Computer System Lab.

NLP – 시소러스(통계기반기법)

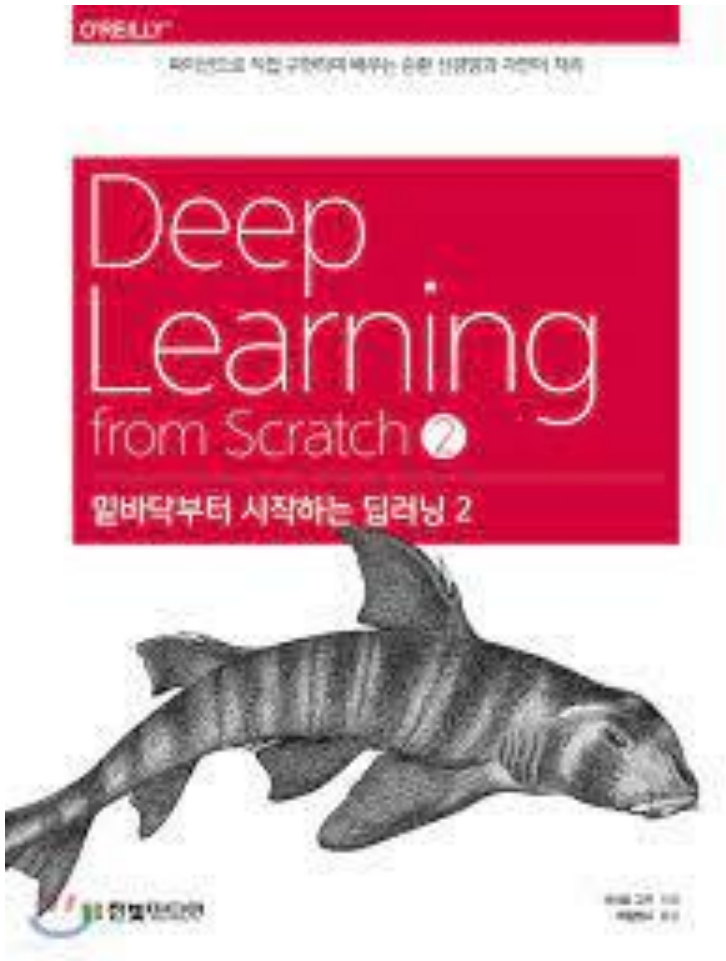
순천향대학교
컴퓨터시스템연구실

이인규

22.10.06

SCH University

레퍼런스



밑바닥부터 시작하는 딥러닝2

- 자연어 처리(NLP)
- 순환신경망(RNN)
- LSTM

…

자연어 처리 NLP

- Natual Language Processing
- 우리말(자연어)을 컴퓨터에게 이해시키기 위한 기술

시소러스 Thesaurus

- 뜻 : 유의어 사전
- 비슷한 단어들끼리 묶어놓은 사전
- 장점 : 컴퓨터가 비슷한 단어를 구분할 수 있다.
  ex) 연관 검색어
- 단점 : 사람이 직접 분류해야 한다.
- 그 해결책으로 통계기반기법을 제시

통계 기반 기법

◆ 분포 가설 Distributional Hypothesis에 기초

-> 단어의 의미는 주변 단어에 의해 형성된다.

아는 단어 : beer(술), apple(사과)

모르는 단어 : wine(와인), banana(바나나)

I drink beer. 나는 술을 마신다.

I eat apple. 나는 사과를 먹는다.

I drink wine. 나는 와인을 마신다.

I eat banana. 나는 바나나를 먹는다.

마실 것: beer, wine | 먹는 것: eat, banana

통계 기반 기법 과정

You say goodbye and I say hello.

◆ 1. 단어마다 split

```
text = text.lower()
text = text.replace('.', ' .')
words = text.split(' ')
```

통계 기반 기법 과정

You say goodbye and I say hello.

◆ 2. window size 결정, 동시 발생 행렬 생성

-> numpy 배열 사용

	you	say	goodbye	and	i	hello	.
you	0	1	0	0	0	0	0
say	1	0	1	0	1	1	0
goodbye	0	1	0	1	0	0	0
and	0	0	1	0	1	0	0
i	0	1	0	1	0	0	0
hello	0	1	0	0	0	0	1
.	0	0	0	0	0	1	0

통계 기반 기법 과정

◆3. 양의 점별 상호정보량 PPMI

–> Positive Pointwise Mutual Information

$$PMI(x,y) = \log_2\left(\frac{P(x,y)}{P(x)P(y)} + eps\right)$$

$$PPMI(x,y) = \max(\, PMI(x,y), 0\,)$$

```
M = np.zeros_like(C, dtype=np.float32)
N = np.sum(C)
S = np.sum(C, axis=0)
total = C.shape[0] * C.shape[1]
cnt = 0

for i in range(C.shape[0]):
    for j in range(C.shape[1]):
        pmi = np.log2(C[i, j] * N / (S[j]*S[i]) + eps)
        M[i, j] = max(0, pmi)
```

	you	say	goodbye	and	i	hello	.
you	0	1	0	0	0	0	0
say	1	0	1	0	1	1	0
goodbye	0	1	0	1	0	0	0
and	0	0	1	0	1	0	0
i	0	1	0	1	0	0	0
hello	0	1	0	0	0	0	1
.	0	0	0	0	0	1	0

통계 기반 기법 과정

◆ 4. PPMI 계산 후 유사도 출력

	you	say	goodbye	and	i	hello	.
you	0	1.807	0	0	0	0	0
say	1.807	0	0.807	0	0.807	0.807	0
goodbye	0	0.807	0	1.807	0	0	0
and	0	0	1.807	0	1.807	0	0
i	0	0.807	0	1.807	0	0	0
hello	0	0.807	0	0	0	0	2.807
.	0	0	0	0	0	2.807	0

통계 기반 기법의 개선 방향

◆ 문장이 많아지면 연산량이 매우 많아진다.

-> 단어의 개수가 10개 늘면 연산량도 10배 많다.

◆ 문장의 노이즈에 약하다.

-> 잘못된 문장이 다른 유사도 계산시에 영향을 준다.

Question?

Please contact :

이인규
순천향대학교 컴퓨터학부
멀티미디어관 M606

Email : dldlsrb1414@naver.com