Assignment 5

1. a) Raw term frequency gives too much reward to documents with high term frequency and document frequency. This is why it might be better to use log(TF) and also account for inverse document frequency. This ensures that rare terms get a higher reward than common words.

b.) Adding another instance of d to the corpus will reduce the IDF for all words in d and increase the IDF for all words not in d.

c.

i.

	precision	recall
+	1/1	1/16
+	2/3	2/16
−	0	0
+	3/4	3/16
+	4/5	4/16
−	0	0
−	0	0
+	5/8	5/16
−	0	0
−	0	0

precision = $\frac{5}{10} = .5$

recall = $\frac{5}{16}$

average precision = $\frac{1 + 1 + 3/4 + 4/5 + 5/8}{10} = .3781$

F1 score = $\frac{2\left(\frac{5}{10}\right)\left(\frac{5}{16}\right)}{\left(\frac{5}{10}\right) + \left(\frac{5}{16}\right)} = .385$

a.)

		Cumulative Gain	Discounted Cumulative Gain
1	+	1	1
2	+	2	$1 + 1/\log(2)$
3	−	2	$1 + 1/\log(2) + 0$
4	+	3	$1 + 1/\log(4) + 1/\log(5)$
5	+	4	$1 + 1/\log(4) + 1/\log(5) + 1/\log(6)$
6	−	4	
7	−	4	$1 + 1/\log(2) + 1/\log(4) + 1/\log(5) = 2.93$
	+		
	−		
	−		

Cumulative Gain at 7 = 4

$$NDCG \text{ at } 7 = \frac{2.93}{1 + 1/\log(2) + 1/\log(3) + \ldots + 1/\log(7)} = .68$$

2.) a.)

$$\sum_{w \in D} c(w, Q) \log \frac{P_s(w|\theta_D)}{\alpha_D P(w|C)} + |Q| \log \alpha_D$$

Let $\alpha_D = \lambda$ and $P_s(w|\theta_d) = (1-\lambda) P_{ML}(w|\theta_d) + \lambda P(w|C)$

then

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

b/c $|Q| \log \lambda$ does not depend on the document we can remove it

$$= \sum_{w \in D} c(w,q) \log \frac{(1-\lambda) p_{ML}(w|\theta_D) + \lambda p(w|C)}{\lambda p(w|C)}$$

$$= \sum_{w \in D} c(w,q) \log \left(\frac{(1-\lambda) p_{ML}(w|\theta_D)}{\lambda p(w|C)} + \frac{\lambda p(w|C)}{\lambda p(w|C)} \right)$$

$$= \sum_{w \in D} c(w,q) \log \left[\frac{(1-\lambda) p_{ML}(w|\theta_D)}{\lambda p(w|C)} + 1 \right]$$

we only care about words in both doc and query ∴

$$score(Q,D) = \sum_{w \in Q \cap D} c(w,q) \cdot \log \left[1 + \frac{(1-\lambda) p_{ML}(w|\theta_D)}{\lambda p(w|C)} \right]$$

$$p(w|C) = p(w|REF) \text{ and } p_{ML}(w|\theta_D) = \frac{c(w,D)}{|D|}$$

∴

$$score(Q,D) = \sum_{w \in Q \cap D} c(w,q) \cdot \log \left[1 + \frac{(1-\lambda) \cdot c(w,D)}{\lambda p(w|REF)|D|} \right]$$

b) Q is the query vector, D is the document vector. The similarity function is the entire function, TF is captured in $c(w,Q)$ and $c(w,D)$. IDF is captured in $\log \frac{1}{p(w|REF)|D|}$

Document length normalization is $|D|$.

c) $\alpha_D = \frac{\mu}{|D|+\mu}$ $\quad p_s(w|\theta_D) = \frac{c(w,D)+\mu p(w|C)}{|D|+\mu}$

Subbing these values in we get

$$= \sum_{w \in D} c(w,Q) \log \left[\frac{c(w,D)+\mu p(w|C)}{|D|+\mu} \cdot \frac{(|D|+\mu)}{\mu p(w|C)} \right] + |Q| \log \left[\frac{\mu}{|D|+\mu} \right]$$

$$= \sum_{w \in D} c(w,Q) \log \left[\frac{c(w,D)}{\mu p(w|C)} + \frac{\mu p(w|C)}{\mu p(w|C)} \right] + |Q| \log \left[\frac{\mu}{|D|+\mu} \right]$$

$$= \sum_{w \in D} c(w,Q) \log \left[\frac{c(w,D)}{\mu p(w|C)} + 1 \right] + |Q| \log \left[\frac{\mu}{|D|+\mu} \right]$$

$$= \sum_{w \in D} c(w,Q) \log \left[\frac{c(w,D)}{\mu p(w|C)} + 1 \right] + |Q| \log(\mu) - \log(|D|+\mu)$$

$\log(N)$ does not depend on d so we can remove it

$$= \sum_{w \in Q \cap D} c(w,Q) \log\left[1 + \frac{c(w,D)}{N\, p(w|C)}\right] - |Q| \log(|D|+N)$$

d. Q is the query vector D is the document vector, $c(w,Q)$ and $c(w,D)$ are term frequency. $-|Q|\log(|D|+N)$ is document length normalization. IDF is $p(w|C)$.

e.) Jelinek-Mercer

$$\text{Score}(Q,D) = \sum_{w \in Q \cap D} c(w,Q) \log\left(\frac{1 + (1-\lambda)\cdot c(w,D)}{\lambda \cdot p(w|REF)\cdot |D|}\right)$$

$$\text{Score}(Q,D') = \sum_{w \in Q \cap D} c(w,Q) \log\left(\frac{1 + (1-\lambda)\cdot kc(w,D)}{\lambda \cdot p(w|REF)\, k|D|}\right)$$

k cancels out $\therefore$ $\text{Score}(Q,D) = \text{Score}(Q,D')$

Dirichlet prior:

$$Score(Q,D) = \sum_{w \in Q \cap D} c(w,Q) \log\left[1 + \frac{c(w,D)}{\mu P(w|C)}\right] - |Q| \log\left(\frac{|D|+\mu}{\mu}\right)$$

$$Score(Q,D') = \sum_{w \in Q \cap D} c(w,Q) \log\left[1 + \frac{k\,c(w,D)}{\mu P(w|C)}\right] - |Q| \log\left(\frac{k|D|+\mu}{\mu}\right)$$

$$Score(Q,D) \neq Score(Q,D')$$

Therefore Jelinek-Mercer does not overpenalize for a long document but Dirichlet prior does.