Big Data Processing

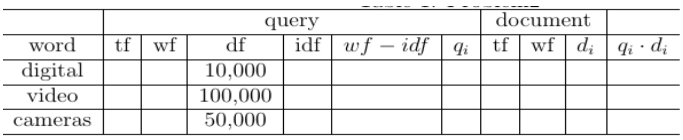
Homework 6

作业

- 完成指定的题目
- 编写报告
- 单人不组队（本次作业都是书后题目，不涉及到代码的编写以及程序的部署，所以不组队）

Exercise 1

Compute the vector space similarity between the query "digital cameras" and the document "digital cameras video cameras" by filling out the empty columns in following table. Assume N = 10, 000, 000, logarithmic term weighting(wf columns) for query and document, idf weighting for the query only. Enter term counts in the tf columns, What is the final similarity score?



	query						document			
word	tf	wf	df	idf	$wf - idf$	q_i	tf	wf	d_i	$q_i \cdot d_i$
digital			10,000							
video			100,000							
cameras			50,000							

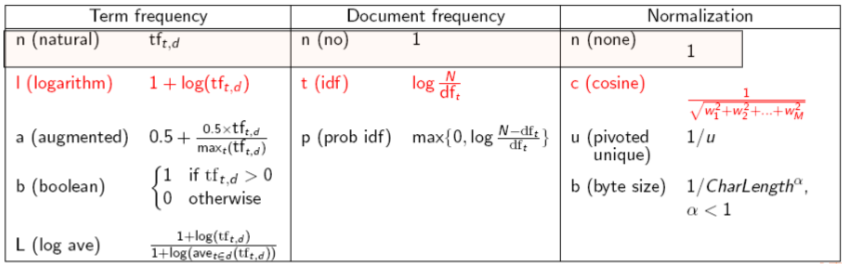
Note:
wf : see slides SearchEngines Page 97
qi: normalization of wf-idf(wf * idf) of query
di: normalization of wf of document
final similarity score calculation: sum of qi * di

Exercise 2

Compute the top scoring documents on the query best car insurance for each of the following weighing schemes:

- nnn.atc (nnn for documents, atc for query)
- ntc.atc (ntc for documents, atc for query)

Various TF-IDF weighting methods



Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(\mathrm{tf}_{t,d})$	t (idf)	$\log \frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5 + \frac{0.5 \times \mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases} 1 & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha < 1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t \in d}(\mathrm{tf}_{t,d}))}$				

Exercise 2

Term frequency and idf of three documents

(a) Term Frequency

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

(b) IDF

term	idf_t
car	1.65
auto	2.08
insurance	1.62
best	1.5

Exercise 2

nnn.atc :

	Query(atc weight)			
Term	tf	idf	tf-idf	atc weight
Car				
Auto				
Insurance				
Best				

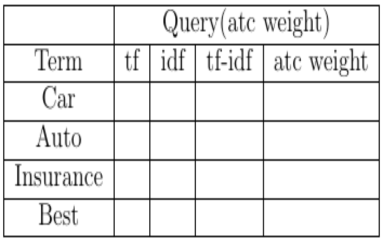
	Doc1/2/3(nnn weight)			
Term	tf	idf	tf-idf	nnn weight
Car		1		
Auto		1		
Insurance		1		
Best		1		

Note:
Score(query, doc) = sum of (atc weight(i) * nnn weight(i))

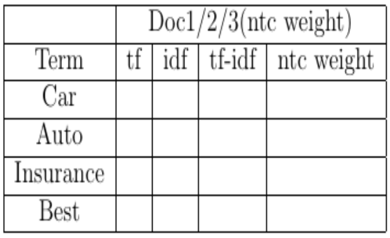
Then rank three documents by the scores

Exercise 2

ntc.atc :



	Query(atc weight)			
Term	tf	idf	tf-idf	atc weight
Car				
Auto				
Insurance				
Best				



	Doc1/2/3(ntc weight)			
Term	tf	idf	tf-idf	ntc weight
Car				
Auto				
Insurance				
Best				

Note:
Score(query, doc) = sum of (atc weight(i) * ntc weight(i))

Then rank three documents by the scores

报告要求

- 使用Word，Pages, LaTeX或者markdown等编写都可以，但最后提交时转成PDF文件格式。

- （本次作业涉及到数学公式的排版，建议采用LaTeX编写、配合markdown使用mathjax、使用word自带的公式编辑或mathtype）

提交

- 作业提交位置
  - ftp://public.sjtu.edu.cn username: shen_yao password: public
  - 提交到ftp中/upload/CS426/hw6/ 目录下
- 作业提交时间
  - ddl: 6月1号23:59:59
  - 晚交惩罚：每超时24小时，该次作业总分扣除20%成绩，不满24小时按照24小时计算，6月4日23:59:59之后提交的作业一概不接收。
  - 时间根据ftp服务器接收到文件的时间为准。
- 作业命名规则
  - 学号_姓名_hw6.pdf

评分标准（满分10分）

- Exercise 1:

  共2分（表1分， 最终score结果1分）

- Exercise 2：
  - nnn.atc（共4分）： 四张表每个0.5分、三个score结果每个0.5分，rank结果0.5分
  - ntc.atc （共4分）：四张表每个0.5分、三个score结果每个0.5分，rank结果0.5分

遇到任何问题，请发邮件到cs_jerrychen@sjtu.edu.cn