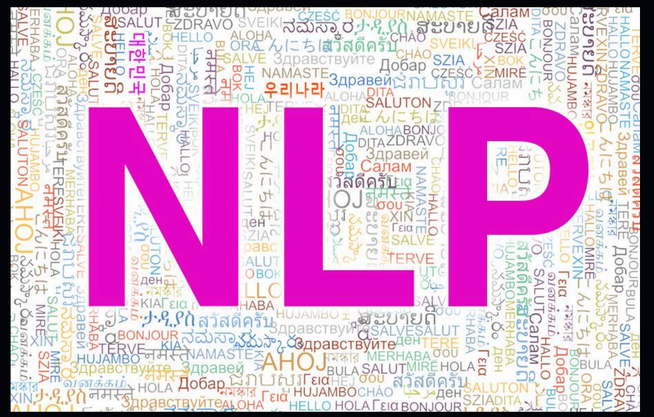
자연어 처리, 임베딩

정규 수업 8차시



AI 여자친구를 만든다고 합시다...

가장 중요한 핵심 기능은 무엇일까요?

자연어 처리란 무엇인가?

네! 바로..
말을 하는 기능입니다!

그래서 우리는 인공지능이
인간의 자언어를 분석하고, 말할 수 있게 하는
방법을 공부합니다

자연어 처리(Natural Language Processing)

컴퓨터가 인간의 자연어를

처리(분석, 판단, 이해?)할 수 있게 하는 분야

자연어 처리의 예시들

1. ChatGPT, LlaMa
2. 파파고, 구글 번역기
3. 단어 연관성 기반 검색 추천 시스템

컴퓨터는 자연어를 이해할 수 없습니다

???? 그럼 자연어 처리는 어케 하는데요?

먼저 벡터가 무엇인지 알아야합니다.

뭣??? 벡터가 뭔가여???

그거 어려운거 아닌가요?

어렵고 무거운 정의를 이야기 할 수도 있지만..
벡터는 엄청나게 많은 의미를 담고 있기에
예시를 들어서 설명하겠습니다.

1. 순서쌍 (x,y) 및 (x,y,z)

각각 2차원 및 3차원 벡터 데이터를 표현

2. 튜플 및 배열 등등

n차원의 벡터 데이터를 표현 가능
ex : (1,0,1,0,0)

사람의 자연어를 벡터로 변환하는 과정을

임베딩

이라고 합니다

1. Word2Vec

단어를 벡터 값으로 변환

2. ELMo, BERT, GPT

문장의 문맥을 바탕으로 벡터화

Word2Vec(워드2벡)

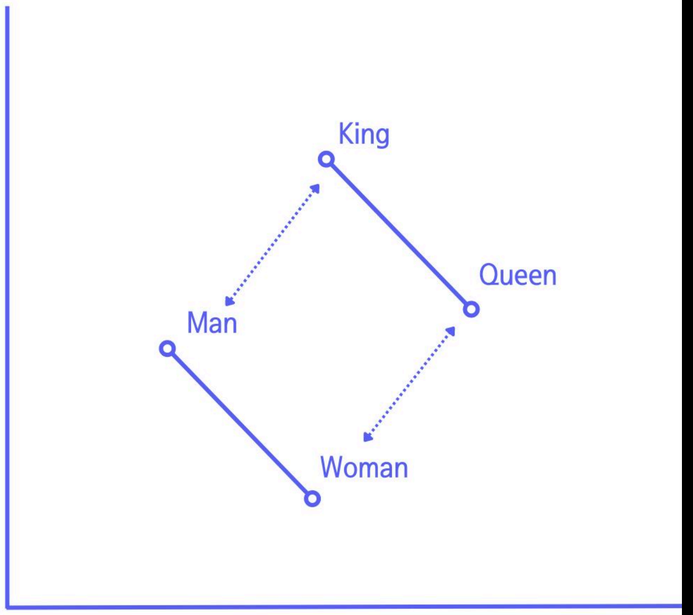
구글에서 만든 단어를 벡터로 변환하는 기법

비슷한 의미를 가진 값일 수록 벡터값이 비슷

Word2vec

2013년 구글에서 발표
자연어 처리(NLP) 기술
워드 임베딩(Word embedding)
모델 학습 및 최적화 제품군

Tomas Mikolov



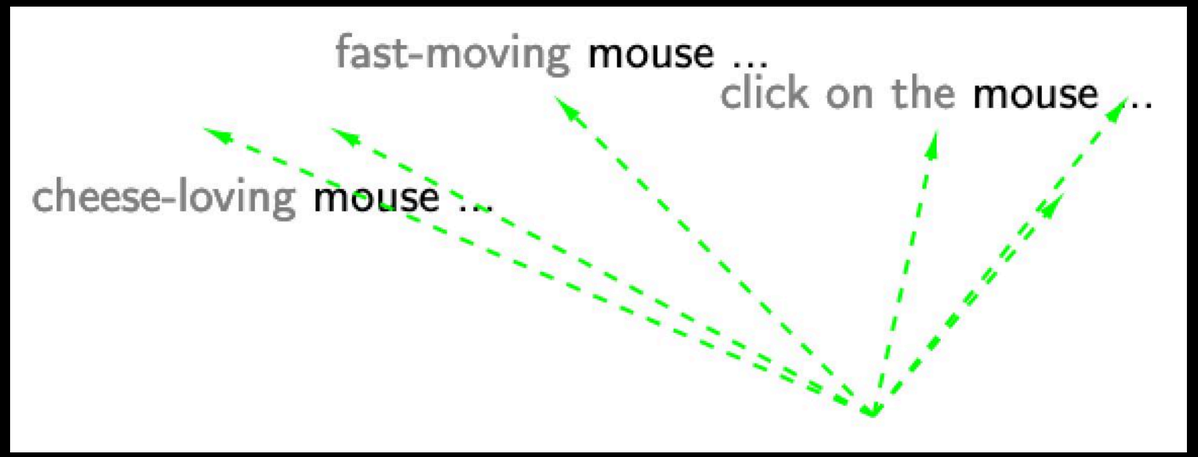
워드 임베딩 예시

참고로 얘네들은 임베딩만 하는 애들이 아닙니다

ELMo, BERT, GPT

문장의 문맥을 분석하고 벡터값 생성

얘네들은 동음이의어도 문맥에 따라 구분 가능



오늘은 문장 임베딩을 활용하여
무언가 만들어볼겁니다.

오늘 만들 프로그램

먼저 사용자로 부터 문장들을 입력받는다.

새로 문장을 입력 받았을 때,

기존에 입력 받은 내용 중에 새로 입력 받은 문장과 가장 비슷한 것을 찾는다.

오늘의 실습

1. 사용자로 부터 문장을 입력 받는다.

2. 입력 받은 문장을 임베딩하여 저장한다.

3. 새로운 문장을 입력 받는다.

4. 새로 받은 문장을 임베딩한다.

5. 새로 받은 문장과 기존 문장의 코사인 유사도를 분석한다.

6. 가장 유사도가 높은 문장을 출력한다.

코사인 유사도(Cosine Similarity)

두 벡터의 사잇값의 코사인 값을 사용하여,

두 값의 유사함을 수치로 나타낸 값

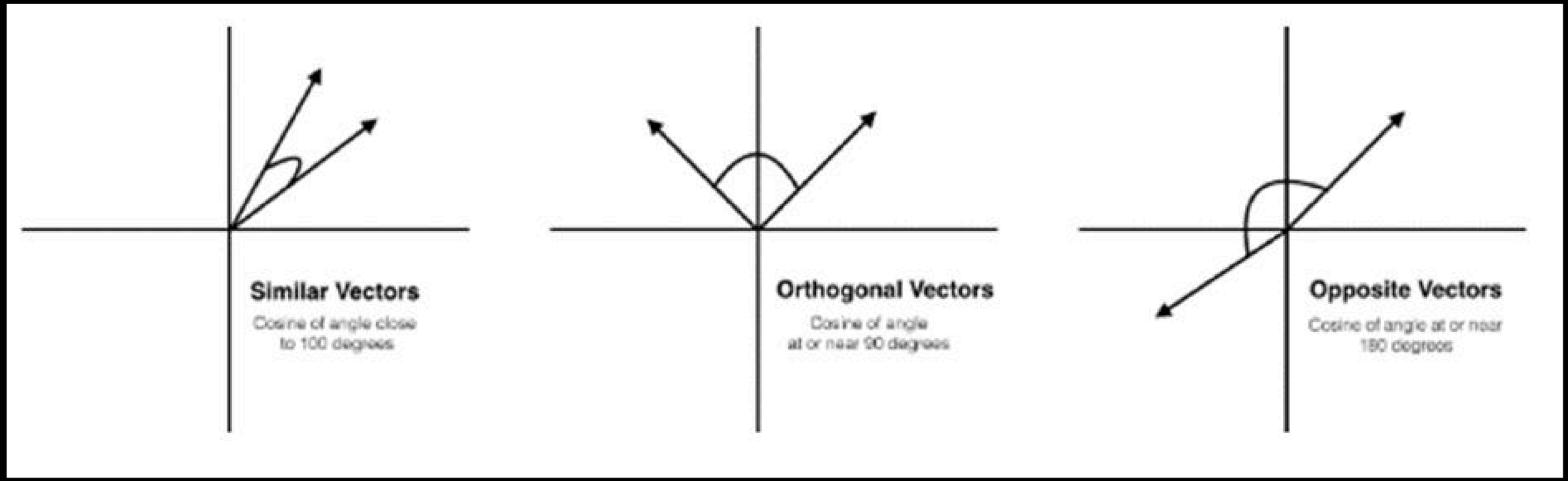
코사인 유사도의 범위와 값에 따른 의미

-1 <= Cosine Similarity <= 1

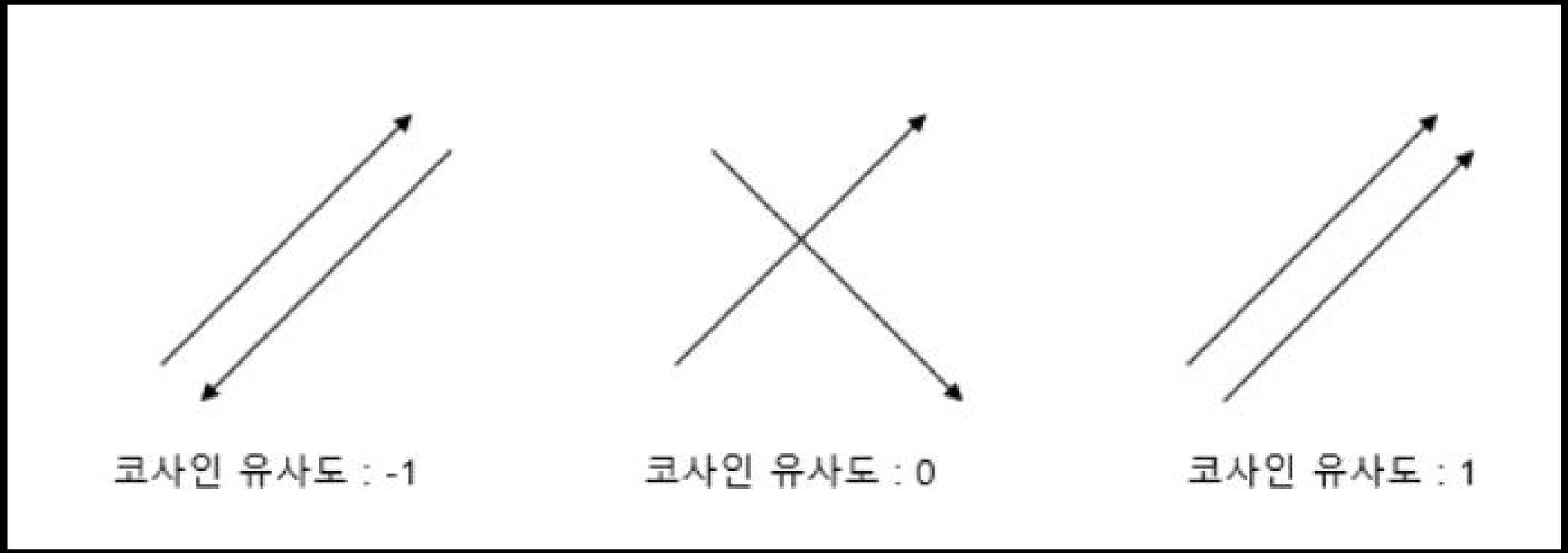
1에 가까울 수록 값이 유사함

-1에 가까울 수록 유사하지 않음

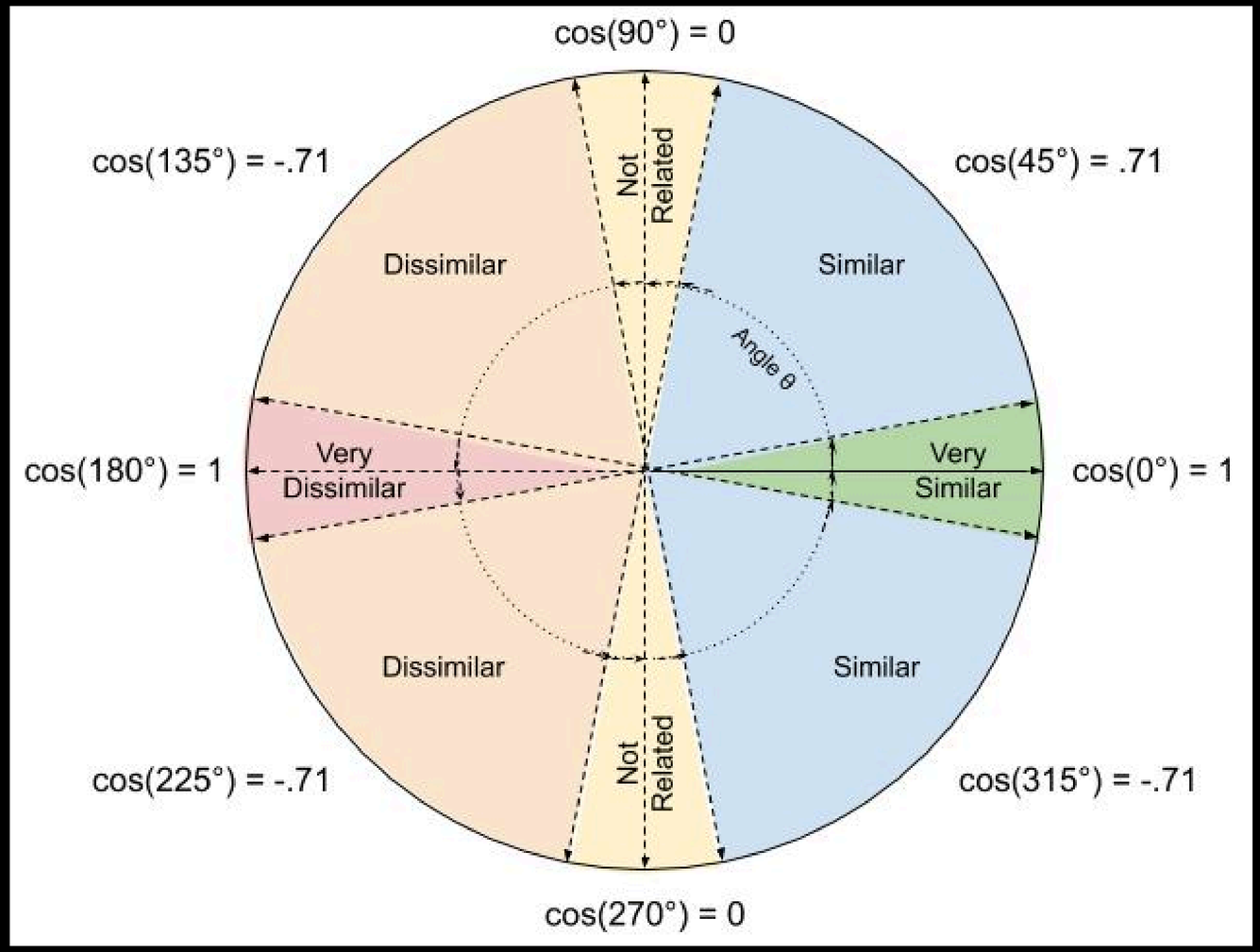
오늘의 실습 - 코사인 유사도



오늘의 실습 – 코사인 유사도



오늘의 실습 – 코사인 유사도



오늘의 실습 – 코사인 유사도