How to smooth a LM

- Key Question: what probability should be assigned to an unseen word?
- Let the probability of an unseen word be proportional to its probability given by a reference LM
- One possibility: Reference LM = Collection 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