4장

4. Language Model

✳ Language Model

— 문서로부터 query 생성.

- 각 문서 d_i로 probabilistic language model M_{d_i} 생성.
- $p(q|M_{d_i})$ 기반으로 document ranking.
- probabilistic ⟷ deterministic (automata)
- 

w	p(w|q_1)
the	0.2
a	0.1
⋮	⋮

표에 의해 확률 계산됨.

— "generative" model.

- MT, spell correction, speech recognition 등에서 활용.

✳ Formulating LM

— Chain rule

- $$p(w_1, w_2, \cdots, w_n) = \prod_i p(w_i | w_1, w_2, \cdots, w_{i-1})$$
- 단점) 너무 많은 가능한 sentence가 있다. 모두 커버X.

  complexity $= O(V^{n^*})$. n^* : maximum sentence length.

— Markov assumption

- 각 단어 발생은 서로 independent 하다.
- $$p(w_1, \cdots, w_n) = p(w_1)p(w_2)\cdots p(w_n)$$: Unigram Model
- 간단, popular한 방법. 문법, 순서등을 고려 X.

— N-gram language models

- 이전 (n-1)개의 단어에 conditioned.
- ex) bigram : $p(w_1, \cdots, w_n) = p(w_2|w_1)p(w_3|w_2)\cdots p(w_n|w_{n-1})$
- n이 높아질수록 spatial, time complexity가 너무 증가함.
- 정확해지긴 하지만, unigram을 보통 씀.

- 확률로 다음 단어 뽑기
  - cdf (누적분포함수)

* Language models for IR

- query 만든 확률이 가장 높은 document 찾기.
- 각 문서는 language model의 basis가 됨.
  - $P(d|q) = \frac{P(q|d)P(d)}{P(q)}$
  - $P(q)$: constant.
    $P(d)$: prior. same for all d or PageRank
  - $P(q|d)$ 와 $P(d|q)$를 동급으로 취급함.

* MLE와 MAP (best θ 찾기)

- MLE (Maximum Likelihood Estimation)
  - $\hat{\theta} = \text{argmax}_\theta P(X|\theta)$ 찾기
  - small sample size일 때, 어려움.
- MAP (Maximum A posterior)
  - $\hat{\theta} = \text{argmax}_\theta P(X|\theta) = \text{argmax}_\theta P(\theta|X)P(X)$ (prior)
  - prior 정의 문제

* Query Likelihood Model

- $P(q|d)$ 로 ranking하기.
- 가정: 각 query 단어는 서로 독립적으로 생성된다.
  - $P(q|d) = P(\langle t_1, \cdots, t_{|q|}\rangle | d) = \prod_{1 \le i \le |q|} P(t_i|d)$

- $q = w_1 w_2 \cdots w_n$ 이라하면

- $f(q,d) = \log p(q|d) = \sum_{w \in q} \log p(w|d) = \sum_{w \in q} c(w,q) \log p(w|d)$

  document LM

  $\left(p(w|d) = \frac{c(w,d)}{|d|} \right)$

  - 문제) 문법, 순서 고려 X

    d에 없는 단어에 대해 $p(w|d) = 0 \rightarrow f(q,d) = 0$. ∴ smoothing 필요.

- Smoothing : 본 단어에 대한 확률을 낮추고, 못본 단어들에게 확률 할당.

✳ Smoothing LM

- Reference LM 사용.

- $$p(w|d) = \begin{cases} p_{seen}(w|d) & \text{if } w \text{ is in } d \\ \alpha_d p(w|C) & \text{otherwise} \end{cases}$$

  Collection based LM

- $$f(q,d) = \log p(q|d) = \sum_{w \in V, c(w,d)>0} c(w,q) \log p_{seen}(w|d) + \sum_{w \in V, c(w,d)=0} c(w,q) \log \alpha_d p(w|C)$$

  $$= \sum_{w \in V} c(w,d) \log \alpha_d p(w|C) - \sum_{w \in V, c(w,d)>0} \text{''}$$

- $$f(q,d) = \sum_{w \in V, c(w,d)>0} c(w,q) \log \frac{p_{seen}(w|d)}{\alpha_d p(w|C)} + \sum_{w \in V} c(w,q) \log \alpha_d p(w|C)$$

  $|q| \log \alpha_d$, $\sum_{w \in V} c(w,d) \log p(w|C)$

- $$\log p(q|d) = \sum_{\substack{w \in d \\ w \in q}} c(w,q) \log \frac{p_{seen}(w|d)}{\alpha_d p(w|C)} + |q| \log \alpha_d + \sum_{w \in V} c(w,q) \log p(w|C)$$

  TF, DL, IDF, DL normalization., ignore

✳ Smoothing

- Additive smoothing
  - $p(w|d) = \frac{c(w,d) + \alpha}{|d| + \alpha|V|}$
    - $|V|$: Vocab 길이
    - $|d|$: d 길이
    - 여기서 $c(w,d)$: d에서 w의 개수
  - 대하기 : Laplace smoothing.
  - 문제 : 각 단어들의 중요성이 다른데, 이를 무시함.
- Bayesian Smoothing
  - $p(w|d) = \frac{c(w,d) + \alpha p(w|M_c)}{|d| + \alpha}$ → 모든 collection 내 term frequency (not document).
  - known as dirichlet prior smoothing.
  - $p(w|d) = \frac{c(w,d) + \alpha p(w|C)}{|d| + \alpha} = \frac{|d|}{|d| + \alpha} \cdot \frac{c(w,d)}{|d|} + \frac{\alpha}{|d| + \alpha} p(w|C)$ $\quad \alpha \in [0, +\infty]$
- JM Smoothing
  - $p(w|d) = (1-\lambda)\, p(w|d) + \lambda\, p(w|C)$ $\quad \lambda \in [0,1]$
  - λ가 작으면 query를 모두 포함한 문서 위주.
    λ가 크면 긴 query에 적합. (긴 query는 $p(w|d)$가 0이 될 가능성 높음)
  - known as JM Smoothing
  - p59 예제 풀기.

✳ Linear Interpolation

- N-1 gram 확률로 N gram 확률 smoothing 하기.
- $p(w_i | w_{i-1}, \cdots, w_{i-n+1}) = \lambda_1 p(w_i | w_{i-1}, \cdots, w_{i-n+1}) + \lambda_2 p(w_i | w_{i-1}, \cdots, w_{i-n+2}) + \cdots + \lambda_n p(w_i | w_i)$

  $\sum_{i=1}^{n} \lambda_i = 1$

✳ Bayesian smoothing 적용

- $p(w|d) = \frac{|d|}{|d|+\alpha} \cdot \frac{c(w,d)}{|d|} + \frac{\alpha}{|d|+\alpha} p(w|C)$

α_d

$\alpha_d p(w|C) = \frac{\alpha}{|d|+\alpha} p(w|C)$

$\Rightarrow \frac{p_{seen}(w|d)}{\alpha_d p(w|C)} = \frac{\frac{c(w,d)+\alpha p(w|C)}{|d|+\alpha}}{\frac{\alpha p(w|C)}{|d|+\alpha}} = 1 + \frac{c(w,d)}{\alpha \cdot p(w|C)}$

- $\log p(q|d) = \sum_{\substack{w \in q \\ w \in d}} c(w,d) \log\left(1 + \frac{c(w,d)}{\alpha p(w|C)}\right) + |q| \log \frac{\alpha}{|d|+\alpha}$

$= \sum_{\substack{w \in q \\ w \in d}} c(w,d) \log\left(1 + \frac{c(w,d)}{\alpha} \cdot \frac{1}{p(w|C)}\right) + |q| \log \frac{\alpha}{|d|+\alpha}$

TF　　IDF

- JM smoothing도 똑같음. ($\frac{1-\lambda}{\lambda}$ term 추가 ; doc length normalization)

→ 기초적인 것은 choose all correct 문제로.

- 계산보다, 이론 문제 많이나옴 (꼼꼼)
- 옳은 것 모두 고르시오.
  - 하나만 고르면 부분점수 있음.
  - 3개 이상은 거의 없음. 1~2개
- 수식의 변형, 이러한 수식에 대한 설명 고르시오.
- 확률 문제 2개 특징.
  - sample of 10 balls : 10개 = 1 sample
- 10문제. 각 10점.