자연어처리 (NLP)

학회: ACL, EMNLP, NAACL

- Low-level parsing - Tokenization, stemming.
  I like math 어미 변화 → 같은 의미
- Word and phrase level
  - Named entity recognition (NER) 고유명사
  - part-of-speech (POS) tagging 동사, 문장성분
- Sentence level — Sentimental analysis, machine translation
- Multi-sentence and paragraph level.
  - Entailment prediction 논리적 내포, 모순
  - question answering
  - dialog systems 챗봇
  - summerization

텍스트마이닝

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

- 텍스트와 문서 데이터에서 유용한 정보, 인사이트 추출 ex) 대량의 뉴스데이터에서 AI 관련 키워드의 트렌드 분석.
- 문서 분류화 ex) 토픽 모델링, 뉴스데이터 분류화 서로 다른 주제로 그룹핑
- Computational social science와 강하게 연관 ex) 소셜미디어 데이터 기반 사람들의 정치적 적극성 발달 분석

정보 검색

학회: SIGIR, WSDM, CIKM, RecSys

Computation social science와 강하게 연관
- 자주는 안보이게 연구되고 있지 않음
- 추천시스템으로 발전, 여전히 활발한 연구 분야

NLP 트렌드

- 텍스트 데이터를 기본적으로 단어의 sequence로 볼 수 있고, 각 단어는 벡터로 표현 → 워드 임베딩 (Word2Vec, Glove)
  순서정보
- RNN 계열 모델 (LSTM, GRU)은 단어 벡터 sequence를 input으로 하는 NLP 작업에서 핵심 아키텍처.
- "Attention is all you need" → self-attention 구조를 가진 Transformer : 큰 성능 상승
  → 기계번역...

자기 지도 학습 → BERT, GPT-3

Bag-of-Words

Step 1. text 데이터셋에서 unique한 단어를 모아 Vocabulary 추출

Step 2. 각각의 unique 단어를 one-hot vector로 표현

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

한 단어쌍에서 유클리드 거리는 $\sqrt{2}$, 코사인유사도는 0

→ 단어의 의미나 우관계, 오두가 동일한 관계

문장 / 문서는 one-hot vector들의 합으로 표현

Naive Bayes Classifier

문서나 Class에 Bayes' rule 적용

문서 d와 class c에서, $c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c|d)$ MAP: maximum a posteriori = most likely class

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

$P(d)$ → constant라고 할 수 있음 (fixed 문서)

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

문서 d에서, 단어 sequence W와 class C로 구성

문서의 확률은 각 단어의 출현 확률을 곱하는 것으로 표현

$P(d|c)P(c) = P(w_1, w_2 \cdots w_n|c)P(c)$

$= P(c) \prod_{w_i \in W} P(w_i|c)$ (각 단어가 등장할 확률이 C가 고정되어 있을 때 독립이라고 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	?

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

$P(C_{NLP}) = \frac{2}{4} = \frac{1}{2}$: prior

각 단어 W_i에 대한 class C_{cv}에 대한 조건부 확률 계산

→ $P(W_k|C_i) = \frac{n_k}{n}$, 토픽이 C_i인 문서 안에 W_k이 나타나는 횟수 n_k

$$\begin{cases} P(W\text{"classification"}|C_{CV}) = \frac{1}{14} \\ P(W\text{"task"}|C_{CV}) = \frac{1}{14} \\ P(W\text{"uses"}|C_{CV}) = \frac{1}{14} \\ P(W\text{"transformer"}|C_{CV}) = \frac{1}{14} \end{cases} \quad \begin{cases} P(W\text{"classification"}|C_{NLP}) = \frac{1}{10} \\ P(W\text{"task"}|C_{NLP}) = \frac{2}{10} \\ P(W\text{"uses"}|C_{NLP}) = \frac{1}{10} \\ P(W\text{"transformer"}|C_{NLP}) = \frac{1}{10} \end{cases}$$

Test 문서 d_5에 대해, 각 class에 대한 조건부 확률 계산 후 최대 확률 class 선택

$$\begin{cases} P(C_{CV}|d_5) = P(C_{CV}) \prod_{w \in W} P(W|C_{CV}) = \frac{1}{2} \times \frac{1}{14} \times \frac{1}{14} \times \frac{1}{14} \times \frac{1}{14} \\ P(C_{NLP}|d_5) = P(C_{NLP}) \prod_{w \in W} P(W|C_{NLP}) = \frac{1}{2} \times \frac{1}{10} \times \frac{2}{10} \times \frac{1}{10} \times \frac{1}{10} \end{cases}$$

Naive Bayes classifier는 class 개수가 3개 이상인 때도 확장 가능

특정 클래스에서 학습 data 내에 특정 단어가 전혀 없는 경우, 해당 단어가 나타난 확률이 0으로 수렴

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

→ 추가적인 regularization 기법들이 적용되기도 함

MLE는 동전 유튜