1. a. The complete function for Dirichlet prior smoothing is:

$$P_{seen}(w|d) = \frac{c(w,d) + \mu \cdot p(w|c)}{|d| + \mu}$$

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

$$= \frac{|d|}{|d|+\mu} \cdot P_{MLE}(w|d) + \frac{\mu}{|d|+\mu} p(w|c)$$

Let,

$\frac{\mu}{|d|+\mu} = \alpha_d$ = controls the amount of probability mass that we assign to reference corpus

(i)

If, $|d| \to \alpha$ then $\alpha_d = 0$

That means we assign 0 weights on reference corpus and all weight(1) to MLE.

So, $P_{seen}(w|d) = \frac{|d|}{|d|+\mu} \cdot P_{MLE}(w|d)$

$= 1. P_{MLE}(w|d) = P_{MLE}(w|d)$

ii) If $\mu \to \infty$, then $\frac{|d|}{|d|+\mu} \to 0$ and $\frac{\mu}{|d|+\mu} \to 1$

So,

$P_{seen}(w|d) = P(w|c)$

= probability of the word in the collection

6Q. The complete function for Jalinek-Mercer smoothing is:

$P_{seen}(w|d) = (1-\lambda) \cdot P_{MLE}(w|d) + \lambda \cdot P(w|c)$

where, λ is a smoothing parameter and $\lambda \in [0,1]$

$P_{MLE}(w|d)$ = MLE estimate of the word in the document

$P(w|c)$ = Probability of the word in the collection

Scanned by CamScanner

The complete function for Katz - Backoff smoothing is:

$$\hat{P}(w_i \mid w_{i-2} w_{i-1}) = \begin{cases} P(w_i \mid w_{i-2} w_{i-1}) & \text{if } C(w_{i-2} w_{i-1} w_i) > 0 \\ \alpha_1 P(w_i \mid w_{i-1}) & \text{if } C(w_{i-2} w_{i-1} w_i) = 0 \\ & \text{and } C(w_{i-1} w_i) > 0 \\ \alpha_2 P(w_i) & \text{otherwise} \end{cases}$$

Katz back-off is a generative smoothing technique for n-gram language model. The main idea is:

- If there are no examples of a particular trigram, $w_{n-2} w_{n-1} w_n$, to compute $P(w_n \mid w_{n-2} w_{n-1})$, we can estimate its probability by using the bigram probability $P(w_n \mid w_{n-1})$

- If there are no examples of the bigram to compute $P(w_n \mid w_{n-1})$, we can me the unigram probability $P(w_n)$

<u>Problems with Katz-Back off</u>

Probability estimates can change suddenly on adding more data when the back-off algorithm selects a different order of n-gram model on which to base the estimate.

For example,

We want to compute $P(c \mid a\ b)$ but

$C(\text{"a b c"}) = 0$ so the method will back off to the bigram and estimate $P(c \mid b)$, which may be too high.

But this doesn't happen in Jelinek-Mercer smoothing as a certain amount of weight is assigned to reference corpus if the count is zero in document.

Scanned by CamScanner

Document: d and |d| = 11

Word	Count	$P_{MLE}(w\|d)$
the	3	3/11
sun	1	1/11
rises	1	1/11
in	2	2/11
east	1	1/11
and	1	1/11
sets	1	1/11
west	1	1/11
a	0	0
from	0	0
retrieval	0	0
BM25	0	0

a. Unigram LM with dirichlet smoothing and μ = 4:

$$P(w|d) = \frac{c(w,d)+\mu\,.p(w|c)}{|d|+\mu}$$

$$P(the|d) = \frac{3+4*0.17}{11+4} = 0.245$$

$$P(sun|d) = \frac{1+4*0.05}{11+4} = 0.08$$

$$P(rises|d) = \frac{1+4*0.04}{11+4} = 0.077$$

$$P(in|d) = \frac{2+4*0.16}{11+4} = 0.176$$

$$P(east|d) = \frac{1+4*0.02}{11+4} = 0.072$$

$$P(and|d) = \frac{1+4*0.16}{11+4} = 0.10933333333333333$$

$$P(sets|d) = \frac{1+4*0.04}{11+4} = 0.07733333333333333$$

$$P(west|d) = \frac{1+4*0.02}{11+4} = 0.072$$

P(a|d) = $\frac{0+4*0.18}{11+4}$ = 0.048

P(from|d) = $\frac{0+4*0.13}{11+4}$ = 0.0346

P(retrieval|d) = $\frac{0+4*0.02}{11+4}$ = 0.0053

P(BM25|d) = $\frac{0+4*0.01}{11+4}$ = 0.002667

b. Unigram LM with dirichlet smoothing and μ = 0.01:

P(w|d) $= \frac{c(w,d)+\mu \, .p(w|c)}{|d|+\mu}$

P(the|d) = $\frac{3+0.01*0.17}{11+0.01}$ = 0.2726

P(sun|d) = $\frac{1+0.01*0.05}{11+0.01}$ = 0.09087

P(rises|d) = $\frac{1+4*0.04}{11+0.01}$ = 0.1053

P(in|d) = $\frac{2+0.01*0.16}{11+0.01}$ = 0.1818

P(east|d) = $\frac{1+0.01*0.02}{11+0.01}$ = 0.0908

P(and|d) = $\frac{1+0.01*0.16}{11+0.01}$ = 0.09097

P(sets|d) = $\frac{1+0.01*0.04}{11+0.01}$ = 0.09086

P(west|d) = $\frac{1+0.01*0.02}{11+0.01}$ = 0.0908

P(a|d) = $\frac{0+0.01*0.18}{11+0.01}$ = 0.000163

P(from|d) = $\frac{0+0.01*0.13}{11+0.01}$ = 0.000118

P(retrieval|d) = $\frac{0+0.01*0.02}{11+0.01}$ = 0.0000181

$P(BM25|d) = \frac{0+0.01*0.01}{11+0.01} = 0.00000908$

a. Unigram LM with dirichlet smoothing and μ = 100:

$$P(w|d) = \frac{c(w,d)+\mu\,.p(w|c)}{|d|+\,\mu}$$

$P(the|d) = \frac{3+100*0.17}{11+100} = 0.18018$

$P(sun|d) = \frac{1+100*0.05}{11+100} = 0.054054$

$P(rises|d) = \frac{1+100*0.04}{11+100} = 0.045045$

$P(in|d) = \frac{2+100*0.16}{11+100} = 0.162162$

$P(east|d) = \frac{1+100*0.02}{11+100} = 0.027027$

$P(and|d) = \frac{1+100*0.16}{11+100} = 0.153153$

$P(sets|d) = \frac{1+100*0.04}{11+100} = 0.045045$

$P(west|d) = \frac{1+100*0.02}{11+100} = 0.027027$

$P(a|d) = \frac{0+100*0.18}{11+100} = 0.162162$

$P(from|d) = \frac{0+100*0.13}{11+100} = 0.117117$

$P(retrieval|d) = \frac{0+100*0.02}{11+100} = 0.018018$

$P(BM25|d) = \frac{0+100*0.01}{11+100} = 0.009$

Comparison:

P(w|d)	μ = 0.01	μ = 4	μ = 100
P(the|d)	0.2726	0.245	0.18018
P(sun|d)	0.09087	0.08	0.054054
P(rises|d)	0.1053	0.077	0.045045
P(in|d)	0.1818	0.176	0.162162
P(east|d)	0.0908	0.072	0.027027
P(and|d)	0.09097	0.109333333333333	0.153153
P(sets|d)	0.09086	0.0773333333333333	0.045045
P(west|d)	0.0908	0.072	0.027027
P(a|d)	0.000163	0.048	0.162162
P(from|d)	0.000118	0.0346	0.117117
P(retrieval|d)	0.0000181	0.0053	0.018018
P(BM25|d)	0.00000908	0.002667	0.009

From the comparison, we can see that when μ is higher P(w|d) is close to the probability of the word in the collection. On the other hand, when μ is lower P(w|d) is close to the MLE of the word from the document. That matches with our intuition.

c. Unigram LM with Jelinek Mercer smoothing and λ = 0.01:

P(w|d) = $(1-\lambda)\frac{c(w,d)}{|d|}+\lambda p(w|c)$

P(the|d) = (1-0.01) $\frac{3}{11}+0.01*$ 0.17 = 0.2717

P(sun|d) = (1-0.01) $\frac{1}{11} + 0.01 * 0.05$ = 0.0905

P(rises|d) = (1-0.01) $\frac{1}{11} + 0.01 * 0.04$ = 0.0904

P(in|d) = (1-0.01) $\frac{2}{11} + 0.01 * 0.16$ = 0.1816

P(east|d) = (1-0.01) $\frac{1}{11} + 0.01 * 0.02$ = 0.0902

P(and|d) = (1-0.01) $\frac{1}{11} + 0.01 * 0.16$ = 0.0916

P(sets|d) = (1-0.01) $\frac{1}{11} + 0.01 * 0.04$ = 0.0904

P(west|d) = (1-0.01) $\frac{1}{11} + 0.01 * 0.02$ = 0.0902

P(a|d) = (1-0.01) $\frac{0}{11} + 0.01 * 0.18$ = 0.0018

P(from|d) = (1-0.01) $\frac{0}{11} + 0.01 * 0.13$ = 0.0013

P(retrieval|d) = (1-0.01) $\frac{0}{11} + 0.01 * 0.02$ = 0.0002

P(BM25|d) = (1-0.01) $\frac{0}{11} + 0.01 * 0.01$ = 0.0001

Unigram LM with Jelinek Mercer smoothing and λ = 0.5:

P(w|d) = $(1 - \lambda)\frac{c(w,d)}{|d|} + \lambda p(w|c)$

P(the|d) = (1-0.5) $\frac{3}{11} + 0.5 * 0.17$ = 0.2213636363636364

P(sun|d) = (1-0.5) $\frac{1}{11} + 0.5 * 0.05$ = 0.07045

P(rises|d) = (1-0.5) $\frac{1}{11} + 0.5 * 0.04$ = 0.0654545454545455

P(in|d) = (1-0.5) $\frac{2}{11} + 0.5 * 0.16$ = 0.170909

P(east|d) = (1-0.5) $\frac{1}{11} + 0.5 * 0.02$ = 0.0554545

P(and|d) = (1-0.5) $\frac{1}{11} + 0.5 * 0.16$ = 0.12545

P(sets|d) = (1-0.5) $\frac{1}{11} + 0.5 * 0.04$ = 0.06545

P(west|d) = (1-0.5) $\frac{1}{11} + 0.5 * 0.02$ = 0.0554545

P(a|d) = (1-0.5) $\frac{0}{11} + 0.5 * 0.18$ = 0.09

P(from|d) = (1-0.5) $\frac{0}{11} + 0.5 * 0.13$ = 0.065

P(retrieval|d) = (1-0.5) $\frac{0}{11} + 0.5 * 0.02$ = 0.01

P(BM25|d) = (1-0.5) $\frac{0}{11} + 0.5 * 0.01$ = 0.005

Unigram LM with Jelinek Mercer smoothing and λ = 0.9:

P(w|d) = $(1 - \lambda)\frac{c(w,d)}{|d|} + \lambda p(w|c)$

P(the|d) = (1-0.9) $\frac{3}{11} + 0.9 * 0.17$ = 0.1802727

P(sun|d) = (1-0.9) $\frac{1}{11} + 0.9 * 0.05$ = 0.05409

P(rises|d) = (1-0.9) $\frac{1}{11} + 0.9 * 0.04$ = 0.04509

P(in|d) = (1-0.9) $\frac{2}{11} + 0.9 * 0.16$ = 0.1621818

P(east|d) = (1-0.9) $\frac{1}{11} + 0.9 * 0.02$ = 0.02709

P(and|d) = (1-0.9) $\frac{1}{11} + 0.9 * 0.16$ = 0.1530909

P(sets|d) = (1-0.9) $\frac{1}{11} + 0.9 * 0.04$ = 0.0450909

P(west|d) = (1-0.9) $\frac{1}{11} + 0.9 * 0.02$ = 0.02709

P(a|d) = (1-0.9) $\frac{0}{11} + 0.9 * 0.18$ = 0.162

P(from|d) = (1-0.9) $\frac{0}{11} + 0.9 * 0.13$ = 0.117

P(retrieval|d) = (1-0.9) $\frac{0}{11} + 0.9 * 0.02$ = 0.018

P(BM25|d) = (1-0.9) $\frac{0}{11} + 0.9 * 0.01$ = 0.009

Comparison:

P(w|d)	λ = 0.01	λ = 0.5	λ = 0.9
P(the|d)	0.2717	0.2213	0.1802727
P(sun|d)	0.0905	0.07045	0.05409
P(rises|d)	0.0904	0.065454	0.04509
P(in|d)	0.1816	0.170909	0.1621818
P(east|d)	0.0902	0.0554545	0.02709
P(and|d)	0.0916	0.12545	0.1530909
P(sets|d)	0.0904	0.06545	0.0450909
P(west|d)	0.0902	0.0554545	0.02709
P(a|d)	0.0018	0.09	0.162
P(from|d)	0.0013	0.065	0.117
P(retrieval|d)	0.0002	0.01	0.018
P(BM25|d)	0.0001	0.005	0.009

From the comparison, we can see that when λ is higher P(w|d) is close to the probability of the word in the collection. On the other hand, when λ is lower P(w|d) is close to the MLE of the word from the document. That matches with our intuition. Also, both λ and μ resolves the sparsity problem where λ $\in [0, 1]$ and μ $\in [0, \infty]$

3.a.

$$\text{score}(Q, D) = \frac{P(R=1 \mid Q, D)}{P(R=0 \mid Q, D)}$$

$$\approx \frac{P(D \mid Q, R=1)}{P(D \mid Q, R=0)}$$

$$= \prod_{i=1}^{k} \frac{P(w_i \mid Q, R=1)}{P(w_i \mid Q, R=0)}$$

Here, we assume a ~~query~~ Document D contains the words, such that $|D| = k$

$$D = w_1, w_2, \ldots, w_k$$

So, $$\text{score}(Q, D) = \prod_{i=1}^{k} \frac{P(w_i \mid Q, R=1)}{P(w_i \mid Q, R=0)}$$

$$= \sum_{i=1}^{k} \log \frac{P(w_i \mid Q, R=1)}{P(w_i \mid Q, R=0)}$$

$$= \sum_{w \in V} c(w, D) \log \frac{P(w \mid Q, R=1)}{P(w \mid Q, R=0)}$$

In the score function, we have a sum over all the possible words in the vocabulary V. and iterate through each word in the Document.

Essentially, we are only considering the word the documents because the word is not in the document, if a word is not in the document, its contribution to the sum would be zero.

Counts of parameter:

$c(w, D) \longrightarrow |V|$ times

$P(w \mid Q, R=1) \longrightarrow |V|$ times

$P(w \mid Q, R=0) \longrightarrow |V|$ times

$3|V|$ times total

(b) $$P_{MLE}(w \mid Q, R=0) = \frac{c(w, Q, R=0)}{c(Q, R=0)}$$

(c) $$P_{MLE}(w \mid Q, R=1) = \frac{c(w, Q, R=1)}{c(Q, R=1)}$$

(d)

$$P(w \mid Q, R=1) = \begin{cases} (1-\lambda)\, P_{MLE}(w \mid Q, R=1) & \text{if word is seen in relevant query} \\ \lambda\, P(w \mid C) & \text{otherwise} \end{cases}$$

Here, $P(w \mid C)$ is the probability of the word in the collection and λ is a smoothing parameter and

$\lambda \in [0, 1]$

So, the Jelinek-Mercer smoothing function will be:

$$P(w \mid Q, R=1) = (1-\lambda) P_{MLE}(w \mid Q, R=1) + \lambda P(w \mid C)$$

$$= (1-\lambda) \frac{c(w, Q, R=1)}{c(Q, R=1)} + \lambda P(w \mid C)$$

(e) $$\text{score}(Q,D) = \sum_{w \in V} c(w,D) \log \frac{P(w|Q,R=1)}{P(w|Q,R=0)}$$

$$= \sum_{w \in V} c(w,D) \log \frac{(1-\lambda) P_{MLE}(w|Q,R=1) + \lambda P(w|C)}{\lambda P(w|C)}$$

$$= \sum_{w \in V} c(w,D) \log \left[1 + \frac{1-\lambda}{\lambda} \cdot \frac{P_{MLE}(w|Q,R=1)}{P(w|C)} \right]$$

The value of the logarithm term is non-negative. We see very clearly the TF weighting in the numerator IDF weighting, which is $P(w|C)$ term in the denominator