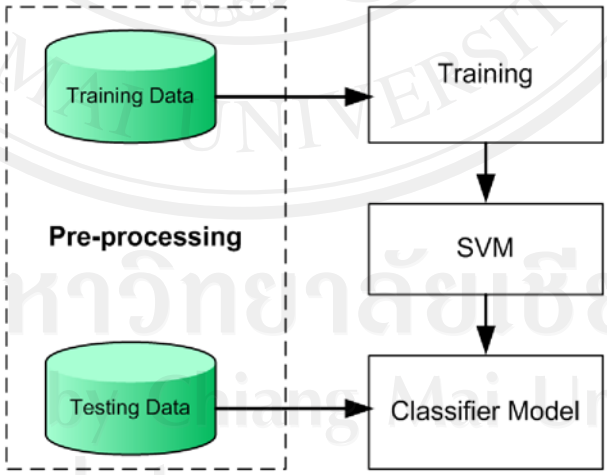
บทที่ 4

การจัดกลุ่มเอกสารด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน

สำหรับบทนี้จะเป็นการกล่าวถึงขั้นตอนการจัดกลุ่มเอกสารโดยใช้ ซัพพอร์ตเวกเตอร์แมชชีน โดยเอกสารได้มาจากการครอว์เลอร์ของเว็บสไปเดอร์และการเตรียมเอกสารในบทที่ 3 ซึ่งขั้นตอนแรกของการวิจัยคือขั้นตอนก่อนการประมวลผล จากนั้นจึงนำเข้าสู่การจัดกลุ่มเอกสารโดยจะใช้อัลกอริทึม ซัพพอร์ตเวกเตอร์แมชชีน ก่อนที่จะสร้างโมเดลสำหรับการจัดกลุ่มเอกสาร

4.1 การจัดกลุ่มเอกสารโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน

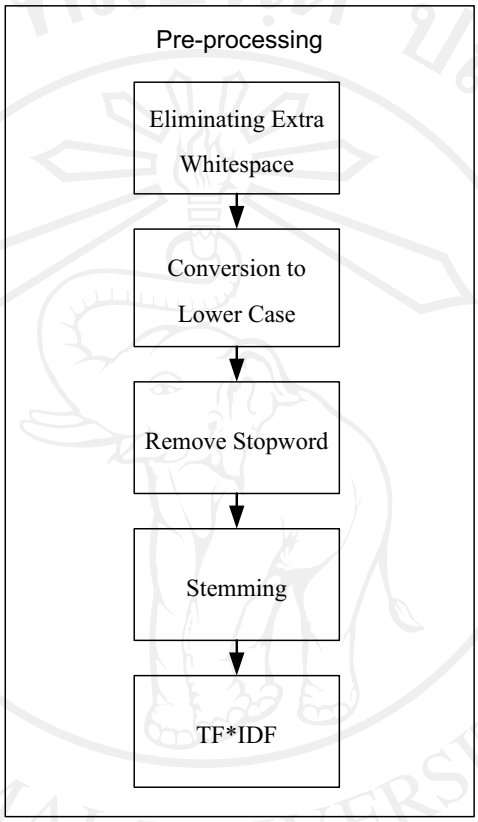
เมื่อได้เอกสารในรูปแบบที่ต้องการจากบทที่ 3 ในขั้นตอนการเตรียมข้อมูลเรียบร้อยแล้ว โดยนำเอกสารมาสร้างโมเดลจำนวน 1,000 เอกสาร และทดสอบจำนวน 500 เอกสาร ทางผู้วิจัยดำเนินการการจัดกลุ่มเอกสารโดยใช้ ซัพพอร์ตเวกเตอร์แมชชีน ซึ่งการจัดกลุ่มเอกสารข่าวจาก CNN ประกอบด้วยเอกสารกลุ่มข่าว 5 กลุ่มคือ 1.ข่าวธุรกิจ 2.ข่าวสุขภาพ 3.ข่าวกีฬา 4.ข่าวท่องเที่ยว 5.ข่าวคอมพิวเตอร์ ซึ่งในการทดสอบสำหรับงานวิจัยนี้ ได้ใช้โปรแกรมอาร์ (R Language) เวอร์ชั่น 2.8.1 แพกเก็จที่ใช้ทดลองคือ แพกเก็จทีเอ็มและแพกเก็จอี 1071 เพื่อใช้ในการทดสอบ ซึ่งขั้นตอนการดำเนินงานวิจัยสามารถแสดงได้ดังรูป 4.1



รูป 4.1 ขั้นตอนการของงานวิจัย

4.1.1 ขั้นตอนก่อนการประมวลผล (Pre-processing)

ขั้นตอนแรกในการสร้างโมเดลสำหรับการจัดกลุ่มเอกสารข่าว คือการตัดคำและการสกัดคำหลัก (Word Segmentation) และการสร้างตัวแทนเวกเตอร์หรือการหาน้ำหนักของคำ (Term word weighting) โดยมีขั้นตอนดังรูป 4.2



รูป 4.2 ขั้นตอนก่อนการประมวลผล

จากรูป 4.2 ขั้นตอนก่อนการประมวลผล เป็นขั้นตอนในการเตรียมเอกสารที่จะนำมาใช้ในการจัดกลุ่มเอกสาร โดยมี 5 ขั้นตอนดังนี้ 1.การตัดคำจากช่องว่าง 2.การเปลี่ยนอักษรตัวพิมพ์ใหญ่ให้เป็นตัวพิมพ์เล็ก 3.การกำจัดคำศัพท์ที่ไม่มีผลกระทบต่อความหมายโดยทั่วไป 4.การตัดคำเพื่อหารากศัพท์ 5.การสร้างตัวแทนเวกเตอร์หรือการหาน้ำหนักของคำ

Marketing your business school - CNN.com LONDON, England (CNN) -- A great deal of attention is paid to the tricky business of how aspirant MBAs can identify and then get into the best courses at top business schools. But there is another side to this, one that is less often remarked on but is equally crucial for the whole process -- how business schools can attract students. At the sharp end of this equation, business schools have to make sure they have sufficient enrollments each year to pay the wages of their professors, not to mention the host of other bills involved in running such an institution.

รูป 4.3 ตัวอย่างหลังการวิเคราะห์คำในประโยคของเอกสารจากบทที่ 3

จากรูป 4.3 ตัวอย่างหลังการวิเคราะห์คำในประโยคของเอกสารจากบทที่ 3 เป็นเอกสารที่ได้ทำการวิเคราะห์คำในประโยคจากเอกสารในรูป 3.5 ตัวอย่างก่อนการวิเคราะห์คำในประโยคของเอกสารในบทที่ 3

1) ทำการกำจัดเครื่องหมายต่างๆ และเปลี่ยนอักษรตัวพิมพ์ใหญ่ให้เป็นตัวพิมพ์เล็ก

marketing your business school - cnn.com london, england (cnn) -- a great deal of attention is paid to the tricky business of how aspirant mbas can identify and then get into the best courses at top business schools. but there is another side to this, one that is less often remarked on but is equally crucial for the whole process -- how business schools can attract students. at the sharp end of this equation, business schools have to make sure they have sufficient enrollments each year to pay the wages of their professors, not to mention the host of other bills involved in running such an institution.

รูป 4.4 ผลที่ได้จากการกำจัดเครื่องหมายต่างๆ และเปลี่ยนอักษรตัวพิมพ์ใหญ่ให้เป็นตัวพิมพ์เล็ก

2) ทำการกำจัดคำศัพท์ที่ไม่มีผลกระทบต่อความหมายโดยทั่วไป (Remove Stopword)

marketing business school - cnn.com london, england (cnn) -- deal attention paid tricky business aspirant mbas identify courses top business schools. this, remarked equally crucial process -- business schools attract students. sharp equation, business schools sufficient enrollments pay wages professors, mention host bills involved running institution.

รูป 4.5 ผลที่ได้จากการกำจัดคำศัพท์

3) การตัดคำเพื่อหารากศัพท์ (Word Stemming)

market busi school - cnn.com london, england (cnn) -- deal attent paid tricki busi aspir mbas identifi cours top busi schools. this, remark equal crucial process -- busi school attract students. sharp equation, busi school suffici enrol pay wage professors, mention host bill involv run institution.

รูป 4.6 ผลที่ได้จากการหารากศัพท์

4) การสร้างคำสำคัญ-เอกสาร(Keyword) จากชุดสำหรับเรียนรู้

[1] "aspir" "attent" "attract" "bill" "busi" "cnn" "com" "cours" "crucial" "deal"
[11] "england" "enrol" "equal" "equation" "host" "identifi" "institution" "involv" "london" "market"
[21] "mbas" "mention" "paid" "pay" "process" "professors" "remark" "run" "school" "schools"
[31] "sharp" "students" "suffici" "this" "top" "tricki" "wage"

รูป 4.7 ผลที่ได้จากการสร้างคำสำคัญ-เอกสาร

จากขั้นตอนทำการกำจัดเครื่องหมายต่างๆ และเปลี่ยนอักษรตัวพิมพ์ใหญ่ให้เป็นตัวพิมพ์เล็กการกำจัดคำศัพท์ที่ไม่มีผลกระทบต่อความหมายโดยทั่วไป การตัดคำเพื่อหารากศัพท์ การสร้างคำสำคัญ-เอกสาร (Keyword) เราจะได้คำสำคัญ -เอกสารในแต่ละกลุ่มเอกสารข่าว โดยเลือกคำสำคัญที่ปรากฏในเอกสารทั้งหมดมากกว่า 4 เอกสาร การเลือกจำนวนคำสำคัญนี้ได้จากการทดลองจัดกลุ่มเอกสาร เพื่อให้สามารถนำไปคำนวณในคอมพิวเตอร์ส่วนบุคคล และให้ประสิทธิภาพการจัดกลุ่มที่ดีในระดับที่ยอมรับได้ ดังนี้

1. เอกสารข้อความข่าวธุรกิจ

averag, baggag, british, business, countri, european, execut, experience, expert, financi, foreign, global, manag, market, mba, media, parliament, partner, pay, peopl, polici, popular, pressur, qualiti, rank, recruit, report, salari, secret, senior, social, staff,..etc

2. เอกสารข้อความข่าวสุขภาพ

afghanistan, age, anger, approv, bargain, behavior, california, cancer, case, chart, concern, contain, critic, doctor, energy, florida, haiti, hand, head, healthi, healthier, heart, inject, institut, journal, louisiana, medic, mediterranean, monitor, panel, peak, pitch, plastic, portion, prize, reduct, restaur, screen, serious, sex, similar, spectrum, spot, spread, strain, surgery, symptom, tongue, treat, vaccin, virus, wear,.. etc

3. เอกสารข้อความข่าวกีฬา

basketbal, boom, club, footbal, goal, hamilton, heavyweight, isinbayeva, jockey, liverpool, nba, park, quarter, rafael, rider, ruptur, score, sea, sebastian, stadium, touchdown, tournament, wba, wbc, weather, welsh, winter,.. etc

4. เอกสารข้อความข่าวท่องเที่ยว

airlines, airport, airway, beach, depart, earthquak, flight, hotel, photo, resort, ticket, tour, tourism, traffic, travel, visa, visitor, volcan, volcano, weather,.. etc

5. เอกสารข้อความข่าวคอมพิวเตอร์

camera, comput, hack, hacker, iphon, livestock, machin, metabolix, micro, microsoft, palm, passage, password, startup, stream, tech, technology, twitter, virtual, warm, window, wireless, xbox, xerox, youtub,.. etc

เมื่อได้คำสำคัญ-เอกสาร (Keyword) เรียบร้อยแล้ว ขั้นตอนการต่อไปคือการให้น้ำหนักของคำในเอกสาร (Term word weighting) เป็นขั้นตอนการที่จะนำมาหาจำนวนเอกสารที่มีคำนั้นๆ ปรากฏอยู่ (Document Frequency: DF) เพื่อนำไปสู่การหารวมถึงค่าส่วนกลับของเอกสาร (Inverse Document Frequency: IDF) ตามสมการ $IDF = 1 - \log(\left|\frac{N}{DF}\right|)$ เพื่อนำค่า IDF ไปใช้ในการคำนวณค่าน้ำหนักของคำเหล่านั้นในแต่ละเอกสารตามสมการของ TF.IDF โดย TF หรือ Term Frequency คือ ค่าความถี่ของคำนั้นๆ ที่ปรากฏอยู่ในแต่ละเอกสาร ซึ่งการคำนวณค่าน้ำหนักของคำสามารถแสดงได้ในตาราง 4.1

การเตรียมเอกสารก่อนนำเข้าสู่การประมวลผลด้วย ซัพพอร์ตเวกเตอร์แมชชีน สมมติว่าในคลังเอกสาร (Document Collection) มีเอกสาร 2 กลุ่มได้แก่ ข่าวคอมพิวเตอร์ และข่าวสุขภาพ

เอกสารข่าวคอมพิวเตอร์

Windows Media Player 10- CNN.com Here to Install Silverlight United States Change | All Microsoft Sites Windows Media Home | Windows Media Worldwide For Home Windows Media Player FAQ Extras Music & Video Cool Devices For Professionals Enterprise & A/V Pros Resources Downloads Help & Support Community Windows Family Warning: Parts of this page require the use of scripts, which your browser does not currently allow. If script is not enabled in your browser, you will not be able to access some parts of the site and some content may not appear. To enable script in Internet Explorer 5.01 or later On the Tools menu, select Internet Options .Under the Security tab, click the Internet Web content zone. Do one of the following: To use the default security setting (Medium), which allows script, click Default Level . To turn on script without changing other security settings, click Custom Level . In the Settings list, scroll down to Active scripting and click Enable . To enable script in other browsers Consult the documentation for your browser for steps to enable script. Refresh this page after you have enabled scripting. Windows Media Player 10 gives you more music and more choices, and for the first time makes it possible to sync high-quality music, video, and photos to the latest portable devices. WHAT'S NEW New Streamlined Design A whole new look giving you quicker and easier access to your favorite digital media. …etc

เอกสารข่าวสุขภาพ

Patients demand: 'Give us our damned data' - CNN.com Patients demand: 'Give us our damned data' By Elizabeth Cohen , CNN Senior Medical Correspondent Artist Regina Holliday gives Jen McCabe a jacket that reads "In emergency break glass ceiling. Demand access to your medical record." STORY HIGHLIGHTS There have been "repeated" complaints to the Department of Health and Human Services Expert encourages patients to get their records before they're admitted to the hospital The HIPAA law gives hospitals and doctors 30 days to respond to a request for medical records Your doctor doesn't have to give you access to everything in your record (CNN) -- For five days as her husband lay in his hospital bed suffering from kidney cancer, Regina Holliday begged doctors and nurses for his medical records, and for five days she never received them. On the sixth day, her husband needed to be transferred to another hospital -- without his complete medical records. "When Fred arrived at the second hospital, they couldn't give him any pain medication because they didn't know what drugs he already had in his system, and they didn't want to overdose him," says Holliday, who lives in Washington. "For six hours he was in pain, panicking, while I ran back to the first hospital and got the rest of the records.…etc

นำคำทั้งหมดที่อยู่ในเอกสารทั้ง 4 เอกสาร มาตัดคำแบบพจนานุกรมด้วยวิธีเทียบคำที่ยาวที่สุด ซึ่งจะได้คำทั้งหมดที่อยู่ในเอกสาร จากนั้นหาความถี่ของคำทั้งหมดที่อยู่ในเอกสารทั้งหมด จะได้ดังตาราง 4.1

ตาราง 4.1 ผลลัพธ์การหาความถี่ของคำที่อยู่ในเอกสารทั้งหมดและส่วนกลับของเอกสาร

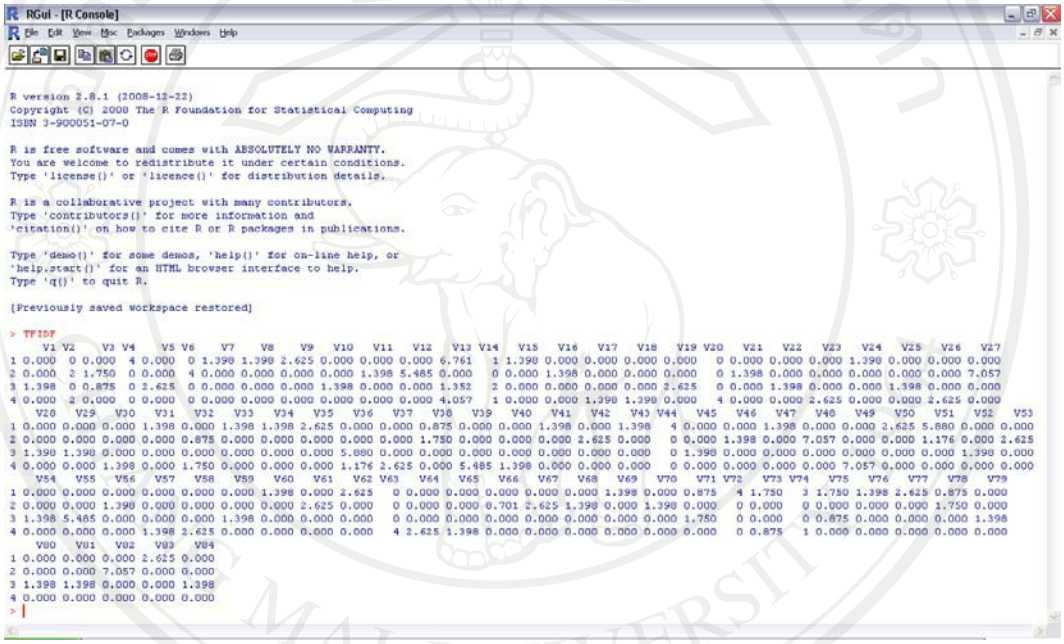
ลำดับของคำ	Term word	TF	IDF	ลำดับของคำ	Term word	TF	IDF
1	accept	2	0.699	43	iphon	2	0.699
2	access	4	1.000	44	keyboard	4	1.000
3	allow	3	0.875	45	lemir	2	0.699
4	arm	4	1.000	46	level	2	0.699
5	bad	3	0.875	47	love	2	0.699
6	browser	4	1.000	48	media	6	1.176
7	button	2	0.699	49	medic	6	1.176
8	call	2	0.699	50	mellon	3	0.875
9	carnegi	3	0.875	51	microsoft	6	1.176
10	champion	2	0.699	52	movement	2	0.699
11	chang	2	0.699	53	music	3	0.875
12	click	5	1.097	54	old	2	0.699
13	cnn	9	1.352	55	overweight	5	1.097
14	com	4	1.000	56	page	2	0.699
15	comment	2	0.699	57	pain	2	0.699
16	content	2	0.699	58	patient	3	0.875
17	damn	2	0.699	59	people	2	0.699
18	data	2	0.699	60	person	2	0.699
19	daughter	3	0.875	61	player	3	0.875
20	day	4	1.000	62	prototyp	3	0.875
21	default	2	0.699	63	record	4	1.000
22	definit	2	0.699	64	records	3	0.875
23	demand	3	0.875	65	regina	2	0.699
24	develop	2	0.699	66	script	7	1.243
25	dictat	2	0.699	67	secur	3	0.875
26	doctor	2	0.699	68	set	2	0.699
27	enabl	6	1.176	69	share	2	0.699
28	fat	2	0.699	70	site	2	0.699
29	fit	2	0.699	71	size	3	0.875
30	five	2	0.699	72	skinput	4	1.000
31	forearm	2	0.699	73	system	3	0.875
32	give	3	0.875	74	them	4	1.000
33	hand	2	0.699	75	topic	3	0.875
34	hands	2	0.699	76	touch	2	0.699
35	harrison	3	0.875	77	univers	3	0.875
36	health	6	1.176	78	video	3	0.875
37	holliday	3	0.875	79	waist	2	0.699
38	home	3	0.875	80	weight	2	0.699
39	hospit	5	1.097	81	well	2	0.699
40	husband	2	0.699	82	window	6	1.176

ตาราง 4.1 ผลลัพธ์การหาความถี่ของคำที่อยู่ในเอกสารทั้งหมดและส่วนกลับของเอกสาร (ต่อ)

41	intern	2	0.699	83	world	3	0.875
42	internet	3	0.875	84	year	2	0.699

จากตาราง 1 ค่า IDF คือ ส่วนกลับของเอกสารได้จากสมการ $IDF = 1 - \log(\left|\frac{N}{DF}\right|)$ โดยอ้างอิงจากทฤษฎีที่กล่าวมาแล้วในบทที่ 2 จากนั้นหาน้ำหนักของคำในเอกสาร (Term word weighting) จากสมการที่ (2.3) ในบทที่ 2 จะได้ผลลัพธ์ดังตาราง 4.2

ตาราง 4.2 ผลลัพธ์ที่ได้จากการหาน้ำหนักของคำในเอกสาร



ลำดับของคำ	1	2	3	4	5	6	7	8
D1	0.000	0.000	0.000	4.000	0.000	0.000	1.398	1.398
D2	0.000	2.000	1.750	0.000	0.000	4.000	0.000	0.000
D3	1.398	0.000	0.875	0.000	2.625	0.000	0.000	0.000
D4	0.000	2.000	0.000	0.000	0.000	0.000	0.000	0.000

ลำดับของคำ	9	10	11	12	13	14	15	16
D1	2.625	0.000	0.000	0.000	6.761	1.000	1.398	0.000
D2	0.000	0.000	1.398	5.485	0.000	0.000	0.000	1.398
D3	0.000	1.398	0.000	0.000	1.352	2.000	0.000	0.000
D4	0.000	0.000	0.000	0.000	4.057	1.000	0.000	0.000

ตาราง 4.2 ผลลัพธ์ที่ได้จากการหาน้ำหนักของคำในเอกสาร (ต่อ)

ลำดับของคำ	17	18	19	20	21	22	23	24
D1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.398
D2	0.000	0.000	0.000	0.000	1.398	0.000	0.000	0.000
D3	0.000	0.000	2.625	0.000	0.000	1.398	0.000	0.000
D4	1.398	1.398	0.000	4.000	0.000	0.000	2.625	0.000

ลำดับของคำ	25	26	27	28	29	30	31	32
D1	0.000	0.000	0.000	0.000	0.000	0.000	1.398	0.000
D2	0.000	0.000	7.057	0.000	0.000	0.000	0.000	0.875
D3	1.398	0.000	0.000	1.398	1.398	0.000	0.000	0.000
D4	0.000	2.625	0.000	0.000	0.000	1.398	0.000	1.750

ลำดับของคำ	33	34	35	36	37	38	39	40
D1	1.398	1.398	2.625	0.000	0.000	0.875	0.000	0.000
D2	0.000	0.000	0.000	0.000	0.000	1.750	0.000	0.000
D3	0.000	0.000	0.000	5.880	0.000	0.000	0.000	0.000
D4	0.000	0.000	0.000	1.176	2.625	0.000	5.485	1.398

ลำดับของคำ	41	42	43	44	45	46	47	48
D1	1.398	0.000	1.398	4.000	0.000	0.000	1.398	0.000
D2	0.000	2.625	0.000	0.000	0.000	1.398	0.000	7.057
D3	0.000	0.000	0.000	0.000	1.398	0.000	0.000	0.000
D4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

ลำดับของคำ	49	50	51	52	53	54	55	56
D1	0.000	2.625	5.880	0.000	0.000	0.000	0.000	0.000
D2	0.000	0.000	1.176	0.000	2.625	0.000	0.000	1.398
D3	0.000	0.000	0.000	1.398	0.000	1.398	5.485	0.000
D4	7.057	0.000	0.000	0.000	0.000	0.000	0.000	0.000

ลำดับของคำ	57	58	59	60	61	62	63	64
D1	0.000	0.000	0.000	1.398	0.000	2.625	0.000	0.000
D2	0.000	0.000	0.000	0.000	2.625	0.000	0.000	0.000
D3	0.000	0.000	1.398	0.000	0.000	0.000	0.000	0.000
D4	1.398	2.625	0.000	0.000	0.000	0.000	4.000	2.625

ตาราง 4.2 ผลลัพธ์ที่ได้จากการหาน้ำหนักของคำในเอกสาร (ต่อ)

ลำดับของคำ	65	66	67	68	69	70	71	72
D1	0.000	0.000	0.000	0.000	1.398	0.000	0.875	4.000
D2	0.000	8.701	2.625	1.398	0.000	1.398	0.000	0.000
D3	0.000	0.000	0.000	0.000	0.000	0.000	1.750	0.000
D4	1.398	0.000	0.000	0.000	0.000	0.000	0.000	0.000

ลำดับของคำ	73	74	75	76	77	78	79	80
D1	1.750	3.000	1.750	1.398	2.625	0.875	0.000	0.000
D2	0.000	0.000	0.000	0.000	0.000	1.750	0.000	0.000
D3	0.000	0.000	0.875	0.000	0.000	0.000	1.398	1.398
D4	0.875	1.000	0.000	0.000	0.000	0.000	0.000	0.000

ลำดับของคำ	81	82	83	84
D1	0.000	0.000	2.625	0.000
D2	0.000	7.057	0.000	0.000
D3	1.398	0.000	0.000	1.398
D4	0.000	0.000	0.000	0.000

4.1.2 การจัดกลุ่มเอกสารโดยจะใช้อัลกอริทึม Support Vector Machines (SVMs)

หลังจากที่ได้หาน้ำหนักของคำในเอกสาร (Term word weighting) ซึ่งจะได้ตัวแทนเอกสาร (Topic Identifier) แล้วจะจัดกลุ่มเอกสารโดยจะใช้อัลกอริทึม Support Vector Machine (SVMs) โดยจะมีขั้นตอนในการจัดกลุ่มเอกสารดังนี้

1. นำเอกสารที่อินพุตเข้ามาคำนวณหาค่า y ซึ่งค่าของ $y \in \{-1,1\}$ และ ค่า $x \in R^n$ จากสมการที่ (2.3) ในบทที่ 2 โดย

ถ้าค่าของ $w^T x + b > 0$ จะกำหนดให้ค่า y = +1 ซึ่งจะจัดอยู่ใน Class 1

ถ้าค่าของ $w^T x + b < 0$ จะกำหนดให้ค่า y = -1 ซึ่งจะจัดอยู่ใน Class 2

2. คำนวณหาเส้นตรงที่แบ่งเอกสารซึ่งเรียกว่า เส้น Optimal Hyperplane จากสมการที่ (2.4) ในบทที่ 2

3. นำค่าที่ได้จากข้อที่ 1 และ 2 ไปเขียนบนเส้นตรงตามแนวแกนตั้งและแกนนอน ดังรูป 2.12 ในบทที่ 2 เพื่อที่จะหาจุดที่ใกล้เส้น Optimal Hyperplane ซึ่งจุดที่อยู่เหนือเส้น Optimal Hyperplane จะเรียกว่า “ขอบล่าง” และใต้เส้น เรียกว่า “ขอบบน”

4. หาระยะทางระหว่างเส้นขอบทั้งสองโดยจะเลือกเอาค่าระยะทางที่ห่างจากเส้นตรง Optimal Hyperplane ที่น้อยที่สุดเป็นตัวแทนในการจัดกลุ่มเอกสาร โดยระยะทาง (d) หรือ maximum margin จากเส้นขอบ ณ จุด x_i ไปยัง Hyperplane หาได้จากสมการที่ (2.5) ในบทที่ 2

5. ในกรณีที่เวกเตอร์ตัวแทนเอกสารไม่สามารถแบ่งกลุ่มด้วยเส้นตรง (Linear Support Vector Machines) ดังรูป 2.6 ในบทที่ 2 ได้ คือข้อมูลนั้นอยู่ในรูปแบบของ Non-Linear Support Vector Machines ซึ่งข้อมูลในลักษณะนี้จะไม่สามารถแบ่งกลุ่มได้เนื่องจากไม่มีจุดที่แน่นอนที่จะใช้เป็นหลักในการแบ่ง Hyperplane ทั้งสองข้าง ดังนั้นจะต้องทำการปรับจุดก่อน โดยใช้ kernel function เพื่อให้การทำงานทำได้ง่ายขึ้นและทำให้การแบ่งข้อมูลที่ได้ถูกต้องยิ่งขึ้น ในการทำให้ข้อมูลเปลี่ยนไปอยู่ในรูปสามารถแบ่งกลุ่มด้วยเส้นตรง Linear Support Vector Machines ได้ โดยใช้ kernel function จากสมการที่ (2.6), (2.7) และ (2.8) ในบทที่ 2 โดยงานวิจัยฉบับนี้เลือกใช้ kernel function แบบ Radial basis function kernel (RBF)

6. หลังจากการทำการปรับจุดด้วย kernel function แล้วจะทำการแบ่งคลาสให้ข้อมูลนั้น โดยหาได้จากสมการ $f(x)=\sum \alpha_i y_i K(x_i, x) : \alpha_i$ คือน้ำหนักของคำ และ y คือ sign {-1,+1}

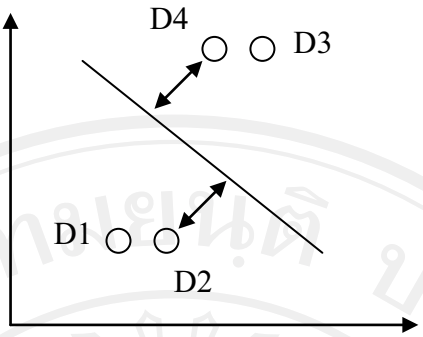
โดยข้อมูลที่ได้จากการคำนวณ จะพิจารณาว่าถ้า $f(x) > 0$ จะจัดอยู่ในคลาสที่ 1 ส่วนถ้า $f(x) < 0$ จะจัดให้อยู่ในคลาสที่ 2 หลังจากนั้นโปรแกรมก็จะทำการรู้จำการทำงานแบบนี้ ดังนั้นเมื่อข้อมูลมีหลายคลาส การทำงานหลายคลาสก็จะอาศัยลักษณะการทำงานของสองคลาสช่วยในการทำงานเบื้องต้นของการแบ่งคลาสและจะทำการแบ่งคลาสหลายคลาสโดยทำการแบ่งที่ละ 2 คลาสต่อไป

แสดงการคำนวณการจัดกลุ่มเอกสารด้วยอัลกอริทึม Support Vector Machines (SVMs) แบบ linear จากสมการ $\sum_{x=1}^{i} w^T x_i + b$

หาความถี่ของคำแต่ละเอกสารเพื่อหาเอกสารที่ตรงตาม คำสำคัญ -เอกสา ร สมมติว่ากำหนดให้คำของเอกสารได้แก่คำว่า => microsoft, video, doctor, fat

ตาราง 4.3 การหาความถี่ของคำ

คำ / เอกสารคำ	microsoft	video	doctor	fat	y
D1	5.880	0.875	0	0	1
D2	1.176	1.750	0	0	1
D3	0	0	0	1.398	-1
D4	0	0	1.398	0	-1
W	1.176	0.875	0.699	0.699	b= -1.803



รูป 4.8 การแบ่งกลุ่มด้วย Maximum Margin Hyperplane

จาก $wx + b = 0$

b_Q $= -wx$

$= - [(1.176\times1.176)+(0.875\times0.875)+(0.699\times0.699)+(0.699\times0.699)]$

$= - [0.138+0.765+0.489+0.489]$

$= - 1.881$

D1 $= wx + b$

$= [(1.175\times5.880)+(0.875\times0.875)+(0.699\times0)+(0.699\times0)]+(-2.157)$

$= 6.909+0.765+(-1.881) = 5.793$ จะได้ $y = 1$

D2 $= wx + b$

$=[(1.176\times1.176)+(0.875\times1.750)+(0.699\times0)+(0.699\times0)]+(2.157)$

$= 1.382+1.531+(-1.881) = 1.032$ จะได้ $y = 1$

D3 $= wx + b$

$= [(1.175\times0)+(0.875\times0)+(0.875\times0)+(0.699\times1.398)]+(-2.157)$

$= 0.977+(-1.881) = -0.904$ จะได้ $y = -1$

D4 $= wx + b$

$= [(1.175\times0)+(0.875\times0)+(0.699\times1.398)+(0.699\times0)]+(-1.803)$

$=0.977+(-1.881) = -0.904$ จะได้ $y = -1$

ในการที่จะเลือกว่าเอกสารอยู่ในกลุ่มใดจะเลือกจากค่า Maximum Margin ที่น้อยที่สุดเอกสารที่มีค่า $y = 1$ จะเลือกเอกสารที่มีค่า $wx + b$ ที่มีค่าน้อยที่สุด

$$\text{D1} = \frac{\left|w^T x_i + b\right|}{\left\|w^T\right\|}$$

$$= \left|\frac{1.032}{1.175 + 0.875 + 0.699 + 0.699}\right|$$

$$= 0.299$$

เอกสารที่มีค่า y = -1 จะเลือกเอกสารที่มีค่า wx + b ที่มีค่ามากที่สุด

$$D3 = \left| \frac{-0.904}{1.175 + 0.875 + 0.699 + 0.699} \right|$$

$$= 0.262$$

ดังนั้น สรุปได้ว่าจากเอกสารทั้ง 4 ชุด จะสามารถแบ่งกลุ่มได้ 2 กลุ่ม คือกลุ่มที่ 1 จะประกอบด้วยเอกสาร D1 และ D2 ขณะที่กลุ่มที่ 2 จะประกอบด้วยเอกสาร D3 และ D4

4.1.3 การสร้างโมเดลสำหรับการจัดกลุ่มเอกสาร

เมื่อทำการสร้างเมตริกซ์ความถี่ของคำสำคัญ -เอกสารของข้อมูลสำหรับเรียนรู้แล้ว ขั้นต่อไปคือการสร้างเมตริกซ์คำสำคัญ-เอกสารของข้อมูลสำหรับทดสอบ โดยทำตามขั้นตอนการสร้างเมตริกซ์ความถี่ของคำสำคัญ -เอกสารของข้อมูลสำหรับเรียนรู้ ในขั้นตอนการตัดคำและการให้น้ำหนักของคำในเอกสาร จากนั้นสร้างชุดข้อมูลสำหรับเรียนรู้ในการสร้างเมตริกซ์ความถี่ของคำสำคัญ-เอกสารของข้อมูลสำหรับทดสอบ เอกสารที่ใช้ในการสร้างโมเดลนั้นได้มาจากการครอว์เลอร์ของเว็บสไปเดอร์ในบทที่ 3 โดยนำเอกสารมาสร้างโมเดลจำนวน 1,000 เอกสาร ซึ่งสามารถดูตัวอย่างตามตาราง 4.4

ตาราง 4.4 ตัวอย่างเอกสารสำหรับการสร้างโมเดลสำหรับการจัดกลุ่มเอกสาร

ลำดับเอกสาร	Term1	Term2	Term3	Term4	Term5	Term6	Term ที่ n	Type
1	1.786	0	18.347	3.133	2.582	0	-	Business
2	1.786	0	0	0	0	0	-	Business
3	0	0	0	0	0	0	-	Business
4	0	0	0	0	0	0	-	Business
5	10.715	0	0	0	0	0	-	Business
6	7.144	0	0	0	0	0	-	Business
7	0	0	0	0	0	0	-	Business
8	0	0	0	0	0	0	-	Business
9	0	0	0	0	0	0	-	Business
10	0	0	0	0	0	0	-	Business
11	0	0	0	0	0	0	-	Computer
12	1.786	0	0	0	0	0	-	Computer
13	0	0	0	0	0	0	-	Computer
14	0	0	0	0	0	0	-	Computer
15	0	0	0	0	5.164	0	-	Computer
16	0	0	0	0	0	0	-	Computer
17	12.501	0	0	0	0	0	-	Computer
18	0	0	0	0	0	0	-	Computer

ตาราง 4.4 ตัวอย่างเอกสารสำหรับการสร้างโมเดลสำหรับการจัดกลุ่มเอกสาร (ต่อ)

19	1.786	2.442	2.293	0	5.164	9.079	-	Computer
20	5.358	2.442	0	12.532	2.582	0	-	Computer
21	0	0	0	0	2.582	0	-	Health
22	0	0	0	0	0	0	-	Health
23	0	0	18.347	0	2.582	0	-	Health
24	0	0	0	0	2.582	0	-	Health
25	0	0	0	0	0	0	-	Health
26	0	0	0	0	0	0	-	Health
27	0	0	0	0	2.582	0	-	Health
28	0	0	0	0	0	0	-	Health
29	1.786	2.442	2.293	0	5.164	4.540	-	Health
30	0	0	0	0	0	0	-	Health
31	5.358	0	18.347	0	0	0	-	Sport
32	1.786	0	0	0	0	0	-	Sport
33	0	0	0	0	0	0	-	Sport
34	0	0	0	0	0	0	-	Sport
35	1.786	0	0	0	0	4.540	-	Sport
36	1.786	0	0	0	0	0	-	Sport
37	0	0	0	0	0	0	-	Sport
38	0	0	0	0	0	0	-	Sport
39	0	0	0	0	0	0	-	Sport
40	0	0	0	0	0	0	-	Sport
41	0	0	0	3.133	0	0	-	Travel
42	3.572	0	0	0	0	0	-	Travel
43	0	0	0	0	0	0	-	Travel
44	0	0	0	0	0	0	-	Travel
45	1.786	2.442	2.293	3.133	5.164	4.540	-	Travel
46	3.572	2.442	2.293	0	5.164	0	-	Travel
47	0	0	0	0	0	0	-	Travel
48	1.786	4.884	0	0	0	0	-	Travel
49	0	0	0	0	0	4.540	-	Travel
50	1.786	0	0	0	0	0	-	Travel

เมื่อเราได้โมเดลที่ต้องการแล้ว จากนั้นจะเข้าขั้นตอนสำหรับทดสอบโมเดลที่ได้สร้างขึ้น ซึ่งจะกล่าวในบทที่ 5 ต่อไป