Big Data Processing: homework 6

凌康伟 5140219295

June 1, 2017

Exercise 1

	query						document			
word	tf	wf	df	idf	wf-idf	q_i	tf	wf	d_i	$q_i \cdot d_i$
digital	1	1	10,000	3	3	0.79	1	1	0.52	0.41
video	0	0	100,000	2	0	0	1	1	0.52	0
cameras	1	1	50,000	2.3	2.3	0.61	2	1.3	0.68	0.41

$$\text{final similarity} = 0.41 + 0.41 = 0.82$$

Exercise 2

nnn.atc

	Query(atc weight)			
Term	tf	idf	tf-idf	atc weight
Car	1	1.65	1.65	0.560
Auto	0.5	2.08	1.04	0.353
Insurance	1	1.62	1.62	0.550
Best	1	1.5	1.5	0.509

	Doc 1(nnn weight)			
Term	tf	idf	tf-idf	nnn weight
Car	27	1	27	27
Auto	3	1	3	3
Insurance	0	1	0	0
Best	14	1	14	14

	Doc 2(nnn weight)			
Term	tf	idf	tf-idf	nnn weight
Car	4	1	4	4
Auto	33	1	33	33
Insurance	33	1	33	33
Best	0	1	0	0

	Doc 3(nnn weight)			
Term	tf	idf	tf-idf	nnn weight
Car	24	1	24	24
Auto	0	1	0	0
Insurance	29	1	29	29
Best	17	1	17	17

$$Sim(Query, Doc1) = 23.310$$
$$Sim(Query, Doc2) = 32.037$$
$$Sim(Query, Doc3) = 38.046$$

Ranking (from high to low): Doc 3, Doc 2, Doc 1

ntc.atc

	Query(atc weight)			
Term	tf	idf	tf-idf	atc weight
Car	1	1.65	1.65	0.560
Auto	0.5	2.08	1.04	0.353
Insurance	1	1.62	1.62	0.550
Best	1	1.5	1.5	0.509

	Doc 1(ntc weight)			
Term	tf	idf	tf-idf	ntc weight
Car	27	1.65	44.55	0.897
Auto	3	2.08	6.24	0.126
Insurance	0	1.62	0	0
Best	14	1.5	21	0.423

	Doc 2(ntc weight)			
Term	tf	idf	tf-idf	ntc weight
Car	4	1.65	6.6	0.076
Auto	33	2.08	68.64	0.787
Insurance	33	1.62	53.46	0.613
Best	0	1.5	0	0

	Doc 3(ntc weight)			
Term	tf	idf	tf-idf	ntc weight
Car	24	1.65	39.6	0.595
Auto	0	2.08	0	0
Insurance	29	1.62	46.98	0.706
Best	17	1.5	25.5	0.383

$$Sim(Query, Doc1) = 0.762$$
$$Sim(Query, Doc2) = 0.657$$
$$Sim(Query, Doc3) = 0.917$$

Ranking (from high to low): Doc 3, Doc 1, Doc 2