NLP 가 무엇이고 왜 사용하는지

자연어 처리 의 대표 예제 = 파파고 , 구글 번역

Github Copilot = 코드를 자연어로 번역 = 지린다,,,

음성인식도 대표적인 예시

맥락을 이해 하는것도 필요하다

즉 음성인식을 할려면은 문장의 이해도 필요하다.

추천 시스템 - 이번주에 할것

영화 추천 시스템 영상 추천 시스템 - 카운트 기반으로 보여줌

-> 주어진 정보의 의미를 이해 하면서 보여준다

어떠한 도구들의 집합 = text 를 이해하는것이다.

NLP 어떻게 쓸건인지 how

의미를 이해하는것 = 단어 혹은 문서의 의미를 나타낼수 있는 표현을 얻는것에서 부터 시작한다.

1.기호로 표현한다(one-hot-vector)

서로직교 할수 밖에 없다

모든 단어가 교유한 의미를 담고 있어서 내적을 하면 전부 0 이된다.

2.vector 자체의 의미를 담자(distribution semantics = 의미 분포론)

어떤 단어의 의미를 알고 싶으면 그 단어의 문장을 파악하자

맥락을 파악하여 단어를 예측 하자

DTM (document term matrix )

구축하여 단어 벡터 or 문서 벡터를 추출한다.

But 문장이 너무 많아지면 0 의 값이 많아져서 메모리 한계 에 부딫힌다.

단점: the bag of words representation

단어의 빈도만 같고 의미가 다른 평서형이나 의문형 은 빈도수로 체크 하면 안된다

3.predict based :예측 기반 알고리즘

단어의 의미를 벡터에 담는 규칙을 스스로 학습 하게 한다.

DTM 보다는 훨씬더 많은 정보를 담을 수있다.

훈련되지 않은 데이터는 예측 불가 데이터 편향이나 특정 단어만 계속 나오는 경우 학습이 잘 안될수 도 있다.

A robot must +obey(예측)

예측 값들과의 오차를 구해 경사 하강법을 통해서 예측 하는것이다

Dense vector = 0 이 아니고 값을 가지고 있는 벡터가 Dense 한 벡터 이다

숙제

DTM 구현

문장 4개