UNIVERSITY of DE

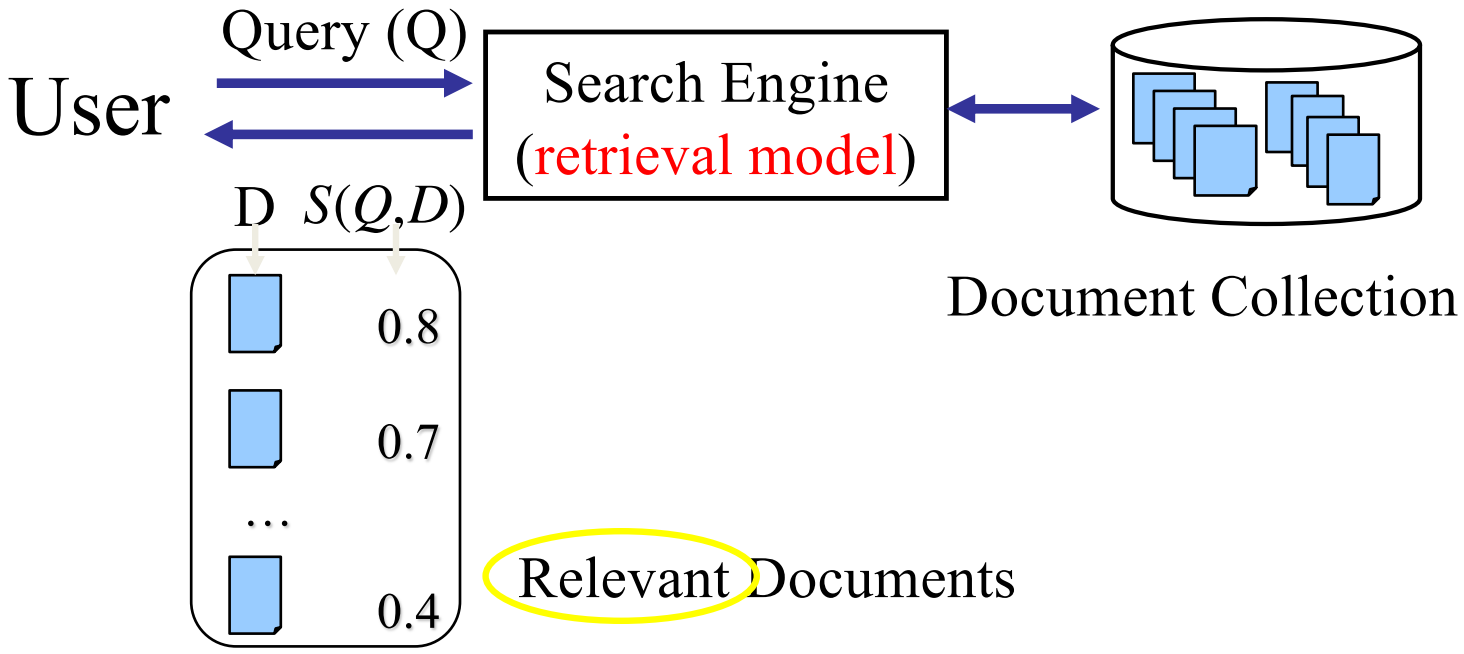
Language Models Approach

(Search and Data Mining)

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

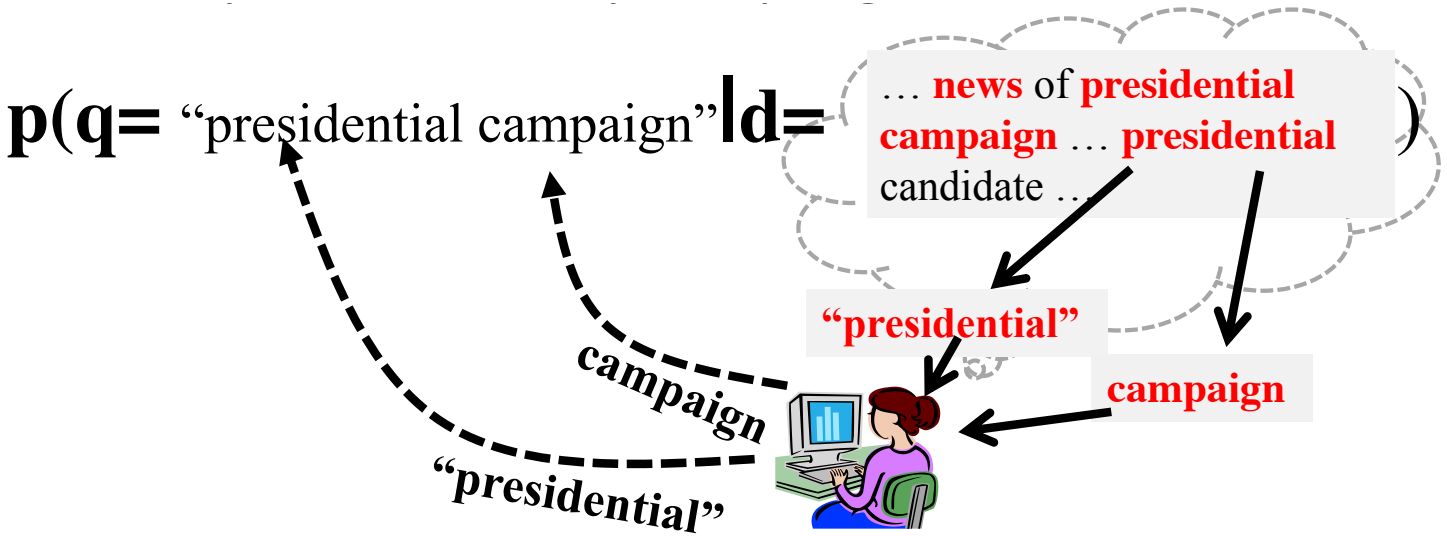
UNIVERSITY of DE

Key Problem: Retrieval Models



How to accurately model relevance, i.e., $S(Q,D)$?

Query Generation by Sampling Words from a Document



If the user is thinking of this document,
how likely would she pose this query?

Unigram Query Likelihood

$$\mathbf{p(q=}\text{“presidential campaign”}\mathbf{|d=}\ \text{… news of presidential campaign … presidential candidate …}\ \mathbf{)}$$

$$= \mathbf{p(}\text{“presidential ”}\mathbf{|d)*p(}\text{“campaign ”}\mathbf{|d)}$$

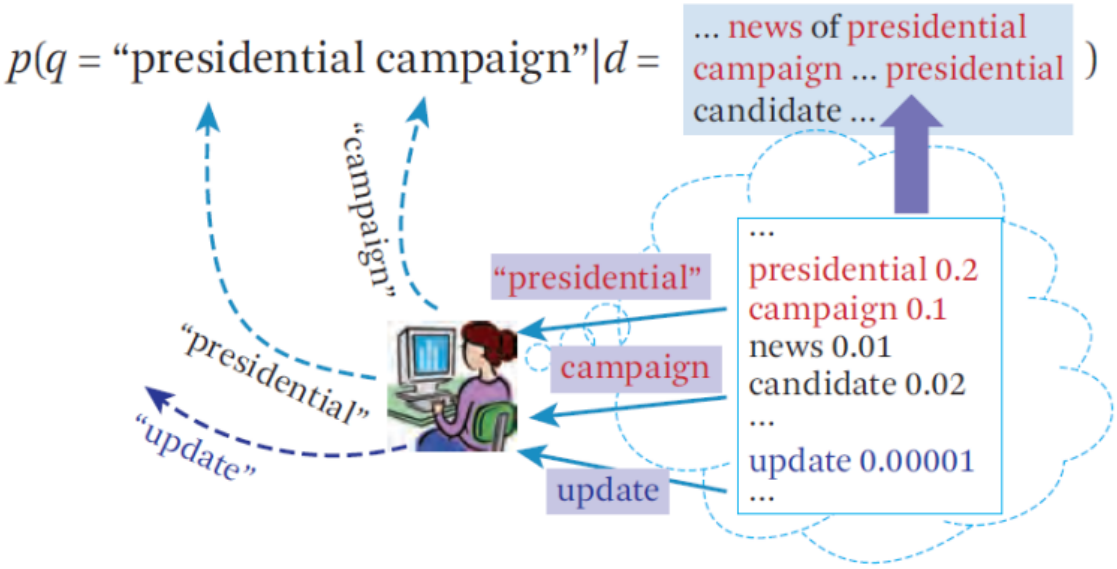
$$= \frac{c(\text{"}presidential\text{"},d)}{|d|} * \frac{c(\text{"}campaign\text{"},d)}{|d|}$$

Assume each word is generated independently.

What if we have a query “presidential campaign update”?

UNIVERSITY of DE

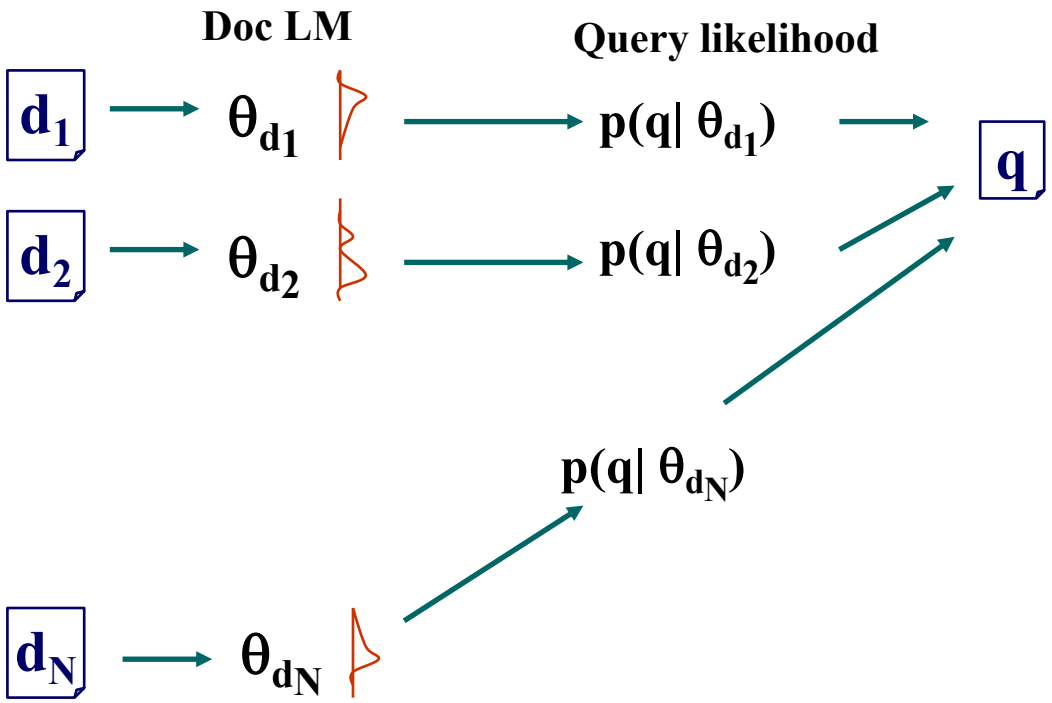
Sampling Words from a Document Model



Computing the probability of a query given a document using a document language model

UNIVERSITY of DE

Ranking Documents based on Query Likelihood



UNIVERSITY of DE

Retrieval as Language Model Estimation

- Rank documents based on *query likelihood*

$$q = w_1, w_2, \ldots, w_n$$

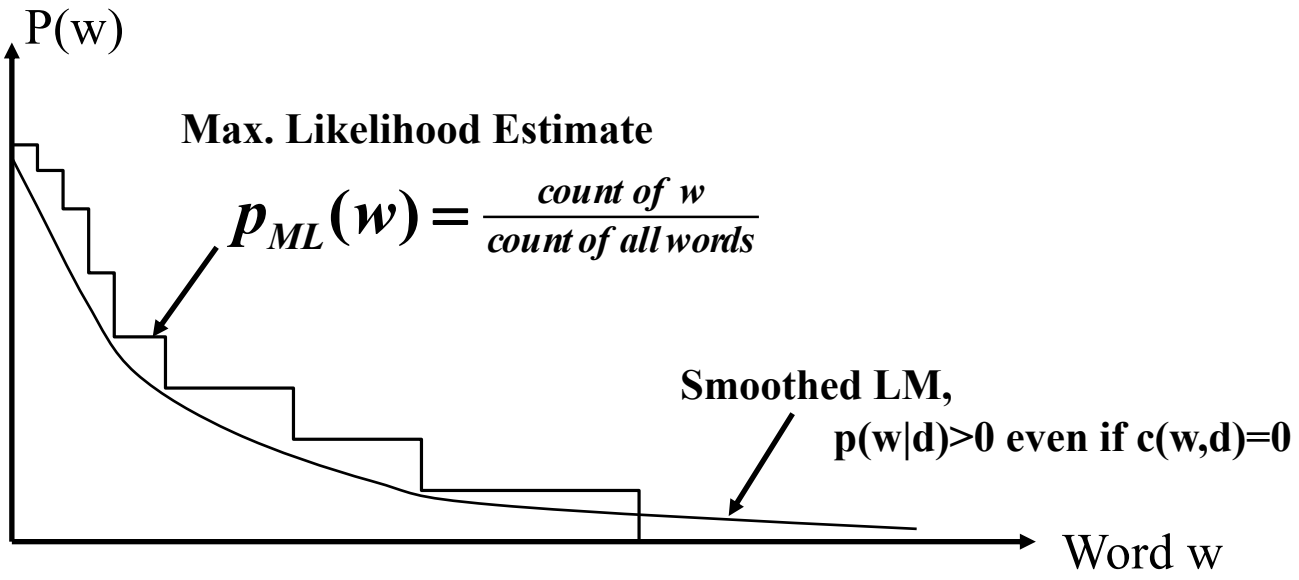
$$p(q \mid d) = p(w_1 \mid d) \times p(w_2 \mid d) \times \ldots \times p(w_n \mid d).$$

$$\text{score}(q, d) = \log p(q \mid d) = \sum_{i=1}^{n} \log p(w_i \mid d) = \sum_{w \in V} c(w, q) \log p(w \mid d)$$

- Retrieval problem ≈ Estimation of *p(w|d)*

UNIVERSITY of DE

How to Estimate p(w|d)?



How to smooth a LM

- Key question of smoothing
  - What probability should be assigned to an unseen word?

- Let the probability of an unseen word be proportionally to its probability given by a reference LM.

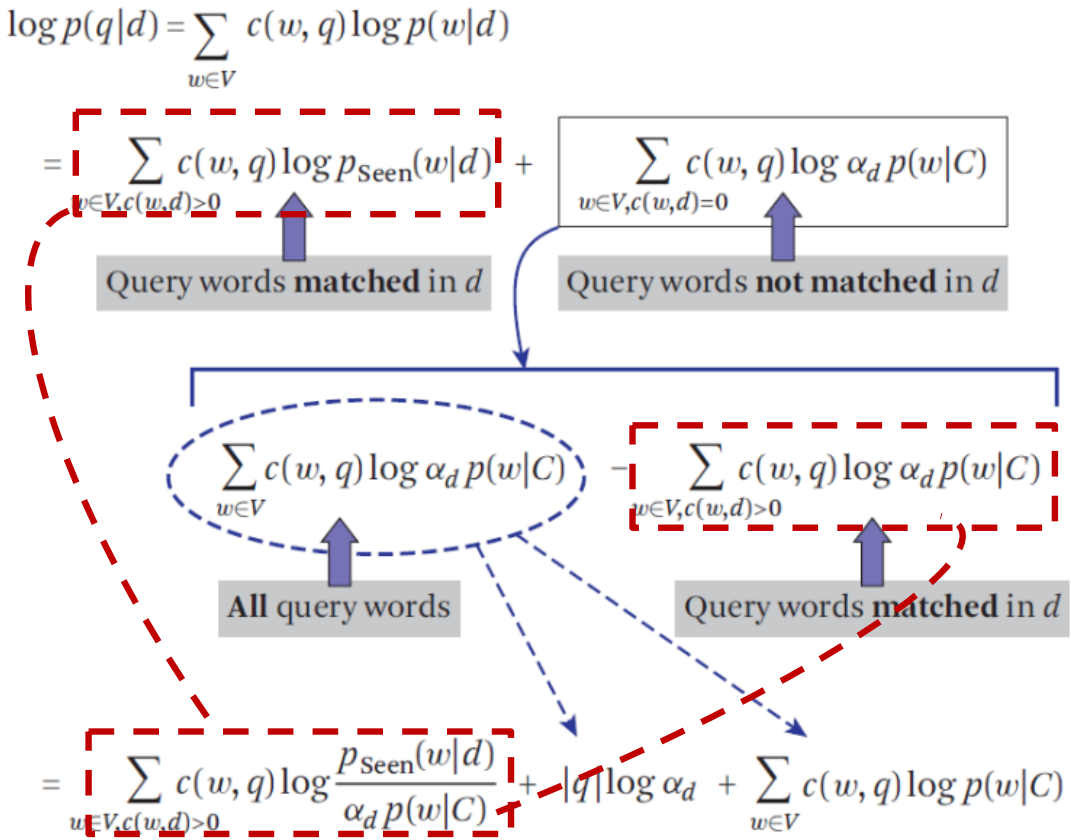
$$p(w \mid d) = \begin{cases} p_{seen}(w \mid d) & \text{if } w \text{ is seen in } d \\ \alpha_d p(w \mid C) & \text{otherwise} \end{cases}$$

Discounted ML estimate

Collection language model

UNIVERSITY of DE

Rewriting the Ranking Function with Smoothing



UNIVERSITY of DE

Benefit of Rewriting

- Efficient computation
- Better understanding of the ranking function
  - Smoothing with $p(w|C)$ ≈ TF-IDF + length norm.

TF weighting

Doc length normalization (long doc is expected to have a smaller α_d)

$$\log p(q \mid d) = \sum_{\substack{w_i \in d \\ w_i \in q}} [\log \frac{p_{seen}(w_i \mid d)}{\alpha_d \, p(w_i \mid C)}] + |q| \log \alpha_d + \boxed{\sum_i \log p(w_i \mid C)}$$

IDF weighting

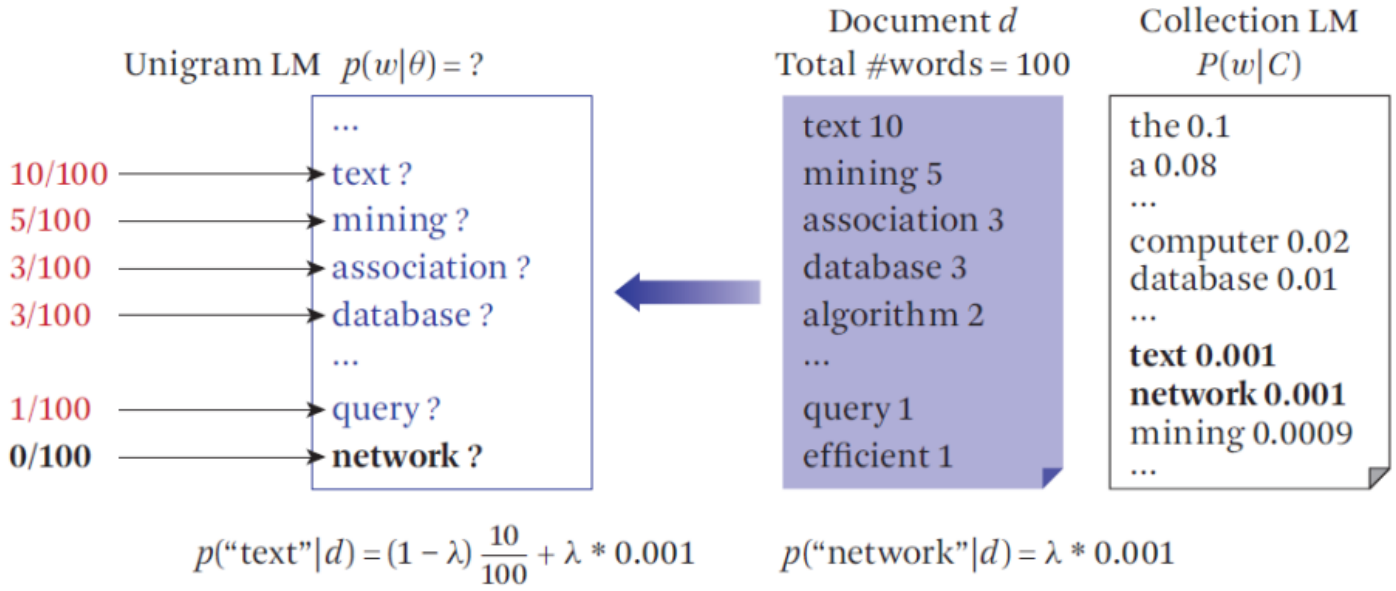
Ignore for ranking

UNIVERSITY of DE

Specific Smoothing Methods (1)

- Jelinek-Mercer (JM) smoothing:

$$p(w|d) = (1-\lambda)\frac{c(w,d)}{|d|} + \lambda p(w|C) \qquad \lambda \in [0,1]$$



$$p(\text{“text”}|d) = (1-\lambda)\frac{10}{100} + \lambda * 0.001 \qquad p(\text{“network”}|d) = \lambda * 0.001$$

Retrieval Function based on JM smoothing

$$\sum_{w \in d,q} c(w,q) \log \left(\frac{p_{\text{seen}}(w \mid d)}{\alpha_d \cdot p(w \mid C)} \right) + |q| \log \alpha_d .$$

- **Jelinek-Mercer (JM) smoothing:**

$$p(w \mid d) = (1-\lambda) p_{ml}(w \mid d) + \lambda p(w \mid C)$$

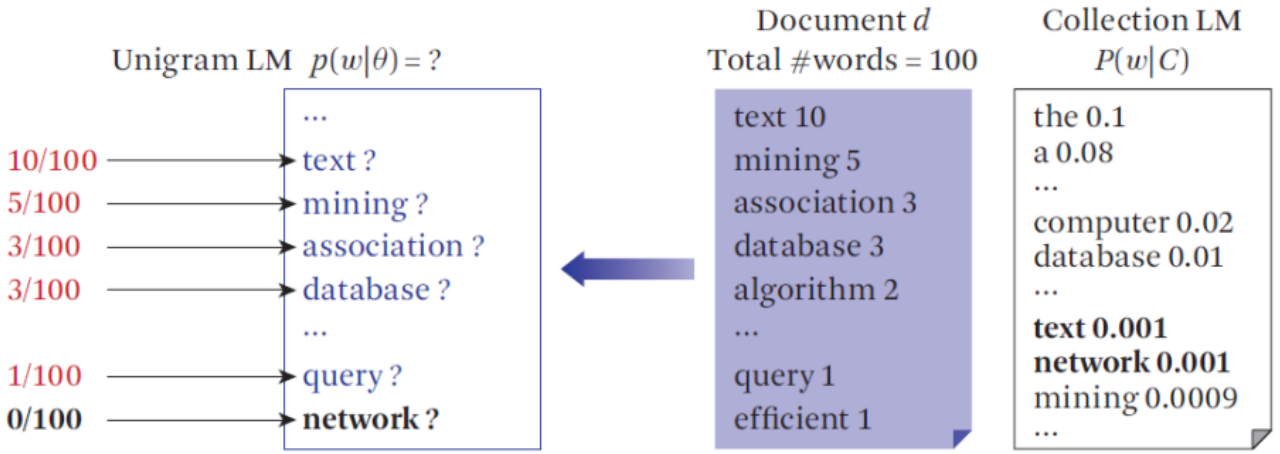
$$\frac{p_{\text{seen}}(w \mid d)}{\alpha_d \cdot p(w \mid C)} = \frac{(1-\lambda) \cdot p_{MLE}(w \mid d) + \lambda \cdot p(w \mid C)}{\lambda \cdot p(w \mid C)} = 1 + \frac{1-\lambda}{\lambda} \cdot \frac{c(w,d)}{|d| \cdot p(w \mid C)}.$$

$$\text{score}_{JM}(q,d) = \sum_{w \in q,d} c(w,q) \log \left(1 + \frac{1-\lambda}{\lambda} \cdot \frac{c(w,d)}{|d| \cdot p(w \mid C)} \right).$$

Specific Smoothing Methods (2)

- Dirichlet Prior smoothing:

$$p(w|d) = \frac{c(w,d) + \mu p(w|C)}{|d| + \mu} = \frac{|d|}{|d| + \mu}\frac{c(w,d)}{|d|} + \frac{\mu}{|d| + \mu}p(w|C) \qquad \mu \in [0, +\infty)$$



$$p(\text{“text”}|d) = \frac{10 + \mu * 0.001}{100 + \mu} \qquad p(\text{“network”}|d) = \frac{\mu}{100 + \mu} * 0.001$$

Dirichlet Prior Function

$$\sum_{w \in d,q} c(w,q) \log \left(\frac{p_{\text{seen}}(w \mid d)}{\alpha_d \cdot p(w \mid C)} \right) + |q| \log \alpha_d.$$

- **Dirichlet Prior Smoothing**

$$p(w \mid d) = \frac{c(w;d) + \mu\, p(w|C)}{|d| + \mu} = \frac{|d|}{|d| + \mu} p_{ml}(w \mid d) + \frac{\mu}{|d| + \mu} p(w \mid C)$$

$$\alpha_d = \frac{\mu}{|d| + \mu}:$$

$$\text{score}_{DIR}(q,d) = \sum_{w \in q,d} c(w,q) \log \left(1 + \frac{c(w,d)}{\mu \cdot p(w \mid C)} \right) + |q| \log \frac{\mu}{\mu + |d|}.$$

UNIVERSITY of DE

- More Info about Course Projects
  - March 13 and 15 (W, F)
    - No lectures
    - Project meeting between the instructor and each group
    - Working on the assignment 1
  - March 25 and 27 (M, W)
    - Oral presentation for the project proposal
  - March 29 (F)
    - Written proposals due