KUGGLE

NLP 자연어 처리

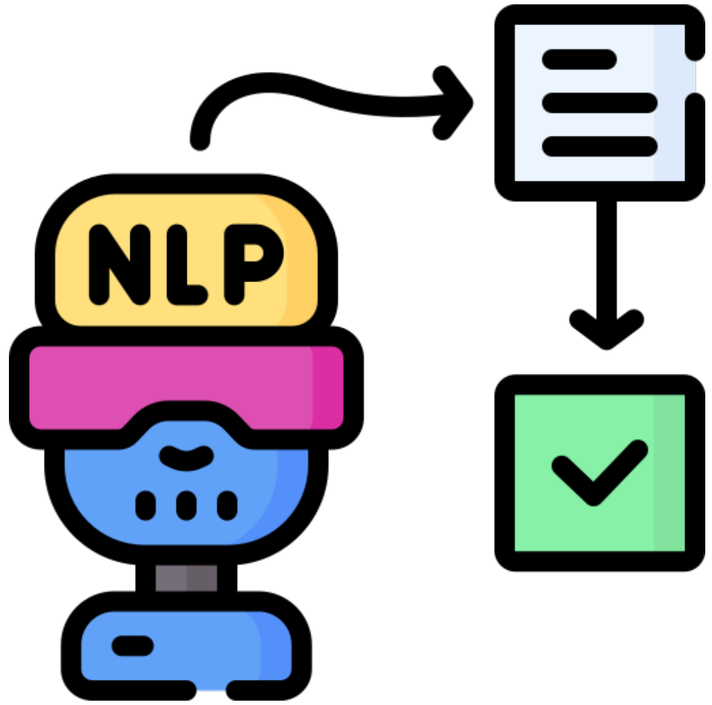
KUGGLE 11기 조동현 김영진

목차

KUGGLE

1. 자연어 처리(NLP)란?

자연어 처리(NLP)란 컴퓨터가 인간의 언어를 이해하고 해석할 수 있도록 하는 기술 및 연구 분야로 감정 분석, 음석 인식, 문장 분류, 기계 번역 등의 활용 분야가 있음



2. 자연어 처리(NLP)의 중요성

자연어 처리(NLP)는 인간 언어인 텍스트 데이터를 기계가 이해하도록 하는 기술로, 정보 분석, 인공지능 등에서 핵심적인 역할을 하며, 디지털 사회의 소통과 자동화를 이끄는 핵심 기술임

3. 자연어 처리(NLP)의 활용

요약

긴 텍스트나 문서에서 핵심 내용을 추출하여 **간결하게 요약**해 주는 기술

챗봇

사람과 기계 간의 대화를 자연스럽게 처리, 사용자의 질문이나 요청에 **적절한 응답 제공**해주는 기술

기계 번역

하나의 언어로 작성된 텍스트를 자동으로 **다른 언어로 변환**하는 과정

감정 분석

텍스트에서 **감정이나 의견**을 자동으로 **식별하고 분류**하는 기술

4. 자연어 처리(NLP) 과정

1

텍스트 수집
(Data collection)

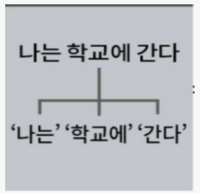
분석할 데이터 수집
예: 뉴스 기사, 리뷰, 문서 등

2

데이터 전처리
(Preprocessing)

- 단어들을 어근 형태로 축소
- 불필요 단어 제거

3



토큰화
(Tokenization)

기계는 텍스트를 직접 이해하지 못하므로, 수치적 표현(벡터)으로 변환

4

벡터화
(Vectorization)

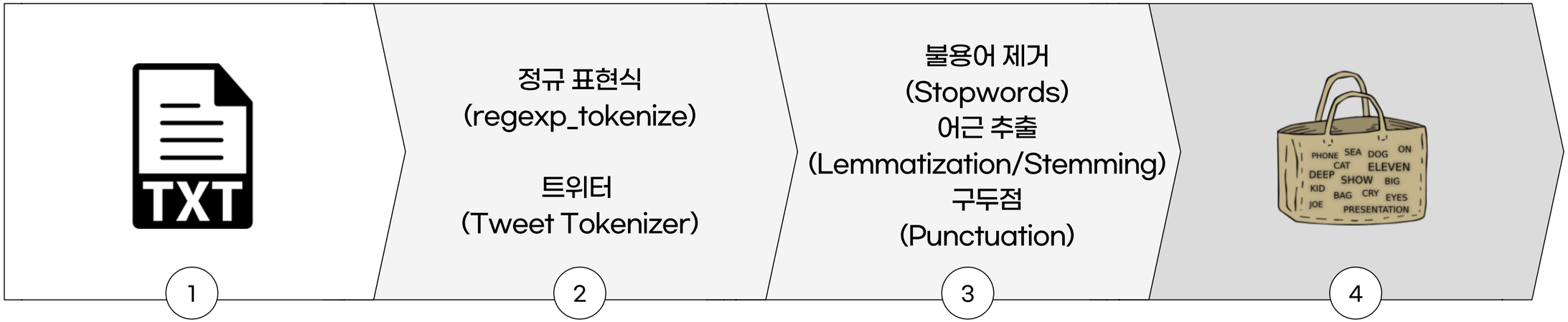
변환된 텍스트 벡터를 활용해 분류, 요약, 생성, 추론 등 다양한 task 수행

5

Task 적용
(Modeling)

변환된 텍스트를 활용하여 감정 분석, 자동 요약, 주제별 분류 등의 다양한 작업을 수행함

데이터 전처리



문서 (Document)

토큰화

(Tokenization)

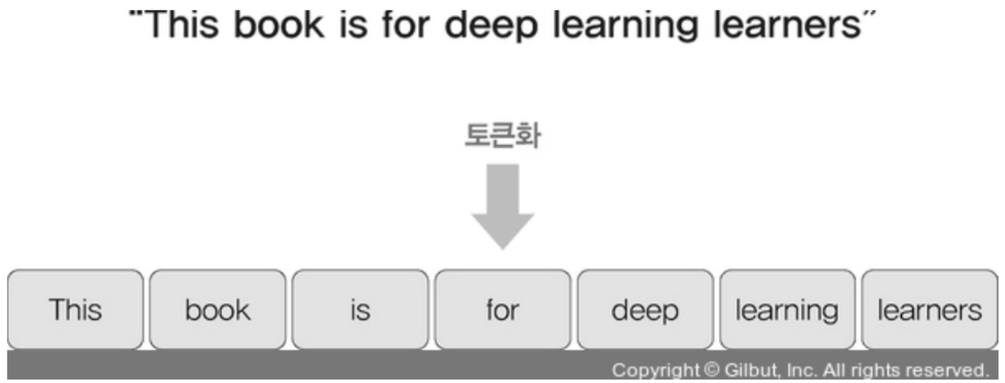
텍스트 전처리

(Preprocessing)

단어 주머니

(Bag of Words)

토큰화(Tokenization)



텍스트 조각을 토큰이라고 하는 더 작은 단위로 분리하는 방법으로 띄어쓰기, 글자, 형태소 등 다양한 방식으로 처리

불용어 제거

불용어 : 문장에서 자주 등장하지만 실제 의미분석에 도움 되지 않는 조사나 접미사
('그리고','이','그','에서' 등)

```
stopwords = ['가입', '기여', '역대','오리지널','회사','최신','소개','관련', '이', '시간', '나오', '있', '가지
'되', '생각하', '수', '그러', '이', '속', '생각', '보', '하나', '않', '집', '없', '살',
'나', '모르', '사람', '적', '주', '월', '아니', '데', '등', '자신', '같', '안', '우리', '어떤',
'때', '내', '년', '가', '경우', '한', '명', '지', '생각', '대하', '시간', '오', '그녀',
'말', '다시', '일', '이런', '그렇', '앞', '위하', '보이', '때문', '번', '그것', '나',
'두', '다른', '특징', '말하', '어떻', '알', '여자', '남자', '그러나', '개', '발', '전',
'못하', '들', '일', '사실', '그런', '이렇', '또', '점', '문제', '싶', '더', '말', '사회', '정도',
'많', '좀', '그리고', '원', '좋', '잘', '크', '통하', '따르', '소리', '중', '놓']
```

불용어를 포함한 채로 모델 학습 시키면 불필요한 패턴을 학습한 위험이 있어 이를 제거해야함

형태소 분석을 통한 불용어 파악 및 제거

형태소 분석기인 KoNLPy의 Okt 형태소 분석기를 사용하여 단어를 형태소 단위로 분리하고, 문장 에서 조사, 접미사, 대명사, 감탄사 등을 파악 및 제거 가능함

Sejong project (ntags=42)		Sim Gwangsub project (ntags=26)		Twitter Korean Text (ntags=19)		Komoran (ntags=42)		Mecab-ko (ntags=43)		Kkma (ntags=10)	Kkma (ntags=30)	Kkma (ntags=56)		Hannanum (ntags=9)		Hannanum (ntags=22)		Hannanum (ntags=26)		Hannanum (ntags=69)		Example
Tag	Description	Tag	Description	Tag	Description	Tag	Description	Tag	Description	Tag	Tag	Tag	Description	Tag	Description	Tag	Description	Tag	Description	Tag	Description	
NNG	일반 명사	NN	명사	Noun	명사 (Nouns, Pronouns, Company Names, Proper Noun, Person Names, Numerals, Standalone, Dependent)	NNG	일반 명사	NNG	일반 명사	N	NN	NNG	보통명사	N	체언	NC	보통명사	NCP	서술성명사	NCPA	동작성 명사	
																				NCPS	상태성 명사	
																		NCN	비서술성명사	NCN	비서술성 명사	어린이집
																				NCR	비서술성-직위 명사	
NNP	고유 명사					NNP	고유 명사	NNP	고유 명사			NNP	고유명사			NQ	고유명사	NQ	고유명사	NQPA	성	손
																				NQPB	이름	예진
																				NQPC	성+이름	손예진
																				NQQ	기타-일반	대한민국
NNB	의존 명사	NX	의존 명사			NNB	의존 명사	NNB	의존 명사			NNB	일반 의존 명사			NB	의존명사	NB	의존명사	NBN	비단위성 의존명사	때문
																				NBS	비단위성 의존명사 - 하다 붙는 것	
		UM	단위 명사					NNBC	단위를 나타내는 명사			NNM	단위 의존 명사							NBU	단위성 의존명사	퍼센트
NR	수사	NU	수사			NR	수사	NR	수사		NR	NR	수사			NN	수사	NN	수사	NNC	양수사	백만
																				NNO	서수사	첫째
NP	대명사	NP	대명사			NP	대명사	NP	대명사		NP	NP	대명사			NP	대명사	NP	대명사	NPP	인칭대명사	자기
																				NPD	지시대명사	
VV	동사	VV	동사	Verb	동사	VV	동사	VV	동사	V	VV	VV	동사	P	용언	PV	동사	PV	동사	PVD	지시 동사	
																				PVG	일반 동사	
VA	형용사	VA	형용사	Adjective	형용사	VA	형용사	VA	형용사		VA	VA	형용사			PA	형용사	PA	형용사	PAD	지시 형용사	
																				PAA	성상 형용사	
VX	보조 용언	VX	보조 동사			VX	보조 용언	VX	보조 용언		VX	VXV	보조 동사			PX	보조 용언	PX	보조 용언	PX	보조 용언	아니하
		AX	보조 형용사									VXA	보조 형용사									
VCP	긍정 지정사	CP	서술격 조사 '이다'			VCP	긍정 지정사	VCP	긍정 지정사		VC	VCP	긍정 지정사, 서술격									
VCN	부정 지정사					VCN	부정 지정사	VCN	부정 지정사			VCN	부정 지정사, 형용사									아니
MM	관형사	DN	수 관형사	Determiner	관형사 (ex: 새, 헌, 첫, 첫, 이, 그, 저)	MM	관형사	MM	관형사	M	MD	MDN	수 관형사	M	수식언	MM	관형사	MM	관형사	MMD	지시 관형사	각종 (분들은)
		DT	일반 관형사									MDT	일반 관형사							MMA	성상 관형사	
MAG	일반 부사	AD	부사	Adverb	부사 (ex: 잘, 걱우, 빨리, 반드시,	MAG	일반 부사	MAG	일반 부사		MA	MAG	일반 부사			MA	부사	MA	부사	MAG	일반 부사	전부, 직접, 미리
MAJ	접속 부사			Conjunction	접속사	MAJ	접속 부사	MAJ	접속 부사			MAC	접속 부사							MAJ	접속 부사	또는, 다만
																				MAD	지시 부사	
IC	감탄사	EX	감탄사	Exclamation	감탄사 (ex: 헐, 어머나, 앞써구)	IC	감탄사	IC	감탄사	I	IC	IC	감탄사	I	독립언	II	감탄사	II	감탄사	II	감탄사	어머나
JKS	주격 조사	JO	조사	Josa	조사 (ex: 의, 에, 에서)	JKS	주격 조사	JKS	주격 조사	J	JK	JKS	주격 조사	J	관계언	JC	격조사	JC	격조사	JCS	주격 조사	
JKC	보격 조사					JKC	보격 조사	JKC	보격 조사			JKC	보격 조사							JCC	보격 조사	
JKG	관형격 조사					JKG	관형격 조사	JKG	관형격 조사			JKG	관형격 조사							JCM	관형격 조사	
JKO	목적격 조사					JKO	목적격 조사	JKO	목적격 조사			JKO	목적격 조사							JCO	목적격 조사	
JKB	부사격 조사					JKB	부사격 조사	JKB	부사격 조사			JKM	부사격 조사							JCA	부사격 조사	으로, 에서, 에게
JKV	호격 조사					JKV	호격 조사	JKV	호격 조사			JKI	호격 조사							JCV	호격 조사	
JKQ	인용격 조사					JKQ	인용격 조사	JKQ	인용격 조사			JKQ	인용격 조사							JCR	인용격 조사	
JC	접속 조사					JC	접속 조사	JC	접속 조사		JC	JC	접속 조사							JCJ	접속격 조사	이나
																				JGT	공동격 조사	
JX	보조사					JX	보조사	JX	보조사		JX	JX	보조사			JX	보조사	JX	보조사	JXC	통용 보조사	부터, 까지
																				JXF	통용 보조사	합니다) 그려,
																JP	서술격 조사	JP	서술격 조사	JP	서술격 조사	
EP	선어말어미	EP	선어말어미	PreEomi	선어말어미 (ex: 았)	EP	선어말어미	EP	선어말어미		EP	EPH	존칭 선어말			EP	선어말어미	EP	선어말어미	EP	선어말어미	
												EPT	시제 선어말									
												EPP	공손 선어말									
												EFN	평서형 종결 어미									
													어말형									

BOW: Bag of Words

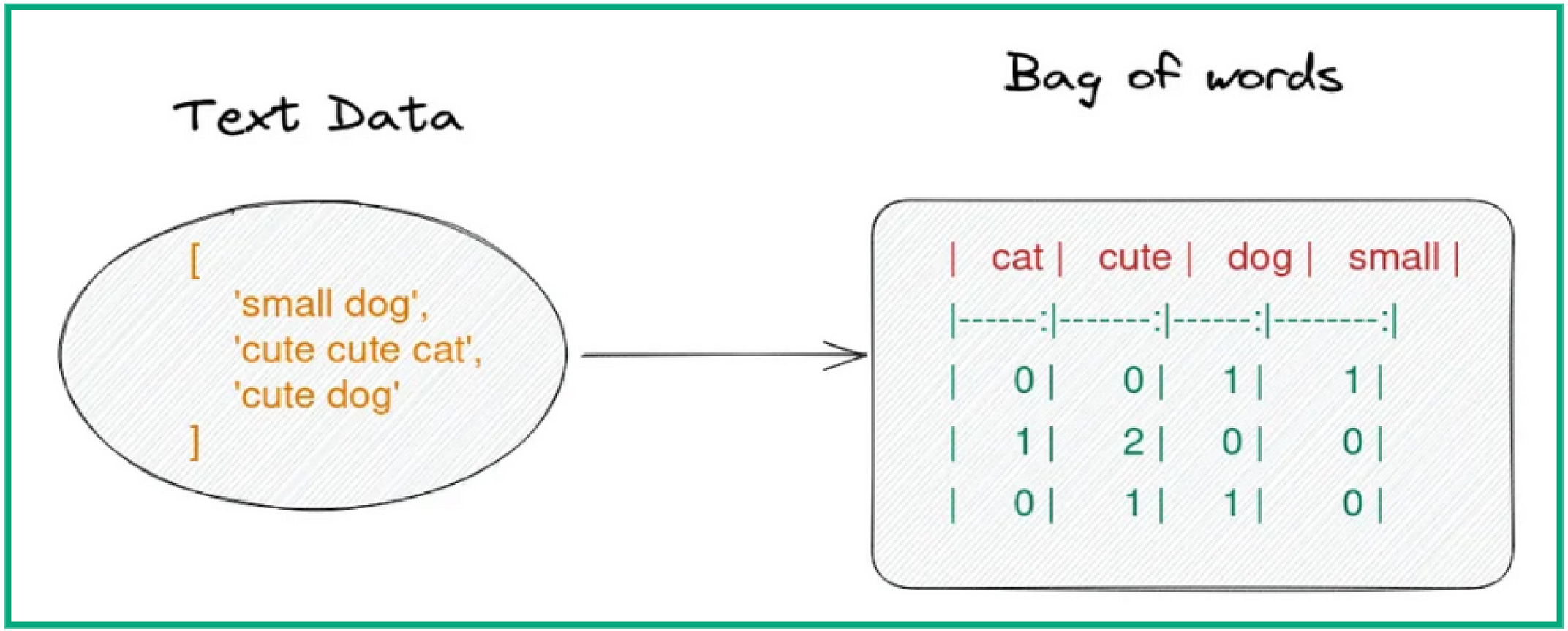
텍스트를 담는 가방

단어들의 순서는 고려하지 않고,
단어들의 출현 빈도에만 집중하는
텍스트 데이터의 수치화 표현 방법

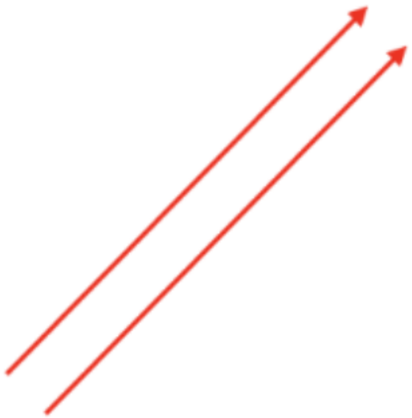
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

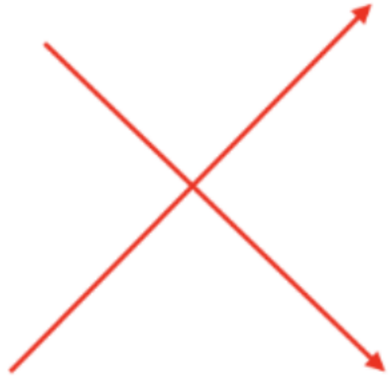
BOW행렬 생성


Text Data
[
'small dog',
'cute cute cat',
'cute dog'
]
Bag of words
cat	cute	dog	small
0	0	1	1
1	2	0	0
0	1	1	0

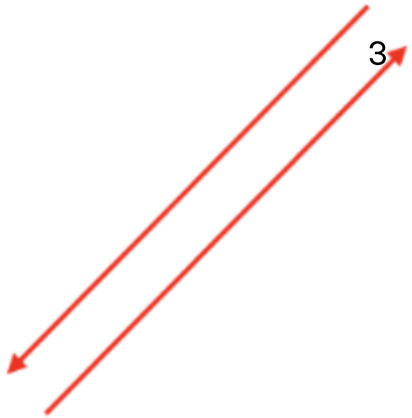
5. 코사인 유사도



코사인 유사도: 1



코사인 유사도: 0



코사인 유사도: -1

두 문서 벡터의 방향이 얼마나 유사한지를 측정

코사인 유사도 정의

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

A·B : 두 벡터의 내적

‖ A ‖ : 벡터 A의 크기

‖ B ‖ : 벡터 B의 크기

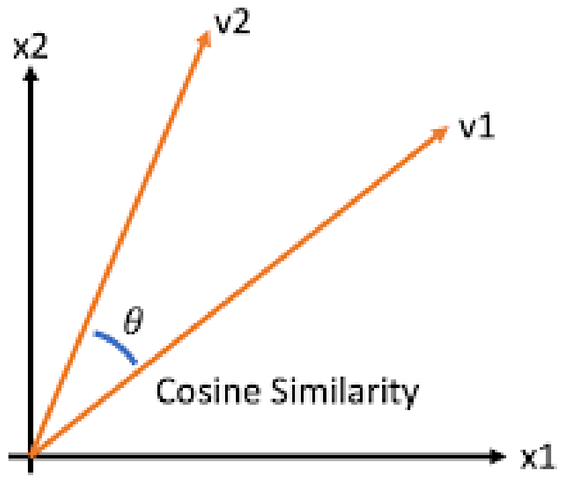
두 벡터의 내적을 각 벡터의 크기의 곱으로 나눈 값

[-1,1] 범위를 가지며, 일반적인 문서 벡터에서는 0~1 사이의 값을 가짐

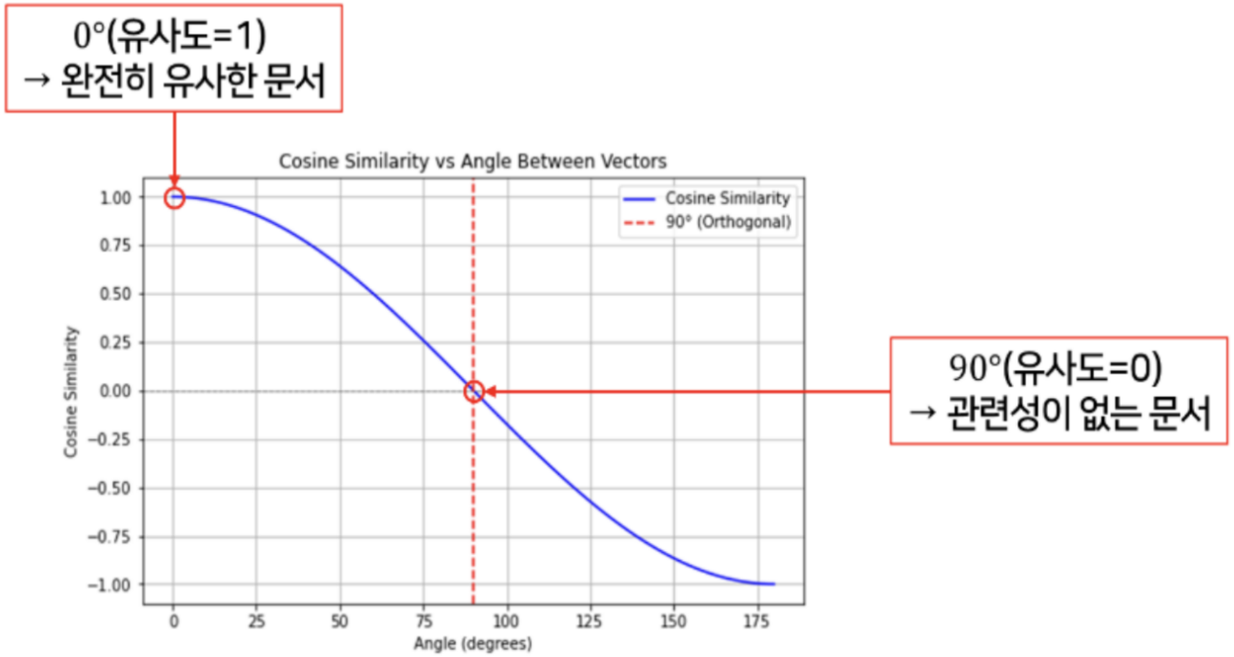
코사인 유사도 특징

1. 크기가 아닌 방향으로 유사도 측정
: 문장의 길이가 다르더라도 내용이 비슷할 경우, 높은 유사도 가질 수 있음

2. 코사인 유사도는 0에서 1 사이의 값을 가짐
: 1에 가까울수록 두 문장은 비슷하고, 0에 가까울수록 다름



두 벡터가 이루는 각도에 따른 코사인 유사도



코사인 유사도 그래프

X축: 두 벡터가 이루는 각도

Y축: 코사인 유사도 값

- 각도가 0° : 두 벡터가 완전히 유사
- 각도가 90° : 관련성 x

예제

문장 A : "나는 파이썬을 좋아합니다."
문장 B : "나는 파이썬과 자바를 공부합니다."

벡터화(Vectorization)
- 단어 출현 수 CountVectorizer 사용

단어	나는	파이썬	좋아합니다	자바	공부합니다
문장 A	1	1	1	0	0
문장 B	1	1	0	1	1

벡터 A = (1, 1, 1, 0, 0)
벡터 B = (1, 1, 0, 1, 1)

벡터 A = (1, 1, 1, 0, 0) / 벡터 B = (1, 1, 0, 1, 1)

코사인 유사도 계산

1	내적	→ $(1\times1)+(1\times1)+(1\times0)+(0\times1)+(0\times1)=1+1+0+0+0=2$
2	벡터의 크기(norm)	→ $\|\mathbf{A}\| = \sqrt{(1^2+1^2+1^2+0^2+0^2)} = \sqrt{3} \approx 1.732$ $\|\mathbf{B}\| = \sqrt{(1^2+1^2+0^2+1^2+1^2)} = \sqrt{4} = 2$
3	코사인 유사도 계산	→ $\cos(\theta) = \frac{1.732 \times 2}{2 \times 2} = \frac{3.464}{4} \approx 0.577$

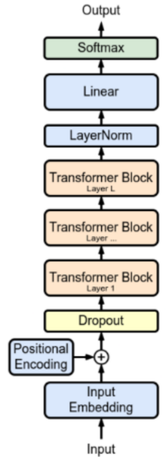
6. 자연어 처리(NLP) 최신 동향

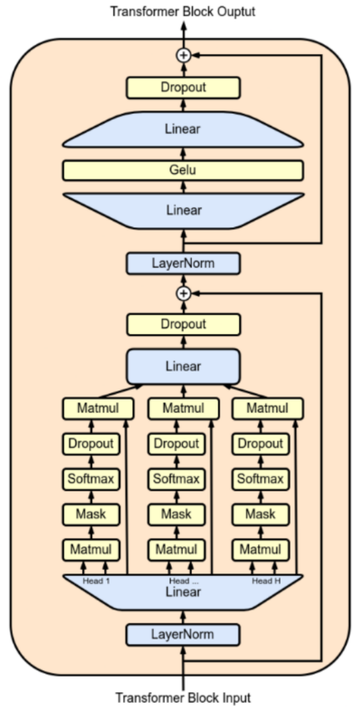
LLM
(Large Langauge Model)

Chat GPT

LLM
(Large Language Models)

대량의 텍스트 데이터를 기반으로 학습한
매우 큰 인공신경망 기반의 언어 모델





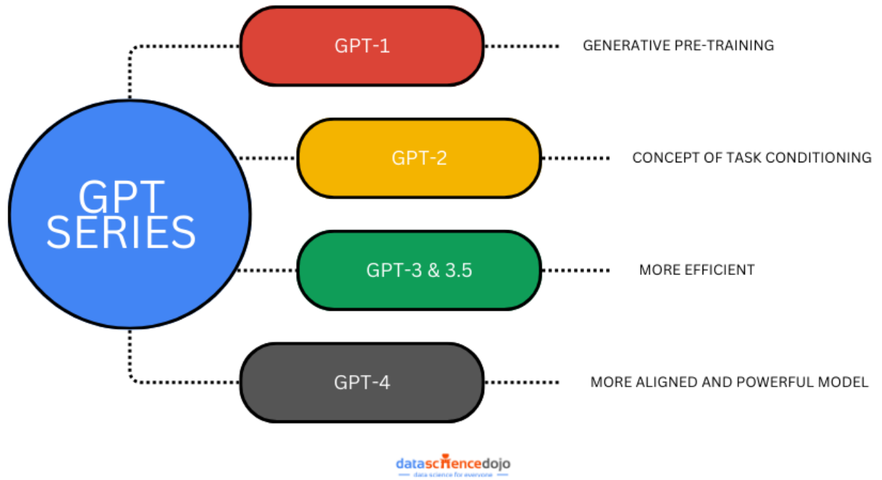
특징) 수많은 파라미터를 가지고 다양한 언어 작업을 하나의 모델로 수행

기술기반) Transformer 아키텍처(Vaswani et al., 2017)

GPT
(Generative Pre-trainedTransformer)

최근 자연어 처리(NLP)의 발전은 LLM중심으로 빠르게 전개되고 있음

특히 GPT(Generative Pre-trained Transformer) 계열의 모델이 등장하면서 언어 생성, 이해, 요약, 번역, 질의응답 등 거의 모든 NLP 작업을 하나의 모델로 통합 수행할 수 있게 됨



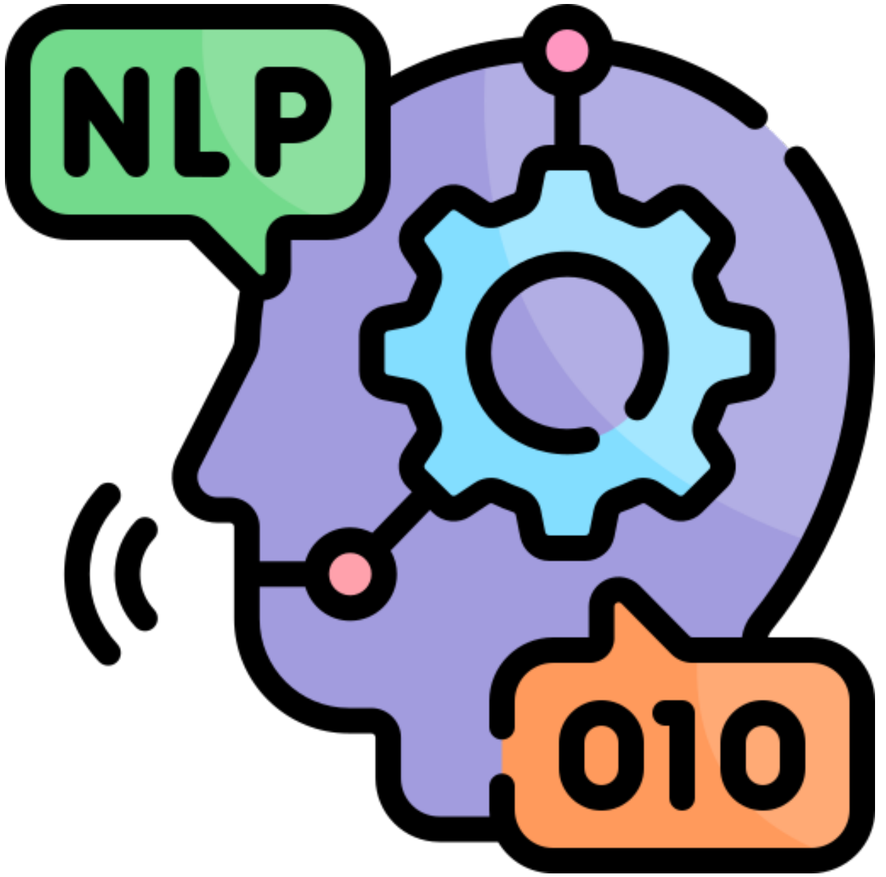
7. 정리

자연어 처리

텍스트 전처리

NLP 기본 예제 : 코사인 유사도

활용예시



KUGGLE

감사합니다