Language Models

- Consider the DIR model (query likelihod with Dirichlet smoothing slide 36 "Probabilistic IR")

$$P(q|\theta_d) = \frac{|q|!}{\prod_{t \in V}\left(\left|w_t^q\right|!\right)} \prod_{t \in V} p(t|\theta_d)^{w_t^q} \propto \prod_{t \in V} p(t|\theta_d)^{w_t^q}$$

- When using logs:

$$P(q|\theta_d) \propto \ln\left(\prod_{t \in V} p(t|\theta_d)^{w_t^q}\right)$$

$$\propto \sum_{t \in V} \ln\left(p(t|\theta_d)^{w_t^q}\right)$$

$$\propto \sum_{t \in V} w_t^q . \ln(p(t|\theta_d))$$

Dirichlet Smoothing Language (DIR) Model

- Theoretical formula, with $\hat{\theta}_d$ the Dirichlet smoothed document language model ($\mu \in R^+$)

$$P(q \mid \hat{\theta}_d) \propto \sum_{w \in V} c(w,q).\ln\left(\frac{c(w,d) + \mu.P(w \mid C)}{|d| + \mu}\right)$$

  - Good from a theoretical point of view, but not from an efficiency point of view as, as not directly implemtable using inverted files

- Actual implementation using inverted files :

$$f(d,q) = \sum_{w \in d \cap q} c(w,q).\ln\left(1 + \frac{c(w,d)}{\mu.P(w \mid C)}\right) + |q|.\ln\left(\frac{\mu}{\mu + |d|}\right)$$

… Not obvious… explanations follow…

DIR model

- If $\hat{\theta}_d$ is the Dirichlet smoothed document language model, $\mu \in R^+$

$$\ln P(q|\hat{\theta}_d) =_{rank} \sum_{w \in V} c(w,q).\ln(\frac{c(w,d)+\mu.P(w|C)}{|d|+\mu})$$

$$=_{rank} \sum_{w \in d \cap q} c(w,q).\ln(\frac{c(w,d)+\mu.P(w|C)}{|d|+\mu}) + \sum_{w \in q, w \notin d} c(w,q).\ln(\frac{\mu.P(w|C)}{|d|+\mu})$$

$$=_{rank} \sum_{w \in d \cap q} c(w,q).\ln(\frac{c(w,d)+\mu.P(w|C)}{|d|+\mu}) + \sum_{w \in q, w \notin d} c(w,q).\ln(\frac{\mu.P(w|C)}{|d|+\mu})$$

$$+ \sum_{w \in q} c(w,q).\ln(\frac{\mu.P(w|C)}{|d|+\mu}) - \sum_{w \in q} c(w,q).\ln(\frac{\mu.P(w|C)}{|d|+\mu})$$

DIR model

$$\ln P(q|\hat{\theta}_d) =_{rank} \sum_{w \in d \cap q} c(w,q).\ln\left(\frac{c(w,d)+\mu.P(w|C)}{|d|+\mu}\right) - \sum_{w \in d \cap q} c(w,q).\ln\left(\frac{\mu.P(w|C)}{|d|+\mu}\right)$$

$$+ \sum_{w \in q} c(w,q).\ln\left(\frac{\mu.P(w|C)}{|d|+\mu}\right)$$

$$=_{rank} \sum_{w \in d \cap q} c(w,q).\ln\left(1+\frac{c(w,d)}{\mu.P(w|C)}\right) + \sum_{w \in q} c(w,q).\ln\left(\frac{\mu}{|d|+\mu}\right)$$

$$+ \sum_{w \in q} c(w,q).\cancel{\log P(w|C)}$$

$$=_{rank} \sum_{w \in d \cap q} c(w,q).\ln\left(1+\frac{c(w,d)}{\mu.P(t|C)}\right) + |q|.\ln\left(\frac{\mu}{|d|+\mu}\right)$$