자연어: 일상 생활에서 자연히 발생하여 쓰이는 언어

NLP: 자연어를 컴퓨터가 처리하고 그 의미를 이해하는 기술

규칙 기반 접근

if (날씨를) ->

규칙적으로 검색해올것

확률적 접근

TF-IDF 키워드 추출

ㄴ인공지능을 위한 수학 책로

전처리

- 개행문자 제거
- 특수문자 제거
- 공백 제거
- 중복 표현제거 (ㅋㅋㅋㅋ)
- 이메일 링크 제거
- 제목 제거
- 불용어 제거
- 조사 제거
- 띄어쓰기, 맞춤법 교정
- 자소 추출

Tokenizing

- 어절 tokenizing
- 형태소 tokenizing
- n-gram tokenizing
- word piece tokenizing

Lexical analysis

- 어휘분석
  - 형태소 분석
  - 개체명 인식
  - 상호 참조

Syntactic analysis

- 구문 분석

Semantic analysis

- 의미분석

One hot encoding

단어 벡터가 sparse해서 단어간 갖는 의미를 벡터공간에 표현못함

Word2Vec

- 단어의 주변단어들을 통해, 그 단어의 의미를 파악
- 자연어의 의미를 벡터 공간에 임베딩
- OOV (out of vocabulary)에 취약 불가능

Sparse representation	Dense representation
· one hot encoding · 차원의 저주	· word embedding · 한정된 차원으로 표현가능 · 의미 관계 유추가능

단어의 의미가 벡터로 표현됨으로써 벡터 연산 가능

Word Embedding 방법

Word Similarity (사람이 하기) 평가 (=) 코사인유사도

analogy test 벡터 연산을 통해 검증

gensim word2vec library

Fasttext

- 단어를 n-gram으로 나누어 학습
- 단어를 n-gram으로 분리를 한 후, 모든 n-gram vector를 합산한 후 평균을 통해 단어 벡터를 획득

orange -> <orange> ->

2-gram	3-gram	4-gram	5-gram	origin
<o,	<or,	<ora,	<oran,	orange
or,	ora,	oran,	orang,	
ra,	ran,	rang,	range,	
an,	ang,	ange,	ange>	
ng,	nge,	nge>		
ge,	ge>			
e>				

앞뒤에서는 작은단위로 ~~묶어서 6개로~~

total n-gram vocabsize 22

그래서 oranges, oranze도 orange와 유사한 벡터를 가짐

오탈자, OOV, 등장 횟수가 적은 단어에 강하다

word2vec < fast text

Word Embedding의 한계

다의어의 학습의 어려움