

CS172: Information Retrieval

UC Riverside

Handout 1

Fall 2019

Assume the following fragments comprise your document collection:

Doc 1: Interest in real estate speculation

Doc 2: Interest rates and rising home costs

Doc 3: Kids do not have an interest in banking

Doc 4: Lower interest rates, hotter real estate market

Doc 5: Feds interest in raising interest rates rising

Assume the following are stopwords: **an, and, do, in, not**

- (a) Construct the term-document matrix for the above documents that can be used in Boolean retrieval. The index terms have already been arranged for you alphabetically in the following table:

Term	Doc1	Doc2	Doc3	Doc4	Doc5
banking					
costs					
estate					
feds					
have					
home					
hotter					
interest					
kids					
lower					
market					
raising					
rates					
real					
rising					
speculation					

- (b) What documents would be returned in response to the following queries?

- interest NOT rates
- (interest AND rates) NOT (rising OR kids)

(c) Construct the vector space term-document matrix for the above documents using tf.idf term weighting. Normalize your vectors. The following blank tables are provided for your convenience. You can use as many or as few of them as you wish. Clearly indicate your final answer. **For these practice problems, I assume you have access to a calculator!**

Term	IDF	TF Doc1	TF Doc2	TF Doc3	TF Doc4	TF Doc5
banking						
costs						
estate						
feds						
have						
home						
hotter						
interest						
kids						
lower						
market						
raising						
rates						
real						
rising						
speculation						

Term	TF-IDF Doc1	TF-IDF Doc2	TF-IDF Doc3	TF-IDF Doc4	TF-IDF Doc5
banking					
costs					
estate					
feds					
have					
home					
hotter					
interest					
kids					
lower					
market					
raising					
rates					
real					
rising					
speculation					

- (d) What is the Inner Product Similarity for the query "interest rate" and document D1?
(e) What is the Cosine Similarity for the query "interest rate" and document D1?