

# Tutorial on Text Mining

Yang Yang

yang.yang@itee.uq.edu.au

DKE Group, 78-625

September 16, 2011

# Exercise on Vector Space Model and TF-IDF

- Given the following collections of words appeared in three different documents:

Document 1: {Agent, Linux, Computer, Information, Agent, Computer},
Document 2: {Agent, Student, Computer, Department, Student},
Document 3: {Agent, School, Information, Student}.

- **Q1:** Calculate DF (Document Frequency) and IDF (Inverse Document Frequency) for each word.
- Formula for computing IDF:

$$IDF(i) = \log \frac{\text{Total Number of documents}}{\text{Number of documents that word } i \text{ appears in}} \quad (1)$$

# Solution to Q1

- Total Number of documents is 3.

Word List	DF	IDF
Agent	3	$\log \frac{3}{3} = 0$
Linux	1	$\log \frac{3}{1} = 1.585$
Computer	2	$\log \frac{3}{2} = 0.585$
Information	2	$\log \frac{3}{2} = 0.585$
Student	2	$\log \frac{3}{2} = 0.585$
Department	1	$\log \frac{3}{1} = 1.585$
School	1	$\log \frac{3}{1} = 1.585$

# Exercise on Vector Space Model and TF-IDF

- Given the following collections of words appeared in three different documents:

Document 1: {Agent, Linux, Computer, Information, Agent, Computer},
---

Document 2: {Agent, Student, Computer, Department, Student},
--

Document 3: {Agent, School, Information, Student}.
--

- **Q2:** Represent each document as a weighted vector by using TF-IDF weight scheme.

# Solution to Q2

- Term Frequency

# Solution to Q2

## ■ Term Frequency

	Agent	Linux	Computer	Information	Student	Department	School
Document 1	2	1	2	1	0	0	0
Document 2	1	0	1	0	2	0	0
Document 3	1	0	0	1	1	0	1

# Solution to Q2

## ■ Term Frequency

	Agent	Linux	Computer	Information	Student	Department	School
<b>Document 1</b>	2	1	2	1	0	0	0
<b>Document 2</b>	1	0	1	0	2	0	0
<b>Document 3</b>	1	0	0	1	1	0	1

## ■ TF-IDF Weighting

	Agent	Linux	Computer	Information	Student	Department	School
<b>Document 1</b>	0	1.585	1.17	0.585	0	0	0
<b>Document 2</b>	0	0	0.585	0	1.17	0	0
<b>Document 3</b>	0	0	0	0.585	0.585	0	1.585

# Solution to Q2

## ■ Term Frequency

	Agent	Linux	Computer	Information	Student	Department	School
<b>Document 1</b>	2	1	2	1	0	0	0
<b>Document 2</b>	1	0	1	0	2	0	0
<b>Document 3</b>	1	0	0	1	1	0	1

## ■ TF-IDF Weighting

	Agent	Linux	Computer	Information	Student	Department	School
<b>Document 1</b>	0	1.585	1.17	0.585	0	0	0
<b>Document 2</b>	0	0	0.585	0	1.17	0	0
<b>Document 3</b>	0	0	0	0.585	0.585	0	1.585

## ■ Normalized by $l_2$ -Norm

	Agent	Linux	Computer	Information	Student	Department	School
<b>Document 1</b>	0	0.7713	0.5693	0.2847	0	0	0
<b>Document 2</b>	0	0	0.4472	0	0.8944	0	0
<b>Document 3</b>	0	0	0	0.3272	0.3272	0	0.8865

# Exercise on Vector Space Model and TF-IDF

- Given the following collections of words appeared in three different documents:

Document 1: {Agent, Linux, Computer, Information, Agent, Computer},
Document 2: {Agent, Student, Computer, Department, Student},
Document 3: {Agent, School, Information, Student}.

- **Q3:** For a query on words of “Student Agent Linux”, which document would be returned as the query result? Use Cosine Similarity.

$$\text{sim}(Q, D) = \vec{Q} \cdot \vec{D} \quad (2)$$

# Solution to Q3

- Normalized TF-IDF Vector for Query:

	Agent	Linux	Computer	Information	Student	Department	School
Query	0	0.9381	0	0	0.3463	0	0

- Similarity Between Query and Documents

$$\text{sim}(\text{Query}, \text{Document1}) = 0.7236$$

$$\text{sim}(\text{Query}, \text{Document2}) = 0.3097$$

$$\text{sim}(\text{Query}, \text{Document3}) = 0.1133$$

- So Document 1 should be returned.