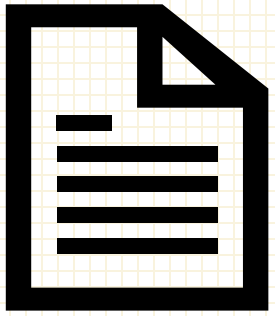
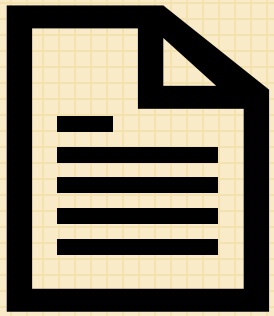


Feature Vectorization

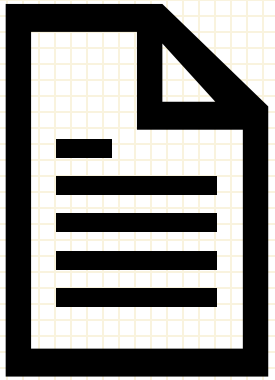
19.04.11 임소현

CONTENT



#1 텍스트 분석 수행 프로세스

Text 문서



Preprocessing

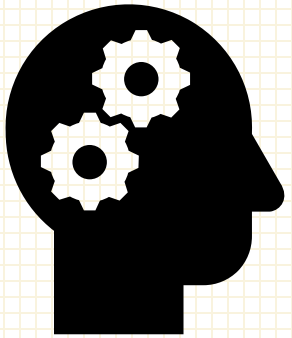
Feature Vectorization

Bag of Words

단어1	단어2	단어3
3	1	4

Feature 기반의
데이터 제공

학습/예측/평가



#2 One-hot encoding

- 단어를 표현하는 가장 기본적인 표현 방법
- 단어 집합의 크기를 벡터의 크기로 하고, 표현하고 싶은 단어의 인덱스에 1의 값을 부여하고, 다른 인덱스에는 0을 부여하는 단어의 벡터 표현 방식
- 단어의 개수가 늘어날수록, 벡터의 차원이 계속 늘어난다. 단어의 유사성을 전혀 표현하지 못한다.

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

#3 Bag of Words - BOW

- 단어들의 순서는 전혀 고려하지 않고, 단어들의 출현 빈도에만 집중하는 텍스트 데이터의 수치화 표현 방법
- 주로 어떤 단어가 얼마나 등장했는지를 기준으로 문서가 어떤 성격의 문서인지를 판단하는 작업에 쓰인다.
  즉, 분류 문제나 여러 문서 간의 유사도를 구하는 문제에 주로 쓰인다.
- 문장 내에서 단어의 문맥적인 해석을 처리하지 못하는 단점

Doc#1
The dog is on the table

Doc#2
The cats are on the table

	are	cats	dog	is	now	on	table	the
Doc#1	0	0	1	1	0	1	1	1
Doc#2	1	1	0	0	0	1	1	1

#4 TF-IDF

- TDM 내의 각 단어들마다 중요한 정도를 가중치로 주는 방법
- 개별 문서에서 자주 나타나는 단어에 높은 가중치를 주되, 모든 문서에서 전반적으로 자주 나타나는 단어에 대해서는 페널티를 주는 방식으로 값을 부여
- 주로 문서의 유사도를 구하는 작업, 검색 시스템에서 검색 결과의 중요도를 정하는 작업, 문서 내에서 특정 단어의 중요도를 구하는 작업 등에 쓰일 수 있다.

$$TFIDF_i = TF_i \times \log \frac{N}{DF_i}$$

TF_i = 개별 문서에서의 단어 i 빈도
DF_i = 단어 i를 가지고 있는 문서 개수
N = 전체 문서 개수

#5 희소 행렬 – COO/CSR 형식

- 희소행렬 = 대규모 행렬의 대부분의 값을 0이 차지하는 행렬
- BOW 모델은 너무 많은 0값이 메모리 공간에 할당되어 많은 메모리 공간이 필요하며,
  연산 시에도 시간이 많이 소모 -> 희소 행렬 변환 필요

COO(Coordinate:좌표) 형식

0이 아닌 데이터만 별도의 데이터 배열에 저장하고,
그 데이터가 가리키는 행과 열의 위치를 별도의 배열로 저장하는 방식

CSR(Compressed Sparse Row)형식

COO 형식이 행과 열의 위치를 나타내기 위해 반복적인 위치 데이터를 사용해야 하는 문제점을 해결한 방식

행 위치 배열 내에 있는 고유한 값의 시작 위치만 다시 별도의 위치 배열로 가지는 변환 방식

#5 희소 행렬 – COO/CSR 형식

원본 데이터

[[0,0,1,0,0,5],
[1,4,0,3,2,5],
[0,6,0,3,0,0],
[2,0,0,0,0,0]]

0이 아닌 데이터 값 배열

[1,5,1,4,3,2,5,6,2]

0이 아닌 데이터 값의 행과 열의 위치

(0,2), (0,5)
(1,0), (1,1), (1,3), (1,4), (1,5)
(2,1), (2,3)
(3,0)

COO 형식

행 위치 배열

[0,0,1,1,1,1,1,2,2,3]

열 위치 배열

[2,5,0,1,3,4,5,1,3,0]

[0,0,1,1,1,1,1,2,2,3]

행 위치 배열의 인덱스

0 1 2 3 4 5 6 7 8 9

CSR 형식

[0,2,7,9,10]

총 항목 개수

감사합니다