Tugas 03 - Text and Web Mining
Gloria Virginia, Ph.D.
1 Maret 2016

1. Hitung *vector space similarity* antara query dan *document* berikut ini dengan cara mengisi Tabel 1:
   - *query (q)*: digital cameras
   - *document (d)*: digital cameras and video cameras

   Asumsi:
   - Total dokumen, N = 10.000.000
   - *tf* untuk *query* dan *document*: jumlah kemunculan *terms*
   - *wf* untuk *query* dan *document*: menggunakan *logarithmic term weighting*
   - *idf weighting* untuk *query*
   - *cosine normalization* untuk dokumen
   - *Stopword*: *and*

Tabel 1. Cosine Similarity

Term	***query***					***document***			$q_i \cdot d_i$	***Similarity***
	tf	***wf***	***df***	***idf***	q_i = ***wf-idf***	***tf***	***wf***	d_i = ***normalized wf***		
digital			10,000							
video			100,000							
cameras			50,000							

2. Hitung *2-top scoring documents* (2 dokumen dengan nilai tertinggi), ketika
   - *query (q)*: best car insurance

   - *term frequency (tf)* pada 3 dokumen (Doc1, Doc2, dan Doc3) dan 4 *terms* adalah sebagai berikut:

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

   - Weighting schemes:
     a). nnn.atc
     b). ntc.atc

Untuk membantu Anda dalam menentukan *2-top scoring documents* di atas, isilah Tabel 2 dan Tabel 3.

Tabel 2. Similarity Score antara query dan Doc1, Doc2, Doc3 dengan weighting scheme nnn.atc

Term	***Weight***				***Similarity Score***		
	query	***Doc1***	***Doc2***	***Doc3***	***query & Doc1***	***query & Doc2***	***query & Doc3***
car							
auto							
insurance							
best							

Tabel 3. Similarity Score antara query dan Doc1, Doc2, Doc3 dengan weighting scheme ntc.atc

Term	*Weight*				*Similarity Score*		
	query	***Doc1***	***Doc2***	***Doc3***	***query & Doc1***	***query & Doc2***	***query & Doc3***
car							
auto							
insurance							
best							