บทที่ 3

วิธีการดำเนินงานวิจัย

ในบทนี้จะกล่าวถึงขอบเขตของการดำเนินงานวิจัย เครื่องมือและซอฟต์แวร์ที่ใช้ในการพัฒนาระบบ แนวคิดและระบบโดยรวม รวมทั้งขั้นตอนการทดลองและวิธีการวัดผลการทดลอง

3.1 ขอบเขตของการดำเนินงานวิจัย

3.1.1 ข้อมูลที่ใช้ในการทำวิจัย

ข้อมูลสำหรับการทดลองในงานวิจัยนี้ ใช้บทความข่าวภาษาไทยจากการวัดเปรียบเทียบสมรรถนะเพื่อพัฒนามาตรฐานการประมวลผลภาษาไทยปี 2009 (Benchmark for Enhancing the Standard of Thai language processing :BEST2009) จำนวน 206,695 คำในการฝึก (training) ระบบเพื่อสร้างโมเดลไฟล์การรู้จำนิพจน์ระบุนาม และใช้บทความข่าวภาษาไทยจากการวัดเปรียบเทียบสมรรถนะเพื่อพัฒนามาตรฐานการประมวลผลภาษาไทยที่เป็นข่าวการเมืองมาทดสอบ (test)

ภาพที่ 3.1

ตัวอย่าง corpus สำหรับฝึกและทดสอบระบบ

http://www.bangkokhealth.com/healthnews _ htdoc/healthnews _
detail.asp?Number=10506|
สงสัย|ติด|หวัด|นก| |อีก|คน|ยัง|น่า|ห่วง|
ตาม|ที่|<NE> นางประนอม ทองจันทร์</NE>| |กับ| |<NE> ด.ช.กิตติพงษ์ แหลมผักแว่น</NE>| |
และ| |<NE> ด.ญ.กาญจนา กรองแก้ว</NE>| |ป่วย|สงสัย|ติด|เชื้อ|ไข้|ขณะ|นี้|ยัง|ไม่|ดี|ขึ้น|
หลัง|เข้า|เยี่ยม|ดู|อาการ|ผู้|ป่วย|แล้ว| |<NE> น.พ.จรัล</NE>|ประชุม|ร่วม|กับ|เจ้าหน้าที่|ทุก|ฝ่าย| |
เพื่อ|สรุป|ผล|การ|ดำเนิน|การ| |รวม|ทั้ง|สอบสวน|โรค|ก่อน|ที่|ผู้|ป่วย|จะ|ถูก|ส่ง|มา|รักษา|ตัว| |จาก|
นั้น|ร่วม|กัน|แถลง|ข่าว| |โดย| |<NE> น.พ.จรัล</NE>|กล่าว|ว่า| |ขณะ|นี้|ผู้|ป่วย|ทั้ง| |3| |ราย| |
อาการ|ยัง|ทรง| |โดย|ใน|ราย|ของ| |<NE> ด.ช.กิตติพงษ์</NE>| |กับ| |<NE> ด.ญ.กาญจนา</NE>|
|ปอด|หาย|เป็น|ปกติ|แล้ว| |คาด|ว่า|จะ|กลับ|บ้าน|ได้|ใน|ไม่|ช้า|นี้| |แต่|ใน|ราย|ของ|<NE> นาง
ประนอม</NE> |อาการ|ยัง|น่า|เป็นห่วง| |ซึ่ง|ทั้ง| |3| |ราย| |ใน|ชั้น|นี้|ถือ|ว่า|เป็น|ผู้|ป่วย|อยู่|ใน|ขั้น|
น่า|สงสัย|อาจ|ติด|เชื้อ|ไข้|หวัด|นก| |เพราะ|ตรวจ|พบ|ผู้|ป่วย|มี|อาการ|ปอด|บวม|ปอด|อักเสบ| |
เนื่อง|จาก|ติด|เชื้อ|ไวรัส| |แต่|ยัง|สรุป|ไม่|ได้|ว่า|ติด|เชื้อ|ไข้|หวัด|นก|แน่ชัด| หรือ|ไม่| |ต้อง|รอ|ผล|

3.1.2 เครื่องมือที่ใช้ในงานวิจัย

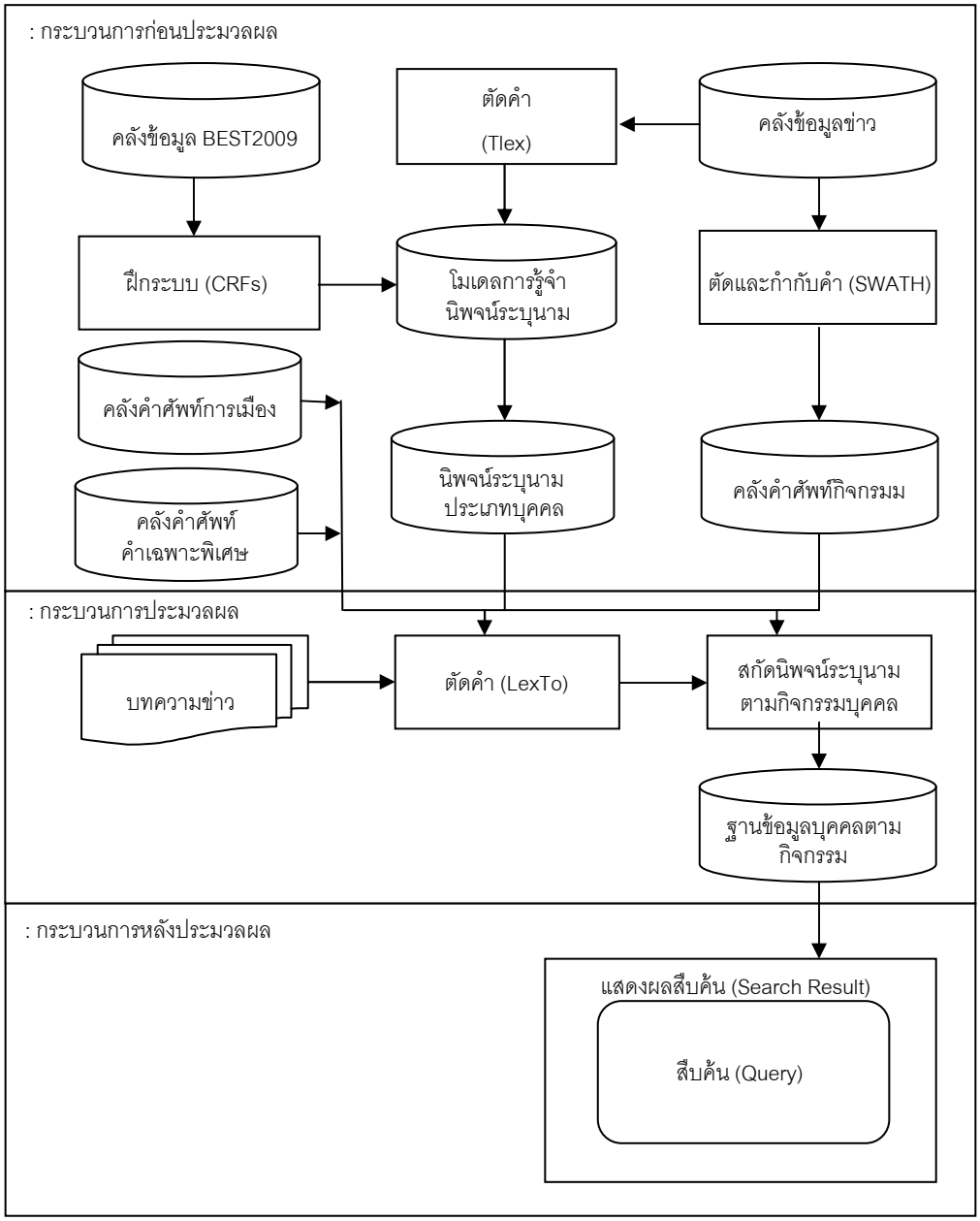
ในส่วนนี้จะกล่าวถึงเครื่องมือและซอฟต์แวร์ ต่างๆ ที่นำมาใช้ในการพัฒนางานวิจัย

ตารางที่ 3.1
เครื่องมือที่ใช้ในงานวิจัย

ฮาร์ดแวร์ (Hardware)	• Intel(R) Core(TM)2 Duo CPU P8700 (2.53 Hz) • Main Memory 4.00 GB • Hard Disk 500 GB 5400 RPM • Graphic Chip ATI Radeon HD4570
ซอฟต์แวร์ (Software)	• ระบบปฏิบัติการ Windows 7 Ultimate • คลังข้อมูลข่าวภาษาไทย (BEST 2009) • Microsoft SQL Server 2005 สำหรับใช้สร้างคลังข้อมูล • Delphi สำหรับใช้สร้างระบบ • Tlex สำหรับใช้ตัดคำ • Swath สำหรับใช้ตัดและกำกับชนิดของคำ • LexTo สำหรับตัดคำ • CRF++-0.53 สำหรับฝึกและทดสอบไฟล์ในการสกัดนิพจน์ระบุนามประเภทบุคคล • Java runtime (jre-6u16-windows-i586)

3.2 แนวคิดและระบบโดยรวม

ภาพที่ 3.2
แนวคิดและระบบโดยรวม



แนวคิดและระบบโดยรวมจะประกอบไปด้วย 3 กระบวนการ ดังนี้

1. กระบวนการก่อนประมวลผล (Preprocessing) เป็นกระบวนการจัดเตรียมข้อมูลแบ่งการทำงานย่อยออกเป็น 3 ส่วน ดังนี้

ส่วนที่ 1 นำบทความข่าวภาษาไทยจากการวัดเปรียบเทียบสมรรถนะเพื่อพัฒนามาตรฐานการประมวลผลภาษาไทยปี 2009 (Benchmark for Enhancing the Standard of Thai language processing:BEST2009) จำนวน 206,695 คำ มาฝึกระบบ(training) ด้วยเทคนิค Conditional Random Fields สำหรับเทคนิคนี้จะใช้คุณสมบัติของคำบ่งชี้ เพื่อใช้ระบุนิพจน์ระบุนามประเภทบุคคล ผลที่ได้จากการฝึกระบบจะเป็นโมเดลไฟล์ที่ใช้สำหรับการทดสอบการรู้จำนิพจน์ระบุนามประเภทบุคคล

ตัวอย่างบทความ : นายอภิสิทธิ์ เวชชาชีวะ นายกรัฐมนตรีเปิดเผยถึงการชุมนุมของกลุ่มคนเสื้อแดงในวันพรุ่งนี้

ตารางที่ 3.2

ตัวอย่างไฟล์สำหรับฝึกระบบ (Training)

คำ	คุณสมบัติ	เงื่อนไขการระบุคำ
นาย	P	B-NE
อภิสิทธิ์	O	I-NE
	SP	I-NE
เวชชาชีวะ	O	I-NE
นายกรัฐมนตรี	O	O
เปิดเผย	O	O
ถึง	O	O
การ	O	O
ชุมนุม	O	O
ของ	O	O
กลุ่มคนเสื้อแดง	O	O
ใน	O	O

ตารางที่ 3.2 (ต่อ)

คำ	คุณสมบัติ	เงื่อนไขการระบุคำ
วัน	O	O
พรุ่ง	O	O
นี้	O	O

จากตารางที่ 3.2 เป็นตัวอย่างไฟล์สำหรับฝึกระบบด้วยเทคนิค Conditional Random Fields (CRF) ด้วยคุณสมบัติของคำบ่งชี้ ประกอบไปด้วย 3 คอลัมน์

คอลัมน์ที่ 1 คือ token หรือ คำที่ได้จากประโยค

คอลัมน์ที่ 2 คือ คุณสมบัติที่จะนำมาใช้สำหรับเทคนิคการเรียนรู้เพื่อระบุนิพจน์ระบุนามประเภทบุคคล

กำหนดให้ P แทนคุณสมบัติของคำบ่งชี้ เช่น นาย , นางสาว , พ.ต.ท. เป็นต้น

SP แทนคำที่เป็นช่องว่าง

คอลัมน์ที่ 3 คือ คอลัมน์ที่ใช้บอกจุดเริ่มต้นและจุดสิ้นสุดของนิพจน์ระบุนามประเภทบุคคล

กำหนดให้ B-NE จะบ่งบอกจุดเริ่มต้นของนิพจน์ระบุนาม

I-NE จะบ่งบอกจุดต่อเนื่องของนิพจน์ระบุนาม

O เป็นคำปกติทั่วไป

ตารางที่ 3.3

ตัวอย่างคำบ่งชี้สำหรับการระบุนิพจน์ระบุนามประเภทบุคคล

คำบ่งชี้
นาย, นาง,นางสาว, น.ส., ม.ร.ว., นพ., นายแพทย์,คุณหญิง, พญ., แพทย์หญิง, ดร., ผู้ช่วยศาสตราจารย์ , รองศาสตราจารย์, สามเณร, พระอธิการ, พระปลัด, พระมหา, บาทหลวง, หม่อมราชวงศ์, เจ้าอธิการ, ,พระครูปลัด, พลเอก, พลโท, พลตรี, พันเอก, พันโท, พันตรี, ร้อยเอก, ร้อยโท, ร้อยตรี, จ่าสิบเอก, จ่าสิบโท, จ่าสิบตรี, สิบเอก, สิบโท, สิบตรี, พลทหาร, พลเรือเอก, พลเรือโท, พลเรือตรี , พลเรือจัตวา, นาวาเอกพิเศษ, นาวาเอก, นาวาโท, นาวาตรี

ส่วนที่ 2 นำบทความข่าวการเมืองจำนวน 1,209 บทความมาทำการตัดและกำกับชนิดของคำ ด้วยเทคนิคสวอธ (Swath) เลือกวิธีการตัดแบบเลือกคำที่ยาวที่สุด (longest matching) ผลที่ได้จากการตัดและกำกับชนิดของคำ จะทำการคัดกรองคำที่กำกับชนิดเป็นคำกริยา(Verb) จากนั้นทำการคัดเลือกคำกริยาที่เกี่ยวข้องกับการเมืองด้วยมือ นำมาสร้างเป็นคลังคำศัพท์กิจกรรม

ตารางที่ 3.4

ตัวอย่างการตัดและกำกับชนิดคำด้วยเทคนิค Swath

นาย@NTTL|อภิสิทธิ์@NCMN| |เวช@NPRP|ชา@NCMN|ชี@NCMN|วะ@NPRP| | นายกรัฐมนตรี@NCMN| |เปิดเผย@VSTA|ถึง@RPRE|การ@FIXN|ชุมนุม@VSTA|ของ@RPRE| กลุ่ม@NCMN|คน@CNIT|เสื้อ@NCMN|แดง@VSTA| ใน@RPRE|วัน@CMTR|พรุ่งนี้@ADVN|

ตารางที่ 3.5

การกำกับชนิดของคำกริยาด้วยโปรแกรม Swath

ชนิดของคำ	คำอธิบาย	ตัวอย่าง
VACT	คำกริยาแสดงถึงการกระทำของตน	ทำงาน, ร้องเพลง, กิน
VSTA	คำกริยาที่อ้างถึงการบอกกล่าว	เห็น, รู้, คือ
VATT	คำกริยาวิเศษณ์	อ้วน, ดี, สาย

ตารางที่ 3.6

ตัวอย่างคำกริยาทางการเมือง

คำกริยาทางการเมือง
ปราศรัย, หาเสียง, กล่าว, เปิดเผย, ชุมนุม, เดินทาง, อภิปราย, ประชุม, ปาฐกถา, แต่งตั้ง, ชี้แจง, ลงมติ, สรรหา, จัดตั้ง, ถอดถอน, ฟ้องร้อง, ตั้งกระทู้, ลงพื้นที่, บันทึกเทป, ฟ้องร้อง, โจมตี, พาดพิง, ประมูล, วิสามัญ, พิสูจน์, คัดเลือก, วินิจฉัย, ลงนาม, สัมมนา, ดำเนินงาน, ถ่ายทอด

ส่วนที่ 3 นำบทความข่าวการเมืองชุดเดียวกันกับส่วนที่สองมาทำการตัดคำด้วยโปรแกรมทีเล็กส์ (Thai Lexeme Analyser: Tlex) โปรแกรมนี้จะใช้การเรียนรู้ของเครื่องด้วยเทคนิค Conditional Random Fields (CRFs) มาตัดคำ หลังจากนั้นนำไฟล์ที่ผ่านกระบวนการตัดคำมาสร้างเป็นไฟล์ชุดทดสอบ และทำการทดสอบกับโมเดลไฟล์การสกัดนิพจน์ระบุนามที่ได้จากส่วนแรก สกัดเอานิพจน์ระบุนามประเภทบุคคลมาสร้างเป็นคลังข้อมูลนิพจน์ระบุนามประเภทบุคคลสำหรับคลังคำศัพท์การเมืองและคลังคำศัพท์คำเฉพาะพิเศษจะใช้วิธีการคัดคำการเมืองและคำเฉพาะพิเศษด้วยมือมนุษย์

ตารางที่ 3.7

ตัวอย่างการตัดคำด้วยเทคนิคทีเล็กส์ (Tlex)

|นายอภิสิทธิ์| |เวชชาชีวะ| |นายก|รัฐมนตรี| |เปิดเผย|ถึง|การ|ชุมนุม|ของ|กลุ่ม|คน|เสื้อ|แดง|ใน|วัน|พรุ่ง|นี้|

ตารางที่ 3.8

ตัวอย่างผลการทดสอบไฟล์โมเดลนิพจน์ระบุนาม

คำ	คุณสมบัติ	เงื่อนไขการระบุคำ	ผลลัพธ์
นาย	P	B-NE	B-NE
อภิสิทธิ์	O	I-NE	I-NE
	SP	I-NE	I-NE
เวชชาชีวะ	O	I-NE	I-NE
นายกรัฐมนตรี	O	O	O
เปิดเผย	O	O	O
ถึง	O	O	O
การ	O	O	O
ชุมนุม	O	O	O
ของ	O	O	O
กลุ่มคนเสื้อแดง	O	O	O
ใน	O	O	O
วัน	O	O	O
พรุ่ง	O	O	O
นี้	O	O	O

จากตารางที่ 3.8 เป็นตัวอย่างผลการทดสอบไฟล์โมเดลนิพจน์ระบุนามด้วยเทคนิค Conditional Random Field (CRF) ประกอบไปด้วย 4 คอลัมน์ ในคอลัมน์ที่ 1-3 จะเหมือนกับไฟล์ฝึกระบบสำหรับคอลัมน์ที่ 4 เป็นผลลัพธ์ที่ได้จากการฝึกระบบโดยเปรียบเทียบความถูกต้องกับคอลัมน์ที่ 3

ตารางที่ 3.9

ตัวอย่างคำโดเมนทางการเมือง

คำโดเมนทางการเมือง
กลุ่มคนเสื้อแดง, กลุ่ม นปช. , กลุ่มคนเสื้อเหลือง, หน่วยเลือกตั้ง, สมาชิกสภาผู้แทนราษฎร, ประธานสภาผู้แทนราษฎร, ถวายฎีกา, คณะรัฐมนตรี, พรรคประชาธิปัตย์, พรรคเพื่อไทย, พรรคภูมิใจไทย, พรรคชาติไทยพัฒนา, คณะกรรมการสมาณฉันท์, รายการเชื่อมั่นประเทศไทยกับนายก ฯ อภิสิทธิ์, รัฐมนตรีประจำสำนักนายกรัฐมนตรี, รองนายกรัฐมนตรีฝ่ายความมั่นคง, สมาชิกวุฒิสภา, รัฐมนตรีว่าการกระทรวงศึกษาธิการ, รัฐมนตรีว่าการกระทรวงวิทยาศาสตร์และเทคโนโลยี, รัฐมนตรีว่าการกระทรวงกลาโหม, รัฐมนตรีว่าการกระทรวงการคลัง, ศาลรัฐธรรมนูญ, ส.ว.สรรหาม , องค์กรปกครองส่วนท้องถิ่น, ประธานกรรมาธิการการแรงงานสภาผู้แทนราษฎร, กรรมการป้องกันและปราบปรามการทุจริตแห่งชาติ, เลขาธิการ กกต., พ.ร.บ., ผู้อำนวยการเลือกตั้งจังหวัด

ตารางที่ 3.10

ตัวอย่างคำเฉพาะพิเศษ

คำเฉพาะพิเศษ
เปิดเผยอีกว่า, มีข่าวว่า, กล่าวด้วยว่า, หลายคนอาทิ, ขณะเดียวกัน, โดยจะรอ

2. กระบวนการประมวลผล (Processing) เป็นกระบวนการสกัดนิพจน์ระบุนามตามกิจกรรมของบุคคล เริ่มต้นจากนำบทความข่าวการเมืองจำนวน 30 บทความ มาทำการตัดคำด้วยเทคนิค เล็กซ์โต (Thai Lexeme Tokenizer : LexTo) เทคนิคนี้จะรวมวิธีการตัดคำแบบ Conditional Random Fields กับพจนานุกรมจากคลังคำศัพท์กิจกรรม, คลังข้อมูลนิพจน์ระบุนามประเภทบุคคล, คลังคำศัพท์การเมืองและคลังคำศัพท์คำเฉพาะพิเศษ เป็นตัวตรวจสอบข้อมูลเพื่อใช้ลดคำที่กำกวมในการตัดคำ ผลลัพธ์จากการตัดคำด้วยโปรแกรมเล็กซ์โต ทำการสกัดนิพจน์ระบุนามตามกิจกรรมของบุคคลด้วยกฎ สร้างเป็นฐานข้อมูลบุคคลตามกิจกรรม

ตารางที่ 3.11

ตัวอย่างการตัดคำด้วยเทคนิคเล็กซ์โต (LexTo)

นาย|อภิสิทธิ์ เวชชาชีวะ| |นายกรัฐมนตรี| |เปิดเผย|ถึง|การ|ชุมนุม|ของ|กลุ่มคนเสื้อแดง|ในวันพรุ่งนี้|

ตารางที่ 3.12

กฎการสกัดนิพจน์ระบุนาม

กฎการสกัดนิพจน์ระบุนาม	
1.	ลำดับการสกัดให้เริ่มจากนิพจน์ระบุนามประเภทบุคคล , กิจกรรม , นิพจน์ระบุนามประเภทบุคคล (อาจมีหรือไม่มีก็ได้) , คำโดเมนการเมือง (อาจมีหรือไม่มีก็ได้)
2.	เป็นนิพจน์ระบุนามประเภทบุคคลที่อยู่ในคลังข้อมูล
3.	เป็นคำกริยาที่อยู่ในคลังคำศัพท์กิจกรรม
4.	ขอบเขตระหว่างนิพจน์ระบุนามประเภทบุคคลกับคำกริยาไม่เกิน 10 คำ
5.	นิพจน์ระบุนามประเภทบุคคลมีได้มากกว่าหนึ่ง
6.	สกัดคำกริยา 2 คำ
7.	หลังคำกิริยาสามารถมีนิพจน์ระบุนามประเภทบุคคล หรือคำโดเมนการเมือง
8.	(กรณีหลังคำกริยา)ขอบเขตระหว่างคำกริยากับนิพจน์ระบุนามประเภทบุคคล ไม่เกิน 10 คำ
9.	(กรณีหลังคำกริยา)ขอบเขตระหว่างคำกริยากับคำโดเมนการเมืองไม่เกิน 10 คำ
10.	กรณีเจอคำเฉพาะพิเศษให้ข้าม 20 คำ แล้วเริ่มลำดับการสกัดใหม่
11.	(กรณีหลังคำกริยา) สกัดคำโดเมนการเมืองถือเป็นจุดสิ้นสุด ให้เริ่มลำดับการสกัดใหม่กรณีที่ข้อมูลยังคงเหลือ

3. กระบวนการหลังประมวลผล (Post-Processing) นำเสนอความสัมพันธ์ที่ได้จากการสกัดนิพจน์ระบุนามตามกิจกรรมของบุคคล ในรูปแบบของการแสดงผลจากการสืบค้น (Search Result)

ภาพที่ 3.3

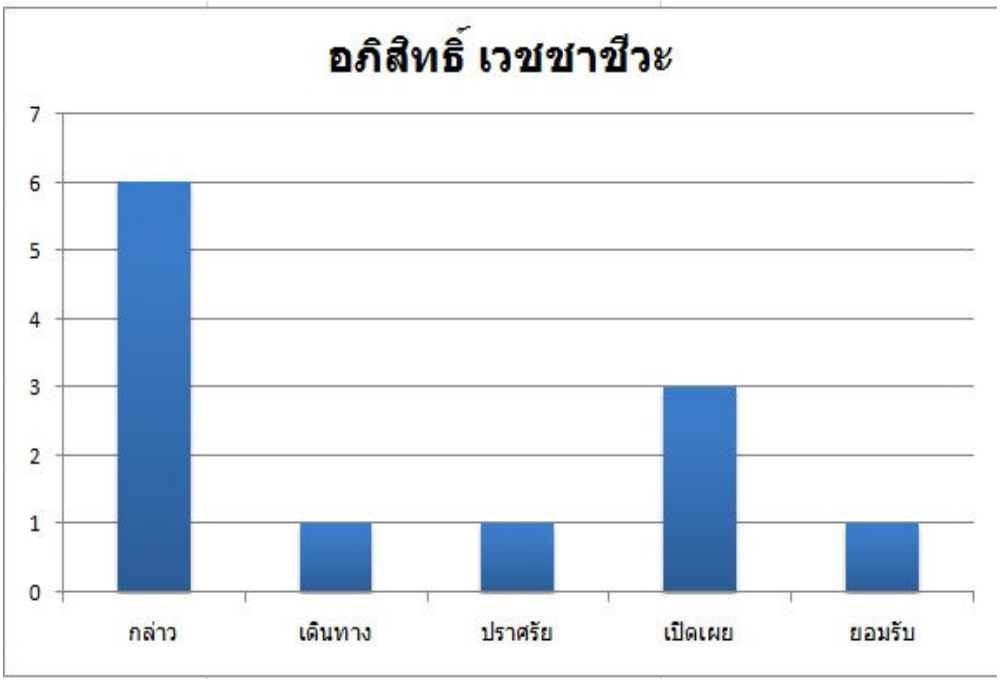
ตัวอย่างความสัมพันธ์จากการสืบค้น

ตารางแสดงความสัมพันธ์ของบุคคลตามกิจกรรม

ข่าวประจำวันที่	บุคคล	กิจกรรม	บุคคลร่วมกิจกรรม	โดเมนการเมือง
01/08/2009	อภิสิทธิ์ เวชชาชีวะ	กล่าวปาฐกถา		วิกฤติเศรษฐกิจ
22/08/2009	อภิสิทธิ์ เวชชาชีวะ	เดินทางลงพื้นที่		เลือกตั้งซ่อม
22/08/2009	อภิสิทธิ์ เวชชาชีวะ	ปราศรัยหาเสียง	ธานี เกือกสุบรรณ	เลือกตั้งซ่อม
23/08/2009	อภิสิทธิ์ เวชชาชีวะ	กล่าวร้องเรียน		ทุจริตโครงการชุมชนพอเพียง
23/08/2009	อภิสิทธิ์ เวชชาชีวะ	กล่าวแก้ไขปัญหา		ประเทศ
23/08/2009	อภิสิทธิ์ เวชชาชีวะ	กล่าวบริหาร		ประเทศ
23/08/2009	ภูชงค์ นุตราวงศ์	กล่าวลงพื้นที่		หน่วยเลือกตั้ง
26/08/2009	สุริยะใส กตะศิลา	แถลงคัดค้าน		พระราชบัญญัตินิรโทษกรรม
28/08/2009	สุเทพ เกือกสุบรรณ	แถลงประชุม		กองอำนวยการรักษาความมั่นคงภายใน
28/08/2009	สรรเสริญ แก้วกำเนิด	เปิดเผยประชุม		กอ.รมน.

ภาพที่ 3.4

กราฟกิจกรรมของ "อภิสิทธิ์ เวชชาชีวะ"



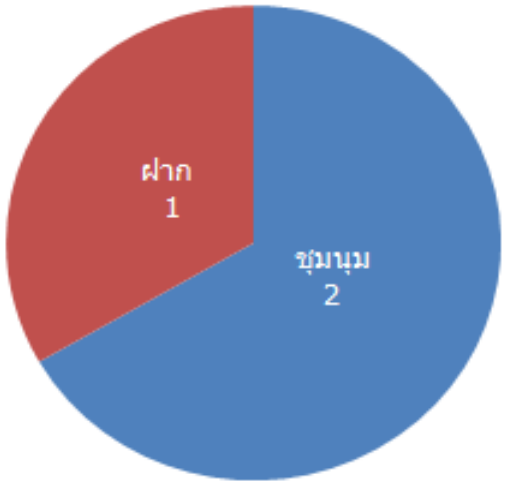
ในภาพที่ 3.4 กำหนดให้แกน X แทนกิจกรรมที่เกิดขึ้นของบุคคลชื่อ "อภิสิทธิ์ เวชชาชีวะ" และแกน Y แทนจำนวนครั้งของกิจกรรมที่ "อภิสิทธิ์ เวชชาชีวะ" ได้กระทำ

ภาพที่ 3.5

กราฟขยายกิจกรรม "กล่าว" ของบุคคลชื่อ "อภิสิทธิ์ เวชชาชีวะ"

ภาพที่ 3.6

กราฟขยายกิจกรรม "เปิดเผย" ของบุคคลชื่อ "อภิสิทธิ์ เวชชาชีวะ"



3.3 ขั้นตอนการทดลอง

ใช้บทความข่าวทางด้านการเมือง จำนวน 30 บทความ มาทดสอบโดยแบ่งการทดสอบออกเป็นสองส่วน ส่วนแรกใช้มนุษย์เป็นผู้ทดสอบจำนวน 5 ท่าน ทำการสกัดเอานิพจน์ระบุนามประเภทบุคคลตามกิจกรรมนำมาเป็นเกณฑ์มาตรฐาน ส่วนที่สองใช้ระบบทำการสกัดเอานิพจน์ระบุนามประเภทบุคคลตามกิจกรรมโดยใช้กฎและคลังคำศัพท์ร่วมกันสกัด นำผลลัพธ์ที่ได้ในส่วนที่สองเปรียบเทียบกับส่วนแรก จากนั้นใช้การประเมินประสิทธิภาพจากค่าความระลึก ค่าความแม่นยำ และค่าประสิทธิภาพโดยรวมของระบบเพื่อประเมินผลการทดลอง

3.4 การประเมินผลการทดลอง

ทำการเปรียบเทียบผลการทดลองการสกัดนิพจน์ระบุนามตามกิจกรรมของบุคคลจากระบบกับการสกัดนิพจน์ระบุนามตามกิจกรรมของบุคคลที่ได้จากมนุษย์

ค่าประสิทธิภาพโดยรวมของระบบ (F-measure)

F-Measure = 2 x Precision x Recall / (Precision + Recall)

ค่าความแม่นยำ(Precision)

Precision = Corr *100 / OutputWord

ค่าความระลึก (Recall)

Recall = Corr * 100 / RefWord

กำหนดให้

Corr = จำนวนคำที่ระบบเลือกมาตรงกับคำมนุษย์คิด

OutputWord = จำนวนคำที่ระบบเลือกออกมาทั้งหมด

RefWord = จำนวนคำที่มนุษย์เลือกออกมาทั้งหมด