Project II

- Problem
  - Given *m* documents, compute the term-term relevance using MapReduce and Spark
  - Input: A text file, each line represents a document
  - Output: A list of term-term pairs sorted by their similarity descending

t1	t2	s1
t3	t4	s2

- Sub-problems:
  - Compute Term Frequency – Inverse Document Frequency (TF-IDF) for each term
    - Output: mxn matrix (m: #documents, n: #terms)
  - Computer and sort term-term relevance between a query term and all terms associated with the TF-IDF matrix
    - Input: a query term *t*
    - Output: term-term relevance between the query term and those terms in the tfidf matrix sorted by the relevance score (descending)

TF-IDF

- Term Frequency – Inverse Document Frequency
  - Relevant to text processing
  - Common web analysis algorithm

The Algorithm, Formally

$$\mathrm{tf_i} = \frac{n_i}{\sum_k n_k}$$

$$\mathrm{idf_i} = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

$$\mathrm{tfidf} = \mathrm{tf} \cdot \mathrm{idf}$$

- | *D* | : total number of documents in the corpus
- $|\{d : t_i \in d\}|$number of documents where the term t_i appears (that is $n_i \neq 0$).

Semantic Similarity

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Example

D1: I like data science
D2: I hate data
D3: want A

	D1	D2	D3
I	1	1	0
like	1	0	0
data	1	1	0
science	1	0	0
hate	0	1	0
want	0	0	1
A	0	0	1

Example

D1: I like data science
D2: I hate data
D3: want A

tf

	D1	D2	D3
I	1/4	1/3	0
like	1/4	0	0
data	1/4	1/3	0
science	1/4	0	0
hate	0	1/3	0
want	0	0	1/2
A	0	0	1/2

Example

D1: I like data science
D2: I hate data
D3: I want A

idf

	D1	D2	D3
I	log(3/2)	log(3/2)	log(3/2)
like	log(3/1)	log(3/1)	log(3/1)
data	log(3/2)	log(3/2)	log(3/2)
science	log(3/1)	log(3/1)	log(3/1)
hate	log(3/1)	log(3/1)	log(3/1)
want	log(3/1)	log(3/1)	log(3/1)
A	log(3/1)	log(3/1)	log(3/1)

Example

D1: I like data science
D2: I hate data
D3: I want A

tf*idf

	D1	D2	D3
I	0.044	0.059	0.0
like	0.119	0.0	0.0
data	0.044	0.059	0.0
science	0.119	0.0	0.0
hate	0.0	0.159	0.0
want	0.0	0.0	0.238
A	0.0	0.0	0.238

Example

I	(0.044,	0.059,	0.0)
A	(0.0,	0.0,	0.238)

$$\text{Similarity (I, A)} = \frac{(0.044*0.0+0.059*0.0+0.0*0.238)}{\text{sqrt}(0.044*0.044+0.059*0.059+0.0*0.0) \text{ x } \text{sqrt}(0.0*0.0+0.0*0.0+0.238*0.238)}$$

Project II

- Rubric
  - Distributed Algorithm design
    - MapReduce (40 points)
    - Report in print (20 points)
      - Must present during your project review!
  - Implementation
    - Local on spark (40 points)

Project II

- Due Date
  - 4:00pm, 12/19
  - No overdue
- Project Review
  - 12/19