Chen Zhang czhang49

CS 598 CXZ Fall 2016 HW1

① a) Because different words have different weights to the meaning of query. e.g. in a search for "news about presidential campaign", about should be given less weights compared with "presidential".

b) $IDF(w) = \log\left(\frac{M+1}{K}\right)$, K being number of docs containing w, M being total number of docs.

If word w is contained in d, $M' = M+1$, $K' = K+1$, since $M > K \Rightarrow IDF(w)$ will decrease.

if word w is not contained in d, $M' = M+1$, $K' = K$, $\Rightarrow$ IDF(w) will increase

c) i) $\text{precision} = \frac{\text{relavant doc \#}}{\text{retrived total doc \#}} = \frac{5}{10} = \frac{1}{2}$

ii) $\text{Recall} = \frac{\text{relavant doc \#}}{\text{total relevant \#}} = \frac{5}{16}$

iii) $F_1 = \frac{2PR}{P+R} = \frac{2 \times \frac{1}{2} \times \frac{5}{16}}{\frac{5}{16} + \frac{1}{2}} = \frac{\frac{5}{16}}{\frac{13}{16}} = \frac{5}{13}$

iv) $\text{Average Precision} = \frac{\frac{1}{1} + \frac{2}{2} + \frac{3}{4} + \frac{4}{5} + \frac{5}{8}}{10} = \frac{\frac{167}{40}}{10} = \frac{167}{400}$