



ดัชนีเพิ่มน้ำหนักคำศัพท์เฉพาะ (Index term weighting)

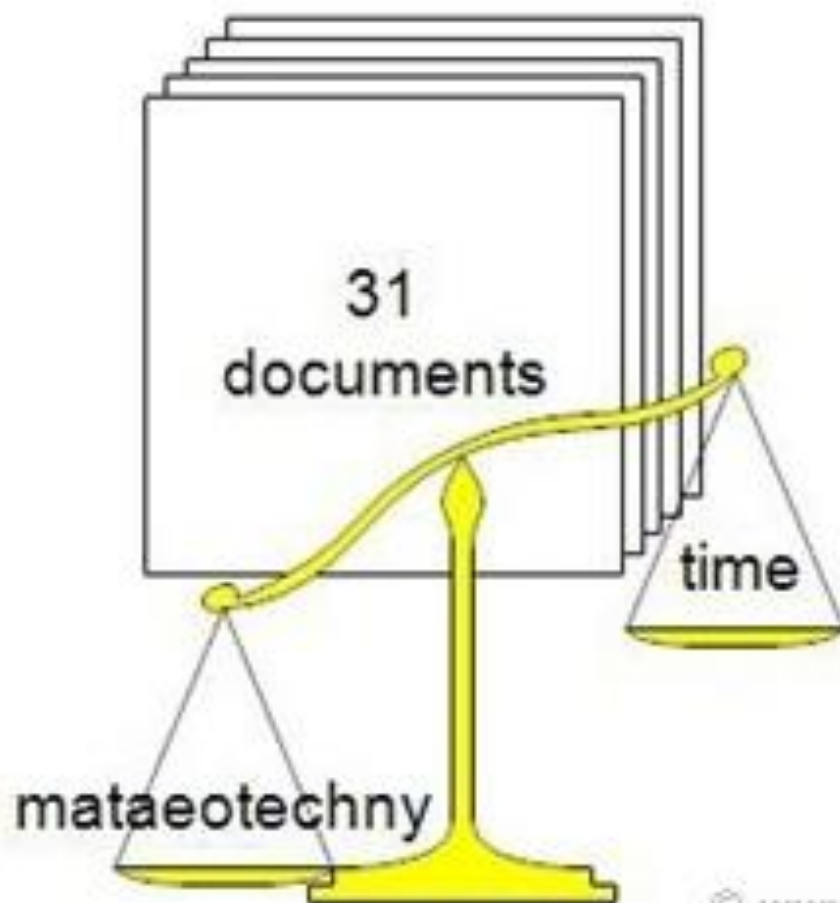
การให้น้ำหนักวรรณคดี ซึ่งอาจกระทำโดย

- ใช้ความถี่ของคำในเอกสารหรือใช้ความยาวของเอกสาร
- ใช้ความถี่ของคำในกลุ่มเอกสารทั้งหมด
- ใช้จำนวนเอกสารที่มีคำๆ นั้นอยู่ในกลุ่มเอกสารทั้งหมด



time for mataeotechny

Search



การสร้างดัชนี (Index term weighting)

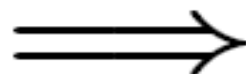
Document	Text
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old



Number	Text	Documents
1	cold	1, 4
2	days	3, 6
3	hot	1, 4
4	in	2, 5
5	it	4, 5
6	like	4, 5
7	nine	3, 6
8	old	3, 6
9	pease	1, 2
10	porridge	1, 2
11	pot	2, 5
12	some	4, 5
13	the	2, 5

Document	Text
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old

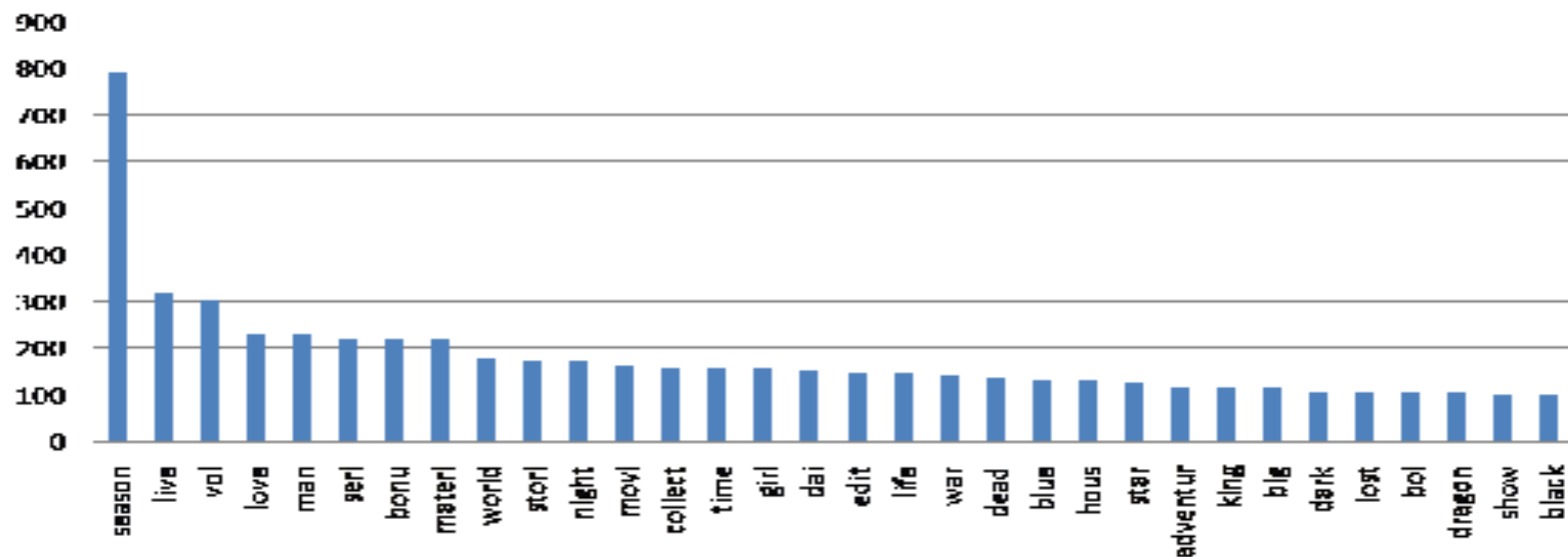
ตำแหน่งของคำในเอกสาร



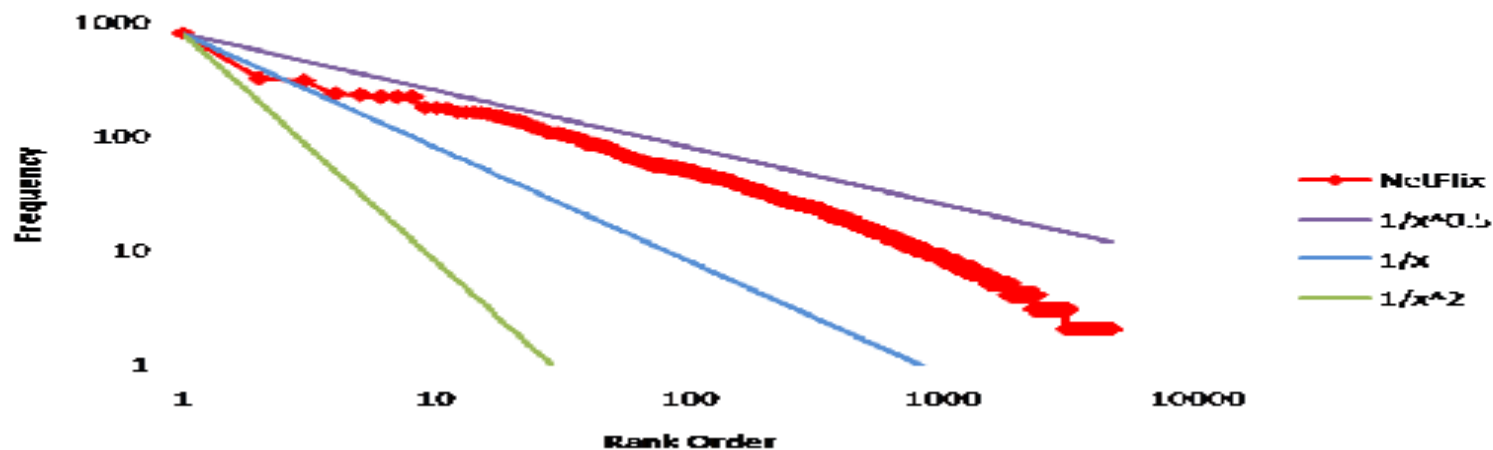
Number	Text	(Document; Word)
1	cold	(1; 6), (4; 8)
2	days	(3; 2), (6; 2)
3	hot	(1; 3), (4; 4)
4	in	(2; 3), (5; 4)
5	it	(4; 3, 7), (5; 3)
6	like	(4; 2, 6), (5; 2)
7	nine	(3; 1), (6; 1)
8	old	(3; 3), (6; 3)
9	pease	(1; 1, 4), (2; 1)
10	porridge	(1; 2, 5), (2; 2)
11	pot	(2; 5), (5; 6)
12	some	(4; 1, 5), (5; 1)
13	the	(2; 4), (5; 5)

ต่อจากนั้นนับความถี่ของคำในเอกสารเพื่อให้น้ำหนักคำแต่ละคำ

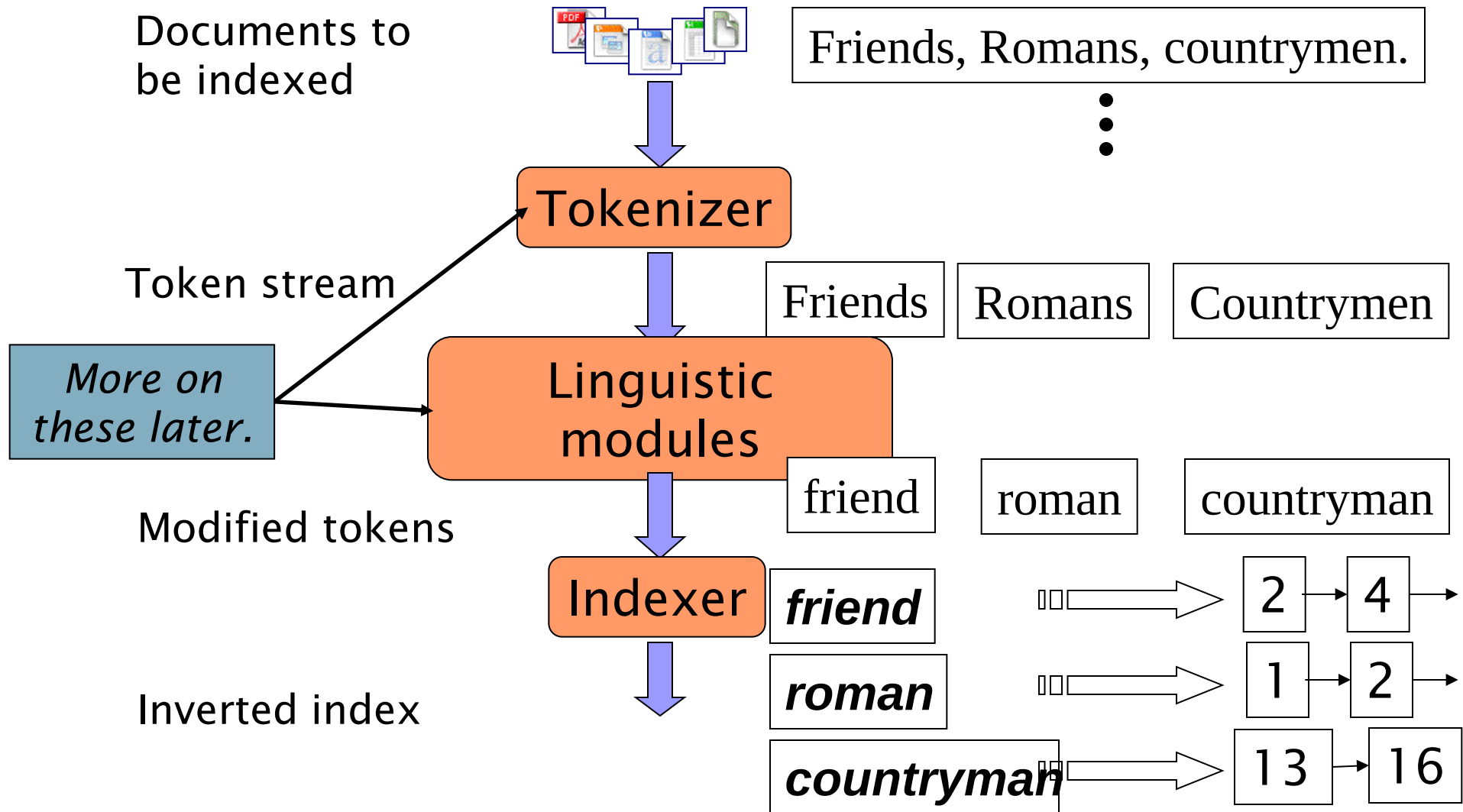
NexFlix Movie Titles Word Freq > 100



NetFlix Movie Titles Word Frequency



Inverted index construction



Indexer steps: Token sequence

Doc 1

I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.

Doc 2

So let it be with
Caesar. The noble
Brutus hath told you
Caesar was ambitious

Term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

Indexer steps: Sort

Term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2



Term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2



Indexer steps: Dictionary & Postings

- Multiple term entries in a single document are merged.
- Split into Dictionary and Postings
- Doc. frequency information is added.

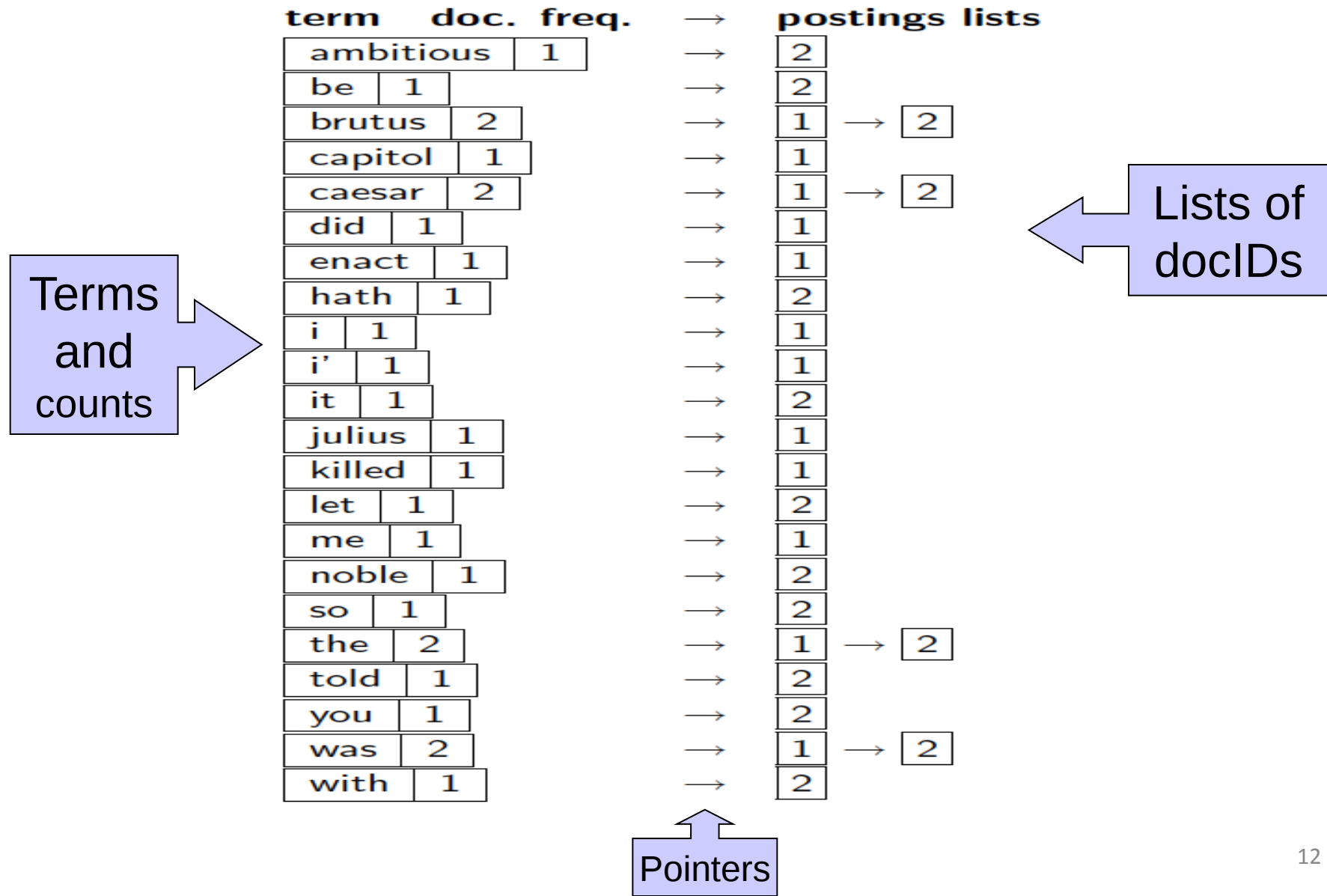
Indexer steps: Dictionary & Postings

Term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2



term	doc. freq.	→	postings lists
ambitious	1	→	2
be	1	→	2
brutus	2	→	1 → 2
capitol	1	→	1
caesar	2	→	1 → 2
did	1	→	1
enact	1	→	1
hath	1	→	2
i	1	→	1
i'	1	→	1
it	1	→	2
julius	1	→	1
killed	1	→	1
let	1	→	2
me	1	→	1
noble	1	→	2
so	1	→	2
the	2	→	1 → 2
told	1	→	2
you	1	→	2
was	2	→	1 → 2
with	1	→	2

Where do we pay in storage?



H.P Luhn -> Zipf

หลักการคือ นำความถี่ คูณด้วย ลำดับของคำ จะได้เป็นค่าคงที่หนึ่งดังนี้

ลำดับ	คำ	ความถี่	ลำดับ*ความถี่
1	the	69,971	69,971
2	of	36,411	72,822
3	and	28,852	86,556

Term Weights: Term Frequency

- More frequent terms in a document are more important, i.e. more indicative of the topic

f_{ij} = frequency of term i in document j

- May want to normalize *term frequency* (tf) across the entire corpus:

$$tf_{ij} = f_{ij} / \max\{f_{ij}\}$$

ความถี่ของคำ (Term Frequency)

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{ij}\}}$$

โดย $f_{i,j}$ เป็นจำนวนความถี่ของคำที่ i ในเอกสาร j

TF-IDF Weighting

- A typical combined term importance indicator is *tf-idf weighting*:

$$w_{ij} = tf_{ij} idf_i = tf_{ij} \log_2 (N / df_i)$$

- A term occurring frequently in the document but rarely in the rest of the collection is given high weight
- Many other ways of determining term weights have been proposed
- Experimentally, *tf-idf* has been found to work well

Term Weights: Inverse Document Frequency

- Terms that appear in many *different* documents are *less* indicative of overall topic

df_i = document frequency of term i

= number of documents containing term i

idf_i = inverse document frequency of term i ,

= $\log_2 (N / df_i)$

(N : total number of documents)

- An indication of a term's *discrimination* power
- Log used to dampen the effect relative to tf

การกำหนดน้ำหนักของเทอม

วัดจาก $tf*idf$ โดย

Term frequency (tf)

Inverse document frequency (idf)

ซึ่งต้องคำนวณน้ำหนักของแต่ละเอกสาร

ถ้าค่าของ **idf** มีค่าสูง ถ้าเป็นเทอมที่หายาก

และค่าของ **idf** มีค่าต่ำ ถ้าเป็นเทอมที่มีความถี่สูง

term	df_t	idf_t
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

► Figure 6.3 Example of idf values. Here we give the idf's of term frequencies in the Reuters collection of 806,791 documents.

ค่าของ idf ของ term car ได้มาจาก $\log(806791/18165) = 1.6475...$

คำนวณน้ำหนักของแต่ละเอกสาร

$$w_{ik} = tf_{ik} * \log(N / n_k)$$

T_k = term k in document D_i

tf_{ik} = frequency of term T_k in document D_i

idf_k = inverse document frequency of term T_k in C

N = total number of documents in the collection C

n_k = the number of documents in C that contain T_k

$$idf_k = \log\left(\frac{N}{n_k}\right)$$

Example: Computing TF-IDF

Given a document containing terms with given frequencies:

A(3), B(2), C(1)

Assume collection contains 10,000 documents and document frequencies of these terms are:

A(50), B(1300), C(250)

Then:

A: $tf = 3/3$; $idf = \log(10000/50) = 5.3$; $tf-idf = 5.3$

B: $tf = 2/3$; $idf = \log(10000/1300) = 2.0$; $tf-idf = 1.3$

C: $tf = 1/3$; $idf = \log(10000/250) = 3.7$; $tf-idf = 1.2$

ตัวอย่าง

- ชุดเอกสารประกอบด้วยเอกสาร D_1 , D_2 , D_3 เอกสารแต่ละฉบับผ่านการตัดคำ (word segmentation) และดึงคำหยุดออกไป ดังนั้นเหลือคำสำคัญดังนี้

D_1 : คอมพิวเตอร์ สารสนเทศ คอมพิวเตอร์ คอมพิวเตอร์

D_2 : อินเทอร์เน็ต คอมพิวเตอร์ อินเทอร์เน็ต ข้อมูล

D_3 : ระบบ อินเทอร์เน็ต

ตัวอย่าง

D_1 : คอมพิวเตอร์ สารสนเทศ คอมพิวเตอร์ คอมพิวเตอร์

D_2 : อินเทอร์เน็ต คอมพิวเตอร์ อินเทอร์เน็ต ข้อมูล

D_3 : ระบบ อินเทอร์เน็ต

จากนั้นหาความถี่ของคำที่ไม่ซ้ำกันในเอกสารแต่ละฉบับ

ตารางที่ 1. ความถี่ของคำในชุดเอกสาร

เอกสาร	คำ	tf
D_1	คอมพิวเตอร์	3
D_1	สารสนเทศ	1
D_2	อินเทอร์เน็ต	2
D_2	คอมพิวเตอร์	1
D_2	ข้อมูล	1
D_3	ระบบ	1
D_3	อินเทอร์เน็ต	1

ตัวอย่าง

ตารางที่ 1. ความถี่ของคำในชุดเอกสาร

เอกสาร	คำ	tf
D_1	คอมพิวเตอร์	3
D_1	สารสนเทศ	1
D_2	อินเทอร์เน็ต	2
D_2	คอมพิวเตอร์	1
D_2	ข้อมูล	1
D_3	ระบบ	1
D_3	อินเทอร์เน็ต	1

ตารางที่ 2. ค่า idf ของคำในชุดเอกสาร

คำ	df	idf
T_1 : คอมพิวเตอร์	2	0.18
T_2 : สารสนเทศ	1	0.48
T_3 : อินเทอร์เน็ต	2	0.18
T_4 : ระบบ	1	0.48
T_5 : ข้อมูล	1	0.48

คำนวณน้ำหนักของแต่ละเอกสาร

$$w_{ik} = tf_{ik} * \log(N / n_k)$$

T_k = term k in document D_i

tf_{ik} = frequency of term T_k in document D_i

idf_k = inverse document frequency of term T_k in C

N = total number of documents in the collection C

n_k = the number of documents in C that contain T_k

$$idf_k = \log\left(\frac{N}{n_k}\right)$$

ตัวอย่าง

ตารางที่ 2. ค่า *idf* ของคำในชุดเอกสาร

คำ	<i>df</i>	<i>idf</i>
T_1 : คอมพิวเตอร์	2	0.18
T_2 : สารสนเทศ	1	0.48
T_3 : อินเทอร์เน็ต	2	0.18
T_4 : ระบบ	1	0.48
T_5 : ข้อมูล	1	0.48

เวกเตอร์เอกสารทั้งหมดแทนด้วยเมตริกซ์
โดยแถวของเมตริกซ์คือเอกสารทั้งหมด
และสดมภ์คือคำที่ไม่ซ้ำกันทั้งหมดในชุดเอกสาร

	T_1	T_2	T_3	T_4	T_5
D_1	0.74	0.67	0	0	0
D_2	0.57	0	0.29	0	0.77
D_3	0	0	0.35	0.94	0

เมตริกซ์เอกสารคำ ใช้เป็น input
ในขั้นตอนการจำกลุ่มเอกสารและ
นำไปหาค่าความคล้ายคลึงของเอกสารคู่หนึ่งๆได้

The Vector-Space Model

- Assume t distinct terms remain after preprocessing; call them index terms or the vocabulary
- These “orthogonal” terms form a vector space.
Dimension = $t = |\text{vocabulary}|$
- Each term, i , in a document or query, j , is given a real-valued weight, w_{ij} .
- Both documents and queries are expressed as t -dimensional vectors:

$$d_j = (w_{1j}, w_{2j}, \dots, w_{tj})$$

Document Collection

- A collection of n documents can be represented in the vector space model by a term-document matrix
- An entry in the matrix corresponds to the “weight” of a term in the document; zero means the term has no significance in the document or it simply doesn’t exist in the document

$$\begin{pmatrix} & T_1 & T_2 & \dots & T_t \\ D_1 & w_{11} & w_{21} & \dots & w_{t1} \\ D_2 & w_{12} & w_{22} & \dots & w_{t2} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ D_n & w_{1n} & w_{2n} & \dots & w_{tn} \end{pmatrix}$$

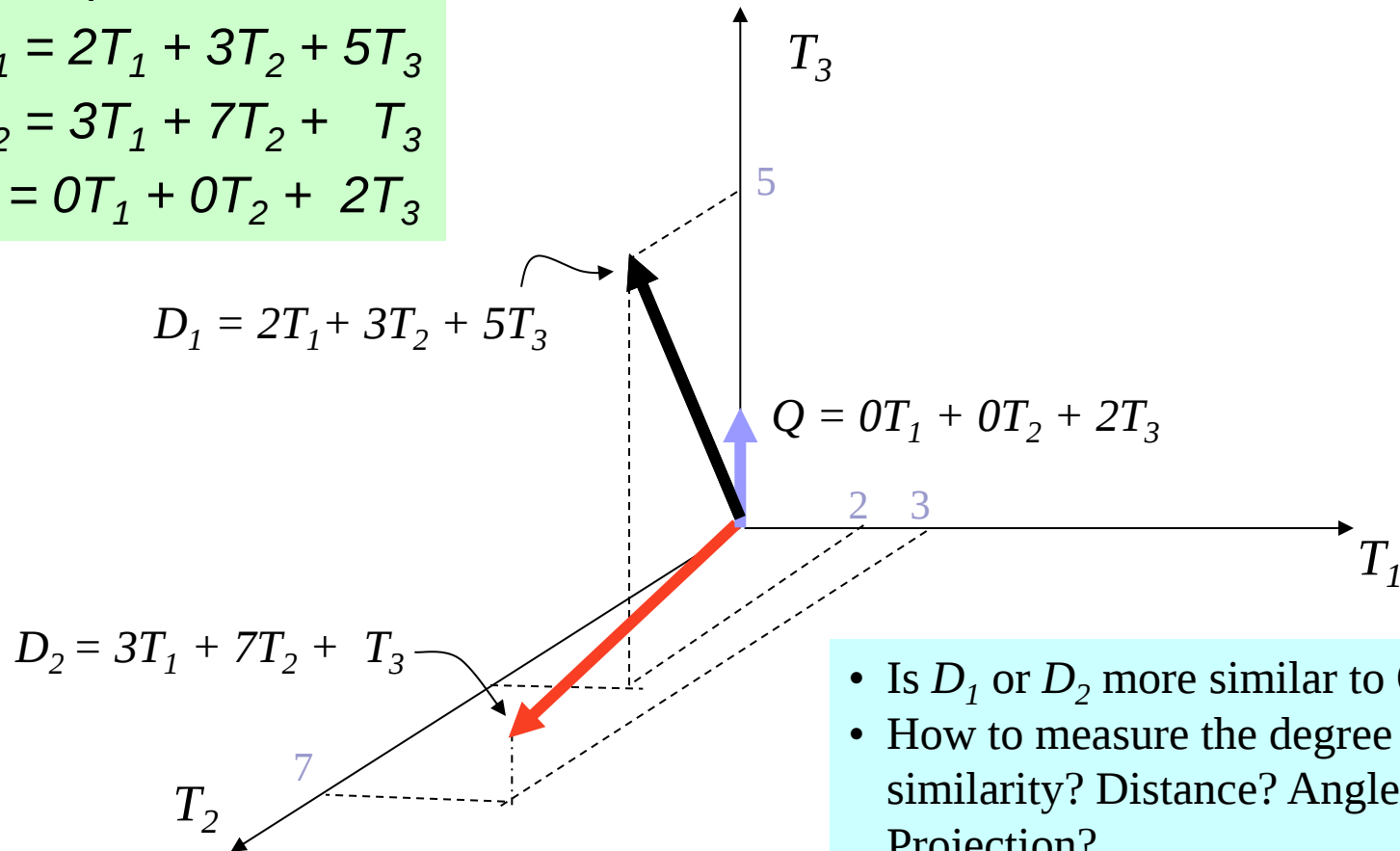
Graphic Representation

Example:

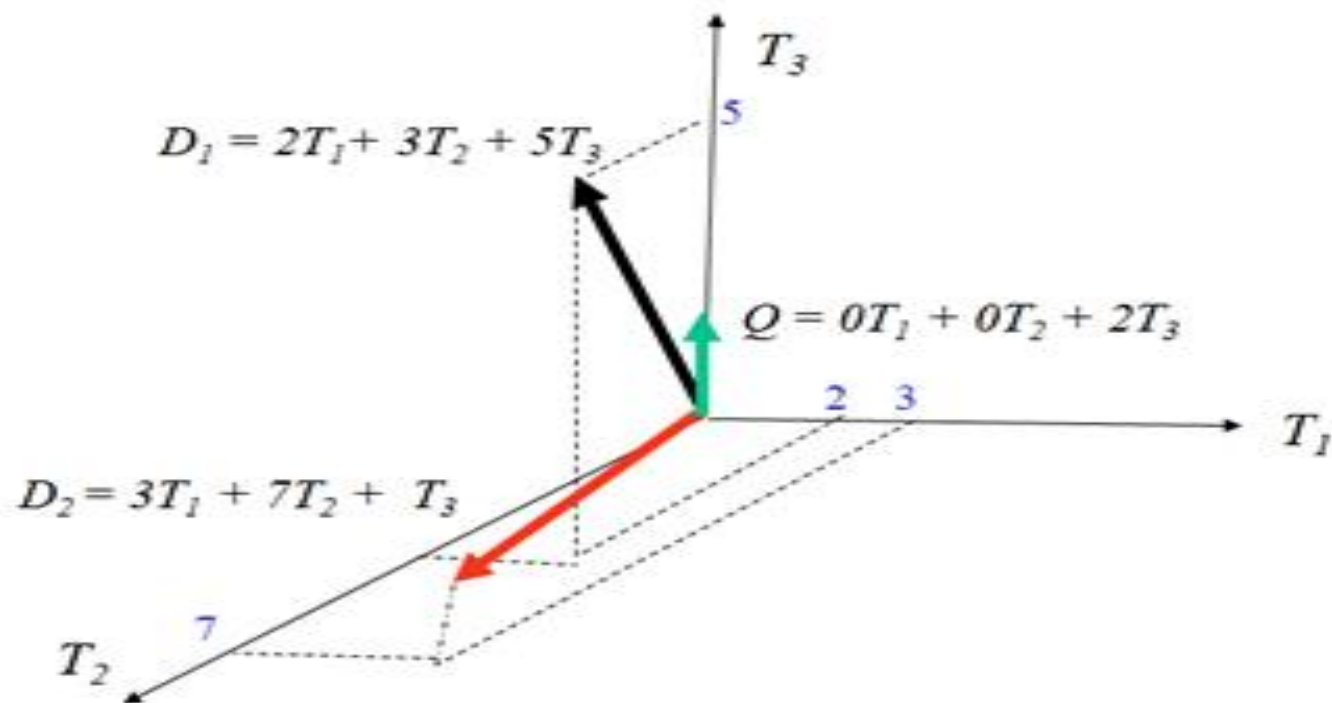
$$D_1 = 2T_1 + 3T_2 + 5T_3$$

$$D_2 = 3T_1 + 7T_2 + T_3$$

$$Q = 0T_1 + 0T_2 + 2T_3$$



- Is D_1 or D_2 more similar to Q ?
- How to measure the degree of similarity? Distance? Angle? Projection?



Term-document matrix:

$$\begin{pmatrix}
 & T_1 & T_2 & \dots & T_t \\
 D_1 & w_{11} & w_{21} & \dots & w_{t1} \\
 D_2 & w_{12} & w_{22} & \dots & w_{t2} \\
 \vdots & \vdots & \vdots & & \vdots \\
 \vdots & \vdots & \vdots & & \vdots \\
 D_n & w_{1n} & w_{2n} & \dots & w_{tn}
 \end{pmatrix}$$

Example I: Inner Product

Binary:

$$\square D = 1, 1, 1, 0, 1, 1, 0$$

$$\square Q = 1, 0, 1, 0, 0, 1, 1$$

Size of vector = size of vocabulary = 7
0 means corresponding term not found in document or query

$$\text{sim}(D, Q) = 3$$

Weighted:

$$D_1 = 2T_1 + 3T_2 + 5T_3 \quad D_2 = 3T_1 + 7T_2 + 1T_3$$

$$Q = 0T_1 + 0T_2 + 2T_3$$

$$\text{sim}(D_1, Q) = 2*0 + 3*0 + 5*2 = 10$$

$$\text{sim}(D_2, Q) = 3*0 + 7*0 + 1*2 = 2$$



Document Length

- Only measuring the inner product has some disadvantage
 - Longer documents are more likely to be relevant, as they are more likely to contain matching terms
 - If two documents have the same score, we would like to prefer the shorter one as it is more focused on the information need
- So, the length of a document has to be integrated in computing the similarity score

Cosine Similarity Measure

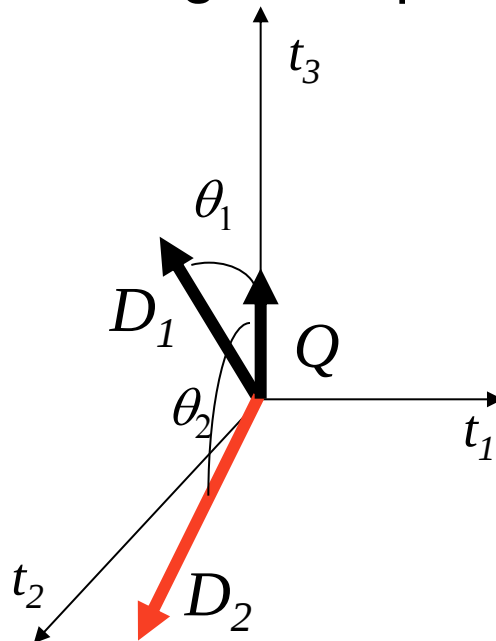
- Cosine similarity measures the cosine of the angle between two vectors.
- Inner product normalized by the vector lengths.

$$\text{CosSim}(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \cdot |\vec{q}|} = \frac{\sum_{i=1}^t (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^t w_{ij}^2 \cdot \sum_{i=1}^t w_{iq}^2}}$$

Example: Cosine Similarity

- $D1 = 2T1 + 3T2 + 5T3$ $\text{CosSim}(D1, Q) = 10 / \sqrt{(4+9+25)(0+0+4)} = 0.81$
- $D2 = 3T1 + 7T2 + 1T3$ $\text{CosSim}(D2, Q) = 2 / \sqrt{(9+49+1)(0+0+4)} = 0.13$
- $Q = 0T1 + 0T2 + 2T3$

$D1$ is 6 times better than $D2$ using cosine similarity but only 5 times better using inner product





Comments on Vector Space Models

- Simple, mathematically based approach
- Considers both local (*tf*) and global (*idf*) word occurrence frequencies
- Provides partial matching and ranked results.
- Tends to work quite well in practice despite obvious weaknesses
- Allows efficient implementation for large document collections

Problems with Vector Space Model

- Missing semantic information (e.g. word sense)
- Missing syntactic information (e.g. phrase structure, word order, proximity information)
- Assumption of term independence (e.g. ignores synonymy)
- Lacks the control of a Boolean model (e.g., *requiring* a term to appear in a document)
 - Given a two-term query “A B”, may prefer a document containing A frequently but not B, over a document that contains both A and B, but both less frequently



References

- Modern Information Retrieval, by Beaza-Yates and Ribeiro-Neto, 1999
- James Allen, University of Massachusetts Amherst
- Raymond J. Mooney, University of Texas
- Christof Monz and Maarten de Rijke, University of Amsterdam

Probabilistic Indexing

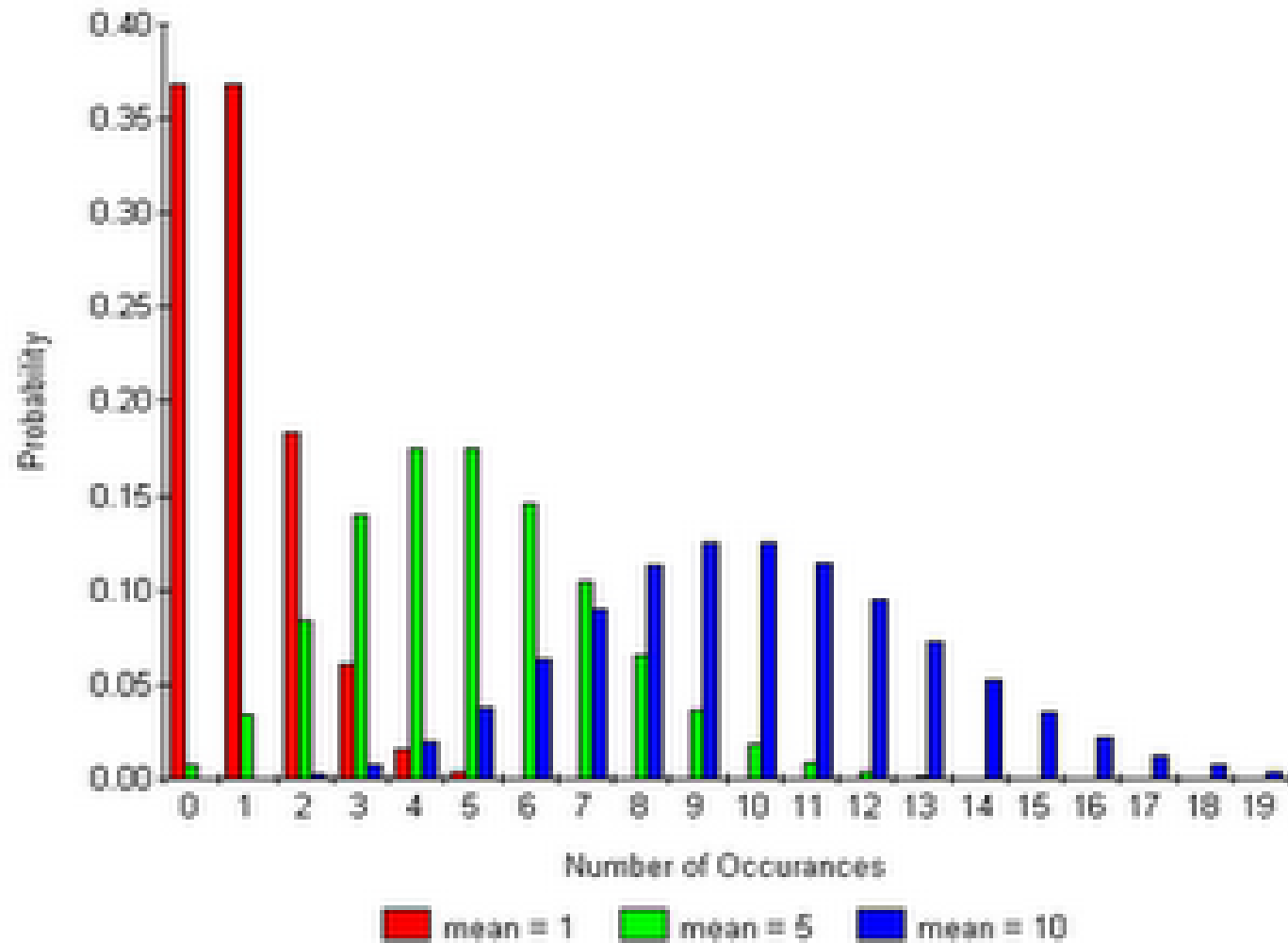
- พื้นฐานของสมมุติฐานทางสถิติมาใช้ในการสร้างดัชนีแบบอัตโนมัติเพื่อแสดงถึงการกระจายของคำในข้อความ

Stone and Rubinoﬀ ,Demeraul and Dennis ได้กล่าวว่า
ฟังก์ชันของคำนั้นจะกระจายไปยังเอกสารทั้งหมด(poisson distribution)
ดังนั้นถ้าต้องการค้นหาการกระจายของคำ W ในข้อความนั้น
ดังนั้นความน่าจะเป็น $f(n)$ เป็นความน่าจะเป็นที่จะมีการเกิดขึ้นของคำ W โดย n
เป็นเหตุการณ์ของฟังก์ชันคำ W

$$f(n) = e^{-x} x^n / n$$

โดย X คือเป็นจำนวนเฉลี่ยของการเกิดขึ้นของคำ W ในข้อความ ซึ่งเปลี่ยนแปลงตามความยาวของข้อความ

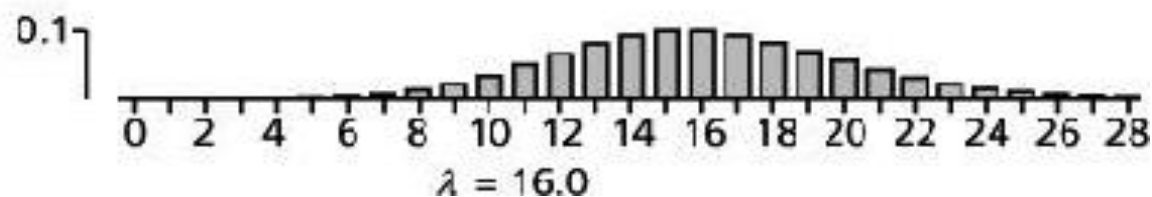
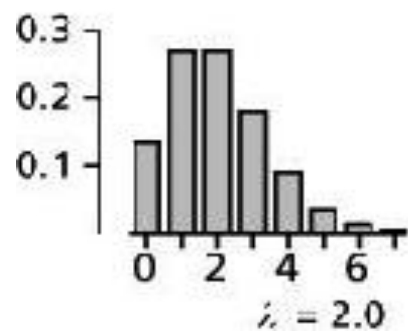
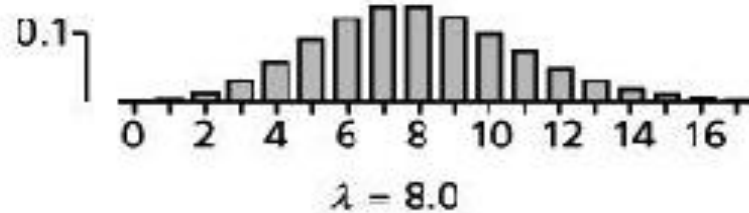
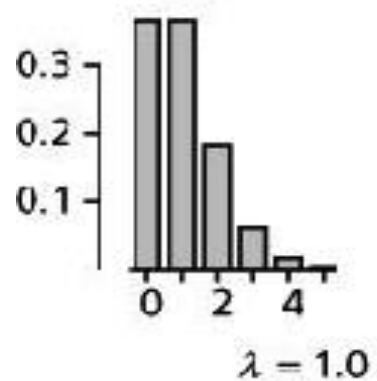
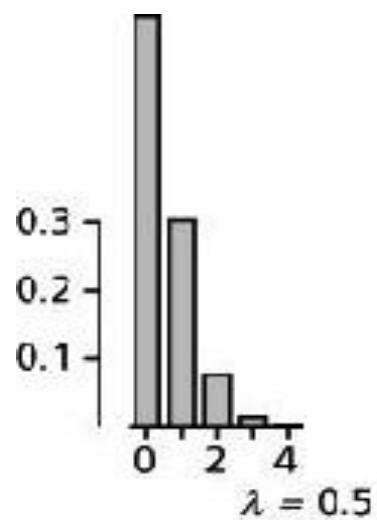
Poisson Distribution



Harter วิจัยและทดลองสมมุติฐานเพื่อหากฎเกณฑ์ สำหรับกำหนดเทอมดัชนี(index term)

$$f(n) = \frac{p_i e^{-x_1} x_1^n}{n} + \frac{(1 - p_i e^{-x_2} x_2^n)}{n}$$

โดย p_i หมายถึงความน่าจะเป็นของการสุ่มเอกสารที่เป็นชุดย่อย
 x_1 และ x_2 เป็นค่าเฉลี่ยของเหตุการณ์ทั้งสอง



ปัจจัยที่สำคัญที่ช่วยควบคุมประสิทธิภาพของดัชนี

indexing exhaustivity

ถูกกำหนดเป็นจำนวนของหัวข้อดัชนีที่แตกต่างกันซึ่งบรรจุดัชนีต่างๆที่ให้
ความหมายครอบคลุมไปทุกๆเรื่องราวสาขาที่มี
อยู่ในกลุ่มของเอกสารทั้งหมด

Exhaustivity สูง

จะให้ค่า recall ที่สูง

และให้ค่า precision ที่ต่ำ

Index Language specificity

เป็นความสามารถของภาษาดัชนีที่บรรยาย
หัวข้ออย่างแม่นยำ บรรจุดัชนีที่มีความหมาย
เจาะจงในวงแคบและถูกต้อง

Specificity ที่สูงจะให้ค่า

Precision สูงด้วย

และได้ค่า recall ที่ต่ำ

ประสิทธิผลของระบบ

HIT

: เอกสารที่เกี่ยวข้องที่ถูกดึงออกมา

WASTED

: เอกสารที่ไม่เกี่ยวข้องที่ถูกดึงออกมา

MISSED

: เอกสารที่เกี่ยวข้องที่ไม่ถูกดึงออกมา

PASSED: เอกสารที่ไม่เกี่ยวข้องที่ไม่ถูกดึงออกมา

Relevant(เกี่ยวข้อง)

Non-relevant(ไม่เกี่ยวข้อง)

MISSED	HIT	WASTED	PASSED
Not Retrieved	Retrieved (ดึงออกมา)		Not Retrieved

Recall = HIT / Relevant

เป็นการวัดความสามารถของระบบในการดึงเอกสารที่เกี่ยวข้องออกมา

Precision = HIT / Retrieved

เป็นการวัดความสามารถของระบบในการจัดเอกสารที่ไม่เกี่ยวข้องออกไป

Main Page - Beagle

File Edit View Go Bookmarks Tools Tabs Help

Back Forward Stop Reload Home History Bookmarks **Find**

http://beaglewiki.org/Main_Page

Using Beagle, you can easily find:

- documents
- emails
- web history
- IM/IRC conversations
- source code**
- images
- music files
- applications
- ... and much more

Find everything fast

Getting Beagle

The most recent version of Beagle is 0.1.4, which was released on December 21, 2005. Get the latest source tarball from [GNOME ftp](#).

Visit [Installing Beagle](#) for source build instructions and links to binary packages (available for most popular distributions). Visit the [Getting Started](#) page for instructions on how to get Beagle up and running on your system.

See Beagle Run

Find: **source** Find Next Find Previous

Beagle Search (beta)

Search terms: beagle **Anywhere** Find

BENCH2.c, in folder code
Last modified 5 days ago, 2:02 PM
CurrentKernel = AIRREL; ... printf("Running AIRREL..."); ... airrel(); ... Times(AIRREL) = EndTime(lastStartTime); ... fprintf(FILEHANDLE, "AIRREL: %.3f, (%.3f ...
Open Send to... Reveal in file manager

cpu25-c.h, in folder code
Last modified 5 days ago, 2:02 PM
extern double airrel(void); ... #define AIRREL 14 ...
Open Send to... Reveal in file manager

AIRREL.c, in folder code
Last modified 5 days ago, 2:02 PM
// airrel.c ... double airrel(void) ...
Open Send to... Reveal in file manager

INIT.c, in folder code
Last modified 5 days ago, 2:02 PM
case (AIRREL): ...
Open Send to... Reveal in file manager

BENCH2.c, in folder code
Last modified 5 days ago, 2:02 PM
CurrentKernel = AIRREL; ... printf("Running AIRREL..."); ... airrel(); ... Times(AIRREL) = EndTime(lastStartTime); ... fprintf(FILEHANDLE, "AIRREL: %.3f, (%.3f ...
Open Send to... Reveal in file manager

Best 5 results of 23 are shown. Show Previous Results Show More Results

 Prev



Next 

Search 

from Dogs to Economic assistance

Dogs

Dolphins

Domestic adjustment assistance

see Trade adjustment assistance .

Domestic animals

see Livestock .

Domestic partners

Domestic policy

see Social policy .

Domestic politics and foreign policy

Domestic relations

Domestic service employees

see Household workers .

Domestic violence

see Family violence .

Donation of organs, tissues, etc.

Doorkeeper of the House

see Congressional officers .

Double employment

see Supplementary employment .

Double isoperity

Double taxation

Down's syndrome

Dowsing

Draft

see Compulsory military service .

ปัญหาในการวิเคราะห์ข้อความ

- คำๆเดียวกันที่บางครั้งมีความสำคัญต่อเนื้อหาใจความ แต่บางครั้งไม่สำคัญ เช่น คำว่า **CAN** เป็นคำกริยาแปลว่า สามารถ แต่ **CAN** เมื่อใช้เป็นคำนาม กลับแปลว่ากระป๋อง
- คำที่เหมือนกันมีความหมายต่างกัน เมื่อใช้ในเรื่องที่ต่างกัน เช่น **Baseball** กับ **Military Base**
- คำที่เขียนต่างกันแต่มีความหมายเหมือนกัน เช่น คำว่า “restaurant “ และ “café”
- คำบางคำไม่ได้ให้ความสำคัญต่อใจความเลย เช่น คำสรรพนาม คำบุพบท คำสันธาน ต้องขจัดออกจากข้อความ

ปัญหาในการวิเคราะห์ข้อความ

- ความหมายของคำในประโยคต่างๆอาจไม่ชัดเจน ต้องพิจารณาจากประโยคก่อนหน้า หรือมีความคลุมเครือ(ambiguous) เช่น
 - คำว่า “bat” ไม่ทราบว่า เป็น baseball หรือ mamal
 - คำว่า “apple” ไม่ทราบว่า เป็น company หรือ fruit
 - คำว่า “bit” ไม่ทราบว่า เป็น unit of data หรือ act of eating
- คำบางคำแปรเปลี่ยนไปตามยุคสมัย เช่น Monitor หรือ Operating System

วิธีการแก้ไขปัญหา

Negative Dictionary

เป็นพจนานุกรมที่เก็บคำต่างๆที่ต้องขจัดออก
จากข้อความ ในการวิเคราะห์ข้อความ เช่น
stop word list , การตัด Suffix

Synonym Dictionary หรือ บัญชีศัพท์สัมพันธ์(Thesaurus)

เป็นพจนานุกรมที่เก็บคำพ้อง(Synonym) ที่แยก
เป็นลำดับชั้น และคำต่างๆที่มีความสัมพันธ์กับคำนั้นใน
หลายๆระดับ เช่น

มีความหมายพ้องกัน (Related Term: RT)

มีความหมายเจาะจงมากกว่า

(Narrower Term : NT)

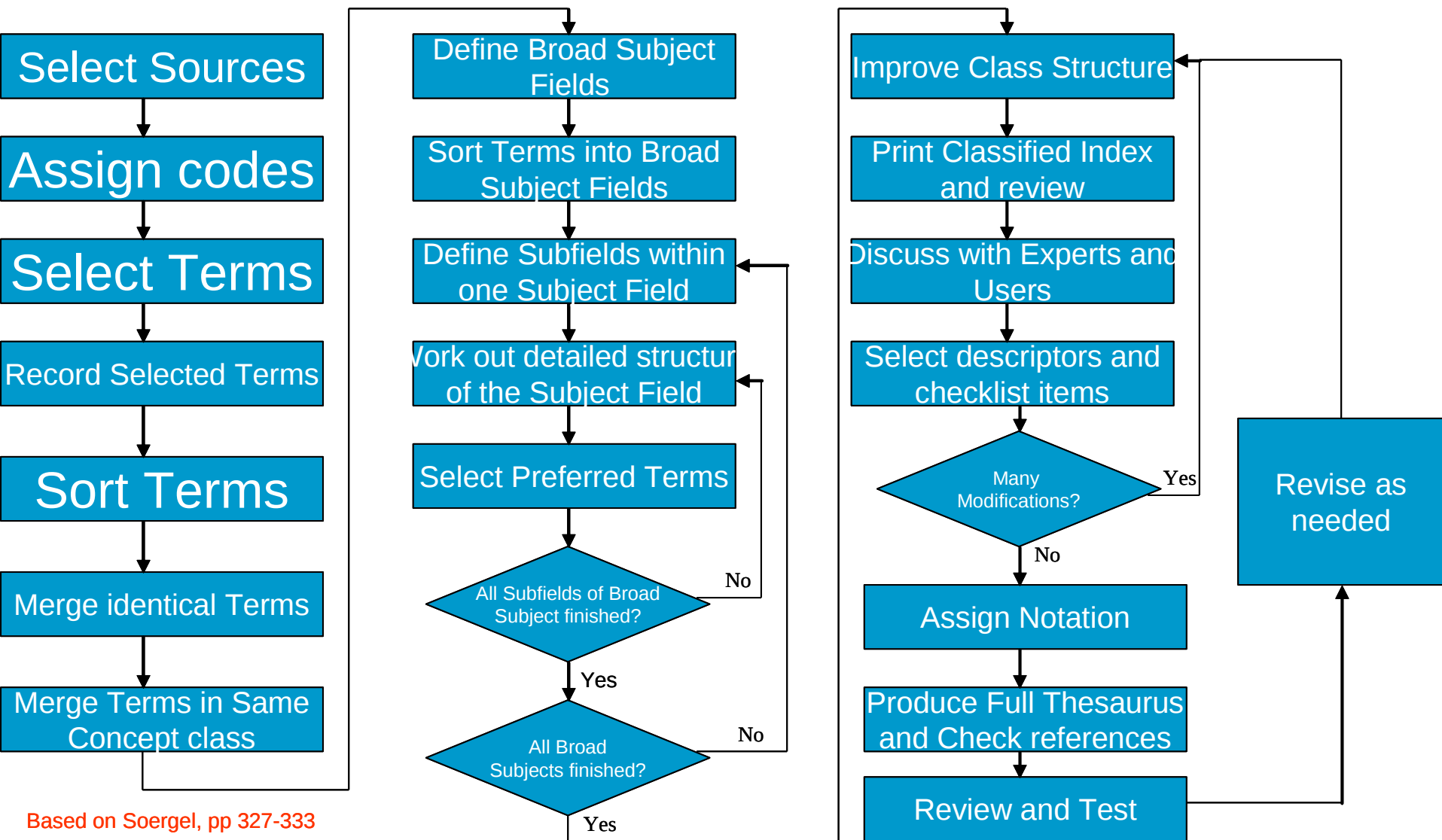
มีความหมายกว้างกว่า(Boarder Term :BT)

โดยแสดงถึงคำศัพท์ที่ถูกควบคุม

หลักเกณฑ์ในการสร้างบัญชีศัพท์สัมพันธ์

- เก็บถ้อยคำในรูปแบบต่างๆที่ไม่ค่อยพบบ่อยนัก
- ไม่เก็บคำที่มีความถี่สูงหรือพบบ่อยมาก เช่น and, in, but , the
- คำที่ไม่แสดงถึงความสำคัญอย่างชัดเจน เช่น hand
- คำที่มีหลายๆความหมาย เก็บความหมายของคำไว้ด้วย

รูป: แสดง Flow of Work in Thesaurus Construction



ตัวอย่าง

บัญชีคำศัพท์คำว่า Secondary Schools (โรงเรียนมัธยม)

จะมีการเก็บ

BT	Education	(การศึกษา)
RT	Primary School	(โรงเรียนประถม)
NT	Classes	(ชั้นเรียน , เวลาเรียน)

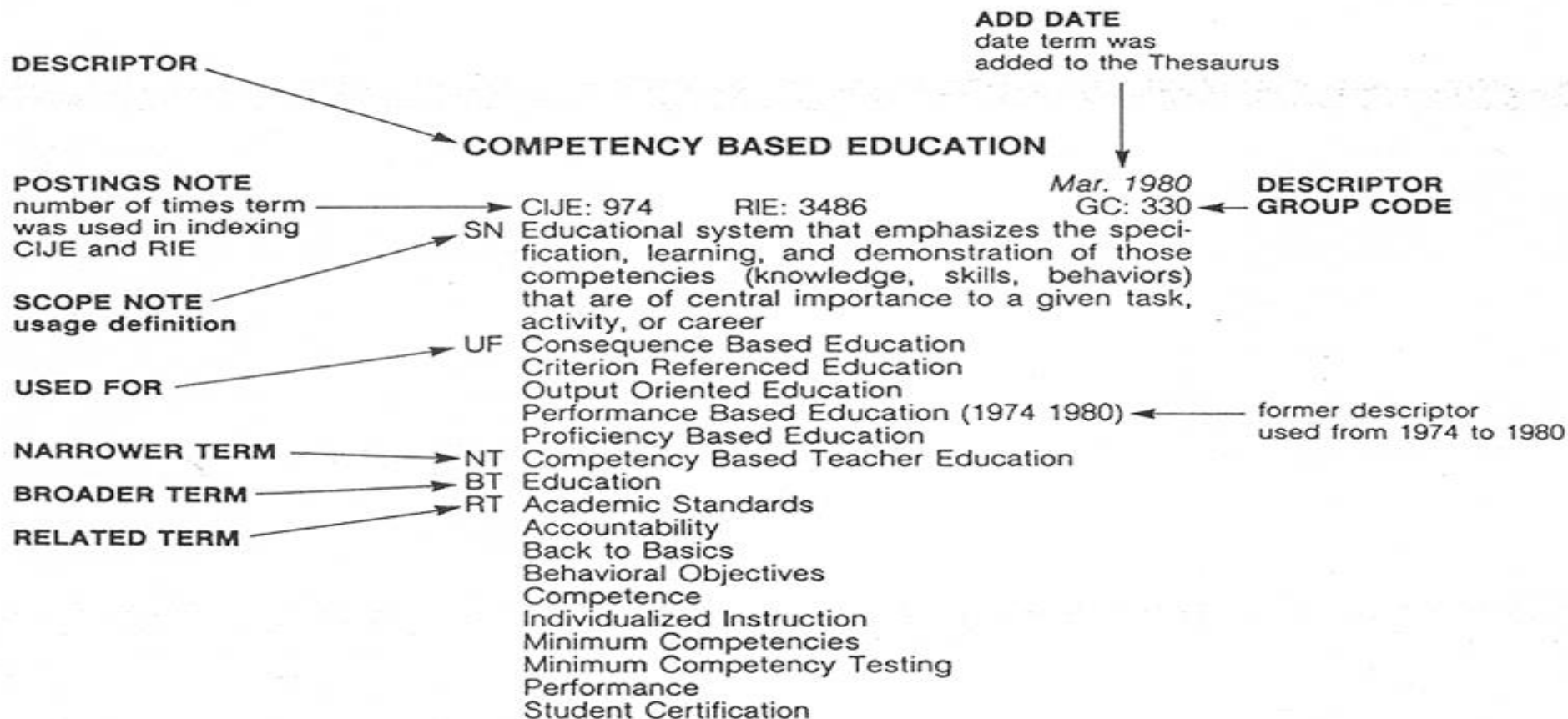


National and International Standards for Thesauri

- ANSI/NISO z39.19-1994 — American National Standard Guidelines for the Construction, Format and Management of Monolingual Thesauri
- ANSI/NISO Draft Standard Z39.4-199x — American National Standard Guidelines for Indexes in Information Retrieval
- ISO 2788 — Documentation — Guidelines for the establishment and development of monolingual thesauri
- ISO 5964 — Documentation — Guidelines for the establishment and development of multilingual thesauri

รูป:แสดง ERIC Thesaurus – Entry

SAMPLE THESAURUS ENTRY ALPHABETICAL DESCRIPTOR DISPLAY



รูป:แสดง ERIC Thesaurus – Groups

PEER INSTITUTIONS
 PRESCHOOL EVALUATION
 PROGRAM ADMINISTRATION
 PROGRAM ATTITUDES
 PROGRAM CONTENT
 PROGRAM DESIGN
 PROGRAM DEVELOPMENT
 PROGRAM EFFECTIVENESS
 PROGRAM EVALUATION
 PROGRAM IMPLEMENTATION
 PROGRAM IMPROVEMENT
 PROGRAM TERMINATION
 PUBLISH OR PERISH ISSUE
 QUALITY CIRCLES
 RECORDKEEPING
 ROTATION PLANS
 SCHEDULING
 SCHOOL ACTIVITIES
 SCHOOL ADMINISTRATION
 SCHOOL BASED MANAGEMENT
 SCHOOL CATALOGS
 SCHOOL CONSTRUCTION
 SCHOOL EFFECTIVENESS
 SCHOOL EXPANSION
 SCHOOL HOLDING POWER
 SCHOOL MAINTENANCE
 SCHOOL ORGANIZATION
 SCHOOL ORIENTATION
 SCHOOL POLICY
 SCHOOL REGISTRATION
 SCHOOL SAFETY
 SCHOOL SECURITY
 SCHOOL SHOPS
 SCHOOL SIZE
 SCHOOL SPACE
 SCHOOL SUPERVISION
 SCIENCE CLUBS
 SCIENCE SUPERVISION
 SEARCH COMMITTEES (PERSONNEL)
 SELECTIVE ADMISSION
 SPECIAL PROGRAMS
 STAFF DEVELOPMENT
 STAFF MEETINGS
 STAFF ORIENTATION
 STAFF ROLE
 STAFF UTILIZATION

→ 330 SOCIETAL PERSPECTIVES

ACADEMIC FREEDOM
 ACCESS TO EDUCATION
 ACCOUNTABILITY
 ACCREDITATION (INSTITUTIONS)
 ACCREDITING AGENCIES
 AGING IN ACADEMIA
 ALTERNATIVE TEACHER CERTIFICATION
 AMERICAN INDIAN EDUCATION
 ARTICULATION (EDUCATION)
 BACK TO BASICS
 BILINGUAL EDUCATION
 BLACK ACHIEVEMENT
 BLACK EDUCATION
 BOARD ADMINISTRATOR RELATIONSHIP
 BOARD OF EDUCATION POLICY
 BOARD OF EDUCATION ROLE
 BOARDS OF EDUCATION
 CERTIFICATION
 CHANGE AGENTS
 CHANGE STRATEGIES
 COEDUCATION
 COLLEGE ATTENDANCE
 COLLEGE CHOICE
 COLLEGE DAY
 COLLEGE OUTCOMES ASSESSMENT
 COLLEGE PREPARATION
 COLLEGE ROLE
 COLLEGE SCHOOL COOPERATION
 COMMUNITY BENEFITS
 COMMUNITY CONTROL
 COMMUNITY INVOLVEMENT
 COMMUNITY SUPPORT
 COMPENSATORY EDUCATION
 COMPETENCY BASED EDUCATION
 COMPETITIVE SELECTION
 COMPULSORY EDUCATION
 COMPUTER LITERACY
 COMPUTER USES IN EDUCATION
 CONSORTIA
 CONTINUATION EDUCATION (1968 1980)
 CONTROVERSIAL ISSUES (COURSE CONTENT)
 COOPERATIVE PROGRAMS

GOVERNMENT SCHOOL RELATIONSHIP
 HIDDEN CURRICULUM
 HOME PROGRAMS
 HOME SCHOOLING
 HOME VISITS
 HONOR SOCIETIES
 INDIVIDUALIZED EDUCATION PROGRAMS
 INSTITUTIONAL AUTONOMY
 INSTITUTIONAL COOPERATION
 INSTITUTIONAL ROLE
 INSTITUTIONAL SURVIVAL
 INTELLECTUAL FREEDOM
 INTERCOLLEGIATE COOPERATION
 INTERCULTURAL COMMUNICATION
 INTERDISTRICT POLICIES
 INTERGROUP EDUCATION
 INTERNATIONAL COMMUNICATION
 INTERNATIONAL EDUCATIONAL EXCHANGE
 INTERSCHOOL COMMUNICATION
 LINKING AGENTS
 LONG RANGE PLANNING
 MEXICAN AMERICAN EDUCATION
 MIGRANT ADULT EDUCATION
 MIGRANT EDUCATION
 MINIMUM COMPETENCIES
 MOBILE EDUCATIONAL SERVICES
 NONCATEGORICAL EDUCATION
 NONFORMAL EDUCATION
 NONSCHOOL EDUCATIONAL PROGRAMS
 NONTRADITIONAL EDUCATION
 OPEN ENROLLMENT
 OUTCOMES OF EDUCATION
 OUTREACH PROGRAMS
 PARENT ASSOCIATIONS
 PARENT CONFERENCES
 PARENT EDUCATION
 PARENT GRIEVANCES
 PARENT PARTICIPATION
 PARENT SCHOOL RELATIONSHIP
 PARENT STUDENT RELATIONSHIP
 PARENT TEACHER CONFERENCES
 PARENT TEACHER COOPERATION
 PARENT WORKSHOPS
 PARENTS AS TEACHERS
 PARTICIPANT CHARACTERISTICS
 POLICE SCHOOL RELATIONSHIP



Educational Resources Information Center

Processing and Reference Facility

[contact us](#) [site map](#) [links](#)

[Home](#)

[Ready
References](#)

[Submitting
Documents](#)

[Reproduction
Release Form](#)

[Products](#)

[Resources](#)

[Back to Thesaurus Search](#)

Term:	Competency Based Education
Record Type:	Main
Scope Note:	Educational system that emphasizes the specification, learning, and demonstration of those competencies (knowledge, skills, behaviors) that are of central importance to a given task, activity, or career
Category:	330
Broader Terms:	Education ;
Narrower Terms:	Competency Based Teacher Education ;
Related Terms:	Academic Standards ; Accountability ; Back to Basics ; Behavioral Objectives ; Competence ; Individualized Instruction ; Minimum Competencies ; Minimum Competency Testing ; Outcome Based Education ; Performance ; Performance Based Assessment ; Student Certification ;
Used For:	Consequence Based Education; Criterion Referenced Education; Output Oriented Education; Performance Based Education (1974 1980); Proficiency Based Education
Use Term:	
Use And:	
Add Date:	03/10/1980

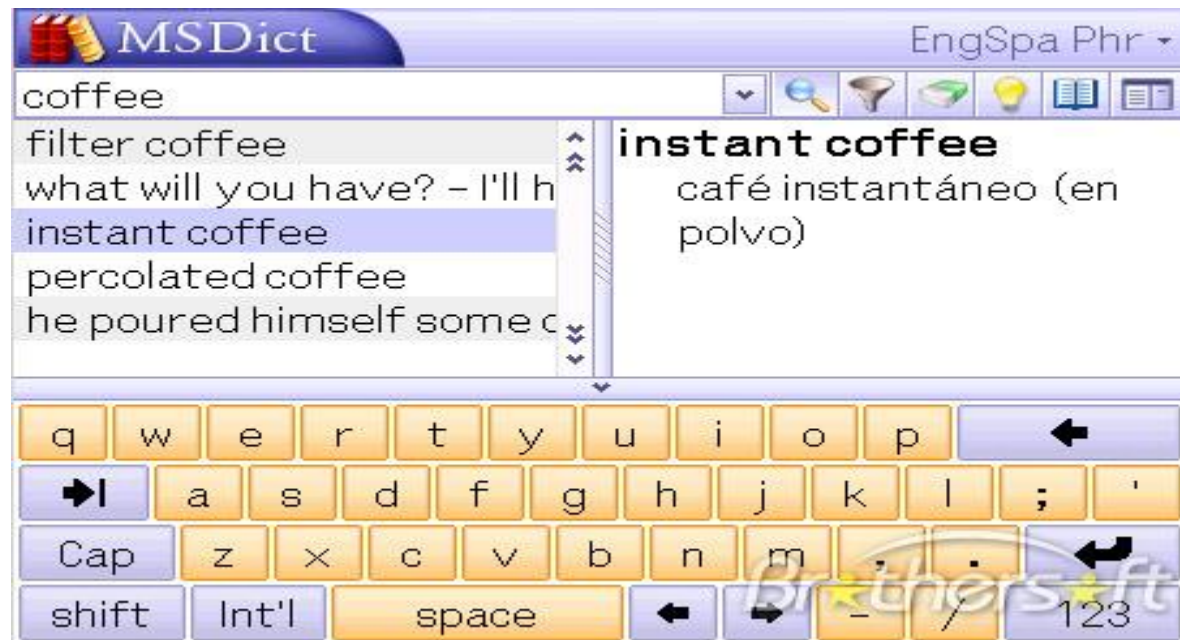
เรานำกลุ่มวรรณิ่ไปใช้ได้ 2 วิธีคือ

- โดยแทนที่แต่ละคีย์เวิร์ด ในตัวแทนเอกสาร หรือข้อคำถาม โดยชื่อของกลุ่มที่คีย์เวิร์ดหรือวรรณิ่ นั้นอยู่ ซึ่งช่วยพัฒนา recall ให้ดีขึ้น ซึ่งเพิ่มจำนวนของเอกสารที่ตรงกับข้อคำถาม(query)ของผู้ใช้

- โดยแทนที่แต่ละคีย์เวิร์ดด้วยทุกๆ คีย์เวิร์ดที่เกิดขึ้นภายในกลุ่มเดียวกันกับคีย์เวิร์ดนั้น เป็นการเพิ่ม precision ช่วยให้เอกสารมีความเจาะจงตรงกับความต้องการของผู้ใช้

Phrase Dictionary

- เป็นพจนานุกรมที่เก็บวลีของคำที่เกิดขึ้นบ่อยๆและแสดงเนื้อหาของเอกสารได้ ซึ่งให้ความหมายเจาะจงมากขึ้นกว่าการใช้ดัชนีเดี่ยวๆ เช่น คำว่า “Programming Language “ วลี จะเป็นการรวบรวมคำ



ระบบ Smart

วิธีการค้นหาวลีที่สำคัญโดยอาศัย **สถิติ**
(Statistical Phrase
Detection) คำนึงเฉพาะคำที่
เกิดขึ้นพร้อมกันบ่อยๆ ในเอกสาร ไม่
สนใจรูปแบบของประโยค

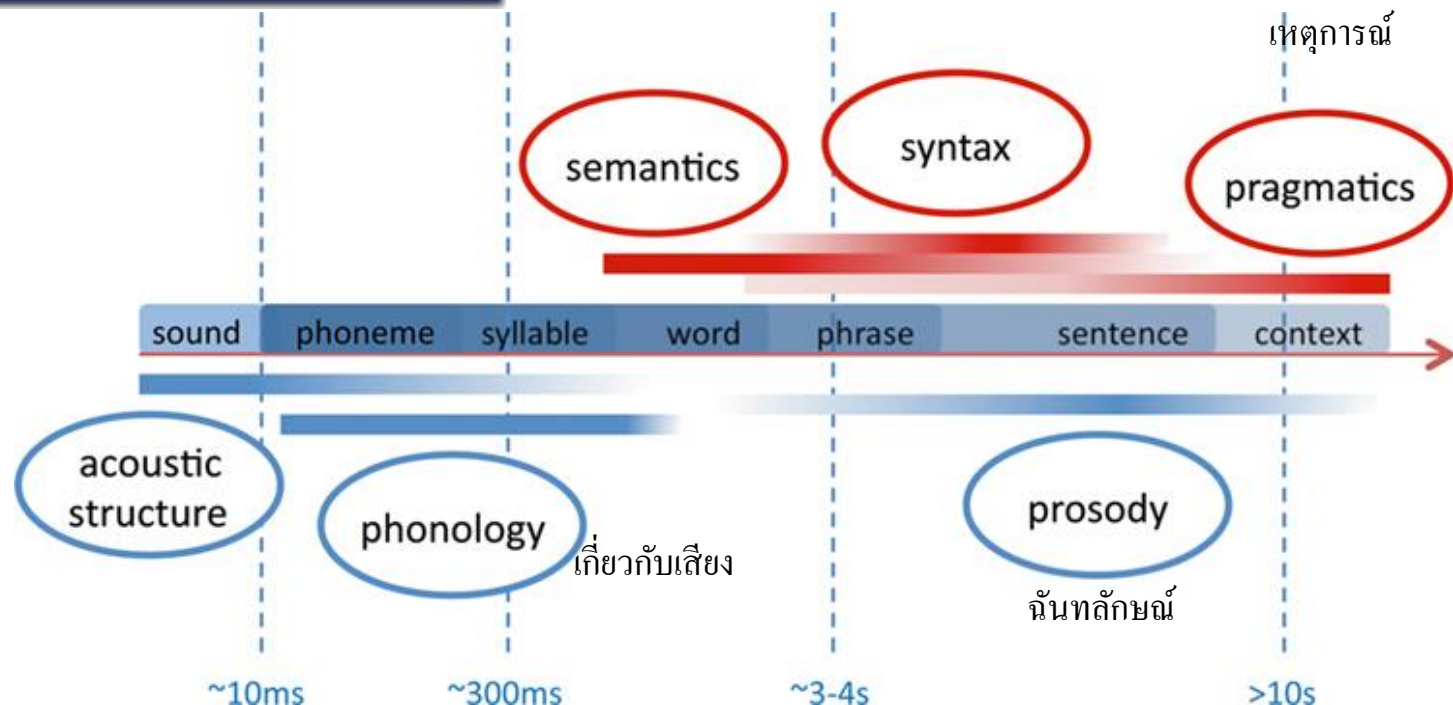
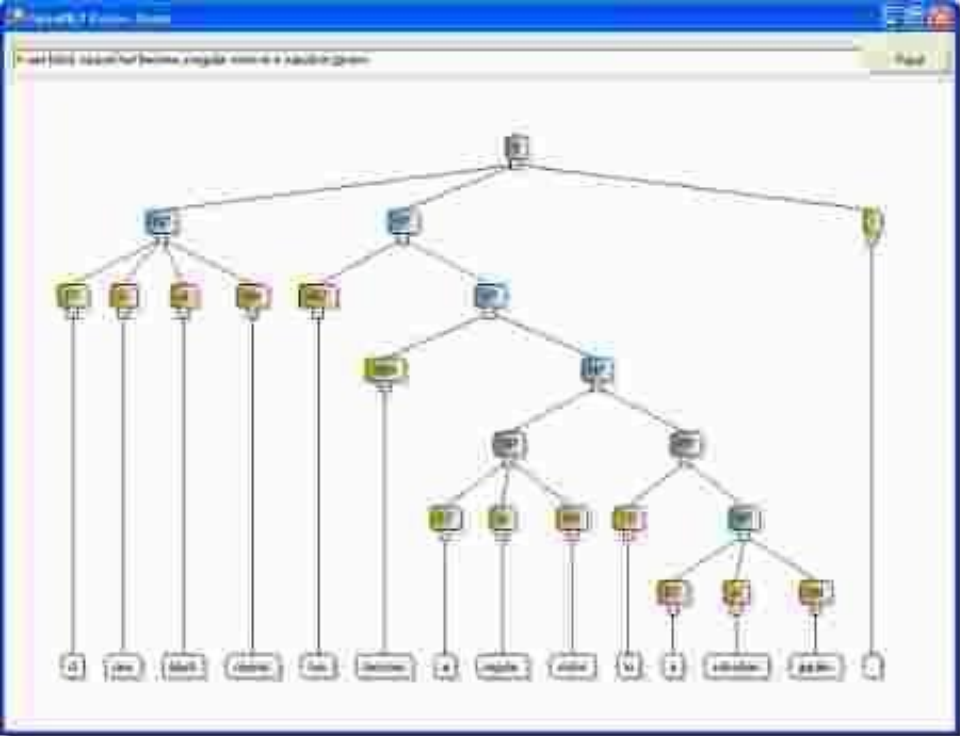
วิธีการค้นหาวลีที่สำคัญโดยอาศัย
รูปแบบ (Syntactical
Phrase Detection)
คำนึงถึงลักษณะเฉพาะส่วนประกอบ
ของวลี โดยเพิ่มข้อจำกัดในการ
ประกอบกันของคำ **โดยนิยามรูปแบบ**
เงื่อนไข ข้อจำกัด

Formulaic phrase selection

Navigate Global scaling Local scaling Other

AB DESCRIBES THE ;
AB DISCUSSES THE ;
AB PAPER PRESENTED AT THE ;
ANNUAL CONFERENCE ;
DATA BASE ;
FOR ABSTRACTS OF ;
IN ORDER TO ;
INFORMATION SCIENCE ;
INFORMATION SERVICES ;
LIBRARY AND INFORMATION ;
LIBRARY ASSOCIATION ;
NUMBERS - ;
PUBLIC LIBRARIES ;
PUBLIC LIBRARY ;
SEE THE FOLLOWING ;
THE DEVELOPMENT OF ;
THE ROLE OF ;

Syntactical Phrase



รูป:แสดง Top Phrases from TIPSTER

65824 United States
61327 Article Type
33864 Los Angeles
18062 Hong Kong
17788 North Korea
17308 New York
15513 San Diego
15009 Orange County
12869 prime minister
12799 first time
12067 Soviet Union
10811 Russian Federation
9912 United Nations
8127 Southern California
7640 South Korea
7620 end recording
7524 European Union
7436 South Africa
7362 San Francisco
7086 news conference
6792 City Council
6348 Middle East
6157 peace process
5955 human rights
5837 White House

5778 long time
5776 Armed Forces
5636 Santa Ana
5619 Foreign Ministry
5527 Bosnia-Herzegovina
5458 words indistinct
5452 international community
5443 vice president
5247 Security Council
5098 North Korean
5023 Long Beach
4981 Central Committee
4872 economic development
4808 President Bush
4652 press conference
4602 first half
4565 second half
4495 nuclear weapons
4448 UN Security Council
4426 South Korean
4219 first quarter
4166 Los Angeles County
4107 State Duma
4085 State Council
3969 market economy
3941 World War II

Hierarchical Dictionary

- เป็นพจนานุกรมที่มีรูปแบบคล้ายๆ กับ Tree และมีการแบ่งประเภท
คำตามลำดับชั้นความสำคัญ

::: METHODS
 :: MEASUREMENT TECHNIQUES
 : TESTING
 COMPARATIVE TESTING

: FUNDAMENTAL CONCEPTS
 COMPENSATION (CONCEPT)

: EXPENDITURES
 COMPENSATION (REMUNERATION)

: EDUCATION
 COMPENSATORY EDUCATION

: ABILITY
 COMPETENCE
 . INTERPERSONAL COMPETENCE
 . MINIMUM COMPETENCIES

: EDUCATION
 COMPETENCY BASED EDUCATION
 . COMPETENCY BASED TEACHER
 EDUCATION

::: EDUCATION
 :: PROFESSIONAL EDUCATION
 : TEACHER EDUCATION
 : EDUCATION
 : COMPETENCY BASED
 EDUCATION
 COMPETENCY BASED TEACHER
 EDUCATION

: BEHAVIOR
 COMPETITION
 . COMPETITIVE SELECTION

: SELECTION
 :: BEHAVIOR
 : COMPETITION
 COMPETITIVE SELECTION

::: BEHAVIOR
 :: COOPERATION
 : COMPARATIVE PSYCHOLOGY

: LINGUISTICS
 COMPUTATIONAL LINGUISTICS
 . MACHINE TRANSLATION

: DESIGN
 COMPUTER ASSISTED DESIGN

:::: METHODS
 :::: EDUCATIONAL METHODS
 :: TEACHING METHODS
 : PROGRAMED INSTRUCTION
 : COMPUTER USES IN
 EDUCATION
 COMPUTER ASSISTED
 INSTRUCTION

:: TECHNOLOGY
 . MANUFACTURING
 COMPUTER ASSISTED
 MANUFACTURING

:::: METHODS
 :: MEASUREMENT TECHNIQUES
 : TESTING
 : COMPUTER USES IN
 EDUCATION
 COMPUTER ASSISTED TESTING

:: FACILITIES
 : RESOURCE CENTERS
 COMPUTER CENTERS

:: ACTIVITIES
 : GAMES
 COMPUTER GAMES

:::: LIBERAL ARTS
 :: HUMANITIES
 :: FINE ARTS
 :: VISUAL ARTS
 : GRAPHIC ARTS
 COMPUTER GRAPHICS

: TECHNOLOGICAL LITERACY
 COMPUTER LITERACY

:: EQUIPMENT
 : ELECTRONIC EQUIPMENT
 COMPUTERS
 . ANALOG COMPUTERS
 . DIGITAL COMPUTERS
 . MICROCOMPUTERS
 . MINICOMPUTERS

:: LIBERAL ARTS
 :: SCIENCES
 : INFORMATION SCIENCE
 COMPUTER SCIENCE
 . PROGRAMING

:: EDUCATION
 : PROFESSIONAL EDUCATION
 : INFORMATION SCIENCE
 EDUCATION
 COMPUTER SCIENCE EDUCATION

:: METHODS
 : SIMULATION
 COMPUTER SIMULATION

:: STANDARDS
 : SPECIFICATIONS
 COMPUTER SOFTWARE
 . COURSEWARE
 . DATABASE MANAGEMENT
 SYSTEMS
 . MENU DRIVEN SOFTWARE

:: DEVELOPMENT
 : MATERIAL DEVELOPMENT
 COMPUTER SOFTWARE
 DEVELOPMENT
 . PROGRAMING

: EVALUATION
 COMPUTER SOFTWARE EVALUATION

: PUBLICATIONS
 COMPUTER SOFTWARE REVIEWS

. SELECTION

CONCEPTUAL SCHEMES (1967
 1980)

:::: INDIVIDUAL
 CHARACTERISTICS
 :: PSYCHOLOGICAL
 CHARACTERISTICS
 : COGNITIVE STYLE
 CONCEPTUAL TEMPO

:: ACTIVITIES
 : MUSIC ACTIVITIES
 CONCERTS

:::: STANDARDS
 :: CRITERIA
 : EVALUATION CRITERIA
 : VALIDITY
 CONCURRENT VALIDITY

CONDITIONING
 . BEHAVIOR MODIFICATION
 . CONTINGENCY MANAGEMENT
 . DESENSITIZATION
 . CLASSICAL CONDITIONING
 . OPERANT CONDITIONING
 . VERBAL OPERANT
 CONDITIONING

:: PUBLICATIONS
 : REPORTS
 CONFERENCE PAPERS

:: PUBLICATIONS
 : SERIALS
 CONFERENCE PROCEEDINGS

CONFERENCE REPORTS (1967
 1980)

CONFERENCES
 . PARENT CONFERENCES
 . PARENT TEACHER
 CONFERENCES

กระบวนการขจัดความซ้ำซ้อน (Normalization) ของข้อความในหลายๆระดับ

- ขจัดเอาคำฟุ่มเฟือยและไม่มีความหมายพิเศษในการแสดงลักษณะเด่นของข้อความออกไป คำที่เหลือก็จะเป็นคำสำคัญ (Significant Words)
- นำเอาคำสำคัญต่างๆที่มีความหมายเดียวกันมาจัดรวมเป็นกลุ่ม และแต่ละกลุ่มจะมีคำสำคัญหนึ่งที่ใช้เป็นชื่อกลุ่ม และใช้ชื่อกลุ่มนี้เป็นดัชนีหรือคีย์เวิร์ด
- การให้น้ำหนักดัชนี ซึ่งอาจกระทำโดย
 - ☐ ใช้ความถี่ของคำในเอกสารหรือใช้ความยาวของเอกสาร
 - ☐ ใช้ความถี่ของคำในกลุ่มเอกสารทั้งหมด
 - ☐ ใช้จำนวนเอกสารที่มีคำๆนั้นอยู่ในกลุ่มเอกสารทั้งหมด

ตัวอย่าง แสดงการสร้างดัชนีหรือคำสำคัญ

- **Original text:**

John Davenport, ~~52 years old~~, was appointed chief executive officer ~~of this~~ international telecommunications concern's ~~U.S.~~ subsidiary, Cable & Wireless North America Inc. ~~Mr.~~ Davenport, who succeeds John Zrno, is currently general manager ~~of the~~ group's operations in Bermuda.

- **One indexing result:**

john davenport appoint chief executive officer international telecommunication concern subsidiary cable wireless north america davenport succeed john zrno current general manager group operation bermuda

- **Another possibility:**

John_Davenport #person 52 years_old #age appoint chief_executive_officer international telecommunication concern #USA subsidiary Cable_&_Wireless_North_America #company Davenport #person succeed John_Zrno #person general_manager group operation Bermuda #foreigncountry

Sample Job Posting

Subject: US-TN-SOFTWARE PROGRAMMER

Date: 17 Nov 1996 17:37:29 GMT

Organization: Reference.Com Posting Service

Message-ID: <56nigp\$mrs@bilbo.reference.com>

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer 5 years or more experience with PC Based Voice Mail, but will consider as little as 2 years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is DOS. May go to OS-2 or UNIX in future.

Please reply to:

Kim Anderson

AdNET

(901) 458-2888 fax

kimander@memphisonline.com

Extracted Job Template

computer_science_job
id: 56nigp\$mrs@bilbo.reference.com
title: SOFTWARE PROGRAMMER
salary:
company:
recruiter:
state: TN
city:
country: US
language: C
platform: PC \ DOS \ OS-2 \ UNIX
application:
area: Voice Mail
req_years_experience: 2
desired_years_experience: 5
req_degree:
desired_degree:
post_date: 17 Nov 1996

Amazon Book Description

....

</td></tr>

</table>

<b class="sans">The Age of Spiritual Machines : When Computers Exceed Human Intelligence

by <a href="/exec/obidos/search-handle-url/index=books&field-author=

Kurzweil%2C%20Ray/002-6235079-4593641">

Ray Kurzweil

List Price: \$14.95

Our Price: \$11.96

You Save: \$2.99

(20%)

<p>



Extracted Book Template

Title: **The Age of Spiritual Machines :**
When Computers Exceed Human Intelligence

Author: **Ray Kurzweil**

List-Price: **\$14.95**

Price: **\$11.96**

:

:

Query processing: AND

■ Consider processing the query:

Brutus AND Caesar

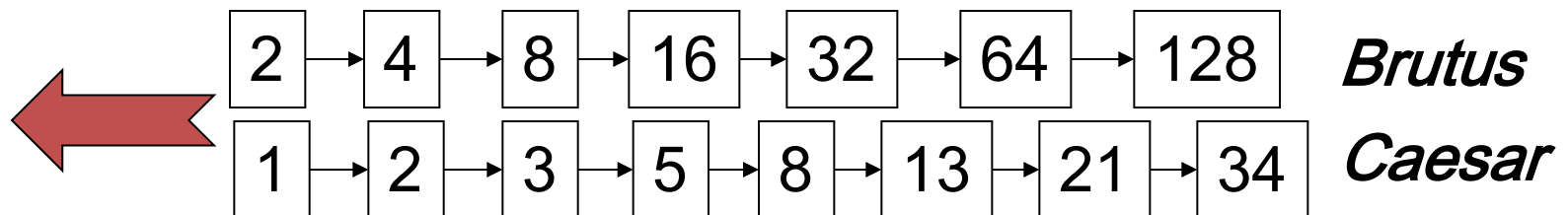
□ Locate ***Brutus*** in the Dictionary;

■ Retrieve its postings.

□ Locate ***Caesar*** in the Dictionary;

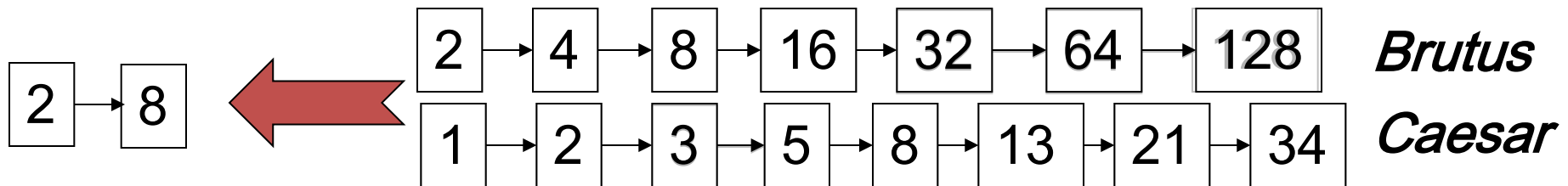
■ Retrieve its postings.

□ “Merge” the two postings:



The merge

- Walk through the two postings simultaneously, in time linear in the total number of postings entries



If list lengths are x and y , merge takes $O(x+y)$ operations.
Crucial: postings sorted by docID.

Intersecting two postings lists (a “merge” algorithm)

INTERSECT(p_1, p_2)

```
1  answer  $\leftarrow \langle \rangle$ 
2  while  $p_1 \neq \text{NIL}$  and  $p_2 \neq \text{NIL}$ 
3  do if  $\text{docID}(p_1) = \text{docID}(p_2)$ 
4      then ADD(answer,  $\text{docID}(p_1)$ )
5           $p_1 \leftarrow \text{next}(p_1)$ 
6           $p_2 \leftarrow \text{next}(p_2)$ 
7      else if  $\text{docID}(p_1) < \text{docID}(p_2)$ 
8          then  $p_1 \leftarrow \text{next}(p_1)$ 
9          else  $p_2 \leftarrow \text{next}(p_2)$ 
10 return answer
```