|  |  |
| --- | --- |
| 교육 제목 | TF-IDF 모델 |
| 교육 일시 | 12월 1일(수) |
| 교육 장소 | 강의실 |
| **교육 내용** | |
| 전체 내용 | * WordFeatrure Vectors * Bag of Words   단어들의 순서는 전혀 고려하지 않고, 단어들의 출현 빈도(frequency)에만 집중하는 텍스트 데이터의 수치화 표현방법  만드는과정 :   1. 각 단어에 고유한 정수 인덱스를 부여한다. 2. 각 인덱스의 위치에 단어 토큰의 등장 횟수를 기록한 벡터를 만든다.  * 문서 단어 행렬(Documenct-Term Matrix,DTM)   다수의 문서에서 등장하는 각 단어들의 빈도를 행렬로 표현한 것  각 문서에 대한 BoW를 하나의 행렬로 만든 것  BoW표현을 다수의 문서에 대해서 행렬로 표현하고 부르는 용어  DTM의 한계   1. 희소 표현   각 문서 벡터의 차원은 원-핫 벡터와 마찬가지로 전체 단어 집합의 크기를 가짐  많은 문서벡터가 대부분의 값이 0을 가질 수도 있음 -> 희소벡터, 희소행렬  희소벡터는 많은 양의 저장공간과 계산을 위한 리소스가 필요   1. 단순 빈도 수 기반 접근   예를 들어 영어에 대한 DTM을 만들었을 때, 불용어인 The는 어떤 문서든 자주 등장한다.  그런데, 유사한 문서인지 비교하고 싶은 문서1, 문서2, 문서3에서 동일하게 the가 빈도수가 높다고 해서 이 문서들이 유사한 문서라고 판단 할 수 없다.   * TF – IDF * 텍스트 마이닝에서 중요하게 사용 * 어떤 단어 w가 문서 d 내에서 얼마나 중요한지 나타내는 수치 * TF(Term Frequency)   단어의 문서 내에 출현한 횟수  숫자가 클수록 문서 내에서 중요한 단어  하지만, ‘the’와 같은 단어도 TF값이 매우 클 것   * IDF(Inverse Document Frequency)   그 단어가 출현한 문서의 숫자의 역수(inverse)  값이 클수록 ‘the’와 같이 일반적으로 많이 쓰이는 단어   * Thesaurus 기반 방식에 비해 코퍼스(or 도메인) 특화된 표현 가능 * 여전히 sparse한 vector로 표현됨   PCA를 통해 차원 축소를 하는 것도 한 방법 |