

บทที่ 3

วิธีการดำเนินงานวิจัย

งานวิจัยนี้เป็นการจำแนกหัวเรื่องด้วยวิธีการแบ่งเว็บเพจเป็นส่วนย่อยและซอฟต์แวร์แมชชีน ซึ่งในส่วนนี้อธิบายรายละเอียดของขอบเขตการทดลอง เครื่องมือและซอฟต์แวร์ที่ใช้ในการพัฒนาระบบ รายละเอียดขั้นตอนของการทดลอง และวิธีการวัดผลทดลอง มีดังต่อไปนี้

3.1 ขอบเขตการทดลอง

ขอบเขตการทดลองสำหรับงานวิจัยชิ้นนี้ประกอบด้วย ข้อมูลที่ใช้ทดลอง และกลุ่มผู้ประเมิน โดยมีรายละเอียดดังนี้

ข้อมูลที่ใช้ทดลอง

เนื่องจากข้อมูลที่ใช้ทดสอบจำแนกหัวเรื่องจากรูปแบบการแสดงผลยังไม่มีข้อมูลที่ใช้ทดสอบที่มีมาตรฐาน แต่ใน (Brudvik, 2008) ได้เสนอชุดข้อมูลเว็บเพจที่นำมาใช้ในทดสอบการจำแนกหัวเรื่อง ได้แก่ 64 เว็บเพจจาก 10 หมวดหมู่ของเว็บไซต์ของ Alexa ได้แก่หมวด Business, Computers, Health, Home, News, Recreation, Science, Society, Sports, World และ 36 เว็บเพจของ 500 เว็บไซต์อันดับสูงสุดจากผลลัพธ์การค้นหาของกูเกิลด้วยคำว่า "Web Accessibility" รวมทั้งหมด 100 เว็บเพจ

กลุ่มผู้ประเมิน

ผู้ประเมินเป็นนักศึกษาปริญญาโท คณะวิทยาศาสตร์และเทคโนโลยี ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยธรรมศาสตร์ จำนวน 5 คน ทำการระบุส่วนที่เป็นหัวเรื่องที่ปรากฏในหน้าเว็บเพจจากรูปแบบการแสดงผล

3.2 เครื่องมือและซอฟต์แวร์ที่ใช้ในการพัฒนาระบบ

เครื่องมือและซอฟต์แวร์ที่นำมาใช้ในการพัฒนาการจำแนกหัวเรื่องจากรูปแบบการแสดงผล แสดงรายละเอียดดังตารางที่ 3.1

ตารางที่ 3.1

แสดงรายละเอียดของเครื่องมือและซอฟต์แวร์ที่ใช้ในการพัฒนาระบบ

ฮาร์ดแวร์ (Hardware)	Intel® Dual Core™ processor T2050 (1.60 GHz) 1024 MB DDR2
ระบบปฏิบัติการ (Operating System)	Window XP Professional
ซอฟต์แวร์ (Software)	SVM ^{light} 6.01, Visual studio 2005, Eclipse

3.3 ขั้นตอนการทดลอง

สำหรับรายละเอียดของขั้นตอนการทดลองแบ่งออกเป็น 4 ขั้นตอน คือ การประมวลผลเบื้องต้น (Pre-processing), การแบ่งเว็บเพจเป็นส่วนย่อยด้วยวิธีไอพีเอสอัลกอริทึม (Web page segmentation using VIPS algorithms), การดึงลักษณะเด่น (Feature Extraction) และการจำแนกหัวเรื่องด้วยซัพพอร์ตเวกเตอร์แมชชีน (Web Heading Classification using Support Vector Machine) แสดงดังภาพที่ 3.1 โดยมีรายละเอียดดังต่อไปนี้

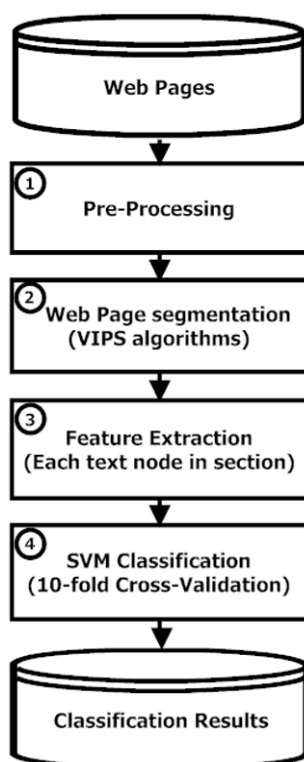
3.3.1 ขั้นตอนการประมวลผลเบื้องต้น (Pre-processing step)

ขั้นตอนการประมวลผลเบื้องต้นเป็นขั้นตอนการทำงานเบื้องต้นก่อนจะนำข้อมูลไปคำนวณในส่วนถัดไป โดยทำการปรับข้อมูลเว็บเพจก่อนที่จะทำการแบ่งเว็บเพจเป็นส่วนย่อย เพื่อแยกแยะการดึงลักษณะเด่นเมื่อเว็บเพจถูกแบ่งออกเป็นส่วนย่อยแล้ว โดยทำการโหลดเว็บเพจเข้ามาในระบบ และใช้คอมเดินทางไปตามโหนดต่างๆของเว็บเพจนั้นเพื่อถ่ายทอดคุณสมบัติที่โหนดพ่อแม่มีไปสู่โหนดลูก เช่น ถ้าโหนดพ่อแม่มีการกำหนดใช้สีตัวอักษรเป็นสีแดง โหนดลูกก็ต้องมีคุณสมบัติเช่นเดียวกับโหนดพ่อแม่ ยกเว้นโหนดลูกมีคุณสมบัติที่เป็นประเภทเดียวกับโหนดพ่อแม่ เช่น ถ้า

โหนดพ่อแม่มีการกำหนดสีตัวอักษรเป็นสีแดง แต่โหนดลูกมีการกำหนดสีตัวอักษรเป็นสีฟ้า ดังนั้น โหนดลูกจะมีสีตัวอักษรเป็นสีฟ้า

ภาพที่ 3.1

แสดงกระบวนการทำงานการจำแนกหัวเรื่อง



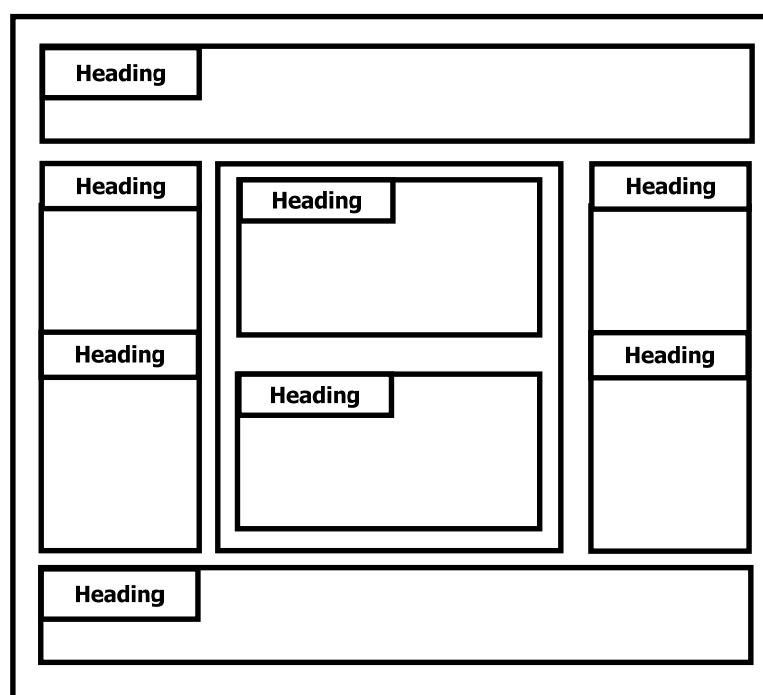
3.3.2 ขั้นตอนการแบ่งเว็บเพจออกเป็นส่วย่อยด้วยวีไอพีเอสอัลกอริทึม (Web page segmentation using VIPS algorithm)

ขั้นตอนการแบ่งเว็บเพจเป็นส่วนย่อยจะนำเว็บเพจจากขั้นตอนการประมวลผลเบื้องต้นมาแบ่งออกเป็นส่วย่อยตามรูปแบบการแสดงผลที่เกี่ยวข้องกัน เนื่องจากผู้ออกแบบเว็บไซต์จะสร้างเว็บเพจโดยแบ่งแยกเนื้อหาที่ต่างกันออกเป็นส่วย่อย และสร้างหัวเรื่องเพื่อแทนข้อมูลโดยสรุปของเนื้อหาภายในส่วย่อยนั้นโดยใช้รูปแบบการแสดงผลที่แตกต่างจากส่วอื่นของส่วย่อยอย่างชัดเจน ตัวอย่างโครงสร้างของเว็บเพจที่แบบแยกเนื้อหาที่ต่างกันเป็นส่วย่อยแสดงดังภาพที่ 3.2 ดังนั้นการแบ่งเว็บเพจเป็นส่วนย่อยทำให้ทราบว่าโหนดข้อความที่จะจำแนกว่าเป็นหัวเรื่องอยู่

ส่วนย่อยใดของเว็บเพจเพื่อนำเสนอข้อความทั้งหมดของส่วนย่อยนั้นมาเปรียบเทียบว่ามีรูปแบบการแสดงผลแตกต่างจากรูปแบบการแสดงผลของข้อความทั้งหมดของส่วนย่อยเป็นเช่นใด

ภาพที่ 3.2

โครงสร้างของเว็บเพจที่แบ่งเนื้อหาที่ต่างกันออกเป็นส่วนย่อย



ในงานวิจัยนี้ใช้วิธีไอพีเอสอัลกอริทึม (Cai, 2003) เป็นอัลกอริทึมที่ใช้ในการแบ่งหน้าเว็บเพจออกเป็นส่วนย่อยซึ่งมาจากการรวมโครงสร้างแบบดอม (DOM Structure) และ โครงสร้างเนื้อหา (Vision-based Content Structure) เพื่อแบ่งหน้าเว็บเพจให้มีลักษณะตามโครงสร้างของเนื้อหาสำคัญตามความหมาย

ตารางที่ 3.2

รายละเอียด 10 ลักษณะเด่นที่นำเสนอในงานวิจัยนี้

ประเภท	ประเภทย่อย	ลักษณะเด่น
การแสดงผล	ขนาดตัวอักษร	1. โหนดข้อความมีขนาดตัวอักษรใหญ่ที่สุดในส่วนย่อย
		2. โหนดข้อความขนาดตัวอักษรใหญ่กว่าค่าเฉลี่ย แต่ไม่ใช่ขนาดตัวอักษรที่มีขนาดใหญ่ที่สุดในส่วนย่อย
		3. โหนดข้อความมีขนาดตัวอักษรเท่ากับค่าเฉลี่ยขนาดตัวอักษรของส่วนย่อย
		4. โหนดข้อความมีขนาดตัวอักษรเล็กกว่าค่าเฉลี่ย แต่ไม่ใช่ตัวอักษรที่มีขนาดเล็กที่สุดในส่วนย่อย
		5. โหนดข้อความมีขนาดตัวอักษรเล็กที่สุดในส่วนย่อย
	สีตัวอักษร	6. โหนดข้อความมีสีตัวอักษรต่างจากสีตัวอักษรหลักของส่วนย่อยหรือไม่
	สีพื้นหลังตัวอักษร	7. โหนดข้อความมีสีพื้นหลังตัวอักษรต่างจากสีพื้นหลังหลักของส่วนย่อยหรือไม่
	ไอคอนหน้าข้อความ	8. โหนดข้อความมีการใช้ไอคอนหน้าข้อความหรือไม่
	การไม่ย่อหน้าของข้อความ	9. โหนดข้อความมีการใช้รูปแบบการย่อหน้าหรือไม่
ตำแหน่ง	ลำดับของเลขที่เอนแวล	10. โหนดข้อความมีลำดับภายในส่วนย่อยมีค่าน้อยกว่า 10 หรือไม่

3.3.3 ขั้นตอนการสกัดลักษณะเด่น (Feature extraction step)

ขั้นตอนการสกัดลักษณะเด่นจะแบ่งออกเป็น 2 ขั้นตอนย่อยได้แก่ การกำหนดลักษณะเด่นของข้อมูลและขั้นตอนการสกัดลักษณะเด่น

3.3.3.1 การกำหนดลักษณะเด่นของข้อมูล (Feature selection)

งานวิจัยนี้แบ่งลักษณะเด่นที่ใช้ในการจำแนกหัวเรื่องจากรูปแบบของการแสดงผลเป็น 2 ประเภท คือ ลักษณะเด่นเกี่ยวกับการแสดงผล และ ลักษณะเด่นเกี่ยวกับตำแหน่ง รวมทั้งหมด 10 ลักษณะเด่น ดังตารางที่ 3.2 ซึ่งมีรายละเอียดดังต่อไปนี้

3.3.3.1.1 ลักษณะเด่นเกี่ยวกับการแสดงผล

ลักษณะเด่นเกี่ยวกับการการแสดงผล ผู้วิจัยนำลักษณะเด่นจากงานวิจัย (Masahiro, 2007) ซึ่งได้จากการทดลองโดยใช้คนในการจำแนกหัวเรื่องของเว็บเพจจากรูปแบบของการแสดงผล ดังนี้

1. ขนาดตัวอักษร (Size of characters): ลักษณะเด่นเกี่ยวกับการแสดงผลของขนาดตัวอักษรของโหนดข้อความจะแยกย่อยได้ 5 ลักษณะเด่นและมีขั้นตอนในการการหาลักษณะเด่นขนาดตัวอักษรของข้อความดังนี้ ขั้นแรกทำการหาค่าเฉลี่ยขนาดตัวอักษรของโหนดทั้งหมดในส่วนย่อย ต่อมาเมื่อได้ค่าเฉลี่ยขนาดตัวอักษรของโหนดข้อความทั้งหมดในส่วนย่อยแล้วจึงนำขนาดตัวอักษรของโหนดข้อความข้อความที่ต้องการจะจำแนกว่าเป็นหัวเรื่องหรือไม่ มาเปรียบเทียบกับค่าสัมพัทธ์อย่างไรกับค่าเฉลี่ยของขนาดตัวอักษรของโหนดข้อความทั้งหมดในส่วนย่อย
 - ถ้าพบว่าขนาดตัวอักษรของโหนดข้อความมีค่ามากกว่าค่าเฉลี่ยของขนาดตัวอักษรของโหนดข้อความทั้งหมดในส่วนย่อยและเป็นขนาดตัวอักษรของโหนดข้อความที่มีขนาดใหญ่ที่สุดในส่วนย่อย ลักษณะเด่นนี้เป็น 1 ถ้าไม่ใช่เป็น 0
 - ถ้าพบว่าขนาดตัวอักษรของโหนดข้อความมีค่ามากกว่าค่าเฉลี่ยของขนาดตัวอักษรของโหนดข้อความทั้งหมดในส่วนย่อยแต่ไม่ได้เป็นขนาดตัวอักษรของ

โหนดข้อความที่มีขนาดที่ใหญ่ที่สุดของส่วนย่อย ลักษณะเด่นนี้เป็น 1 ถ้าไม่ใช่เป็น 0

- ถ้าพบว่าขนาดตัวอักษรของโหนดข้อความมีค่าเท่ากับค่าเฉลี่ยของขนาดตัวอักษรโหนดทั้งหมดในส่วนย่อย ลักษณะเด่นนี้เป็น 1 ถ้าไม่ใช่เป็น 0
- ถ้าพบว่าขนาดตัวอักษรของโหนดข้อความมีค่าน้อยกว่าค่าเฉลี่ยของขนาดตัวอักษรของโหนดข้อความทั้งหมดในส่วนย่อยแต่ไม่ได้เป็นขนาดตัวอักษรของโหนดข้อความที่มีขนาดเล็กที่สุดของส่วนย่อยลักษณะเด่นนี้เป็น 1 ถ้าไม่ใช่เป็น 0
- ถ้าพบว่าขนาดตัวอักษรของโหนดข้อความมีค่าน้อยกว่าค่าเฉลี่ยของขนาดตัวอักษรของโหนดข้อความทั้งหมดในส่วนย่อยและเป็นตัวอักษรของโหนดข้อความที่มีขนาดเล็กที่สุดในส่วนย่อย ลักษณะเด่นนี้เป็น 1 ถ้าไม่ใช่เป็น 0

2. สีตัวอักษร (Color of characters): การหาลักษณะเด่นสีตัวอักษรของโหนดข้อความนั้นต้องทำการหาสีตัวอักษรหลักของโหนดข้อความทั้งหมดส่วนย่อยก่อน โดยใช้วิธีนับจำนวนตัวอักษรของแต่ละสีที่ปรากฏทั้งหมดในส่วนย่อย ถ้าพบว่าสีตัวอักษรที่ปรากฏในส่วนย่อยสีใดมีจำนวนตัวอักษรมากที่สุด ถือว่าสีนั้นเป็นสีตัวอักษรหลักของส่วนย่อย แล้วจึงนำสีตัวอักษรของโหนดข้อความที่ต้องการจะจำแนกว่าเป็นหัวเรื่องหรือไม่ มาเปรียบเทียบกับค่าสีตัวอักษรหลักของส่วนย่อย

- ถ้าสีตัวอักษรของโหนดข้อความมีสีต่างจากสีตัวอักษรหลักของส่วนย่อย ลักษณะเด่นนี้เป็น 1 ถ้าไม่ต่างเป็น 0

3. สีพื้นหลังตัวอักษร (Color of characters' background): การหาลักษณะเด่นสีพื้นหลังตัวอักษรของโหนดข้อความนั้นต้องทำการหาสีพื้นหลังหลักตัวอักษรของโหนดข้อความทั้งหมดของส่วนย่อยก่อน ผู้วิจัยใช้วิธีการเช่นเดียวกันกับการหาสีตัวอักษรหลักของส่วนย่อย โดยใช้วิธีนับจำนวนตัวอักษรของแต่ละสีพื้นหลังที่ปรากฏ

ทั้งหมดในส่วนย่อย ถ้าพบว่าสีพื้นหลังตัวอักษรที่ปรากฏในส่วนย่อยสีใดมีจำนวนตัวอักษรมากที่สุด ถือว่าสีพื้นหลังนั้นเป็นสีพื้นหลังตัวอักษรหลักของส่วนย่อย แล้วจึงนำสีพื้นหลังตัวอักษรของโหนดข้อความที่ต้องการจะจำแนกว่าเป็นหัวเรื่องหรือไม่ มาเปรียบเทียบกับค่าสีพื้นหลังตัวอักษรหลักของส่วนย่อย

- ถ้าสีพื้นหลังตัวอักษรของโหนดข้อความมีสีต่างจากสีพื้นหลังหลักตัวอักษรของส่วนย่อย ลักษณะเด่นนี้เป็น 1 ถ้าไม่ต่างเป็น 0

4. ไอคอนหน้าข้อความ (Icon in front of characters): ไอคอน (Icon) คือรูปภาพขนาดเล็กซึ่งถูกใช้บนจอเพื่อแทนการกระทำหรือเนื้อหาบางอย่าง (Powell, 2002) ไอคอนใช้ป้ายระบุรูปภาพ () ในการกำหนดรูปแบบของการแสดงผล การหาลักษณะเด่นของไอคอนหน้าข้อความได้จากการตรวจสอบว่าโหนดข้อความมีโหนดที่เป็นโหนดพี่น้องก่อนหน้า (Previous sibling node) เป็นโหนดสมาชิก และเป็นป้ายระบุรูปภาพหรือไม่

- ถ้าโหนดข้อความมีโหนดพี่น้องก่อนหน้าเป็นโหนดสมาชิกและเป็นป้ายระบุรูปภาพ ลักษณะเด่นนี้เป็น 1 ถ้าไม่ใช่เป็น 0

5. การไม่ย่อหน้าของข้อความ (Unindent of the word): การหาลักษณะเด่นของการไม่ย่อหน้าของโหนดข้อความได้จากการตรวจสอบว่าโหนดที่เป็นโหนดพ่อแม่ (Parent node) ของโหนดข้อความมีการใช้ป้ายระบุหรือสไตล์ชีทที่ใช้ย่อหน้า (Wilson) ได้แก่ <blockquote>, <dl>, , หรือ มีการใช้สไตล์ชีทได้แก่ text-indent, padding, margin

- ถ้าพบว่าโหนดพ่อแม่ของโหนดข้อความไม่มีการใช้ป้ายระบุหรือสไตล์ชีทดังกล่าวมา ลักษณะเด่นนี้เป็น 1 ถ้าใช้ป้ายระบุหรือสไตล์ชีทดังกล่าวจะเป็น 0

3.3.3.1.2 ลักษณะเด่นเกี่ยวกับตำแหน่ง

1. ลำดับของเฮชทีเอ็มแอล (Index of html): ข้อความที่เป็นหัวเรื่องมักจะปรากฏที่ตำแหน่งด้านบนของส่วนย่อยเสมอเพื่อแสดงว่าส่วนย่อยที่ตามมานั้นมีเนื้อหาเกี่ยวข้องกับอะไร โดยงานวิจัยนี้จะใช้ลำดับของเฮชทีเอ็มแอลที่ได้จากการใช้คอมพิวเตอร์เดินทางไปตามโหนดของเฮชทีเอ็มแอลภายในส่วนย่อย

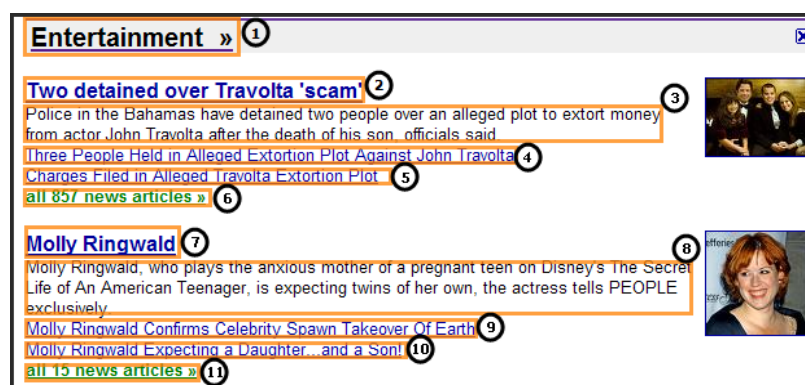
- ถ้าลำดับของโหนดข้อความที่ต้องการจะจำแนกกว่าเป็นหัวเรื่องหรือไม่ มีลำดับภายในส่วนย่อยมีค่าไม่เกิน 10 ลักษณะเด่นนี้จะมีค่า 1 ถ้าเกินกว่าค่าที่กำหนดจะมีค่าเป็น 0

3.3.3.2 การสกัดลักษณะเด่น

ขั้นตอนการดึงลักษณะเด่นจะนำแต่ละส่วนย่อยมาทำการดึงลักษณะเด่นของแต่ละโหนดข้อความภายในส่วนย่อย โดยจะนำโหนดข้อความที่ต้องการดึงลักษณะเด่นมาเทียบโหนดข้อความทั้งหมดของส่วนย่อยนั้น ตัวอย่างแสดงดังภาพที่ 3.3 ถ้าต้องการดึงลักษณะเด่นของโหนดข้อความ 7 ก็จะเปรียบเทียบกับโหนดข้อความที่มีทั้งหมดของส่วนย่อย (1-11) ว่ามีความสัมพันธ์ลักษณะใด เช่น มีขนาดตัวอักษร, สีตัวอักษรต่างกันหรือไม่

ภาพที่ 3.3

แสดงตัวอย่างการพิจารณาว่าโหนดข้อความ (7) เป็นหัวเรื่องหรือไม่ โดยเปรียบเทียบรูปแบบการแสดงผลกับโหนดข้อความทั้งหมด (1-11) ภายในส่วนย่อย



วิทยานิพนธ์นี้ได้เสนออัลกอริทึมในการดึงลักษณะเด่นของแต่ละโหนดข้อความของส่วนย่อยนั้นๆ แสดงรายละเอียดของอัลกอริทึมดังกล่าวที่ 3.4

ภาพที่ 3.4

แสดงอัลกอริทึมของการสกัดลักษณะเด่นของแต่ละโหนดข้อความของส่วนย่อยนั้น

```

Algorithm: Section based Feature Extraction.
Input: the set of section S.
Output: the set of feature of each text node F.

BEGIN
1.   For each  $s_i \in S$  Do
2.       Compute the average font size  $m_{avgFont}$  of  $s_i$ 
3.       Find the major font color  $m_{color}$  of  $s_i$ 
4.       Find the major background color  $m_{bgColor}$  of
         font of  $s_i$ 
5.       For each text node  $t_i \in s_i$  Do
6.           Compare font size of  $t_i$  and  $m_{avgFont}$ 
              and store in F
7.           Compare font color of  $t_i$  and  $m_{color}$ 
              and store in F
8.           Compare font background color of
               $t_i$  and  $m_{bgColor}$  and store in F
9.           Check icon in front of  $t_i$  and store in F
10.          Check unindent of  $t_i$  and store in F
11.          Compute index of  $t_i$  in  $s_i$  and store in F
12.       End.
13.   End.
END.

```

3.3.4 ขั้นตอนการจำแนกหัวเรื่องด้วยซัพพอร์ทเวกเตอร์แมชชีน

ขั้นตอนนี้จะนำลักษณะเด่นที่สกัดมาทำการสอน (Training) และทดสอบ (Testing) หาผลลัพธ์การจำแนกหัวเรื่องด้วยซัพพอร์ทเวกเตอร์แมชชีนซึ่งเป็นอัลกอริทึมการเรียนรู้ที่สามารถนำมาใช้ในการจำแนกรูปแบบหรือกลุ่มของข้อมูล (Vapnik, 1998) โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ ซึ่งมีประสิทธิภาพในด้านความเร็ว, ความสามารถในการแยกแยะกลุ่มข้อมูล และสามารถจำแนกประเภทข้อมูลแบบหลายมิติได้ดี (Kotsiantis, 2006) ทำให้ซัพพอร์ทเวกเตอร์แมชชีนนำมาจำแนกโหนดข้อความที่เป็นหัวเรื่องจากลักษณะการแสดงผลได้ดีเพราะหัวเรื่องจะประกอบด้วย

รูปแบบการแสดงผลหลายประเภท เช่น ตัวอักษรที่มีขนาดใหญ่และใช้พื้นหลังตัวอักษรที่แตกต่างบริเวณรอบข้าง

ในการเตรียมข้อมูลสำหรับการเรียนรู้และทดสอบด้วยซัพพอร์ตเวกเตอร์แมชชีน จะใช้วิธีการทดสอบแบบจับคู่ไขว้ข้าม 10 กลุ่ม (10 Fold Cross validation) ซึ่งขั้นตอนการทำงานจะแบ่งชุดลักษณะเด่นซึ่งได้จากขั้นตอนการสกัดลักษณะเด่นจากเว็บเพจออกเป็น 10 ชุดข้อมูล ชุดข้อมูลละ 10 เว็บเพจ หลังจากนั้นใช้ 9 ชุดข้อมูลในการเรียนรู้ และที่เหลืออีก 1 ชุดข้อมูลใช้ในการทดสอบ และในการทดสอบจะทำซ้ำ 10 ครั้งแล้วหาค่าเฉลี่ยจึงจะได้ผลลัพธ์ของการจำแนกหัวเรื่อง

3.4 วิธีการวัดผลการทดลอง

งานวิจัยนี้มีวัตถุประสงค์ที่จะศึกษาเปรียบเทียบวิธีจำแนกหัวเรื่องของวิธีการที่นำเสนอในวิทยานิพนธ์นี้ (HClassifier) กับวิธีการ Brudvik (HeadingHunter: Brudvik, 2008) ในการเปรียบเทียบค่าทั้งสองวิธีจะใช้ค่าความแม่นยำ (Precision หรือ P) โดยเป็นค่าการประเมินความถูกต้องของระบบในการเลือกหมวดข้อความที่เป็นหัวเรื่องได้ถูกต้อง, ค่าความระลึก (Recall หรือ R) โดยเป็นค่าการประเมินประสิทธิภาพของระบบในการเลือกหมวดข้อความที่เป็นหัวเรื่องจากหมวดข้อความที่เป็นหัวเรื่องทั้งหมด, ค่า F-measure ในการวัดประสิทธิภาพโดยรวมของระบบ จากนั้นจะใช้วิธีวัดประสิทธิภาพการจำแนกหัวเรื่องว่าสูงขึ้นหรือลดลงที่ระดับนัยสำคัญทางสถิติด้วยวิธีทดสอบที่แบบจับคู่ (paired t-test)

- การหาค่าความแม่นยำ (Precision) สามารถหาได้ดังสมการที่ 3.1

$$P = \frac{|P_{man} \cap P_{sys}|}{|P_{sys}|} \quad (3.1)$$

- การหาค่าความระลึก (Recall) สามารถหาได้ดังสมการที่ 3.2

$$R = \frac{|P_{man} \cap P_{sys}|}{|P_{man}|} \quad (3.2)$$

- การหาค่า F-measure สามารถหาได้ดังสมการที่ 3.3

$$F = \frac{2PR}{P+R} \quad (3.3)$$

โดยกำหนดให้ P_{man} เป็นโหนดข้อความที่มนุษย์เลือกหัวเรื่อง และ P_{sys} เป็นโหนดข้อความที่ระบบเลือกเป็นหัวเรื่อง

ในการเปรียบเทียบวิธีจำแนกหัวเรื่องของวิธีการที่นำเสนอในวิทยานิพนธ์นี้ (HClassifier) กับวิธีการ Brudvik นั้น ผู้วิจัยจะทำการศึกษาผลการทดลองดังต่อไปนี้

1. ผลการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องจากลักษณะเด่นซึ่งนำเสนอในวิทยานิพนธ์นี้: ทำการศึกษาผลการเปรียบเทียบลักษณะเด่นซึ่งนำเสนอในวิทยานิพนธ์นี้ (HClassifier) กับลักษณะเด่นจากวิธีการของ Brudvik (HeadingHunter) ในการจำแนกหัวเรื่องของเว็บเพจ
 - 1) วัตถุประสงค์ของการทดลอง เพื่อศึกษาว่าลักษณะเด่นซึ่งนำเสนอในวิทยานิพนธ์นี้สามารถปรับปรุงผลการจำแนกหัวเรื่องของเว็บเพจให้มีประสิทธิภาพสูงขึ้น
 - 2) วิธีการทดลอง นำลักษณะเด่นซึ่งนำเสนอในวิทยานิพนธ์นี้และลักษณะเด่นจากวิธีการของ Brudvik (HeadingHunter) มาทำการศึกษาผลการจำแนกหัวเรื่องโดยใช้ต้นไม้ตัดสินใจเป็นวิธีการเรียนรู้ด้วยเครื่องจักร
2. ผลการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องโดยใช้ซอฟต์แวร์เคเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักร: ทำการศึกษาผลการเปรียบเทียบการใช้การเรียนรู้ด้วยเครื่องจักรทั้งซอฟต์แวร์เคเตอร์แมชชีนและต้นไม้ตัดสินใจด้วยลักษณะเด่นซึ่งนำเสนอในวิทยานิพนธ์ (HClassifier) นี้กับลักษณะเด่นจากวิธีการของ Brudvik (HeadingHunter)
 - 1) วัตถุประสงค์ของการทดลอง เพื่อศึกษาว่าการจำแนกหัวเรื่องของเว็บเพจโดยใช้ซอฟต์แวร์เคเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักรซึ่งนำเสนอในวิทยานิพนธ์นี้สามารถปรับปรุงผลการจำแนกหัวเรื่องของเว็บเพจให้มีประสิทธิภาพสูงขึ้น
 - 2) วิธีการทดลอง นำลักษณะเด่นซึ่งนำเสนอในวิทยานิพนธ์นี้และลักษณะเด่นจากวิธีการของ Brudvik (HeadingHunter) มาทำการศึกษาผลการจำแนกหัวเรื่องของเว็บเพจโดยใช้ซอฟต์แวร์เคเตอร์แมชชีนและต้นไม้ตัดสินใจเป็นวิธีการเรียนรู้ด้วยเครื่องจักร

3. ผลการปรับปรุงประสิทธิภาพการจำแนกหัวเรื่องด้วยวิธีการแบ่งเว็บเพจเป็นส่วนย่อยและชัฟฟอร์ตเวกเตอร์แมชชีน: ศึกษาผลการเปรียบเทียบผลการจำแนกหัวเรื่องของเว็บเพจด้วยวิธีนำเสนอในวิทยานิพนธ์นี้ (HClassifier) กับ วิธีของ Brudvik (HeadingHunter)
 - 1) วัตถุประสงค์ของการทดลอง เพื่อศึกษาว่าวิธีการจำแนกหัวเรื่องของเว็บเพจซึ่งนำเสนอในวิทยานิพนธ์นี้สามารถปรับปรุงผลการจำแนกหัวเรื่องของเว็บเพจให้มีประสิทธิภาพสูงขึ้น
 - 2) วิธีการทดลอง นำลักษณะเด่นซึ่งนำเสนอในวิทยานิพนธ์นี้มาศึกษาผลการจำแนกหัวเรื่องของเว็บเพจโดยใช้ชัฟฟอร์ตเวกเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักร และนำลักษณะเด่นจากวิธีการของ Brudvik (HeadingHunter) มาศึกษาผลการจำแนกหัวเรื่องโดยใช้ต้นไม้ตัดสินใจเป็นวิธีการเรียนรู้ด้วยเครื่องจักร
4. ผลการศึกษาลักษณะเด่นในการส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจ:

ทำการศึกษาผลการเปรียบเทียบว่าลักษณะเด่นซึ่งนำเสนอในวิทยานิพนธ์นี้ (HClassifier) ลักษณะเด่นใดส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจด้วยความสำคัญจากมากไปน้อย

 - 1) วัตถุประสงค์ของการทดลอง เพื่อศึกษาว่าลักษณะเด่น 10 ลักษณะเด่นซึ่งนำเสนอในวิทยานิพนธ์นี้ ลักษณะเด่นใดส่งผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจด้วยความสำคัญจากมากไปน้อย
 - 2) วิธีการทดลอง นำลักษณะเด่นจำนวน 10 ลักษณะเด่น ซึ่งนำเสนอในงานวิจัยนี้มาทำการศึกษาผลการจำแนกหัวเรื่องของเว็บเพจโดยใช้ชัฟฟอร์ตเวกเตอร์แมชชีนเป็นวิธีการเรียนรู้ด้วยเครื่องจักร โดยนำลักษณะเด่นที่ต้องการศึกษาผลกระทบต่อประสิทธิภาพการจำแนกหัวเรื่องของเว็บเพจออกไป และทำการเรียนรู้ด้วยเครื่องจักรโดย 9 ลักษณะเด่นที่เหลือ และเปรียบเทียบว่าประสิทธิภาพการจำแนกหัวเรื่องว่าลดลงเป็นเท่าใดเมื่อเทียบกับการใช้ลักษณะเด่นจำนวน 10 ลักษณะเด่น