บทที่ 4

การทดลองและผลการทดลอง

ในบทนี้จะกล่าวถึงการออกแบบการทดลองการกำหนดหัวข้อข่าวให้กับเอกสาร โดยการใช้เทคนิคการจัดกลุ่มเอกสารเพื่อเปรียบเทียบประสิทธิภาพ โดยคำนึงถึงปัจจัยต่าง ๆ ที่เกี่ยวข้องกับการจัดกลุ่มเอกสาร ได้แก่ จำนวนเอกสาร จำนวนคุณลักษณะ และค่า Threshold ที่ใช้ในการเลือกคำที่ทำหน้าที่เป็นคุณลักษณะ และวัดประสิทธิภาพการจัดกลุ่มโดยใช้ค่าความเที่ยงตรง (Precision) ค่าความระลึก (Recall) และ ค่าเอฟ (F-measure)

4.1 ข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่ใช้ในงานวิจัยนี้รวบรวมมาจากเว็บไซต์ข่าวภาษาอังกฤษ ได้แก่ ได้แก่ Yahoo (http://news.yahoo.com), Accuweather (http://www.accuweather.com), New York Time (http://www.nytimes.com), CNN (http://www.cnn.com), News Week (http://www.newsweek.com) เป็นต้น ในช่วงเดือน ธันวาคม 2553-มีนาคม 2554 แบ่งออกเป็น 6 ประเภทข่าว ได้แก่ การเมือง พยากรณ์อากาศ สุขภาพ กีฬา บันเทิง และข่าวธุรกิจ ประเภทข่าวละ 1,000 เอกสาร รวมทั้งหมดเป็น 6,000 เอกสาร ใช้เป็นข้อมูลที่ใช้ในฝึกสอนและทดสอบ โดยจะเลือกเฉพาะข้อความข่าวเท่านั้น

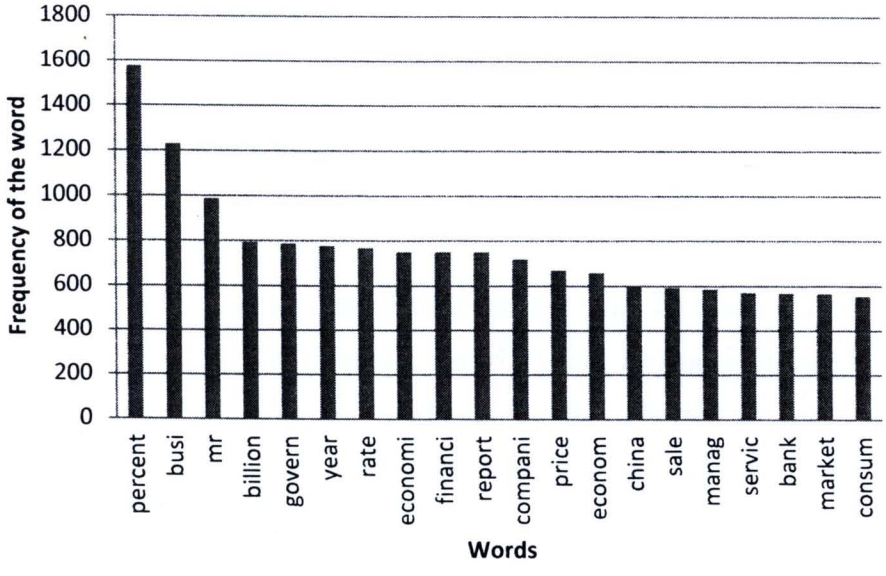
ตารางที่ 4.1 แสดงจำนวนคำในเอกสารแต่ละประเภทที่ตัดคำฟุ่มเฟือยและแปลงให้อยู่ในรูปรากศัพท์ และจำนวนคำไม่ซ้ำ พบว่าเอกสารข่าวที่รวบรวมนั้น ข่าวธุรกิจมีจำนวนคำทั้งหมดมากที่สุด ดังนั้นข่าวประเภทนี้ผู้เขียนข่าวจะมีการอธิบายรายละเอียดและนำเสนอข้อมูลที่มากกว่าข่าวประเภทอื่น ข่าวการเมือง สุขภาพ กีฬา พยากรณ์อากาศ และบันเทิงตามลำดับ และข่าวประเภทการเมืองมีการใช้คำซ้ำ ๆ เป็นจำนวนมากกว่าข่าวประเภทอื่น ๆ และข่าวบันเทิงมีการใช้คำที่หลากหลายมากกว่าข่าวประเภทอื่น

ตารางที่ 4.1 แสดงจำนวนคำทั้งหมดและจำนวนคำที่ไม่ซ้ำในแต่ละประเภทเอกสาร

ประเภทเอกสาร	จำนวนคำที่ไม่ซ้ำ	จำนวนคำทั้งหมด	ลดลง (เท่า)
การเมือง	11,218	179,478	16
พยากรณ์อากาศ	9,259	118,882	13
สุขภาพ	12,068	160,426	13
กีฬา	11,494	119,357	10
บันเทิง	12,574	104,479	8
ธุรกิจ	15,178	206,823	14
รวม	71,791	889,445	

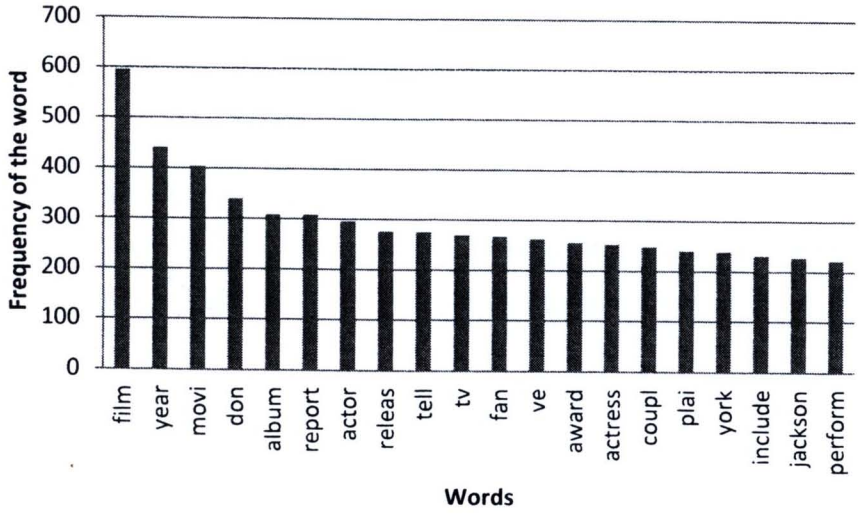
การกระจายของข้อมูลในประเภทข่าวต่าง ๆ หลังจากตัดคำฟุ่มเฟือยและแปลงให้อยู่ในรูปรากศัพท์ คำทั้งหมดในเอกสารที่ไม่ซ้ำกัน โดยคำเหล่านั้นเป็นคำที่ไม่ซ้ำกันในแต่ละประเภทเอกสาร มีทั้งหมด 71,791 คำ แต่คำเหล่านั้นอาจเป็นคำที่ซ้ำกันกับคำในเอกสารประเภทอื่น ดังนั้นจึงตัดคำซ้ำออกไปจึงเหลือคำทั้งหมดที่ไม่ซ้ำในเอกสารทั้งหมด 6 ประเภท ซึ่งมีทั้งหมด 33,700 คำ

แต่ละประเภทเอกสารสามารถแสดงการกระจายของคำได้ดังรูปที่ 4.1, 4.2, 4.3, 4.4, 4.5 และ 4.6 โดยแกน y แสดงความถี่ของคำ และแกน x แสดงคำที่ปรากฏในแต่ละประเภทเอกสาร 20 คำแรกที่มีความถี่สูงสุด



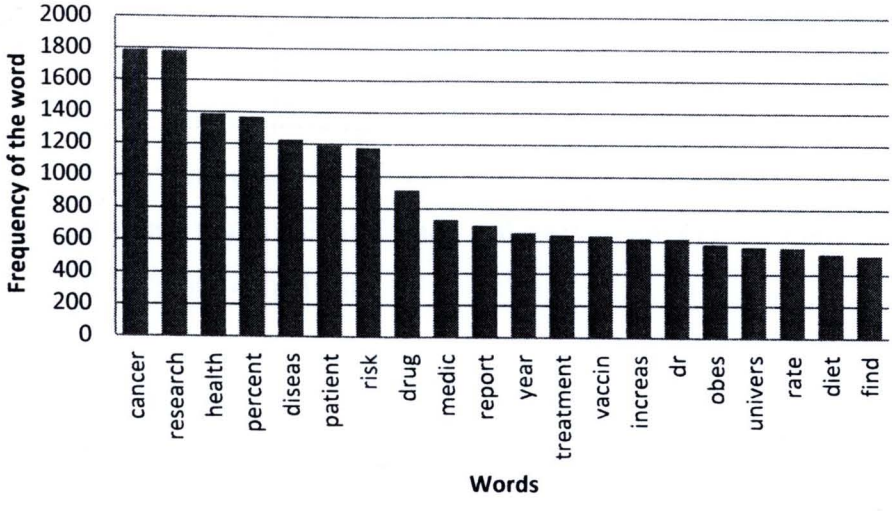
รูปที่ 4.1 การกระจายของคำ 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวธุรกิจ

จากรูปที่ 4.1 ในข่าวธุรกิจ 10 คำแรกที่พบมากที่สุดได้แก่ "percent" มีความถี่เท่ากับ 1,577 ครั้ง "busi" มีความถี่เท่ากับ 1,229 ครั้ง "mr" มีความถี่เท่ากับ 987 ครั้ง "billion" มีความถี่เท่ากับ 793 ครั้ง "govern" มีความถี่เท่ากับ 788 ครั้ง "year" มีความถี่เท่ากับ 777 ครั้ง "rate" มีความถี่เท่ากับ 769 ครั้ง "economi" มีความถี่เท่ากับ 752 ครั้ง "financi" มีความถี่เท่ากับ 752 ครั้ง และ "report" มีความถี่เท่ากับ 752 ครั้ง ซึ่งจำนวนคำที่มีความถี่เท่ากับ 1 คือคำนี้ปรากฏเพียงหนึ่งครั้งในเอกสารประเภทนี้ มีจำนวนคำเท่ากับ 5,481 คำ และจำนวนคำที่มีความถี่เท่ากับ 2 คือ คำที่ปรากฏสองครั้งในเอกสารประเภทนี้ โดยมีจำนวนคำเท่ากับ 2,359 คำ คิดเป็นร้อยละ 36 และ 15 ตามลำดับ



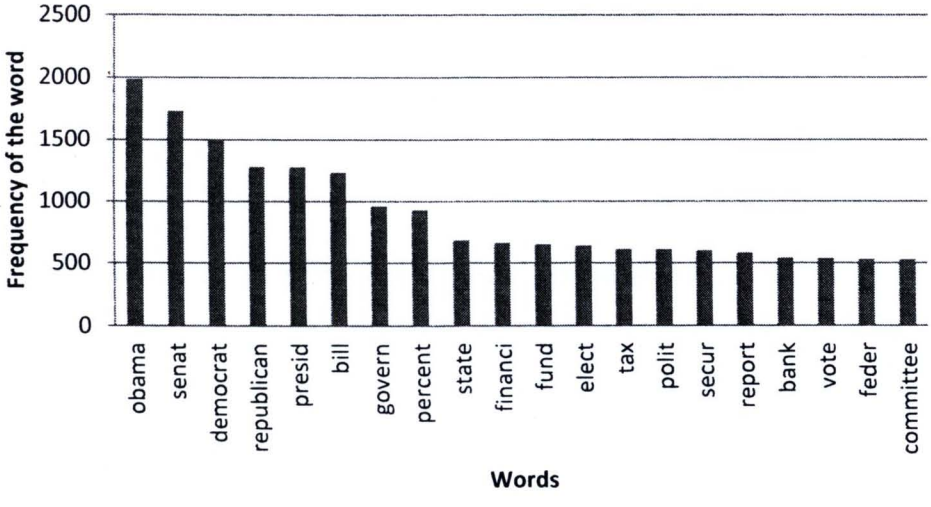
รูปที่ 4.2 การกระจายของคำ 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวบันเทิง

จากรูปที่ 4.2 ในข่าวประเภทบันเทิง 10 คำแรกที่พบมากที่สุดได้แก่ "film" มีความถี่เท่ากับ 595 ครั้ง "year" มีความถี่เท่ากับ 441 ครั้ง "movi" มีความถี่เท่ากับ 405 ครั้ง "don" มีความถี่เท่ากับ 340 ครั้ง "album" มีความถี่เท่ากับ 309 ครั้ง "report" มีความถี่เท่ากับ 309 ครั้ง "actor" มีความถี่เท่ากับ 296 ครั้ง "releas" มีความถี่เท่ากับ 277 ครั้ง "tell" มีความถี่เท่ากับ 276 ครั้ง และ "tv" มีความถี่เท่ากับ 270 ครั้ง โดยจำนวนคำที่มีความถี่เท่ากับ 1 มี 4,742 คำ และจำนวนคำที่มีความถี่เท่ากับ 2 มี 1,966 คำ คิดเป็นร้อยละ 37 และ 15 ตามลำดับ



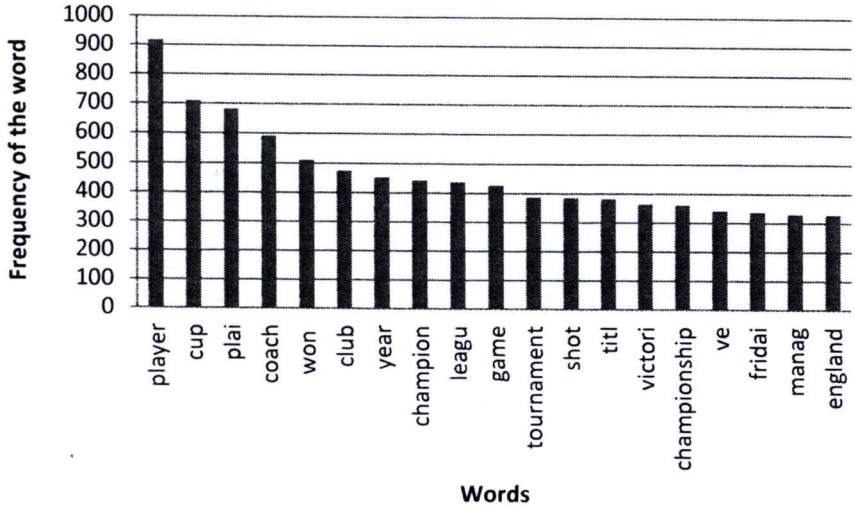
รูปที่ 4.3 การกระจายของคำ 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวสุขภาพ

จากรูปที่ 4.3 ในข่าวประเภทข่าวสุขภาพ 10 คำแรกที่พบมากที่สุดได้แก่ “cancer” มีความถี่เท่ากับ 1,785 ครั้ง “research” มีความถี่เท่ากับ 1,784 ครั้ง “health” มีความถี่เท่ากับ 1,388 ครั้ง “percent” มีความถี่เท่ากับ 1,369 ครั้ง “diseas” มีความถี่เท่ากับ 1,226 ครั้ง “patient” มีความถี่เท่ากับ 1,192 ครั้ง “risk” มีความถี่เท่ากับ 1,177 ครั้ง “drug” มีความถี่เท่ากับ 917 ครั้ง “medic” มีความถี่เท่ากับ 733 ครั้ง และ “report” มีความถี่เท่ากับ 698 ครั้ง โดยจำนวนคำที่มีความถี่เท่ากับ 1 มี 4,160 คำ และจำนวนคำที่มีความถี่เท่ากับ 2 มี 1,893 คำ คิดเป็นร้อยละ 34 และ 15 ตามลำดับ



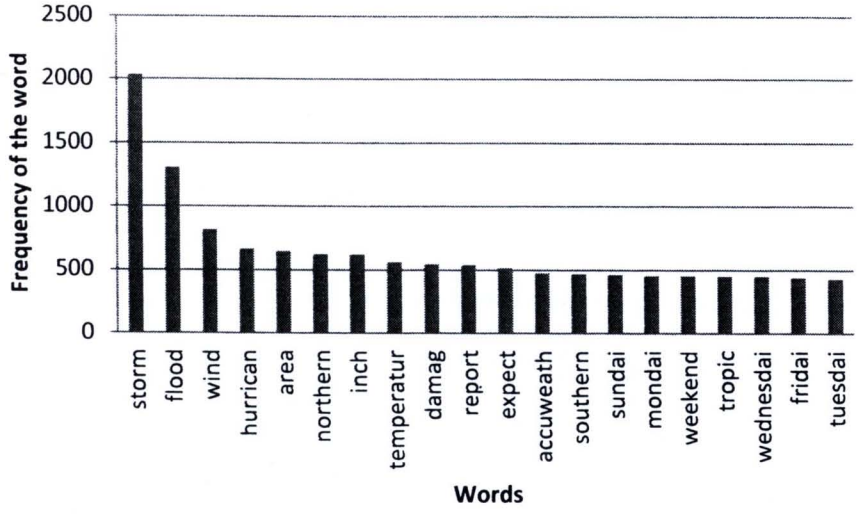
รูปที่ 4.4 การกระจายของคำ 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวการเมือง

จากรูปที่ 4.4 ในข่าวประเภทการเมือง 10 คำแรกที่พบมากที่สุดได้แก่ “obama” มีความถี่เท่ากับ 1,987 ครั้ง “senat” มีความถี่เท่ากับ 1,731 ครั้ง “democrat” มีความถี่เท่ากับ 1,489 ครั้ง “republican” มีความถี่เท่ากับ 1,281 ครั้ง “presid” มีความถี่เท่ากับ 1,278 ครั้ง “bill” มีความถี่เท่ากับ 1,235 ครั้ง “govern” มีความถี่เท่ากับ 961 ครั้ง “percent” มีความถี่เท่ากับ 931 ครั้ง “state” มีความถี่เท่ากับ 686 ครั้ง และ “financi” มีความถี่เท่ากับ 663 ครั้ง โดยจำนวนคำที่มีความถี่เท่ากับ 1 มี 3,635 คำ และจำนวนคำที่มีความถี่เท่ากับ 2 มี 1,623 คำ คิดเป็นร้อยละ 32 และ 14 ตามลำดับ



รูปที่ 4.5 การกระจายของคำ 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวกีฬา

จากรูปที่ 4.5 ในข่าวประเภทกีฬา 10 คำแรกที่พบมากที่สุดได้แก่ "cup" มีความถี่เท่ากับ 709 ครั้ง "plai" มีความถี่เท่ากับ 682 ครั้ง "coach" มีความถี่เท่ากับ 592 ครั้ง "won" มีความถี่เท่ากับ 510 ครั้ง "club" มีความถี่เท่ากับ 475 ครั้ง "year" มีความถี่เท่ากับ 453 ครั้ง "champion" มีความถี่เท่ากับ 444 ครั้ง "leagu" มีความถี่เท่ากับ 438 ครั้ง และ "game" มีความถี่เท่ากับ 428 ครั้ง โดยจำนวนคำที่มีความถี่เท่ากับ 1 มี 4,018 คำ และจำนวนคำที่มีความถี่เท่ากับ 2 มี 1,675 คำ คิดเป็นร้อยละ 35 และ 14 ตามลำดับ



รูปที่ 4.6 การกระจายของคำ 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทพยากรณ์อากาศ

จากรูปที่ 4.6 ในข่าวประเภทพยากรณ์อากาศ 10 คำแรกที่พบมากที่สุดได้แก่ "storm" มีความถี่เท่ากับ 2,034 ครั้ง "flood" มีความถี่เท่ากับ 1,306 ครั้ง "wind" มีความถี่เท่ากับ 817 ครั้ง "hurricane" มีความถี่

เท่ากับ 664 ครั้ง “area” มีความถี่เท่ากับ 648 ครั้ง “northern” มีความถี่เท่ากับ 622 ครั้ง “inch” มีความถี่เท่ากับ 620 ครั้ง “temperatur” มีความถี่เท่ากับ 562 ครั้ง “damag” มีความถี่เท่ากับ 546 ครั้ง และ “report” มีความถี่เท่ากับ 537 ครั้ง โดยจำนวนคำที่มีความถี่เท่ากับ 1 มี 3,260 คำ และจำนวนคำที่มีความถี่เท่ากับ 2 มี 1,496 คำ คิดเป็นร้อยละ 35 และ 16 ตามลำดับ

จากรูปทั้ง 6 รูป (รูปที่ 4.1-4.6) ข้างต้นพบว่าคำที่มีความถี่น้อยหรือคำที่ปรากฏในเอกสารน้อยครั้งนั้นมีเป็นจำนวนมากหรือมีประมาณเกือบร้อยละ 40 ของคำที่ไม่ซ้ำในเอกสารที่ปรากฏหนึ่งครั้งและประมาณร้อยละ 15 ของคำที่ไม่ซ้ำในเอกสารที่ปรากฏสองครั้ง ในทางกลับกันคำที่มีความถี่มากมักจะเป็นมีเพียงคำเดียวดังอธิบายข้างต้น ทำให้เราสามารถกล่าวอ้างได้ว่าคำที่มีความถี่ต่ำมาก ๆ จะเป็นคำที่น่าจะมีความสำคัญน้อย และในทำนองเดียวกันคำที่มีความถี่มาก ๆ ในประเภทเอกสารหลาย ๆ ประเภทนั้นน่าจะเป็นคำที่มีความสำคัญน้อยด้วยเช่นกัน จากกราฟพบว่า “report” ปรากฏในประเภทเอกสารสภาพอากาศ สุขภาพ ธุรกิจและบันเทิง ซึ่งมีความถี่สูงในประเภทเอกสารเหล่านี้ นั้นแสดงว่า “report” ไม่ควรจะถูกกำหนดให้เป็นคำสำคัญสำหรับใช้ระบุประเภทเอกสาร

จากแต่ละประเภทเอกสาร ได้ทำการเลือกตัวอย่างคำที่ปรากฏมากที่สุดในแต่ละประเภทเอกสาร 40 คำแรก ซึ่งจะพบว่าแต่ละคำนั้นมีความหมายที่เกี่ยวข้องกับแต่ละประเภทเอกสาร แสดงดังตารางที่ 4.2

ตารางที่ 4.2 แสดงตัวอย่างคำที่ปรากฏมากที่สุดในแต่ละประเภทเอกสาร 40 คำแรก

ประเภทเอกสาร	คำที่ปรากฏมากที่สุด 40 คำแรก
การเมือง	obama senat democrat republican presid bill govern percent state financi fund elect tax polit secur report bank vote feder committe unit offici issu american administr congress leader afghanistan includ call propos court regul lawmak legisl reform washington year nation rate
พยากรณ์อากาศ	storm flood wind hurrican area northern inch temperatur damag report expect accuweath southern sundai mondai weekend tropic wednesdai fridai tuesdai offici condit forecast saturdai includ atlant central thursdai mile dai part nation todai eastern thunderstorm meteorologist home western airport road
สุขภาพ	cancer research health percent diseas patient risk drug medic report year treatment vaccin increas dr obes univers rate diet find american studi develop brain includ prevent effect result test state breast clinic nation hospit cell medicin gene call higher journal
กีฬา	player cup plai coach won club year champion leagu game tournament shot titl goal victori championship ve fridai manag england jame score in win saturdai start final germani don didn footbal return ad sign point contract team run

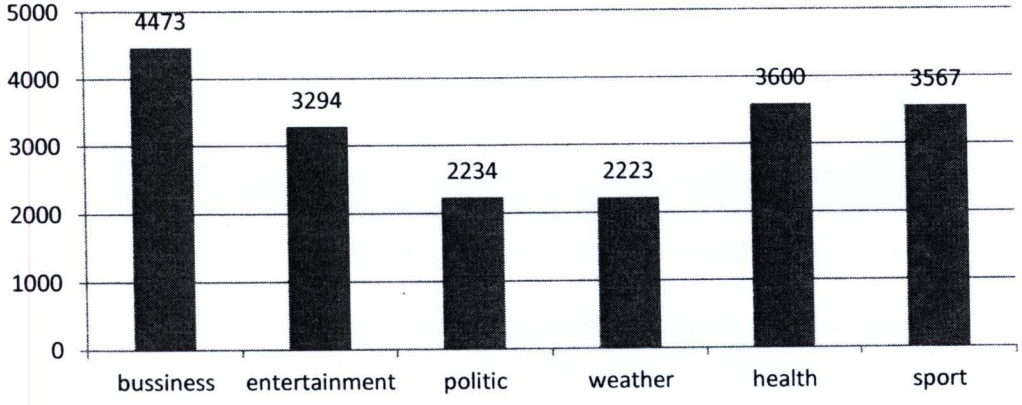
	wimbledon nada
บันเทิง	film year movi don album report actor releas tell tv fan ve award actress coupl plai york includ jackson perform star work recent singer michael show hollywood june kid video call celebr make seri didn angel lo dai thing tour
ธุรกิจ	percent busi mr billion govern year rate economi financi report compani price econom china sale manag servic bank market consum invest quarter growth tax increas american execut expect includ credit presid stock don firm make recent investor product site global

แต่อย่างไรก็ตามจากกลุ่มคำในตารางที่ 4.2 จะพบว่า 'includ', 'report' และ 'year' ปรากฏทั้ง 5 กลุ่มเอกสาร ขณะที่ 'call', 'don', 'percent' และ 'rate' ปรากฏใน 3 กลุ่มเอกสาร และ 'dai', 'didn', 'expect', 'financi', 'fridai', 'govern', 'increas', 'make', 'mange', 'offici', 'plai' เป็นต้น ซึ่งปรากฏใน 2 กลุ่มเอกสาร ดังนั้นจึงพบว่าการใช้คำเพียงอย่างเดียวในการระบุกลุ่มเอกสารนั้นไม่เพียงพอ จึงต้องใช้อัลกอริทึมการจัดกลุ่มเอกสารมาช่วยการทำนายว่าแต่ละเอกสารที่ประกอบด้วยกลุ่มคำเหล่านี้ควรจะจัดอยู่ในประเภทเอกสารใดจึงจะเหมาะสม

สำหรับคำที่ปรากฏในเอกสารทั้ง 6 ประเภทเอกสาร มีทั้งหมด 2,964 คำ แต่ละคำก็จะมีจำนวนความถี่ในแต่ละประเภทเอกสารที่แตกต่างกันโดยจะยกตัวอย่างบางคำที่ปรากฏต่อไปนี้

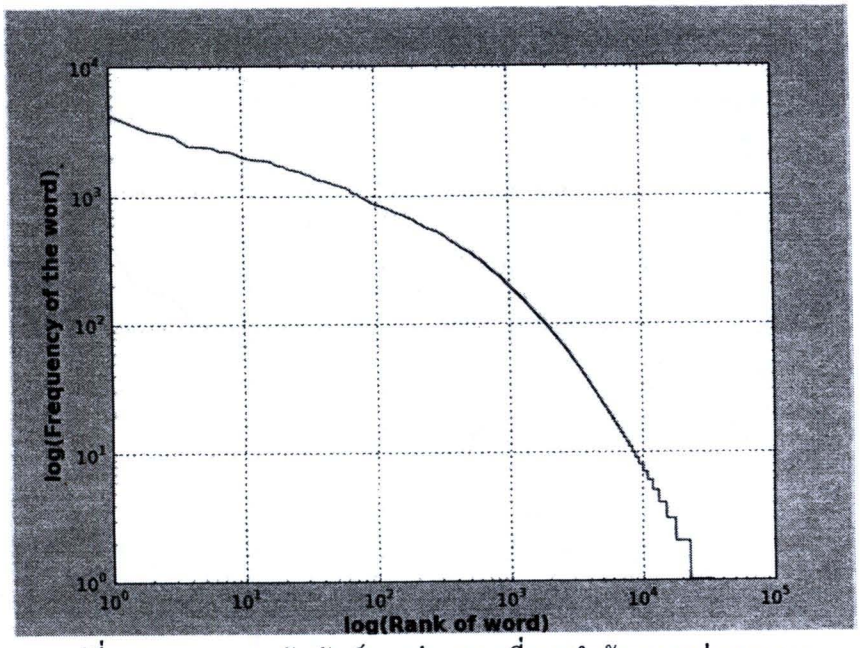
'bear', 'foul', 'protest', 'sleep', 'upsid', 'climb', 'hate', 'request', 'accus', 'accur', 'swai', 'edward', 'pride', 'worth', 'digit', 'risk', 'rise', 'voic', 'tenni', 'loom', 'jack', 'govern', 'affect', 'vast', 'disturb', 'wooden', 'ignit', 'huddl', 'correct', 'wednesdai', 'miller', 'direct', 'histor', 'enjoi', 'consequ', 'second', 'street', 'supervis', 'hide', 'wreck', 'neg', 'calcul', 'asia', 'spokesman', 'toll', 'new', 'net', 'succumb', 'liberti', 'specialist', 'elimin', 'hero', 'avert', 'carv', 'lodg', 'met', 'voter', 'china', 'aftermath', 'enrol', 'interpret', 'incom', 'deterior', 'forum', 'militari', 'anymor', 'loos', 'precis', 'jame', 'smoke', 'permit', 'studi', 'controversi', 'counti', 'golden', 'volunt', 'carl', 'campaign', 'newspap', 'julia', 'mitchel', 'thrust', 'attitud', 'moral', 'total', 'unit', 'highli', 'plot', 'describ', 'prescript', 'overshadow', 'insult', 'concret', 'call', 'telegraph', 'recommend', 'strike', 'indiana', 'type', 'tell', 'relax', 'relat', 'award', 'hurt', 'warn', 'phone', 'connecticut', 'exce', 'adult', 'wari', 'midst', 'hold', 'shoot', 'accid', 'join', 'room', 'henri', 'work', 'wors', 'era', 'ms', 'mr', 'root', 'advocaci', 'shook', 'climat', 'give', 'household', 'dolphin', 'india', 'indic', 'caution', 'refus', 'want', 'basebal', 'david', 'attract', 'vanish', 'end', 'quot', 'polic', 'travel', 'faulti', 'ceremoni', 'recoveri', 'answer', 'gate', 'negoti', 'perspect', 'confid', 'grown', 'recogn', 'lai', 'mess', 'chines', 'lag', 'lab', 'badli', 'modest', 'beauti', 'law', 'demonstr', 'domin', 'third', 'amid', 'grant', 'greet', 'think', 'perform', 'dispar', 'environ', 'reloc', 'enter', 'exclus', 'worst', 'order', 'wind', 'oper', 'offici', ...

ในขณะเดียวกันแต่ละประเภทเอกสารก็จะมีคำที่ไม่ซ้ำกันเลย แสดงดังรูปที่ 4.7 โดยแกน x คือ ประเภทเอกสาร และแกน y คือจำนวนคำที่ไม่ซ้ำ พบว่ากลุ่มเอกสารประเภทพยากรณ์อากาศ (weather) จะมีคำที่ไม่ซ้ำน้อยกว่าข่าวประเภทอื่น และข่าวประเภทธุรกิจ (business) จะมีคำที่ไม่ซ้ำมากกว่าข่าวประเภทอื่น ดังนั้นจากกราฟจึงอธิบายได้ว่าข่าวประเภทธุรกิจจะมีคำที่มีลักษณะเฉพาะสำหรับข่าวประเภทนั้นค่อนข้างมากกว่าข่าวประเภทอื่น ๆ รองลงมาได้แก่ ข่าวสุขภาพ (health) กีฬา (sport) บันเทิง (entertainment) การเมือง (politic) และ พยากรณ์อากาศ (weather) ตามลำดับ

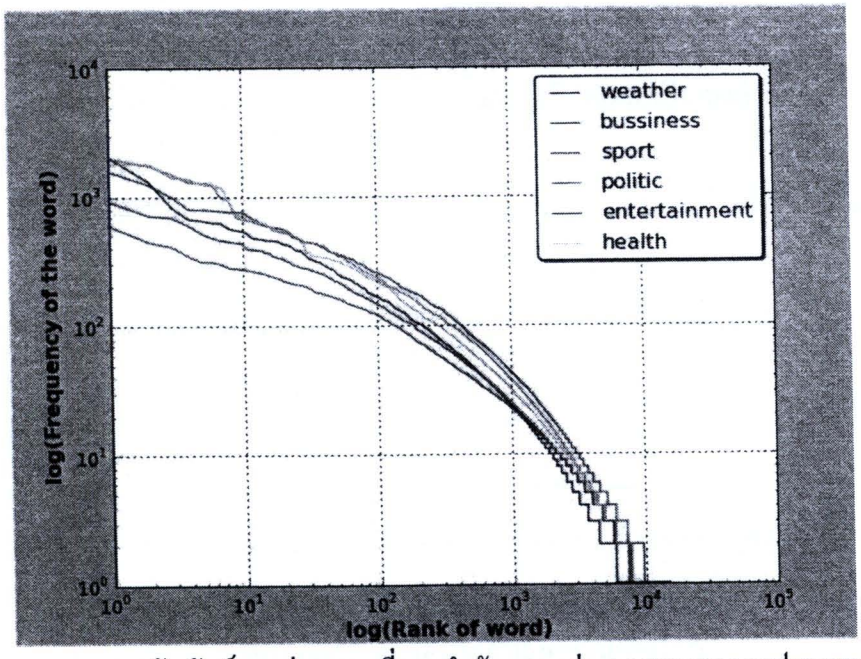


รูปที่ 4.7 แสดงจำนวนคำที่ไม่ซ้ำในแต่ละประเภทเอกสาร

การกระจายของความถี่ของคำในกลุ่มเอกสารตามกฎของ Zipf นั้นกล่าวว่าความถี่ของคำจะแปรผกผันกับลำดับของคำ ซึ่งความถี่ของคำสามารถถูกใช้เพื่อนำมาวัดความสำคัญของคำที่ใช้แทนเอกสารหนึ่งได้ โดยให้ f เป็นความถี่ของคำที่ปรากฏในกลุ่มเอกสาร และ r เป็นลำดับความสำคัญของคำนั้น ๆ ความสัมพันธ์ของ f และ r จะกล่าวได้ว่า เมื่อค่า f สูงจะส่งผลให้ r มีค่าต่ำ ดังที่กล่าวข้างต้น และนำความสัมพันธ์ระหว่าง f และ r มาสามารถแสดงได้ดังกราฟในรูปที่ 4.8 และ 4.9 เมื่อใช้ข้อมูลจากเอกสารที่รวบรวมมาได้



รูปที่ 4.8 แสดงความสัมพันธ์ระหว่างความถี่และลำดับของกลุ่มเอกสาร



รูปที่ 4.9 แสดงความสัมพันธ์ระหว่างความถี่และลำดับของกลุ่มเอกสารแยกตามประเภทเอกสาร

รูปที่ 4.8 แสดงแสดงความสัมพันธ์ระหว่างความถี่และลำดับของกลุ่มเอกสารทั้งหมด และรูปที่ 4.9 แสดงความสัมพันธ์ระหว่างความถี่และลำดับของกลุ่มเอกสารแยกตามประเภทเอกสาร จากรูปทั้งสองนี้คำที่มีความถี่มากจะอยู่ในลำดับที่ต่ำ และกราฟค่อย ๆ ลาดลงโดยในช่วงกลางของกราฟนั้นจะมีลักษณะโค้งค่อนข้างมากและส่วนปลายเส้นกราฟจะมีลักษณะการซ้ำกันของความถี่ค่อนข้างชัดเจน นั้นแสดงว่าคำศัพท์ในส่วนลำดับท้าย ๆ จะมีความถี่ต่ำมาก ๆ และมีจำนวนคำศัพท์เป็นในระดับความถี่นี้เป็นจำนวนมาก

4.2 การออกแบบการทดลองและผลการทดลอง

งานวิจัยนี้ใช้วิธีการทดสอบแบบการตรวจสอบแบบไขว้ (Cross-Validation) โดยแบ่งข้อมูลออกเป็น 5 กลุ่ม แล้วทำการใช้ 1 กลุ่มมาเป็นข้อมูลทดสอบ (Testing set) ส่วนที่เหลือจำนวน 4 กลุ่มจะเป็นข้อมูลฝึกสอน (Training set) แล้วทำการวน 5 ครั้ง ซึ่งจะเปลี่ยนกลุ่มทดสอบไปเรื่อย ๆ ตามลำดับจนครบข้อมูลทั้งหมด

ในการทดสอบนี้แบ่งกลุ่มข้อมูลออกเป็น 3 ชุดข้อมูลทดสอบ ประกอบด้วยข้อมูลจำนวน 3,000 เอกสาร 4,200 เอกสาร และ 6,000 เอกสาร ซึ่งในแต่ละกลุ่มก็จะแบ่งเอกสารออกเป็นเอกสารที่ใช้ในการฝึกสอนและเอกสารที่ใช้ในการทดสอบ ดังรายละเอียดข้างต้น โดยในการอธิบายนี้จะแทนแต่ละกลุ่มเอกสารดังนี้

ชุดเอกสารทดสอบที่ 1 ได้แก่ เอกสารจำนวน 3,000 เอกสาร

ชุดเอกสารทดสอบที่ 2 ได้แก่ เอกสารจำนวน 4,200 เอกสาร

ชุดเอกสารทดสอบที่ 3 ได้แก่ เอกสารจำนวน 6,000 เอกสาร

การออกแบบการทดลองในงานวิจัยนี้ ได้ตั้งสมมติฐานเกี่ยวกับการกำหนดหัวข้อข่าวกับปัจจัยต่าง ๆ ที่มีผลกระทบต่อการจัดกลุ่มได้ดังนี้

1. ค่า Threshold ที่ใช้ในการเลือกคำที่กำหนดที่มีความถี่ต่ำสุดมาเป็นคุณลักษณะ ด้วยการคำนวณวิธี TFIDF มีผลอย่างไร
2. จำนวนคำที่กำหนดเป็นคุณลักษณะมีผลต่อการกำหนดหัวข้อข่าวหรือไม่ (Effect of feature size on performance)
3. จำนวนเอกสารที่ใช้ในการฝึกสอนมีผลต่อการกำหนดหัวข้อข่าวหรือไม่ (Effect of training set size on Topic identification)
4. เปรียบเทียบวิธีการคำนวณค่าน้ำหนักระหว่าง TFIDF, TFICF, IG และ CHI (Performance comparison between four approaches) ให้กับกลุ่มคำสำคัญ

ในการวัดประสิทธิภาพของแต่ละสมมติฐานนั้นจะใช้ค่าความเที่ยงตรง ค่าความระลึก และค่าเอฟ (รายละเอียดอธิบายในบทที่ 2) ซึ่งค่าเหล่านี้จะคำนวณโดย นำเอกสารในกลุ่มฝึกสอนมาสร้างโมเดล และทดสอบโมเดลนี้โดยใช้เอกสารในกลุ่มทดสอบเพื่อวัดประสิทธิภาพในค่าต่าง ๆ

ดังนั้นในการออกแบบการทดลองในงานวิจัยนี้จึงต้องสามารถตอบคำถามข้อมูลข้างต้นได้ ซึ่งสามารถกำหนดได้เป็น

การทดลองที่ 1: ค่า Threshold ที่ใช้กำหนดความถี่ต่ำสุดในการเลือกคำที่กำหนดเป็นคุณลักษณะด้วยการคำนวณวิธี TFIDF มีผลอย่างไร

ในการทดลองนี้เป็นการเลือก Threshold ที่เหมาะสม ได้ใช้กฎของ Zipf สร้างค่า Threshold เพื่อกำหนดจุดที่เหมาะสมในการเลือกกลุ่มคำที่ไม่สำคัญหรือคำที่ปรากฏค่อนข้างน้อยครั้งในเอกสาร โดยถ้าค่า

น้ำหนักน้อยกว่าค่า Threshold จะถือว่าเป็นกลุ่มคำที่มีความสำคัญ ซึ่งในการเลือกค่า Threshold จะอ้างอิงที่ค่าความถี่ของการปรากฏของคำในกลุ่มเอกสาร

ตารางที่ 4.3 แสดงจำนวนคำที่กำหนดคุณลักษณะของกลุ่มเอกสารทดสอบเมื่อกำหนดค่า Threshold = 3, 4, 5 และ 6

ชุดเอกสารทดสอบ	จำนวนคำที่กำหนดคุณลักษณะ			
	Threshold=3	Threshold=4	Threshold=5	Threshold=6
1	10,997	9,565	8,521	7,752
2	11,828	10,337	9,168	8,294
3	14,349	12,615	11,284	10,310

จากตารางที่ 4.3 เป็นการแสดงจำนวนคำที่กำหนดเป็นคุณลักษณะของเอกสารกลุ่มทดสอบแยกตามค่า Threshold ที่กำหนด เมื่อ กำหนดให้ค่า Threshold มีค่าเพิ่มขึ้นจำนวนคำก็จะมีจำนวนลดลงอยู่ระหว่าง 9-13 %

เมื่อกำหนดจำนวนคำที่เป็นคุณลักษณะแต่ละ Threshold จึงทำการเรียนรู้จากเอกสารฝึกสอนเพื่อสร้างโมเดลและทดสอบโมเดลจากเอกสารทดสอบในแต่ละชุดเอกสารทดสอบ โดยใช้อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน ตามวิธีการที่กล่าวข้างต้น และทำในทุก ๆ ชุดเอกสารทดสอบ เพื่อวัดประสิทธิภาพการระบุประเภทเอกสาร ด้วยตัววัดความเที่ยงตรง ความระลึกและค่าเอฟ ผลการทดลองแสดงดังตารางที่ 4.4

ในการคำนวณเพื่อวัดประสิทธิภาพนั้นคำนวณได้จาก

Category		Expert Judgment	
		True	False
Classifier Judgment	True	TP	FP
	False	FN	TN

ค่าความเที่ยงตรงนั้นคำนวณได้จาก (จำนวนเอกสารที่โมเดลสามารถทำนายได้ถูกต้อง) / (จำนวนเอกสารที่โมเดลสามารถทำนายได้ถูกต้อง + จำนวนเอกสารที่โมเดลทำนายว่าอยู่ในประเภทนี้ แต่ในความเป็นจริงแล้วไม่ใช่)

ค่าความระลึกคำนวณได้จาก (จำนวนเอกสารที่โมเดลสามารถทำนายได้ถูกต้อง) / (จำนวนเอกสารที่โมเดลสามารถทำนายได้ถูกต้อง + จำนวนเอกสารที่โมเดลทำนายว่าเอกสารไม่อยู่ในประเภทนี้ แต่ในความเป็นจริงแล้วใช่)

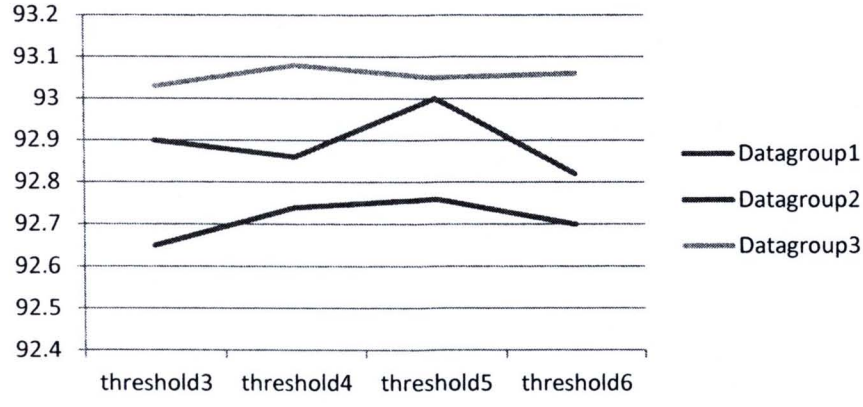
ค่าเอฟคำนวณได้จาก 2*(ค่าความเที่ยงตรง*ค่าความระลึก)/ (ค่าความเที่ยงตรง+ค่าความระลึก)

จากตัววัดทั้ง 3 ตัววัดนั้น ในการระบุว่าโมเดลที่ใช้ในการระบุประเภทเอกสารดีสามารถระบุเอกสารได้ถูกต้องมากที่สุดจะต้องมีค่าความเที่ยงตรง ค่าความระลึกและค่าเอฟที่สูง

ตารางที่ 4.4 แสดงค่าความเที่ยงตรง ค่าความระลึก และค่าเอฟ (เปอร์เซ็นต์) ในแต่ละค่า Threshold และ แต่ละชุดเอกสารทดสอบ

ชุดเอกสารทดสอบ	Threshold=3			Threshold=4			Threshold=5			Threshold=6		
	P	R	F	P	R	F	P	R	F	P	R	F
1	92.99	92.81	92.90	92.94	92.98	92.86	93.07	92.92	93.00	92.90	92.75	92.82
2	92.73	92.57	92.65	92.82	92.67	92.74	92.82	92.69	92.76	92.77	92.64	92.70
3	93.09	92.96	93.03	93.13	93.02	93.08	93.11	93.00	93.05	93.11	93.00	93.06

หมายเหตุ P=ค่าความเที่ยงตรง R=ค่าความระลึก F=ค่าเอฟ



รูปที่ 4.10 แสดงค่าเอฟตามค่า Threshold และกลุ่มเอกสารชุดทดสอบ

จากตารางที่ 4.4 สามารถแสดงในรูปแบบกราฟดังในรูปที่ 4.10 โดยที่ กราฟสีเขียวแสดงชุดเอกสารทดสอบที่ 1 สีน้ำเงินแสดงชุดเอกสารทดสอบที่ 2 และสีแดงแสดงชุดเอกสารทดสอบที่ 3 พบว่าค่า Threshold ที่มีค่าเท่ากับ 5 มีผลให้ประสิทธิภาพดีกว่าค่า Threshold อื่น ๆ ในกลุ่มเอกสารทดสอบชุดที่ 1 และ 2 แต่ค่า Threshold เท่ากับ 4 จะให้ผลดีกว่าค่าอื่น ๆ ในกลุ่มเอกสารทดสอบชุดที่ 3 ดังนั้นในการเปรียบเทียบประสิทธิภาพในการทดลองถัดไปจึงเลือกค่า Threshold เท่ากับ 5 ในวิธีการคำนวณค่าน้ำหนักด้วย TFIDF

ดังนั้นจึงกล่าวได้ว่า เมื่อกำหนดค่า Threshold เท่ากับ 5 แล้วจะทำให้การระบุประเภทเอกสารดีกว่ากำหนดด้วยค่า Threshold อื่นๆ ที่กำหนดในการทดลอง

การทดลองที่ 2: จำนวนคำที่กำหนดเป็นคุณลักษณะและจำนวนเอกสารมีผลต่อการจัดกลุ่มหรือไม่

ในการทดลองนี้เป็นการเปรียบเทียบประสิทธิภาพวิธีการคำนวณค่าน้ำหนักต่าง ๆ โดยมีเงื่อนไขคือจำนวนเอกสารที่แตกต่างกัน และจำนวนคำที่กำหนดเป็นคุณลักษณะที่แตกต่างกันในแต่ละวิธีการคำนวณค่าน้ำหนัก โดยคุณลักษณะที่ได้นั้นจะเป็นคำที่ปรากฏในเอกสาร ในการเลือกจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะนั้นจะเป็นไปตามเงื่อนไขที่กำหนด

การทดลองที่ 2.1 เปรียบเทียบประสิทธิภาพของการวิธีการคำนวณค่าน้ำหนักต่าง ๆ โดยปราศจากการเลือกคำที่ทำหน้าที่เป็นคุณลักษณะ ซึ่งในการทดลองนี้จะมีจำนวนคำเพื่อใช้ในการจัดกลุ่มเป็นจำนวนมาก แสดงดังตารางที่ 4.5 และตารางที่ 4.6 แสดงผลการทดลองการระบุประเภทเอกสาร โดยแสดงค่าความเที่ยงตรงและค่าความระลึก ตามลำดับ และตารางที่ 4.7 แสดงค่าเอฟ

ตารางที่ 4.5 ตารางแสดงผลรวมของจำนวนคำที่ไม่ซ้ำที่กำหนดเป็นคุณลักษณะของแต่ละชุดเอกสารทดสอบ

ชุดเอกสารทดสอบ	รวมจำนวนคำที่ไม่ซ้ำในเอกสาร 6 ประเภท			
	TFIDF	TFICF	IG	CHI
1	8,521	23,217	25,032	25,098
2	9,168	24,492	26,528	26,615
3	11,284	28,747	31,016	31,478

จากตารางที่ 4.5 แสดงจำนวนคุณลักษณะหรือคำที่ถูกกำหนดให้ทำหน้าที่เป็นตัวแทนของเอกสารในการระบุประเภทเอกสาร โดยจะพบว่าแต่ละชุดเอกสารทดสอบจะมีจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะอยู่ระหว่าง 8,000 – 32,000 คำ ซึ่งในการคำนวณค่าน้ำหนักด้วยวิธี CHI จะมีคำที่ถูกกำหนดเป็นคุณลักษณะมากที่สุด แต่อย่างไรก็ตามจะมีจำนวนคำที่ไม่แตกต่างจากวิธี IG มากนัก ส่วนการคำนวณค่าน้ำหนักด้วยวิธี TFICF นั้นก็จะมีจำนวนน้อยกว่ากลุ่มแรกที่กล่าว เพราะว่าถ้าคำใดปรากฏในทุก ๆ ประเภทเอกสารคำนั้นจะไม่ถูกเลือกมากำหนดเป็นคุณลักษณะ ในขณะที่วิธี TFIDF จะมีจำนวนคุณลักษณะน้อยที่สุด เนื่องจากว่าถูกกำหนดด้วยว่าความถี่ของคำจะต้องมีค่ามากกว่าค่า Threshold จากการทดลองที่ 1

ตารางที่ 4.6 แสดงค่าความเที่ยงตรงและค่าความระลึกตามวิธีการคำนวณค่าน้ำหนักแยกตามจำนวนเอกสาร

ชุดเอกสารทดสอบ	ค่าความเที่ยงตรงและค่าความระลึกแบ่งตามวิธีการคำนวณค่าน้ำหนัก							
	TFIDF		TFICF		IG		CHI	
	P	R	P	R	P	R	P	R
1	93.07	92.92	87.80	86.72	89.62	89.44	90.20	90.11
2	92.82	92.69	87.76	87.17	89.45	89.29	89.83	89.73
3	93.11	93.00	86.40	83.04	90.09	89.96	90.57	90.50

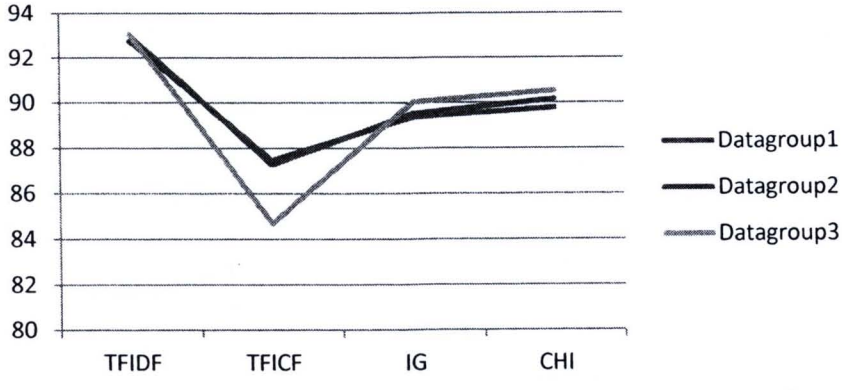
หมายเหตุ P=ค่าความเที่ยงตรง R=ค่าความระลึก

จากตารางที่ 4.6 ค่าความเที่ยงตรงในแต่ละการทดลองจะมีค่ามากกว่าค่าความระลึก แต่อย่างไรก็ตามค่าทั้งสองก็มีค่าที่สูง โดยค่าความเที่ยงตรงมีค่าอยู่ระหว่าง 86-93% และค่าความระลึกมีค่าอยู่ระหว่าง 86-93% เช่นเดียวกัน ซึ่งถือว่ามีประสิทธิภาพในการระบุประเภทเอกสารที่ค่อนข้างดี จากนั้นคำนวณค่าเอฟจากค่าความเที่ยงตรงและค่าความระลึก แสดงดังตารางที่ 4.7

ตารางที่ 4.7 แสดงค่าเอฟตามวิธีการคำนวณค่าน้ำหนักและจำนวนเอกสาร

ชุดเอกสารทดสอบ	ค่าเอฟแบ่งตามวิธีการคำนวณค่าน้ำหนัก			
	TFIDF	TFICF	IG	CHI
1	93.00	87.26	89.53	90.16
2	92.76	87.46	89.37	89.78
3	93.05	84.68	90.03	90.54

จากตารางที่ 4.7 นำมาสร้างกราฟได้ดังรูปที่ 4.11



รูปที่ 4.11 แสดงค่าเอฟตามกลุ่มเอกสารชุดทดสอบ และวิธีการคำนวณค่าน้ำหนัก

จากรูปที่ 4.11 กำหนดให้ Datagroup1 คือ ชุดเอกสารทดสอบที่ 1 Datagroup2 คือ ชุดเอกสารทดสอบที่ 2 และ Datagroup3 คือ ชุดเอกสารทดสอบที่ 3 อธิบายได้ว่า ในการคำนวณค่าน้ำหนักด้วยวิธี TFIDF พบว่าเอกสารที่มีจำนวนทดสอบมากที่สุด (ชุดเอกสารทดสอบที่ 3) จะให้ประสิทธิภาพในการระบุประเภทเอกสารดีที่สุด ซึ่งมีค่าเอฟเท่ากับ 93.05% ซึ่งหมายความว่าสามารถระบุประเภทเอกสารได้ถูกเมื่อเทียบเป็นเปอร์เซ็นต์ได้ 93.05% การคำนวณค่าน้ำหนักด้วยวิธี TFICF นั้นเอกสารทดสอบที่ 2 จะให้ค่าเอฟมากที่สุด เท่ากับ 87.46% การคำนวณค่านำหนักด้วยวิธี IG และ CHI นั้น เอกสารทดสอบที่ 3 จะให้ค่าเอฟมากที่สุด เท่ากับ 90.03% และ 90.54% ตามลำดับ ดังนั้นจึงอาจกล่าวได้ว่า ยิ่งมีจำนวนเอกสารที่ใช้ในการเรียนรู้มากเท่าไหร่แล้วก็ยิ่งทำให้ประสิทธิภาพระบุประเภทเอกสารมีเพิ่มมากขึ้นไปด้วย นอกจากนั้นแล้ว

จากค่าเอฟที่คำนวณได้จากการใช้วิธีการคำนวณค่าน้ำหนักที่แตกต่างกันนั้น การใช้วิธี TFIDF (Threshold=5) นั้นจะให้ประสิทธิภาพการระบุประเภทเอกสารได้ดีกว่าวิธี CHI IG และ TFICF ตามลำดับ หรือสามารถบอกได้ว่าเมื่อใช้วิธีการคำนวณค่าน้ำหนักด้วย TFIDF โดยกำหนดค่า Threshold เท่ากับ 5 นั้น สามารถระบุประเภทเอกสารได้ถูกต้องมากกว่าคำนวณค่านำหนักด้วยวิธีอื่นที่เปรียบเทียบแสดงดังข้างต้น

การทดลองที่ 2.2 เปรียบเทียบประสิทธิภาพวิธีการคำนวณค่าน้ำหนักต่าง ๆ โดยมีเงื่อนไขคือจำนวนเอกสารที่แตกต่างกันและจำนวนคำที่กำหนดเป็นคุณลักษณะที่แตกต่างกัน ในการเลือกจำนวนคำนั้นก็คือคำที่จะใช้เป็นตัวแทนของแต่ละประเภทเอกสารนั้น จะกำหนดโดยเลือกคำที่มีค่าน้ำหนักมากที่สุดของแต่ละประเภทเอกสารจำนวน n คำ แล้วนำมารวมกันเป็นตัวแทนของกลุ่มเอกสาร เช่น กำหนดให้เลือกคำจากแต่ละประเภทเอกสารประเภทละ 500 คำ (n=500) จะได้เซตของคำทั้งหมดที่ทำหน้าที่เป็นตัวแทนประเภทเอกสารจำนวน 3,000 คำ จาก 6 ประเภทเอกสาร ซึ่งในเซตของคำนี้จะมีคำซ้ำกัน เนื่องจากคำ ๆ หนึ่งสามารถเป็นตัวแทนของประเภทเอกสารได้หลายประเภท จึงตัดคำที่ซ้ำกันออก ดังนั้น จะได้จำนวนคำที่เป็นตัวแทนประเภทเอกสารเท่ากับ 1,507 คำ จากวิธีการคำนวณค่าน้ำหนักด้วย TFICF 1,932 คำ จากวิธี IG และ 2,161 คำ จากวิธี CHI และในวิธี TFIDF จะเป็นการคำนวณค่านำหนักของคำในแต่ละเอกสาร ไม่ได้แยกตามประเภทเอกสาร ทำให้ไม่มีคำที่ซ้ำกันจึงเลือกคำทั้งหมดนั้นกำหนดเป็นตัวแทนของกลุ่มเอกสาร แสดงในตารางที่ 4.8

เมื่อกำหนดค่า n=1,000 จะได้เซตของคำที่ทำหน้าที่เป็นตัวแทนจำนวน 6,000 คำ จาก 6 ประเภทเอกสาร เมื่อตัดคำที่ซ้ำกันจะได้คำเท่ากับ 2,937 คำ จากวิธีการคำนวณค่าน้ำหนักด้วย TFICF 3,819 คำ จากวิธี IG และ 4,158 คำ จากวิธี CHI

ในการทดลองนี้แบ่งออกเป็นการทดลองย่อยแยกตามชุดเอกสารทดสอบ จำนวน 3 ชุดเอกสาร ซึ่งจะประกอบด้วยรายละเอียดต่อไปนี้คือ จำนวนคำที่ทำหน้าที่เป็นคุณลักษณะ แยกตามวิธีการคำนวณค่าน้ำหนักต่าง ๆ ประสิทธิภาพของแต่ละวิธี ด้วยตัววัดค่าความเที่ยงตรง ค่าความระลึก และค่าเอฟ

ชุดเอกสารทดสอบ กลุ่มที่ 1 ซึ่งมีเอกสารจำนวน 3,000 เอกสาร แล้วทดลองตามจำนวนที่กำหนดเป็นตัวแทนที่แตกต่างกันไป โดยกำหนดจากค่า n ซึ่งแสดงในตารางที่ 4.8

ตารางที่ 4.8 ตารางแสดงจำนวนคำไม่ซ้ำในเอกสารแต่ละประเภทตามจำนวนคำที่เลือกของเอกสารแต่ละประเภท ของเอกสารจำนวน 3,000 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	รวมจำนวนคำที่ไม่ซ้ำในเอกสาร 6 ประเภท			
	TFIDF	TFICF	IG	CHI
500	3,000	1,507	1,932	2,161

1,000	6,000	2,937	3,819	4,158
1,500	8,251	4,168	5,584	5,504
2,000	-	5,821	6,975	6,836
4,000	-	13,652	12,328	10,394

ตารางที่ 4.9 แสดงค่าความเที่ยงตรงและค่าความระลึกตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท ของเอกสารจำนวน 3,000 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	ค่าความเที่ยงตรงและค่าความระลึกแบ่งตามวิธีการคำนวณค่าน้ำหนัก							
	TFIDF		TFICF		IG		CHI	
	P	R	P	R	P	R	P	R
500	91.55	91.22	80.14	75.86	68.85	67.89	69.99	67.81
1,000	92.92	92.78	80.90	77.00	74.58	74.25	76.62	75.61
1,500	93.07	92.92	83.01	79.97	78.91	78.42	79.82	79.39
2,000	-	-	84.90	82.22	79.47	78.97	80.34	79.89
4,000	-	-	86.32	84.25	81.30	80.58	86.53	86.19

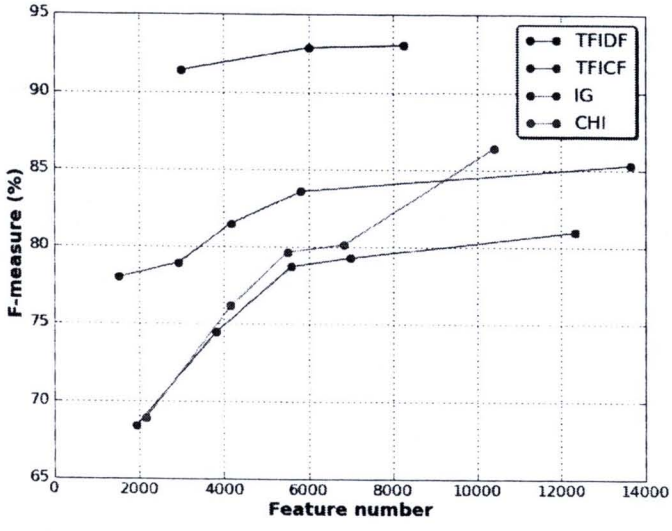
หมายเหตุ P=ค่าความเที่ยงตรง R=ค่าความระลึก

จากตารางที่ 4.9 พบว่าค่าความเที่ยงตรงในแต่ละการทดลองจะมีค่ามากกว่าค่าความระลึก โดยเมื่อจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะเพิ่มมากขึ้นค่าความเที่ยงตรงและค่าความระลึกก็จะมีค่าสูงขึ้น นั้นหมายความว่า การระบุประเภทเอกสารมีความถูกต้องเพิ่มมากขึ้น จากนั้นคำนวณค่าเอฟจากค่าความเที่ยงตรงและค่าความระลึก แสดงดังตารางที่ 4.10

ตารางที่ 4.10 แสดงค่าเอฟตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท โดยที่จำนวนเอกสารเท่ากับ 3,000 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	ค่าเอฟแบ่งตามวิธีการคำนวณค่าน้ำหนัก				ค่าเอฟเฉลี่ยทุกวิธีการคำนวณค่าน้ำหนัก
	TFIDF	TFICF	IG	CHI	
500	91.38	77.94	68.37	68.88	76.64
1,000	92.85	78.90	74.42	76.11	80.57
1,500	93.00	81.46	78.66	79.60	83.18
2,000	-	83.54	79.22	80.11	80.97
4,000	-	85.27	80.94	86.36	84.19

จากตารางที่ 4.10 นำค่าเอฟดังกล่าวมาสร้างกราฟได้ดังรูปที่ 4.12



รูปที่ 4.12 แสดงค่าเอฟแยกตามจำนวนคุณลักษณะและวิธีการคิดค่าน้ำหนัก ของเอกสารชุดทดสอบที่ 1

จากตารางที่ 4.10 และ รูปที่ 4.12 พบว่าเมื่อจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะเพิ่มมากขึ้น ค่าเอฟจะเพิ่มมากขึ้นด้วยในทุก ๆ วิธีการคำนวณค่าน้ำหนัก ได้แก่ ในการวิธีคำนวณค่าน้ำหนักด้วย TFIDF ค่าเอฟจะเพิ่มขึ้นจาก 91.38%, 92.85% และ 93.00% เมื่อกำหนดคุณลักษณะให้ n = 500, 1000 และ 1,500 ตามลำดับ เมื่อใช้วิธี TFICF ค่าเอฟจะเพิ่มขึ้นจาก 77.94% เมื่อกำหนดคุณลักษณะให้ n=500 เป็น 85.27% เมื่อกำหนดคุณลักษณะให้ n=4,000 เมื่อใช้วิธี IG ค่าเอฟจะเพิ่มขึ้นจาก 68.37% เมื่อกำหนดคุณลักษณะให้ n=500 เป็น 80.94% เมื่อกำหนดคุณลักษณะให้ n=4,000 เมื่อใช้วิธี CHI ค่าเอฟจะเพิ่มขึ้นจาก 68.88% เมื่อกำหนดคุณลักษณะให้ n=500 เป็น 86.36% เมื่อกำหนดคุณลักษณะให้ n=4,000

ค่าเอฟเฉลี่ยของทุกวิธีการคำนวณค่าน้ำหนักแยกตามจำนวนคำที่กำหนดให้เป็นคุณลักษณะจะมีแตกต่างกัน เมื่อกำหนดให้ n มีค่า เท่ากับ 500, 1000, 1500, 2000 และ 4000 จะมีค่าเอฟเฉลี่ยเท่ากับ 76.64%, 80.57%, 83.18%, 80.97% และ 84.19% ตามลำดับ โดยเมื่อจำนวนคำที่กำหนดเป็นคุณลักษณะมีค่าเพิ่มขึ้นแล้วค่าเอฟเฉลี่ยจะมีค่าเพิ่มขึ้นด้วย ยกเว้นเมื่อเพิ่ม n จาก 1,500 เป็น 2,000

จากข้อมูลข้างต้นอธิบายได้ว่าเมื่อจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะเพิ่มมากขึ้นในชุดเอกสารทดสอบ กลุ่มที่1 ประสิทธิภาพในการระบุประเภทเอกสารจะเพิ่มมากขึ้น หมายความว่ามีการระบุประเภทเอกสารได้ถูกต้องมากขึ้น โดยพิจารณาจากค่าเอฟที่สูงขึ้น โดยการคำนวณค่าน้ำหนักด้วยวิธี TFIDF จะให้ค่าเอฟที่มากที่สุด

ชุดเอกสารทดสอบ กลุ่มที่ 2 ซึ่งมีเอกสารจำนวน 4,200 เอกสาร แล้วทดลองตามจำนวนคำที่กำหนดเป็นคุณลักษณะที่แตกต่างกันไป โดยกำหนดจากค่า n ซึ่งแสดงในตารางที่ 4.11

ตารางที่ 4.11 ตารางแสดงจำนวนคำไม่ซ้ำในเอกสารแต่ละประเภทตามจำนวนคำที่เลือกของเอกสารแต่ละประเภท ของเอกสารจำนวน 4,200 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	รวมจำนวนคำที่ไม่ซ้ำในเอกสาร 6 ประเภท			
	TFIDF	TFICF	IG	CHI
500	3,000	1,440	1,883	2,188
1,000	6,000	2,853	3,745	4,187
1,500	9,000	4,340	5,563	5,655
2,000	-	5,549	6,976	7,005
4,000	-	13,506	12,328	26,615

จากตารางที่ 4.11 เมื่อกำหนดให้เลือกคำที่ทำหน้าที่เป็นคุณลักษณะจากแต่ละประเภทเอกสารประเภทละ 500 คำ (n=500) จะได้เซตของคำทั้งหมดที่ทำหน้าที่เป็นตัวแทนจำนวน 3,000 คำ จาก 6 ประเภทเอกสาร ซึ่งในเซตของคำนี้จะมีคำซ้ำกัน เนื่องจากคำ ๆ หนึ่งสามารถเป็นตัวแทนของประเภทเอกสารได้หลายประเภท ทำให้ต้องตัดคำที่ซ้ำกันออก ดังนั้น จะได้จำนวนคำเท่ากับ 1,440 คำ จากวิธีการคำนวณค่าน้ำหนักด้วย TFICF 1,883 คำ จากวิธี IG และ 2,188 คำ จากวิธี CHI และในวิธี TFIDF จะเป็นการคำนวณค่านำหนักของคำในแต่ละเอกสาร ไม่ได้แยกตามประเภทเอกสาร ทำให้ไม่มีคำที่ซ้ำกันจึงเลือกคำทั้งหมดนั้นกำหนดเป็นคุณลักษณะ

เมื่อกำหนดค่า n=1,000 จะได้เซตของคำที่ทำหน้าที่เป็นคุณลักษณะจำนวน 6,000 คำ จาก 6 ประเภทเอกสาร เมื่อตัดคำที่ซ้ำกันจะได้คุณลักษณะเท่ากับ 2,853 คำ จากวิธีการคำนวณค่าน้ำหนักด้วย TFICF 3,745 คำ จากวิธี IG และ 4,187 คำ จากวิธี CHI

ตารางที่ 4.12 แสดงค่าความเที่ยงตรงและค่าความระลึกตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท ของเอกสารจำนวน 4,200 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	ค่าความเที่ยงตรงและค่าความระลึกแบ่งตามวิธีการคำนวณค่าน้ำหนัก							
	TFIDF		TFICF		IG		CHI	
	P	R	P	R	P	R	P	R
500	90.93	90.52	76.04	73.55	64.91	64.29	68.33	67.81
1,000	92.64	92.48	79.33	76.07	74.43	74.19	74.94	74.52
1,500	92.87	92.74	81.75	79.67	77.61	77.33	79.32	79.21
2,000	-	-	83.38	81.33	79.54	79.33	81.06	80.98
4,000	-	-	85.01	83.81	83.76	83.50	84.83	84.69

หมายเหตุ P=ค่าความเที่ยงตรง R=ค่าความระลึก

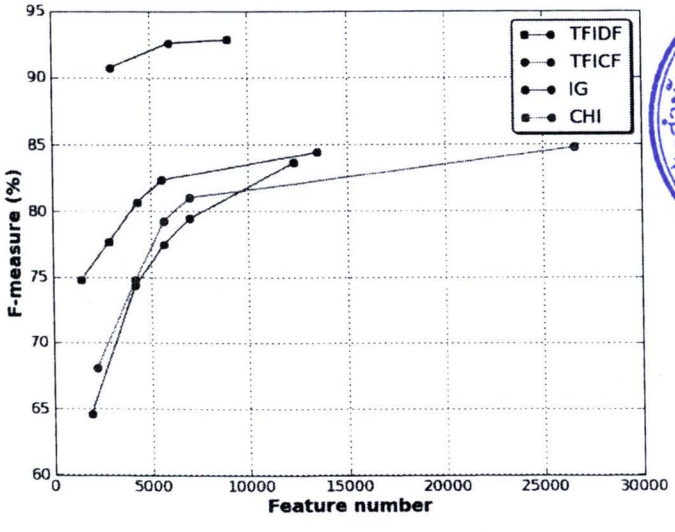
จากตารางที่ 4.12 พบว่าค่าความเที่ยงตรงในแต่ละการทดลองจะมีค่ามากกว่าค่าความระลึก โดยเมื่อจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะเพิ่มมากขึ้นค่าความเที่ยงตรงและค่าความระลึกก็จะมีค่าสูงขึ้น ในชุดเอกสารทดสอบที่ 2 โดยการคำนวณด้วยวิธี TFIDF จะให้ค่าความเที่ยงตรงและค่าความระลึกสูงที่สุด ในขณะเดียวกันเมื่อคำนวณค่าน้ำหนักด้วยวิธี IG จะให้ค่าความเที่ยงตรงและค่าความระลึกน้อยที่สุด

โดยค่าความเที่ยงตรงและค่าความระลึกมีค่าสูง หมายความว่า มีประสิทธิภาพการระบุประเภทเอกสารได้มาก โดยสามารถระบุประเภทเอกสารได้ถูกต้องเป็นจำนวนมาก จากนั้นคำนวณค่าเอฟจากค่าความเที่ยงตรงและค่าความระลึก แสดงดังตารางที่ 4.13

ตารางที่ 4.13 แสดงค่าเอฟตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท โดยที่จำนวนเอกสารเท่ากับ 4,200 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	ค่าเอฟแบ่งตามวิธีการคำนวณค่าน้ำหนัก				ค่าเอฟเฉลี่ยทุกวิธีการคำนวณค่าน้ำหนัก
	TFIDF	TFICF	IG	CHI	
500	90.73	74.77	64.60	68.07	74.54
1,000	92.56	77.67	74.31	74.73	79.82
1,500	92.80	80.69	77.47	79.26	82.53
2,000	-	82.34	79.43	81.02	80.93
4,000	-	84.41	83.63	84.76	84.27

จากตารางที่ 4.13 นำมาสร้างกราฟได้ดังรูปที่ 4.13



รูปที่ 4.13 แสดงค่าเอฟแยกตามจำนวนคำและวิธีการคิดค่าน้ำหนัก ของเอกสารชุดทดสอบที่ 2

จากตารางที่ 4.13 และ รูปที่ 4.13 พบว่าเมื่อจำนวนคุณลักษณะเพิ่มมากขึ้น ค่าเอฟจะเพิ่มมากขึ้นด้วย ในทุก ๆ วิธีการคำนวณค่าน้ำหนักของชุดเอกสารทดสอบที่ 2 ดังรายละเอียดต่อไปนี้ ในวิธีคำนวณค่าน้ำหนักด้วยวิธีต่าง ๆ เมื่อกำหนดคุณลักษณะให้ n = 500, 1,000, 1,500, 2,000 และ 4,000 ตามลำดับ เมื่อค่า n มากขึ้นจำนวนคุณลักษณะก็จะมากขึ้นตามไปด้วย โดยเมื่อใช้วิธี TFICF ค่าเอฟจะมีค่าเท่ากับ 74.77%, 77.67%, 80.69%, 82.34% และ 84.41% ตามลำดับ สำหรับวิธี IG ค่าเอฟจะมีค่าเท่ากับ 64.60%, 74.31%, 77.47%, 79.43 และ 83.63% ตามลำดับ และสำหรับวิธี CHI ค่าเอฟจะมีค่าเท่ากับ 68.07%, 74.73%, 79.26%, 81.02% และ 84.76% ตามลำดับ จากค่าเอฟเหล่านี้อธิบายได้ว่าเมื่อจำนวนคำที่กำหนดเป็นคุณลักษณะเพิ่มมากขึ้นในชุดเอกสารทดสอบกลุ่มที่ 2 ประสิทธิภาพในการระบุประเภทเอกสารจะเพิ่มมากขึ้น หมายความว่ามีการระบุประเภทเอกสารได้ถูกต้องมากขึ้น โดยพิจารณาจากค่าเอฟที่สูงขึ้น

ในทำนองเดียวกันการคำนวณค่าน้ำหนักด้วยวิธี TFIDF นั้น ค่าเอฟจะมีค่าเพิ่มสูงขึ้นเมื่อจำนวนคุณลักษณะเพิ่มมากขึ้น โดยที่เมื่อกำหนดให้ n=500, 1,000 และ 1,500 ค่าเอฟมีค่าเท่ากับ 90.73%, 92.56% และ 92.80%

ค่าเอฟเฉลี่ยของทุกวิธีการคำนวณค่าน้ำหนักแยกตามจำนวนคำที่กำหนดให้เป็นคุณลักษณะจะมีแตกต่างกัน เมื่อกำหนดให้ n มีค่า เท่ากับ 500, 1000, 1500, 2000 และ 4000 จะมีค่าเอฟเฉลี่ยเท่ากับ 74.54%, 79.82%, 82.53%, 80.93% และ 84.27% ตามลำดับ โดยเมื่อจำนวนคำที่กำหนดเป็นคุณลักษณะมีค่าเพิ่มขึ้นแล้วค่าเอฟเฉลี่ยจะมีค่าเพิ่มขึ้นด้วย ยกเว้นเมื่อเพิ่ม n จาก 1,500 เป็น 2,000

จากข้อมูลข้างต้นอธิบายได้ว่าเมื่อจำนวนคำที่กำหนดเป็นคุณลักษณะเพิ่มมากขึ้นในชุดเอกสารทดสอบกลุ่มที่ 2 ประสิทธิภาพในการระบุประเภทเอกสารจะเพิ่มมากขึ้น หมายความว่ามีการระบุประเภทเอกสารได้ถูกต้องมากขึ้น โดยพิจารณาจากค่าเอฟที่สูงขึ้น และการคำนวณค่าน้ำหนักด้วยวิธี TFIDF จะให้ค่าเอฟที่มากกว่าการคำนวณค่าน้ำหนักด้วยวิธีอื่นที่ทำการเปรียบเทียบ ดังนั้นเมื่อใช้วิธี TFIDF เมื่อกำหนดค่า Threshold เท่ากับ 5 ในการกำหนดคำที่เป็นคุณลักษณะจะให้การระบุประเภทเอกสารมีความถูกต้องมากกว่าวิธีอื่น

ชุดเอกสารทดสอบ กลุ่มที่ 3 ซึ่งมีเอกสารจำนวน 6,000 เอกสาร แล้วทดลองตามจำนวนคำที่กำหนดเป็นคุณลักษณะที่แตกต่างกันไป โดยกำหนดจากค่า n ซึ่งแสดงในตารางที่ 4.14

ตารางที่ 4.14 ตารางแสดงจำนวนคำไม่ซ้ำในเอกสารแต่ละประเภทตามจำนวนคำที่เลือกของเอกสารแต่ละประเภท ของเอกสารจำนวน 6,000 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	รวมจำนวนคำที่ไม่ซ้ำในเอกสาร 6 ประเภท			
	TFIDF	TFICF	IG	CHI
500	3,000	1,453	1,870	2,235

1,000	6,000	2,851	3,729	4,260
1,500	9,000	4,523	5,565	6,014
2,000	-	5,626	7,360	7,410
4,000	-	12,453	12,601	11,685

จากตารางที่ 4.14 เมื่อกำหนดให้เลือกคำที่ทำหน้าที่เป็นคุณลักษณะจากแต่ละประเภทเอกสาร ประเภทละ 500 8e (n=500) จะได้เซตของคำทั้งหมดที่ทำหน้าที่เป็นคุณลักษณะจำนวน 3,000 คำ จาก 6 ประเภทเอกสาร ซึ่งในเซตของคำนี้จะมีคำซ้ำกัน เนื่องจากคำ ๆ หนึ่งสามารถเป็นตัวแทนของประเภทเอกสารได้หลายประเภท ทำให้ต้องตัดคำที่ซ้ำกันออก ดังนั้น จะได้จำนวนคำเท่ากับ 1,453 คำ จากวิธีการคำนวณค่าน้ำหนักด้วย TFICF 1,870 คำ จากวิธี IG และ 2,235 คำ จากวิธี CHI และในวิธี TFIDF จะเป็นการคำนวณค่านำหนักของคำในแต่ละเอกสาร ไม่ได้แยกตามประเภทเอกสาร ทำให้ไม่มีคำที่ซ้ำกันจึงเลือกคำทั้งหมดนั้นกำหนดเป็นคุณลักษณะ

เมื่อกำหนดค่า n=1,000 จะได้เซตของคำที่ทำหน้าที่เป็นคุณลักษณะจำนวน 6,000 คำ จาก 6 ประเภทเอกสาร เมื่อตัดคำที่ซ้ำกันจะได้คุณลักษณะเท่ากับ 2,851 คำ จากวิธีการคำนวณค่าน้ำหนักด้วย TFICF 3,729 คำ จากวิธี IG และ 4,260 คำจากวิธี CHI

ตารางที่ 4.15 แสดงค่าความเที่ยงตรงและค่าความระลึกตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท ของเอกสารจำนวน 6,000 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	ค่าความเที่ยงตรงและค่าความระลึกแบ่งตามวิธีการคำนวณค่าน้ำหนัก							
	TFIDF		TFICF		IG		CHI	
	P	R	P	R	P	R	P	R
500	91.34	90.95	72.41	67.86	62.66	61.71	66.03	65.77
1,000	92.74	92.61	77.60	75.43	71.92	70.88	73.24	72.88
1,500	93.04	92.93	80.60	75.60	77.33	76.71	77.15	76.40
2,000	-	-	82.07	78.43	79.74	79.45	79.22	78.80
4,000	-	-	84.70	83.30	83.06	82.95	85.14	85.06

หมายเหตุ P=ค่าความเที่ยงตรง R=ค่าความระลึก

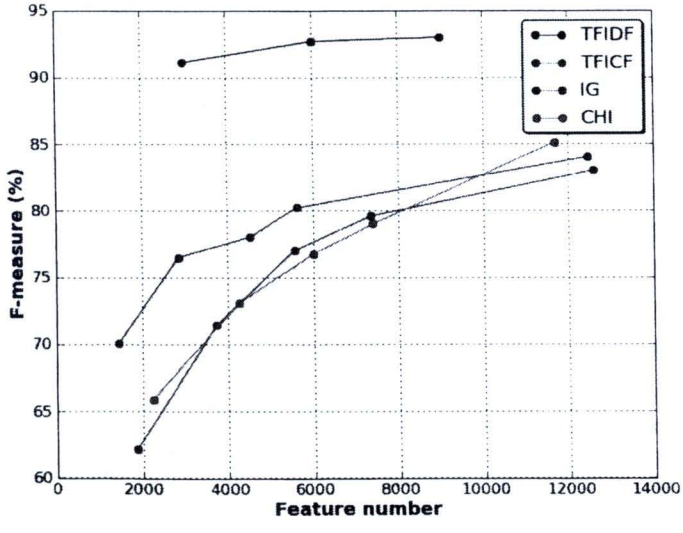
จากตารางที่ 4.15พบว่าค่าความเที่ยงตรงในแต่ละการทดลองจะมีค่ามากกว่าค่าความระลึก โดยเมื่อจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะเพิ่มมากขึ้นค่าความเที่ยงตรงและค่าความระลึกก็จะมีค่าสูงขึ้นในชุดเอกสารทดสอบที่ 3 โดยการคำนวณด้วยวิธี TFIDF ยังคงให้ค่าความเที่ยงตรงและค่าความระลึกสูงที่สุด ซึ่งหมายความว่า การคำนวณค่าน้ำหนักด้วยวิธี TFIDF จะให้ประสิทธิภาพในการระบุประเภทเอกสารดีที่สุด

ในขณะเดียวกันเมื่อคำนวณค่าน้ำหนักด้วยวิธี IG ก็ยังคงให้ค่าความเที่ยงตรงและค่าความระลึกน้อยที่สุด หมายความว่า การคำนวณค่าน้ำหนักด้วยวิธี IG จะให้ประสิทธิภาพในการระบุประเภทเอกสารน้อยที่สุด หรือจะกล่าวว่าระบุประเภทเอกสารให้กับเอกสารได้ถูกต้องน้อยที่สุดเมื่อเปรียบเทียบกับการคำนวณค่าน้ำหนักด้วยวิธีอื่น ๆ จากนั้นคำนวณค่าเอฟจากค่าความเที่ยงตรงและค่าความระลึก ได้ดังตารางที่ 4.16

ตารางที่ 4.16 แสดงค่าเอฟตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท โดยที่จำนวนเอกสารเท่ากับ 6,000 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	ค่าเอฟแบ่งตามวิธีการคำนวณค่าน้ำหนัก				ค่าเอฟเฉลี่ยทุกวิธีการคำนวณค่าน้ำหนัก
	TFIDF	TFICF	IG	CHI	
500	91.14	70.06	62.19	65.90	72.32
1,000	92.68	76.50	71.39	73.06	78.41
1,500	92.98	78.02	77.02	76.77	81.20
2,000	-	80.21	79.59	79.00	79.60
4,000	-	84.00	83.00	85.10	84.03

จากตารางที่ 4.16 นำมาสร้างกราฟได้ดังรูปที่ 4.14



รูปที่ 4.14 แสดงค่าเอฟแยกตามจำนวนคำที่กำหนดเป็นคุณลักษณะและวิธีการคิดค่าน้ำหนัก ของเอกสารชุดทดสอบที่ 3

จากตารางที่ 4.16 และรูปที่ 4.14 พบว่าเมื่อมีจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะเพิ่มขึ้น ค่าเอฟก็จะมากขึ้นด้วยในทุก ๆ วิธีการคำนวณค่าน้ำหนักของชุดเอกสารทดสอบที่ 3 ดังรายละเอียดต่อไปนี้ ในวิธีคำนวณค่าน้ำหนักด้วยวิธีต่าง ๆ เมื่อค่า n มากขึ้นจำนวนคำที่กำหนดเป็นคุณลักษณะก็จะมากขึ้นตามไปด้วย โดยกำหนดคุณลักษณะให้ n = 500, 1000, 1500, 2000 และ 4000 ตามลำดับ เมื่อใช้วิธี TFICF ค่าเอฟจะมีค่าเท่ากับ 70.06%, 76.50%, 78.02%, 80.21% และ 84.00% ตามลำดับ สำหรับวิธี IG ค่าเอฟจะมีค่าเท่ากับ 64.60%, 71.39%, 76.77%, 79.59% และ 83.00% ตามลำดับ และสำหรับวิธี CHI ค่าเอฟจะมีค่าเท่ากับ 65.90%, 73.06%, 79.26%, 79.00% และ 85.10% ตามลำดับ จากค่าเอฟเหล่านี้อธิบายได้ว่าเมื่อจำนวนคำที่กำหนดเป็นคุณลักษณะเพิ่มมากขึ้นในชุดเอกสารทดสอบกลุ่มที่ 3 ประสิทธิภาพในการระบุประเภทเอกสารจะเพิ่มมากขึ้น หมายความว่ามีการระบุประเภทเอกสารได้ถูกต้องมากขึ้น โดยพิจารณาจากค่าเอฟที่สูงขึ้น

ในทำนองเดียวกันการคำนวณค่าน้ำหนักด้วยวิธี TFIDF นั้น ค่าเอฟจะมีค่าเพิ่มสูงขึ้นเมื่อจำนวนคุณลักษณะเพิ่มมากขึ้น โดยที่เมื่อกำหนดให้ n=500, 1000 และ 1500 ค่าเอฟมีค่าเท่ากับ 91.14%, 92.68% และ 92.98%

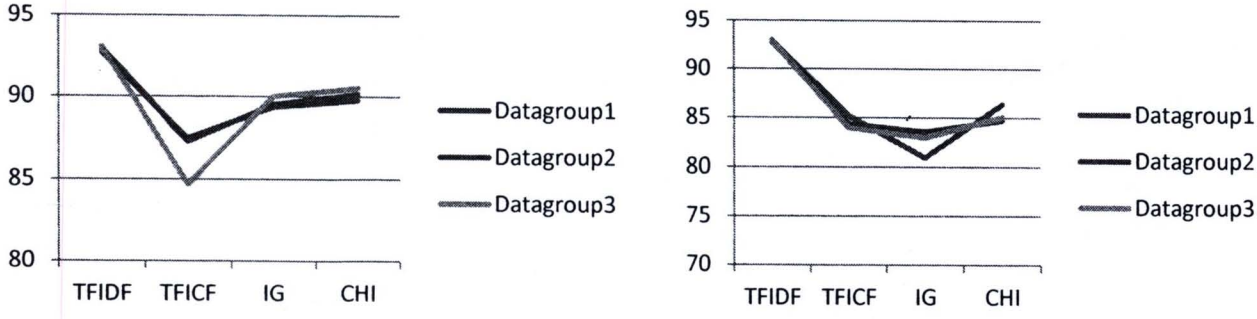
ค่าเอฟเฉลี่ยของทุกวิธีการคำนวณค่าน้ำหนักแยกตามจำนวนคำที่กำหนดให้เป็นคุณลักษณะจะมีแตกต่างกัน เมื่อกำหนดให้ n มีค่า เท่ากับ 500, 1000, 1500, 2000 และ 4000 จะมีค่าเอฟเฉลี่ยเท่ากับ 72.32%, 78.41%, 81.20%, 79.60% และ 84.03% ตามลำดับ โดยเมื่อจำนวนคำที่กำหนดเป็นคุณลักษณะมีค่าเพิ่มขึ้นแล้วค่าเอฟเฉลี่ยจะมีค่าเพิ่มขึ้นด้วย ยกเว้นเมื่อเพิ่ม n จาก 1,500 เป็น 2,000

จากข้อมูลข้างต้นอธิบายได้ว่าเมื่อจำนวนคำที่เป็นคุณลักษณะเพิ่มมากขึ้นในชุดเอกสารทดสอบกลุ่มที่ 3 ประสิทธิภาพในการระบุประเภทเอกสารจะเพิ่มมากขึ้น หมายความว่ามีการระบุประเภทเอกสารได้ถูกต้องมากขึ้น โดยพิจารณาจากค่าเอฟที่สูงขึ้น และการคำนวณค่าน้ำหนักด้วยวิธี TFIDF ก็ยังคงให้ค่าเอฟที่มากกว่าการคำนวณค่าน้ำหนักด้วยวิธีอื่นที่ทำการเปรียบเทียบเหมือนกับการทดลองกับชุดเอกสารทดสอบกลุ่มที่ 1 และ 2 ดังนั้นเมื่อใช้วิธี TFIDF ในการกำหนดคุณลักษณะจะให้การระบุประเภทเอกสารมีความถูกต้องมากกว่าวิธีอื่น

4.3 อธิบายผลการทดลอง

จากการทดลองข้างต้นสามารถสรุปตามสมมติฐานดังนี้

1. จำนวนเอกสารที่ใช้ในการฝึกสอนมีผลต่อการระบุประเภทเอกสารหรือไม่ (Effect of training set size on performance)



(a) (b)

รูปที่ 4.15 แสดงความสัมพันธ์ระหว่างค่าเอฟกับจำนวนเอกสารที่ใช้ในการทดสอบ
(a) ไม่มีการเลือกคุณลักษณะ (b) เลือกคุณลักษณะ

จากรูปที่ 4.15 Datagroup1 (ชุดเอกสารทดสอบที่ 1) ใช้เอกสารทดสอบจำนวน 3,000 เอกสาร Datagroup2 (ชุดเอกสารทดสอบที่ 2) ใช้เอกสารทดสอบจำนวน 4,200 เอกสาร Datagroup3 (ชุดเอกสารทดสอบที่ 3) ใช้เอกสารทดสอบจำนวน 6,000 เอกสาร อธิบายได้ว่า ในกรณีที่ไม่มีการเลือกจำนวนคำที่กำหนดเป็นคุณลักษณะ (รูปที่ 4.15(a)) แสดงว่าทุกคำจะถูกกำหนดให้เป็นคุณลักษณะ เมื่อจำนวนเอกสารทดสอบเพิ่มมากขึ้น ค่าเอฟก็จะสูงขึ้นด้วย ซึ่งค่าเอฟสูงนั้นหมายความว่าการระบุประเภทเอกสารมีความถูกต้องมาก เมื่อกำหนดค่าน้ำหนักด้วยวิธี TFIDF แต่ในวิธี IG และ CHI ค่าเอฟจะลดลงใน Datagroup2 และเพิ่มขึ้นใน Datagroup3 ขณะที่ TFICF ค่าเอฟจะเพิ่มขึ้นใน Datagroup2 และลดลงใน Datagroup3

ในกรณีที่มีการเลือกคำที่กำหนดเป็นคุณลักษณะ (รูปที่ 4.15(b)) อธิบายได้ว่า ในวิธี TFIDF และ CHI ค่าเอฟจะลดลงใน Datagroup2 และเพิ่มขึ้นใน Datagroup3 ขณะที่ TFICF ค่าเอฟจะลดลงเมื่อจำนวนเอกสารทดสอบเพิ่มมากขึ้น และ IG ค่าเอฟจะเพิ่มขึ้นใน Datagroup2 และลดลงใน Datagroup3

ดังนั้นจึงสามารถสรุปได้ดังนี้คือ ความแตกต่างของจำนวนเอกสารที่ใช้ในการเรียนรู้หรือฝึกสอนนั้นส่งผลต่อประสิทธิภาพการระบุประเภทเอกสาร โดยถ้าเราไม่เลือกจำนวนคุณลักษณะทุกวิธีการคำนวณค่าน้ำหนักยกเว้น TFICF จะมีประสิทธิภาพดีในการระบุประเภทเอกสาร นั้นหมายความว่าจำนวนเอกสารมากก็ยิ่งทำให้การระบุประเภทเอกสารดีขึ้น

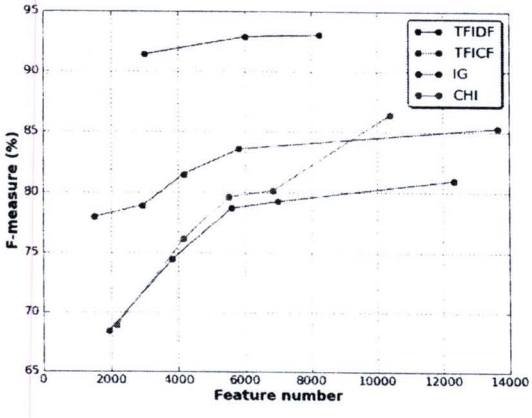
แต่เมื่อมีการลดจำนวนคำที่กำหนดเป็นคุณลักษณะเพื่อลดขนาดของเวกเตอร์แต่ละเอกสารจะมีความไม่แน่นอนในประสิทธิภาพของการระบุประเภทของเอกสารขึ้นอยู่กับแต่ละวิธีการคำนวณค่าน้ำหนัก อย่างไรก็ตามการคำนวณค่าน้ำหนักด้วย TFIDF และ CHI มีแนวโน้มไปทางเดียวกัน คือจำนวนเอกสารที่ใช้ในการทดสอบมีจำนวนมากจะมีผลทำให้ประสิทธิภาพของการระบุประเภทเอกสารเพิ่มมากขึ้นด้วย

2. จำนวนคุณลักษณะมีผลต่อการจัดกลุ่มหรือไม่ (Effect of feature size on performance)

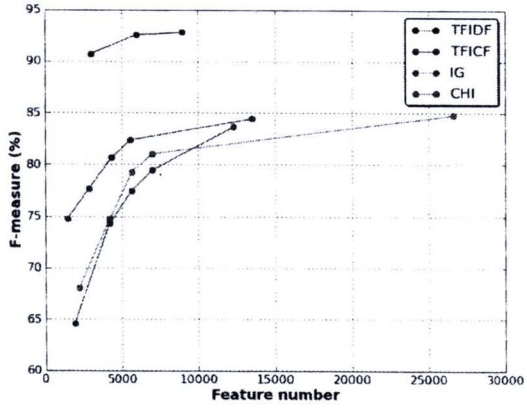
จากรูปที่ 4.16 อธิบายได้ว่าในทุก ๆ วิธีการคำนวณค่าน้ำหนักและทุก ๆ กลุ่มเอกสารทดสอบ เมื่อจำนวนคำที่กำหนดเป็นคุณลักษณะเพิ่มมากขึ้นมีผลทำให้ค่าเอฟมีค่าเพิ่มมากขึ้นตามไปด้วย ดังนั้นจึง

สามารถสรุปได้ว่าจำนวนคุณลักษณะมีผลต่อประสิทธิภาพการจัดกลุ่มเอกสาร โดยเมื่อค่าเอฟสูงหมายความว่าการระบุประเภทเอกสารมีความถูกต้องมาก ในทางตรงกันข้ามถ้าค่าเอฟมีค่าต่ำแสดงว่าการระบุเอกสารนั้นมีความผิดพลาดค่อนข้างมาก โดยระบุประเภทเอกสารไม่ถูกต้อง

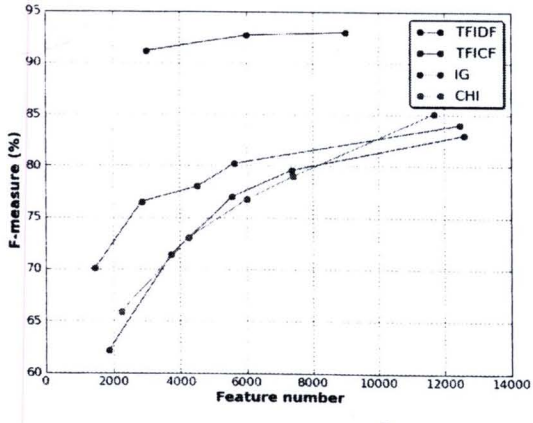
จากกราฟสามารถอธิบายเพิ่มได้ว่า ในทุกชุดเอกสารทดสอบการคำนวณค่านำหนักด้วยวิธี IG และ CHI จะให้ค่าเอฟ (F-measure) ที่ใกล้เคียงกัน คือเป็นค่าที่ระบุความถูกต้องของการระบุประเภทเอกสารที่ไม่ค่อยดีนัก ซึ่งมีการกำหนดจำนวนคำที่กำหนดปืนคุณลักษณะที่ไม่มาก คืออยู่ระหว่าง 2,000-7,000 คำ แต่เมื่อจำนวนเพิ่มจำนวนคุณลักษณะ (มากกว่า 8,000 คำ) แล้วจะมีความแตกต่างของค่าเอฟอย่างชัดเจน แสดงดังกราฟ แต่อย่างไรก็ตามวิธี TFIDF (โดยกำหนด Threshold=5) นั้นให้ค่าเอฟที่สูงในทุก ๆ ชุดเอกสารทดสอบ แสดงว่ามีความถูกต้องในการระบุประเภทเอกสารที่มีประสิทธิภาพ



(a) ชุดเอกสารทดสอบที่ 1



(b) ชุดเอกสารทดสอบที่ 2



(c) ชุดเอกสารทดสอบที่ 3

รูปที่ 4.16 แสดงความสัมพันธ์ระหว่างค่าเอฟกับจำนวนคำที่กำหนดเป็นคุณลักษณะ ในแต่ละชุดเอกสารทดสอบ

เมื่อคำนวณค่าเฉลี่ยของค่าเอฟทุกวิธีการคำนวณค่าน้ำหนักแยกตามจำนวนคำที่เลือกในเอกสารแต่ละประเภท ตามตารางที่ 4.17 พบว่าเมื่อเพิ่มจำนวนคำที่เลือกเพื่อกำหนดเป็นคำสำคัญมากขึ้นจาก 500, 1000, 1500, 2000 และ 4000 ในเอกสารชุดทดสอบที่ 1 ค่าเฉลี่ยค่าเอฟมีค่าเท่ากับ 76.64%, 80.57%, 83.18%, 80.97% และ 84.19% ตามลำดับ และค่าเฉลี่ยค่าเอฟมีการเพิ่มขึ้นเป็น 3.93%, 2.61%,-2.21% และ 3.22% ตามลำดับ โดยเฉลี่ยมีความถูกต้องเพิ่มขึ้น 1.88% ในเอกสารชุดทดสอบที่ 2 ค่าเฉลี่ยค่าเอฟมีค่าเท่ากับ 74.54%, 79.82%, 82.53%, 80.93% และ 84.27% ตามลำดับ และค่าเฉลี่ยค่าเอฟมีการเพิ่มขึ้นเป็น 5.28%, 2.71%, -1.60% และ 3.34% ตามลำดับ โดยเฉลี่ยมีความถูกต้องเพิ่มขึ้น 2.43% ในเอกสารชุดทดสอบที่ 3 ค่าเฉลี่ยค่าเอฟมีค่าเท่ากับ 72.32%, 78.41%, 81.20%, 79.60% และ 84.03% ตามลำดับ และค่าเฉลี่ยค่าเอฟมีการเพิ่มขึ้นเป็น 6.09%, 2.79%, -1.60% และ 4.43% ตามลำดับ โดยเฉลี่ยมีความถูกต้องเพิ่มขึ้น 2.93% จากการทดลองในทุกชุดเอกสารทดสอบแล้วค่าเฉลี่ยค่าเอฟจะมีค่าเพิ่มขึ้นยกเว้น เมื่อกำหนดจำนวนคำจากเดิมเท่ากับ 1500 เป็น 2000 ค่าเฉลี่ยค่าเอฟจะมีค่าลดลง

ค่าเฉลี่ยของผลรวมค่าเฉลี่ยเอฟของทุกชุดเอกสารทดสอบ เมื่อกำหนดให้จำนวนคำที่เลือกเป็นคำสำคัญเท่ากับ 500, 1000, 1500, 2000 และ 4000 นั้น ผลรวมของค่าเฉลี่ยของค่าเฉลี่ยเอฟเท่ากับ 74.50%, 79.60%, 82.30%, 80.50% และ 84.16% ตามลำดับ และค่าเฉลี่ยมีการเพิ่มขึ้นเป็น 5.10%, 2.70%, -1.80% และ 3.66% ตามลำดับ โดยเฉลี่ยมีความถูกต้องเพิ่มขึ้น 2.41%

ตารางที่ 4.17 แสดงค่าเฉลี่ยค่าเอฟแบ่งตามจำนวนคำที่เลือกในเอกสารแต่ละประเภท และจำนวนเอกสารตามชุดเอกสารทดสอบที่ 1, 2 และ 3

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	ค่าเฉลี่ยค่าเอฟแบ่งตามชุดเอกสารทดสอบ						ค่าเฉลี่ยของผลรวมค่าเฉลี่ยเอฟของทุกชุดเอกสารทดสอบ	
	ชุดเอกสารทดสอบที่ 1 (3,000 เอกสาร)		ชุดเอกสารทดสอบที่ 2 (4,200 เอกสาร)		ชุดเอกสารทดสอบที่ 3 (6,000 เอกสาร)			
	ค่าเฉลี่ย	ผลต่าง	ค่าเฉลี่ย	ผลต่าง	ค่าเฉลี่ย	ผลต่าง	ค่าเฉลี่ย	ผลต่าง
500	76.64	-	74.54	-	72.32	-	74.50	-
1,000	80.57	3.93	79.82	5.28	78.41	6.09	79.60	5.10
1,500	83.18	2.61	82.53	2.71	81.20	2.79	82.30	2.70
2,000	80.97	-2.21	80.93	-1.60	79.60	-1.60	80.50	-1.80
4,000	84.19	3.22	84.27	3.34	84.03	4.43	84.16	3.66
เพิ่มขึ้นโดยเฉลี่ย (%)	1.88		2.43		2.93		2.41	

3. ค่า Threshold ที่ใช้ในการเลือกคำที่กำหนดเป็นคุณลักษณะด้วยการคำนวณวิธี TFIDF มีผลอย่างไร

ในการกำหนดค่า Threshold ในงานวิจัยนี้หมายถึง การกำหนดค่าต่ำสุดของความถี่ของคำที่จะนำมาพิจารณาเพื่อกำหนดเป็นคุณลักษณะโดยเลือกค่า Threshold ในการทดลองจำนวน 4 ค่า ได้แก่ 3, 4, 5 และ 6 สามารถอธิบายได้ว่า กำหนดให้ Threshold เท่ากับ 3 หมายความว่า เลือกคำ (ที่ผ่านการตัดคำฟุ่มเฟือยและแปลงให้อยู่ในรูปรากศัพท์) ที่มีความถี่มากกว่า 3 นำมาคำนวณค่าน้ำหนักด้วยวิธี TFIDF

จากการทดลองที่ 1 ค่า Threshold ที่เหมาะสมที่ทำให้การจัดกลุ่มเอกสารมีประสิทธิภาพดีที่สุด ได้แก่ ค่า Threshold เท่ากับ 5 นั้นหมายความว่า คำที่มีความถี่มากกว่า 5 ครั้งนั้นคาดว่าน่าจะคำที่มีความสำคัญที่สามารถนำมาใช้เป็นตัวแทนของเอกสาร โดยสามารถระบุประเภทเอกสารได้ดีกว่าการกำหนดด้วยค่า Threshold อื่น ๆ

4. เปรียบเทียบวิธีการคำนวณค่าน้ำหนักระหว่าง TFIDF, TFICF, IG และ CHI (Performance comparison between four approaches)

จากรูปที่ 1.14 และ 4.15 อธิบายได้ว่า การจัดกลุ่มโดยการใช้วิธีการคำนวณค่าน้ำหนักด้วย TFIDF ในงานวิจัยนี้ ให้ประสิทธิภาพการระบุประเภทเอกสารดีที่สุด นั้นคือมีการระบุประเภทเอกสารที่ถูกต้องค่อนข้างมากเมื่อเทียบกับวิธีอื่น ๆ โดยพิจารณาจากค่าเอฟที่มากกว่าวิธีอื่น โดยค่าเอฟที่มากที่สุดที่คำนวณด้วยวิธีการคำนวณค่าน้ำหนัก TFIDF ที่ Threshold เท่ากับ 5 นั้นมีค่าเท่ากับ 93.05% โดยที่การคำนวณค่าน้ำหนักด้วยวิธีอื่น ๆ นั้นมีค่าอยู่ระหว่าง 80-90% นั้นหมายความว่าการใช้วิธีการคำนวณค่าน้ำหนักด้วยวิธี TFIDF ที่ Threshold เท่ากับ 5 มีความถูกต้องสูงกว่าการคำนวณด้วยวิธีอื่น