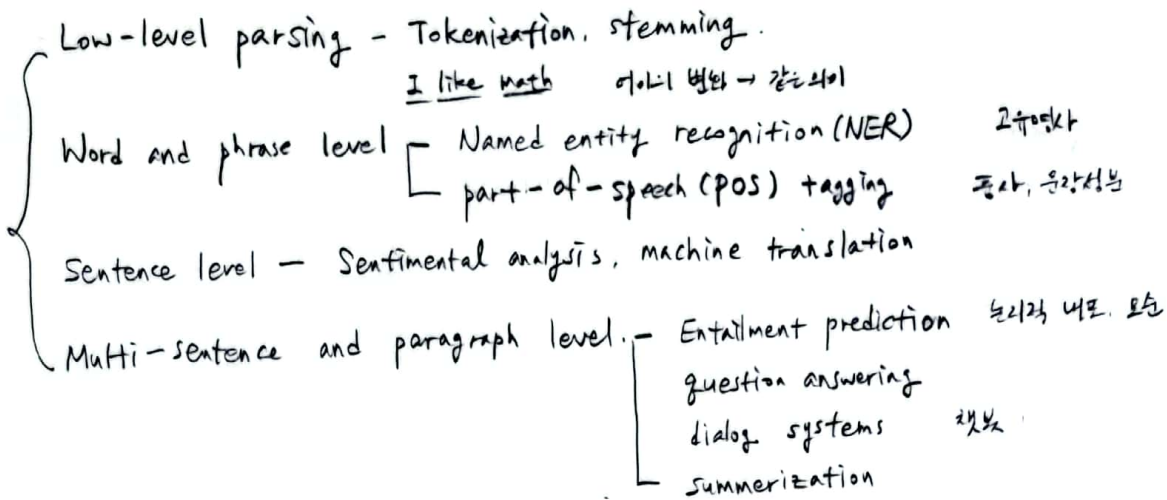


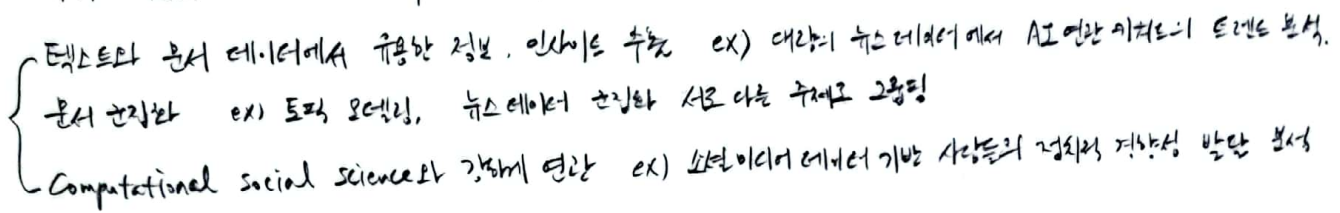
자연어처리 (NLP)

학회: ACL, EMNLP, NAACL



텍스트마이닝

학회: KDD, The WebConf (formerly, WWW), WSDM, CIKM, ICWSM



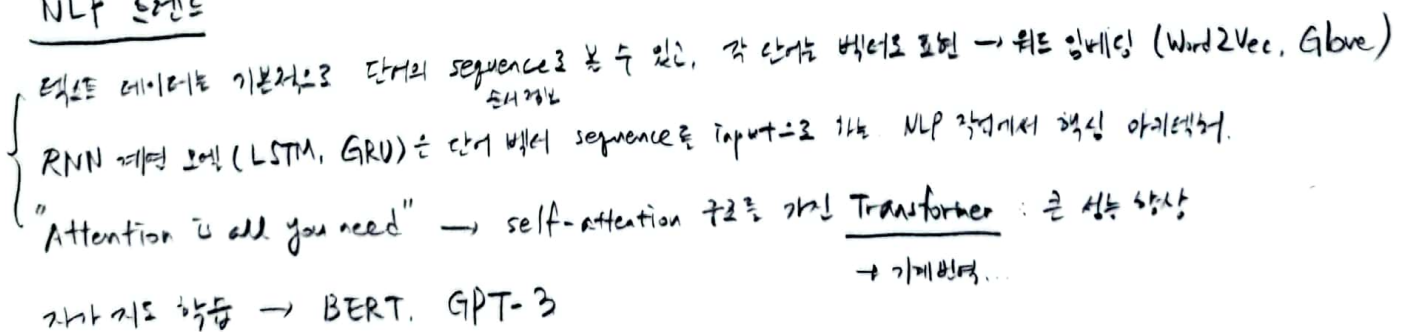
정보 검색

학회: SIGIR, WSDM, CIKM, Rec Sys

Computation social science와 강하게 연관

- 자율 학습 방식으로 연구되고 있지 않음
- 추천시스템으로 발전, 여전히 추천한 연구 분야

NLP 트렌드



Bag-of-Words

Step 1. text 데이터셋에서 unique한 단어를 모아 Vocabulary 구축

Step 2. 각각의 unique 단어를 one-hot vector로 표현

John: [1 0 0 0 0 0 0], really: [0 1 0 0 0 0 0], ...

한 단어씩에서 유클리드 거리는 $\sqrt{2}$, 코사인 유사도는 0

→ 단어의 의미나 상관관계, 유사도 등인한 관계

문장/문서는 one-hot vector들의 합으로 표현

Naïve Bayes Classifier

문서나 Class에 Bayes' rule 적용

문서 d가 class C에서, $C_{\text{map}} = \underset{C \in C}{\text{argmax}} P(C|d)$ MAP: maximum a posteriori = most likely class

$$= \underset{C \in C}{\text{argmax}} \frac{P(d|C)P(C)}{P(d)} \quad \text{Bayes' rule}$$

constant라고 할 수 있음 (fixed value)

$$= \underset{C \in C}{\text{argmax}} P(d|C)P(C)$$

문서 d에서, 단어 sequence W와 class C로 구성

문서의 확률은 각 단어의 확률 곱셈으로 계산하는 것으로 표현

$$P(d|C)P(C) = P(W_1, W_2, \dots, W_n | C)P(C)$$

$$= P(C) \prod_{w_i \in W} P(w_i | C) \quad (\text{각 단어가 등장할 확률이 독립되어 있다고 가정})$$

예제)	Doc(d)	Document (words, W)	Class (C)
Training	1	Image recognition uses convolutional neural networks	CV
	2	Transformer can be used for image classification task	CV
	3	Language modeling uses transformer	NLP
	4	Document classification task is language task	NLP
Test	5	classification task uses transformer	?

$$P(C_{cv}) = \frac{2}{4} = \frac{1}{2}$$

$$P(C_{nlp}) = \frac{2}{4} = \frac{1}{2} \quad : \text{prior}$$

각 단어 w_i 에 대해 class c_i 에 대한 조건부 확률 계산

$$\rightarrow P(w_i | c_i) = \frac{n_{ik}}{n_i}, \text{ 따라서 } c_i \text{의 하위 단어 } w_k \text{이 나타나는 횟수인 } n_{ik}$$

$$\begin{cases} P(w \text{ "classification"} | C_{cv}) = \frac{1}{14} \\ P(w \text{ "task"} | C_{cv}) = \frac{1}{14} \\ P(w \text{ "uses"} | C_{cv}) = \frac{1}{14} \\ P(w \text{ "transformer"} | C_{cv}) = \frac{1}{14} \end{cases} \begin{cases} P(w \text{ "classification"} | C_{nlp}) = \frac{1}{10} \\ P(w \text{ "task"} | C_{nlp}) = \frac{2}{10} \\ P(w \text{ "uses"} | C_{nlp}) = \frac{1}{10} \\ P(w \text{ "transformer"} | C_{nlp}) = \frac{1}{10} \end{cases}$$

Test 문서 d_s 에 대해, 각 class에 대한 조건부 확률 계산 후 최대 확률 class 선택

$$\begin{cases} P(C_{cv} | d_s) = P(C_{cv}) \prod_{w \in w} P(w | C_{cv}) = \frac{1}{2} \times \frac{1}{14} \times \frac{1}{14} \times \frac{1}{14} \times \frac{1}{14} \\ P(C_{nlp} | d_s) = P(C_{nlp}) \prod_{w \in w} P(w | C_{nlp}) = \frac{1}{2} \times \frac{1}{10} \times \frac{2}{10} \times \frac{1}{10} \times \frac{1}{10} \end{cases}$$

Naive Bayes classifier는 class 개수가 3개 이상일 때도 확률 가능

특정 클래스에서 학습 data 내에 특정 단어가 전혀 없는 경우, 해당 단어가 나타날 확률이 0으로 처리

이 경우, 그 단어가 포함된 문장이 주어지면, 그 class가 될 확률이 0으로 계산

\rightarrow 추가적인 regularization 기법들이 적용되기도 함

MLE를 통해 유도