성	3	Log(3/3) = 0
호재	1	Log(3/1) = 0.48
주가	3	Log(3/3) = 0
매수세력	1	Log(3/1) = 0.48
매도세력	2	Log(3/2) = 0.18
경제	2	Log(3/2) = 0.18

I TF IDF 모형

- 단어 빈도 (TF)와 역문서 빈도 (IDF)를 조합한 TF IDF 모형:

⇒ 문서에 있어서 “의미있는” 단어란:

→ 문서 안에서 여러 번 발생한다 (TF 큼).

→ 말뭉치 안에서 가끔씩 희소성 있게 발생한다 (IDF 큼).

⇒ 그러므로 TF행렬에 IDF를 가중치로 적용한 TF IDF 행렬을 구할 수 있다:

$$TF\ IDF = TF\ *\ IDF$$

I TF IDF 모형

- 다음과 같은 세개의 문서 (문장)을 가정한다.

문서 #1 : “삼성 호재, 주가 삼성 매수세력” ⇒ 길이 = 5

문서 #2 : “주가 매도세력 삼성 경제” ⇒ 길이 = 4

문서 #3 : “주가 매도세력, 주가 삼성, 주가 경제” ⇒ 길이 = 6

TF IDF =

단어	문서 #1	문서 #2	문서 #3
삼성	0.4	0.25	0.17
호재	0.2	0	0
주가	0.2	0.25	0.5
매수세력	0.2	0	0
매도세력	0	0.25	0.17
경제	0	0.25	0.17

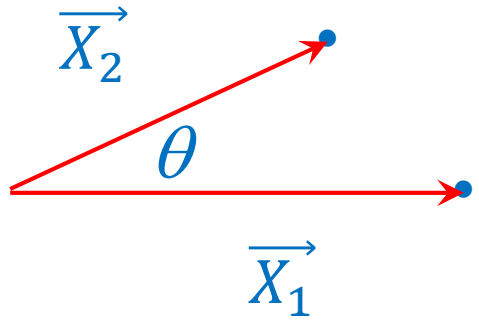
×

단어	IDF
삼성	0
호재	0.48
주가	0
매수세력	0.48
매도세력	0.18
경제	0.18

=

단어	문서 #1	문서 #2	문서 #3
삼성	0	0	0
호재	0.095	0	0
주가	0	0	0
매수세력	0.095	0	0
매도세력	0	0.044	0.03
경제	0	0.044	0.03

I 코사인 유사도:



코사인 유사도는 $Cos(\theta)$이다: $Cos(\theta) = \frac{\overrightarrow{X_1} \cdot \overrightarrow{X_2}}{|\overrightarrow{X_1}||\overrightarrow{X_2}|}$

FAST CAMPUS
ONLINE
장순용 강사.

I 코사인 유사도:

문서 #1 : “destruction of forest caused by deforestation”

문서 #2 : “men causes deforestation by agriculture”

문서 #3 : “destruction of forest initiated by men”

						TF			TF-IDF		
Terms	문서 #1	문서 #2	문서 #3	DF	IDF	문서 #1	문서 #2	문서 #3	문서 #1	문서 #2	문서 #3
destruction	1	0	1	2	0.176	0.167	0.000	0.167	0.029	0.000	0.029
forest	1	0	1	2	0.176	0.167	0.000	0.167	0.029	0.000	0.029
cause	1	1	0	2	0.176	0.167	0.200	0.000	0.029	0.035	0.000
deforestation	1	1	0	2	0.176	0.167	0.200	0.000	0.029	0.035	0.000
men	0	1	1	2	0.176	0.000	0.200	0.167	0.000	0.035	0.029
agriculture	0	1	0	1	0.477	0.000	0.200	0.000	0.000	0.095	0.000
initiate	0	0	1	1	0.477	0.000	0.000	0.167	0.000	0.000	0.080
총 단어수	6	5	6								

I 코사인 유사도:

문서 #1 : “destruction of forest caused by deforestation”

문서 #2 : “men causes deforestation by agriculture”

문서 #3 : “destruction of forest initiated by men”

$$\text{코사인 유사도: } \cos(\theta) = \frac{\overrightarrow{X_1} \cdot \overrightarrow{X_2}}{\left|\overrightarrow{X_1}\right|\left|\overrightarrow{X_2}\right|}$$

I 코사인 유사도:

문서 #1 : “destruction of forest caused by deforestation”

문서 #2 : “men causes deforestation by agriculture”

문서 #3 : “destruction of forest initiated by men”

문서 #1 ~ 문서 #2 유사도: 0.311

문서 #1 ~ 문서 #3 유사도: 0.311

문서 #2 ~ 문서 #3 유사도: 0.097

I 지도학습에 의한 문서의 분류 예측:

- TF IDF 행렬을 바탕으로 텍스트 데이터의 정형화를 이루었다.

⇒ 여러 방식의 분류형 머신러닝 알고리즘을 적용할 수 있다.

⇒ 미리 정의된 카테고리와 카테고리별 주요 키워드 정보가 필요하다 (학습).

⇒ 준비된 데이터로 학습하여 새로운 문서의 카테고리를 알아 맞출 수 있다 (지도학습).

끝.

감사합니다.

FAST CAMPUS
ONLINE
장순용 강사.