Solutions

1.4

a. O(x+y)

→ Just add all documents containing Brutus and skip a document if you see it in Caesar's list.

→ You just need to change the **if** condition around.

b. O(n)

→ All documents which don't have Caesar.

→ Loop through 1 to N and skip if it's in postings[Caesar].

1.7

1) Tangerine OR Trees = 46653 + 316812 = 363465
2) marmalade OR skies = 107913 + 271658 = 379571
3) kaleidoscope OR eyes = 87009 + 213312 = 300321

Order 3,1,2

Q2.

a) 2,4,7

b) (2,4,7) and (4) = (4)

Q3.

$$idf = \log\left(\frac{N_{total}}{N_{occur}}\right) = log(1) = 0 \; [with\ smoothening\ it\ will\ be\ close\ to\ 0]$$

IF stop words are removed, the dimensionality of the vectors decrease hence saving computational cost

Q4.

≪ Used $\log_{10}$ ≫

Word	tf	wf	df	idf	wf-idf	tf	wf	Norm wf	qi.di
digital	1	1	10,000	3	3	1	1	0.52	1.56
video	0	0	100,000	2	0	1	1	0.52	0
camera	1	1	50,000	2.3	2.3	2	1.3	0.68	1.56
								$\left(\frac{wf}{\sqrt{\Sigma w f_i^2}}\right)$	

Similarity = 3.12

Q5

min_sup = 60 % = 3 documents

1) C E K M Y
   3 4 5 3 3

2) CE – 2, EK – 4, KM – 3, MY – 2, YC – 2
   CK – 3, EM – 2, KY – 3, MC – 2, EY – 2
   EK, CK, KM, KY

3) ECK – 2, EKM – 2, EKY – 2, CKM – 1, CKY – 1
   MKY – 2, EMY – 1, EWY – 1, ECM – 1, KYM – 1
   No new sets

Rules:

$$E \to K \left(conf = \frac{4}{4} = 1\right); K \to E \left(conf = \frac{4}{5} = 0.8\right)$$

$$C \to K \left(conf = \frac{3}{3} = 1\right); K \to C \left(conf = \frac{3}{5} = 0.6\right)$$

$$K \to M \left(conf = \frac{3}{5} = 0.6\right); M \to K \left(conf = \frac{3}{3} = 1\right)$$

$$K \to Y \left(conf = \frac{3}{5} = 0.6\right); Y \to K \left(conf = \frac{3}{3} = 1\right)$$