COMP90042

Workshop Week 10

QA project

- Due
  - 11pm, Sunday 27th May
- Submit your Kaggle team name and list of member names and emails to https://goo.gl/forms/Ge6Chc0RyvbbeEls2

Syllabus

1	Introduction and Preprocessing	Text classification
2	Lexical semantics	Distributional semantics
3	Part of Speech Tagging	Hidden Markov Models
4	Unsupervised Hidden Markov Models	Context-Free Grammars
5	Probabilistic Parsing	Dependency parsing
	Easter holiday break	
6	N-gram language models	Neural language models
7	Information Extraction	Question Answering
8	Topic Models	*ANZAC day holiday*
9	Information Retrieval -- indexing and querying in the vector space model	Index compression and efficient query processing
10	Efficient indexing	Query completion and query expansion
11	IR evaluation and learning to rank	Machine Translation (word based)
12	Machine translation (phrase based) and neural encoder-decoder	Subject review

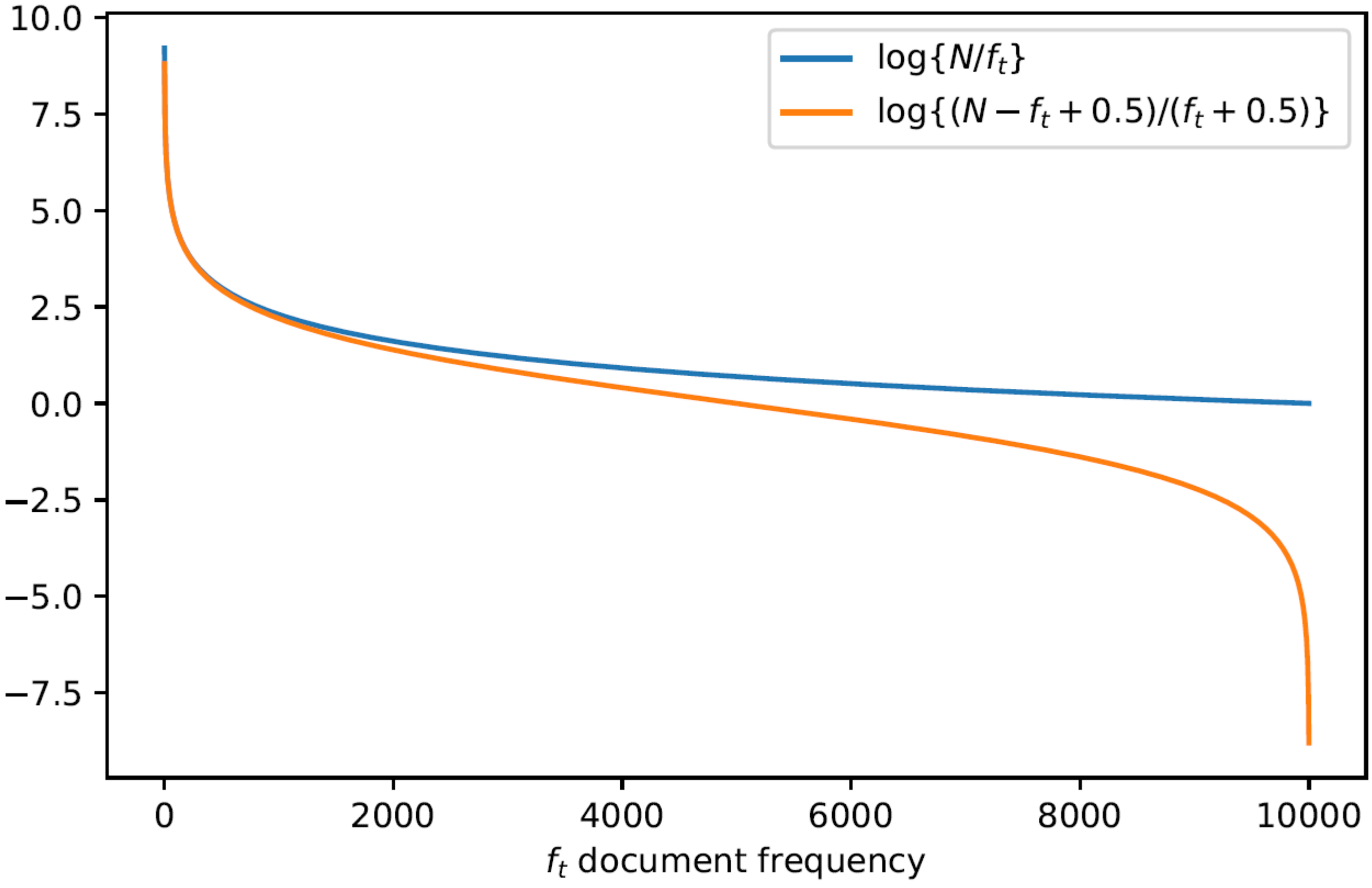
- BM25
- Top-k retrieval

BM25

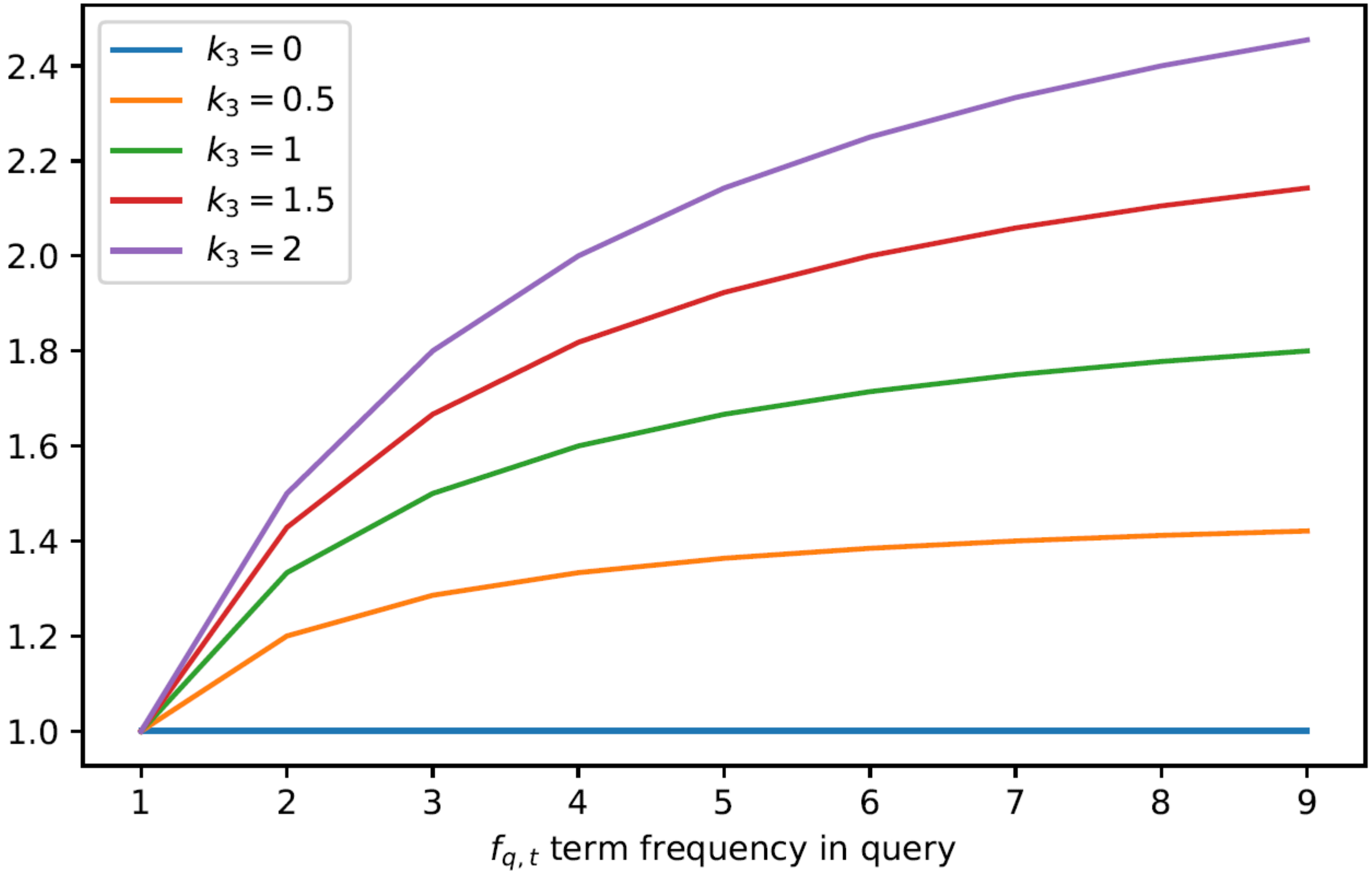
- w_t is

$$\log \frac{N - f_t + 0.5}{f_t + 0.5} \times \frac{(k_1 + 1) f_{d,t}}{k_1 \left((1 - b) + b L_d / L_{avg} \right) + f_{d,t}} \times \frac{(k_3 + 1) f_{q,t}}{k_3 + f_{q,t}}$$

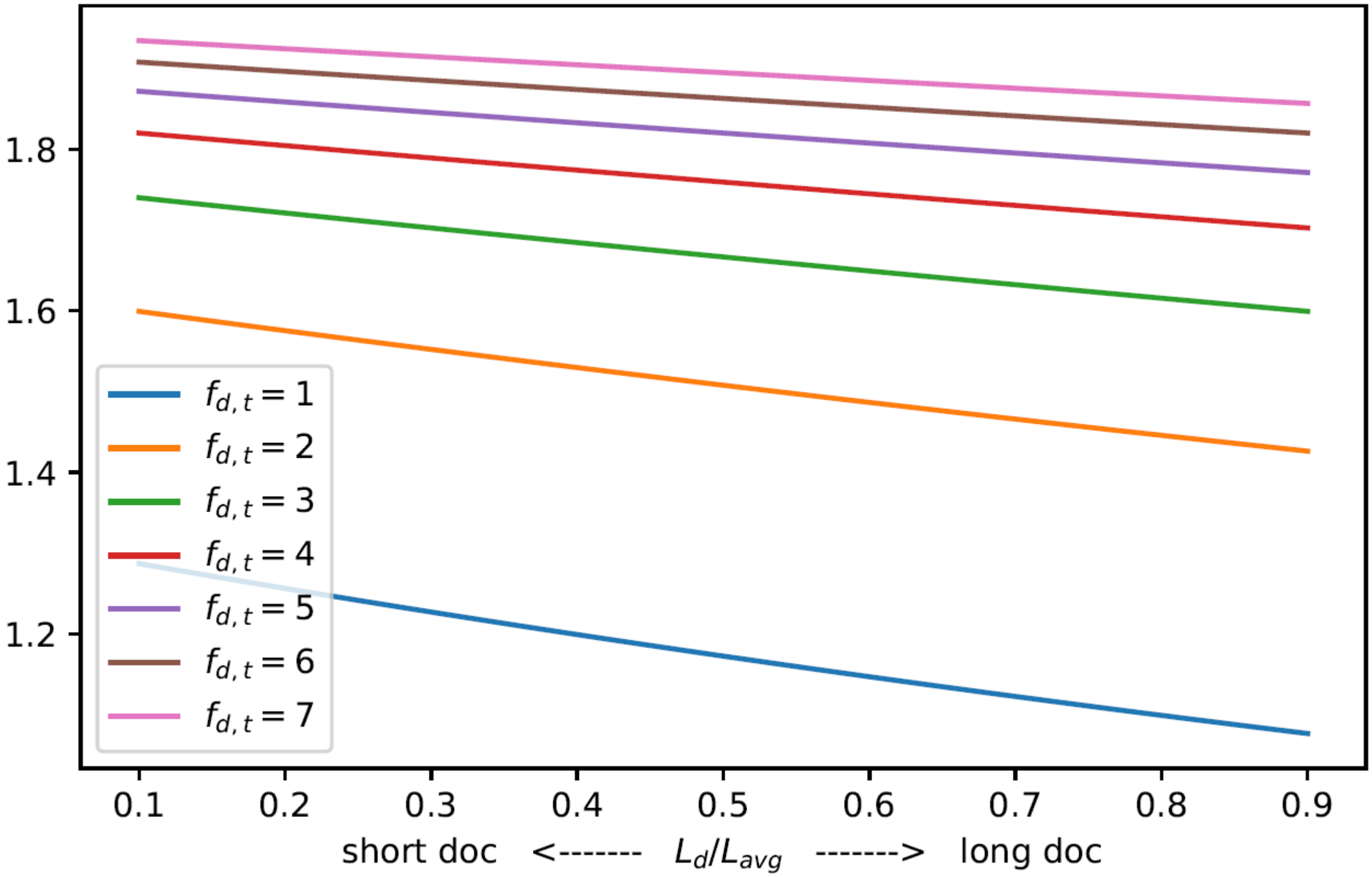
idf term



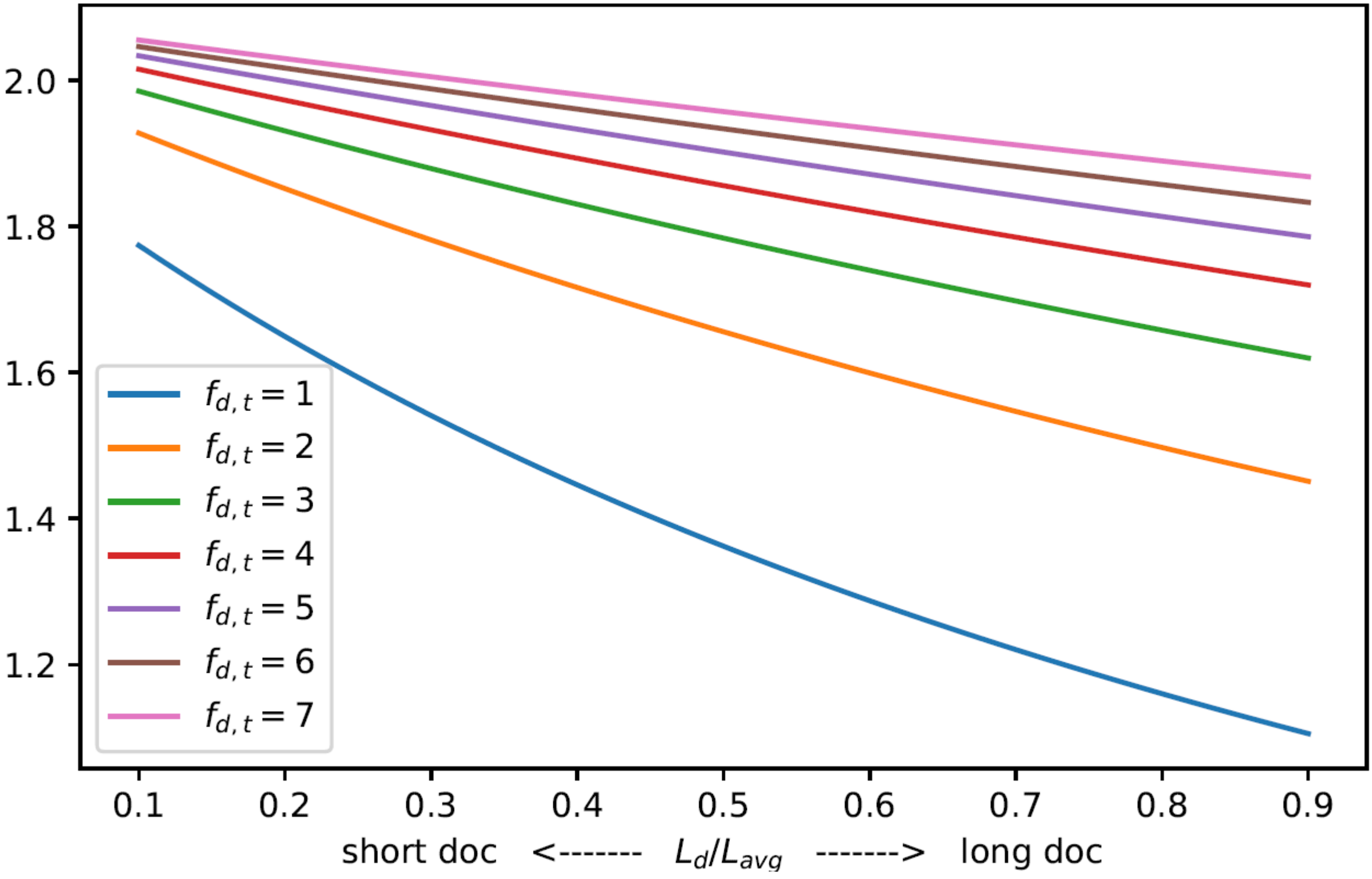
term frequency in query $\frac{(k_3+1)f_{q,t}}{k_3+f_{q,t}}$



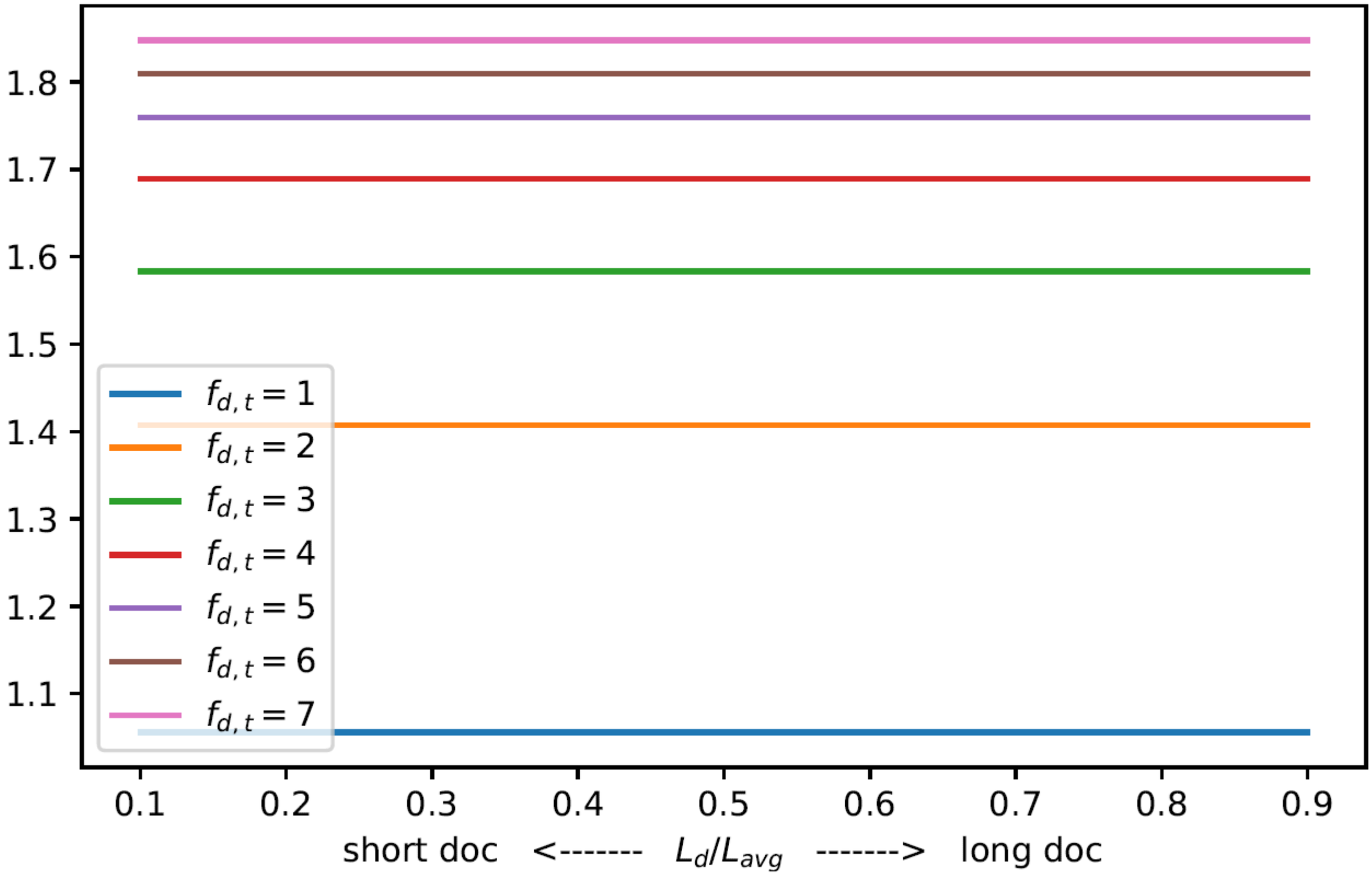
tf in doc $\frac{(k_1+1)f_{d,t}}{k_1\left((1-b)+bL_d/L_{avg}\right)+f_{d,t}}$, $k_1 = 0.9$, $b = 0.4$



tf in doc $\frac{(k_1+1)f_{d,t}}{k_1\left((1-b)+bL_d/L_{avg}\right)+f_{d,t}}$, $k_1 = 0.9$, $b = 0.9$



tf in doc $\frac{(k_1+1)f_{d,t}}{k_1\left((1-b)+bL_d/L_{avg}\right)+f_{d,t}}$, $k_1 = 0.9$, $b = 0$



Top-k retrieval

- A top-1 example

List 1	**List 2**	**List 3**
Doc17 : 0.8	Doc25 : 0.7	Doc83 : 0.9
Doc78 : 0.2	Doc38 : 0.5	Doc17 : 0.7
.	Doc14 : 0.5	Doc61 : 0.3
.	Doc83 : 0.5	.
.	.	.
.	Doc17 : 0.2	.

Round 1 (SA on 1,2,3)

Doc17 : [0.8 , 2.4]
Doc25 : [0.7 , 2.4]
Doc83 : [0.9 , 2.4]
unseen: ≤ 2.4

Round 2 (SA on 1,2,3)

Doc17 : [1.5 , 2.0]
Doc25 : [0.7 , 1.6]
Doc83 : [0.9 , 1.6]
unseen: ≤ 1.4

Round 3 (SA on 2,2,3!)

Doc17 : [1.5 , 2.0]
Doc83 : [1.4 , 1.6]
unseen: ≤ 1.0

Round 4 (RA for Doc17)

Doc17 : 1.7
all others < 1.7
done!

- Figure from http://www.vldb.org/conf/2006/p475-bast.pdf

List 1	List 2	List 3

List 1	**List 2**	**List 3**
Doc17 : 0.8	Doc25 : 0.7	Doc83 : 0.9

Round 1 (SA on 1,2,3)

Doc17 : [0.8 , 2.4]
Doc25 : [0.7 , 2.4]
Doc83 : [0.9 , 2.4]
unseen: ≤ 2.4

List 1	**List 2**	**List 3**
Doc17 : 0.8	Doc25 : 0.7	Doc83 : 0.9
Doc78 : 0.2	Doc38 : 0.5	Doc17 : 0.7

Round 1 (SA on 1,2,3)

Doc17 : [0.8 , 2.4]
Doc25 : [0.7 , 2.4]
Doc83 : [0.9 , 2.4]
unseen: ≤ 2.4

Round 2 (SA on 1,2,3)

Doc17 : [1.5 , 2.0]
Doc25 : [0.7 , 1.6]
Doc83 : [0.9 , 1.6]
unseen: ≤ 1.4

List 1	**List 2**	**List 3**
Doc17 : 0.8	Doc25 : 0.7	Doc83 : 0.9
Doc78 : 0.2	Doc38 : 0.5	Doc17 : 0.7
.	Doc14 : 0.5	Doc61 : 0.3
.	Doc83 : 0.5	.
.	.	.

Round 1 (SA on 1,2,3)

Doc17 : [0.8 , 2.4]
Doc25 : [0.7 , 2.4]
Doc83 : [0.9 , 2.4]
unseen: ≤ 2.4

Round 2 (SA on 1,2,3)

Doc17 : [1.5 , 2.0]
Doc25 : [0.7 , 1.6]
Doc83 : [0.9 , 1.6]
unseen: ≤ 1.4

Round 3 (SA on 2,2,3**!**)

Doc17 : [1.5 , 2.0]
Doc83 : [1.4 , 1.6]
unseen: ≤ 1.0

List 1	List 2	List 3
Doc17 : 0.8	Doc25 : 0.7	Doc83 : 0.9
Doc78 : 0.2	Doc38 : 0.5	Doc17 : 0.7
.	Doc14 : 0.5	Doc61 : 0.3
.	Doc83 : 0.5	.
.	.	.
.	Doc17 : 0.2	.
.	.	.

Round 1 (SA on 1,2,3)

Doc17 : [0.8 , 2.4]
Doc25 : [0.7 , 2.4]
Doc83 : [0.9 , 2.4]
unseen: ≤ 2.4

Round 2 (SA on 1,2,3)

Doc17 : [1.5 , 2.0]
Doc25 : [0.7 , 1.6]
Doc83 : [0.9 , 1.6]
unseen: ≤ 1.4

Round 3 (SA on 2,2,3!)

Doc17 : [1.5 , 2.0]
Doc83 : [1.4 , 1.6]
unseen: ≤ 1.0

Round 4 (RA for Doc17)

Doc17 : 1.7
all others < 1.7
done!