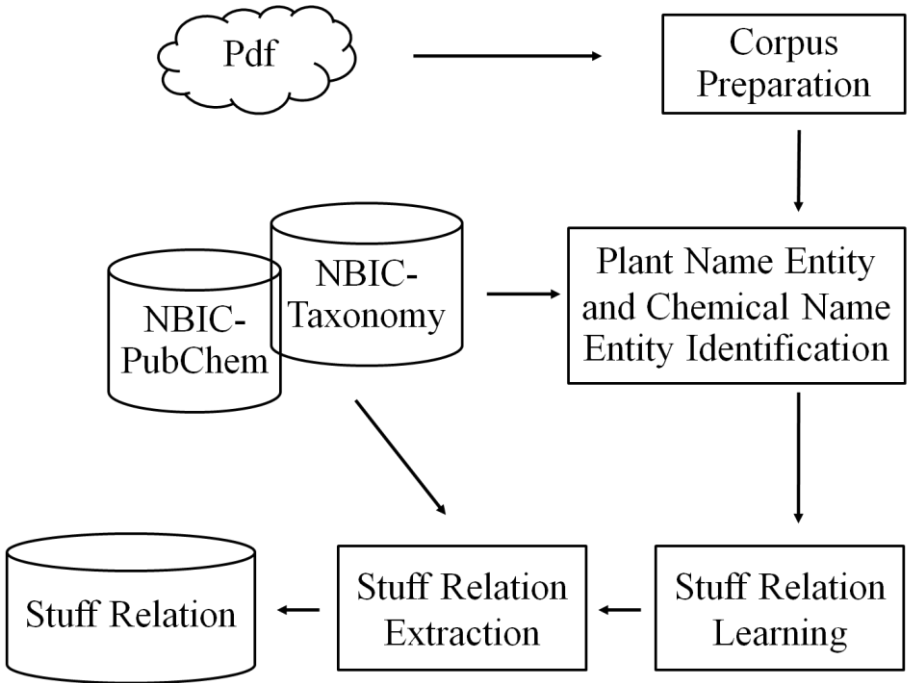
บทที่ 3
วิธีการดำเนินการวิจัยและเครื่องมือ

ในบทที่ 3 นี้จะเป็นบทที่อธิบายถึง วิธีดำเนินการ วิจัยและเครื่องมือ โดยจะแบ่งออกเป็นสองส่วนคือ ส่วนของวิธีการดำเนินการวิจัย เป็นการแสดงลำดับขั้นตอนของ การสกัดความสัมพันธ์แบบสตัฟฟ์และส่วนของเครื่องมือที่ใช้ โดยมีรายละเอียดขั้นตอนดังต่อไปนี้

3.1 วิธีดำเนินการวิจัย

ขั้นตอนของการสกัดความสัมพันธ์แบบสตัฟฟ์ประกอบไปด้วย 4ขั้นตอน ดังนี้คือ 1.การเตรียมคลังข้อมูล 2.การระบุชื่อของพืชและสารเคมี 3.การเรียนรู้ความสัมพันธ์แบบสตัฟฟ์และ 4. การสกัดความสัมพันธ์แบบสตัฟฟ์ดังแสดงในภาพที่3.1



ภาพที่ 3.1 แสดงภาพรวมของระบบการสกัดความสัมพันธ์แบบสตัฟฟ์

1. การเตรียมคลังข้อมูล (Corpus Preparation)

เอกสารที่ใช้ในงานวิจัยนี้เป็นเอกสารทางวิชาการด้านวิทยาศาสตร์ในโดเมนสารเคมีที่เป็นผลิตภัณฑ์จากธรรมชาติ ซึ่งดาวน์โหลดมาจากสำนักพิมพ์ ACS เป็นไฟล์รูปแบบ PDF ซึ่งต้องนำมาแปลงให้อยู่ในรูปของข้อความ text โดยใช้ PDFTextStream (http://snowtide.com) เอกสารที่ผ่านการแปลงแล้วจะนำมาทำเหมืองข้อมูลจำนวน20,000 ประโยค โดยวิธี 10 folds cross-validation

ในการศึกษาพฤติกรรมทางภาษาทำให้ได้เซตคำศัพท์ที่อยู่ระหว่างแนวคิดของสารเคมีกับแนวคิดของพืช(ตาราง 3.1) และจะทำการกำกับประโยคที่มีความสัมพันธ์แบบสตัฟฟ์ เป็นClass = “Yes” ด้วย ตัวกำกับสตัฟฟ์ (Stuff Tag) ดังแสดงในภาพที่ 3.2แล้วจะถูกแยกเอา Stop word Set ออก

ตารางที่ 3.1 แสดงความถี่ของคำศัพท์ที่น่าสนใจ

คำศัพท์ที่อยู่ระหว่างแนวคิดของสารเคมีกับแนวคิดของพืช	ความถี่
isolated	939
extract	418
isolation	228
parts	170
leaves	167
aerial	153
roots	149
species	88
bark	81
seeds	58
obtained	55

<stuff_relation class="yes">Four new flavonoids (1-4), along with 13 known compounds, were isolated from the heartwood of *Dalbergia louvelii* by following their potential to inhibit in vitro the growth of *Plasmodium falciparum*.</ stuff_relation>
<stuff_relation class="no">Of these, the ethyl acetate extract obtained from the heartwood of *Dalbergia louvelii* R. Viguier (Fabaceae).</stuff_relation>
<stuff_relation class="yes">Although several isoflavonoids have been obtained from roots of *P. floribundum*, none of the abovementioned compounds have been isolated previously from this species.</stuff_relation>
<stuff_relation class="no">The cytotoxicity of the isolates obtained herein from *P. floribundum* has been evaluated against a small panel of cancer cell lines.</stuff_relation>

ภาพที่ 3.2 แสดงการกำกับ stuff relation class "YES/NO" แต่ละประโยค

2. การระบุชื่อของพืชและสารเคมี (Plant name entity and chemical name entity identification)

Four new sesquiterpenes, (8R*)-8-bromo-10-epi-â-snyderol (1), (8S*)-8-bromo-α-snyderol (2), 5-bromo-3-(3′-hydroxy-3′-methylpent-4′-enylidene)-2,4,4-trimethylcyclohexanone (3), and the epoxide (4), have been isolated from the chloroform-methanol extract of Laurencia obtusa, together with the three known compounds R-snyderol (5), R-snyderol acetate (6), and stigmasterol.

W_1 W_2 W_3 W_4 W_5 W_6 W_7 W_8 W_9 W_{10} W_{11} W_{12} W_{13} W_{14} W_{15} W_{16} W_{17} W_{18} W_{19} W_{20}

ภาพที่ 3.3 แสดงประโยคตัวอย่างของการระบุชื่อของสารเคมี

โดยเริ่มจาก

1. ระบุตำแหน่งของคำที่เป็นแนวคิดของพืชที่อยู่ในแต่ละประโยคกับ NCBI-Taxonomy

2. ระบุคำที่เป็นแนวคิดของสารเคมีโดยใช้คำที่อยู่ในกรอบหน้าต่างซึ่งมีขนาดตั้งแต่ 1, 2,.., n (nคือจำนวนคำที่อยู่ด้านหน้าของตำแหน่งของคำที่เป็นแนวคิดของพืช) แล้วทำการเปรียบเทียบคำที่อยู่ในกรอบหน้าต่างกับ NBCI-PubChemที่ระดับกรอบหน้าต่างขนาดต่างๆ โดยเลื่อนกรอบหน้าต่างด้วยระยะทางครั้งละ 1 คำ

3. ทำเช่นเดียวกับข้อ 2 แต่ n คือจำนวนคำที่อยู่ด้านหลังของตำแหน่ งของคำที่เป็นแนวคิดของพืช

3. การเรียนรู้ความสัมพันธ์แบบสตัฟฟ์ (Stuff Relation Learning)

ขั้นตอนการเรียนรู้ของเครื่องด้วย Naïve Bayes Classifier โดย ใช้ โดยเครื่องมือวีก้า (Weka Tool) (Mark Hall, 2009) หา Conditional Probabilities ของฟีเจอร์ (Feature) ต่างๆที่แบ่งออกเป็น 3 กรณีศึกษาดังนี้

กรณีศึกษาที่ 1 ฟีเจอร์ที่ประกอบด้วย 3 กลุ่มพร้อมกับ Class ของประโยคที่เป็นความสัมพันธ์แบบสตัฟฟ์เมื่อ Class ={“yes”, “no”} จาก หัวข้อ 2.1.4 ฟีเจอร์กลุ่มแรกคือคำหรือสมาชิก (Element) ของเซต C ซึ่งเป็น เซตของแนวคิดของสาร เคมี กลุ่มที่ 2 คือคำหรือสมาชิก ของเซต S ซึ่งเป็นเซตของแนวคิดของพืชที่เป็น Natural Sourceในระดับดิวิชัน และกลุ่มที่ 3 คือ คำหรือสมาชิกต่างๆใน เซต A ซึ่งเป็นเซตของคำที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช โดยจะใช้คำทั้งหมดเหล่านี้ a_1, a_2,.., a_rที่มีความถี่มากที่สุด ภายใต้ขนาดกรอบหน้าต่าง r ของคำต่างๆ ระหว่างแนวคิดสารเคมีกับแนวคิดของพืช ซึ่งได้แสดงตัวอย่างของความน่าจะเป็นของ a_i (เมื่อ i = 1, 2, .., r) ภายใต้ขนาดกรอบหน้าต่าง r ดังตารางที่ 3.2 ถึงตารางที่ 3.6 สำหรับ r = 5 ตารางที่ 3.7 ถึงตารางที่ 3.10 สำหรับ r = 4 ตารางที่ 3.11ถึงตารางที่ 3.13 สำหรับ r = 3

ตารางที่ 3.2 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 5 ในกรณีศึกษาที่ 1

a_1	Class = ‘YES’	Class = ‘NO’
heartwood	0.00136519	0.00074627
isolated	0.51262799	0.08208955
extract	0.05119454	0.04104478
addition	0.00204778	0.00074627
obtained	0.00204778	0.00298507
…	…	…

ตารางที่ 3.3 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 5 ในกรณีศึกษาที่ 1

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.11083123	0.00956938
roots	0.02518892	0.00683527
parts	0.02078086	0.01025290
leaves	0.01700252	0.00751880
seeds	0.00881612	0.00410116
…	…	…

ตารางที่ 3.4 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 5 ในกรณีศึกษาที่ 1

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.03947368	0.00452196
aerial	0.02033493	0.00968992
parts	0.01614833	0.00129199
leaves	0.01555024	0.00129199
roots	0.01435407	0.00129199
…	…	…

ตารางที่ 3.5 แสดงความน่าจะเป็นของคำที่ 4 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO”เมื่อ r = 5 ในกรณีศึกษาที่ 1

a_4	Class = ‘YES’	Class = ‘NO’
aerial	0.01375358	0.00123381
parts	0.01088825	0.00061690
fruits	0.00916905	0.00123381
leaves	0.00744986	0.00061690
stems	0.00401146	0.00061690
…	…	…

ตารางที่ 3.6 แสดงความน่าจะเป็นของคำที่ 5 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO”เมื่อ r = 5 ในกรณีศึกษาที่ 1

a_5	Class = ‘YES’	Class = ‘NO’
parts	0.01528014	0.0006079
aerial	0.01075269	0.0006079
leaves	0.00962083	0.0006079
stems	0.00509338	0.0006079
bark	0.00509338	0.0006079
…	…	…

ตารางที่ 3.7 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 4 ในกรณีศึกษาที่ 1

a_1	Class = ‘YES’	Class = ‘NO’
leaves	0.01012146	0.01920236
isolated	0.50067476	0.07754801
extract	0.04453441	0.03545052
investigated	0.00067476	0.00295421
roots	0.01079622	0.00590842
…	…	…

ตารางที่ 3.8 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 4 ในกรณีศึกษาที่ 1

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.11407316	0.00810263
bark	0.00557967	0.00540176
parts	0.01673900	0.00810263
leaves	0.01487911	0.00607698
stems	0.00247985	0.00202566
…	…	…

ตารางที่ 3.9 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO”เมื่อ r = 4 ในกรณีศึกษาที่ 1

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.03256365	0.00448430
aerial	0.01539372	0.00768738
obtained	0.00296033	0.00064061
leaves	0.01480166	0.00128123
roots	0.01835406	0.00128123
…	…	…

ตารางที่ 3.10 แสดงความน่าจะเป็นของคำที่ 4 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO”เมื่อ r = 4 ในกรณีศึกษาที่ 1

a_4	Class = ‘YES’	Class = ‘NO’
aerial	0.01534963	0.00122624
parts	0.01421262	0.00061312
roots	0.00397953	0.00674433
leaves	0.01193860	0.00061312
stems	0.00397953	0.00061312
…	…	…

ตารางที่ 3.11 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 3 ในกรณีศึกษาที่ 1

a_1	Class = ‘YES’	Class = ‘NO’
heartwood	0.00133869	0.00074349
isolated	0.48728246	0.08104089
extract	0.04350736	0.04163569
addition	0.00200803	0.00074349
obtained	0.00200803	0.00371747
…	…	…

ตารางที่ 3.12 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ r = 3 ในกรณีศึกษาที่ 1

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.09588199	0.00949153
stems	0.00184388	0.00203390
pigment	0.00184388	0.00067797
leaves	0.01536570	0.00610169
fruit	0.00491703	0.00067797
…	…	…

ตารางที่ 3.13 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO”เมื่อ r = 3 ในกรณีศึกษาที่ 1

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.03627853	0.00448430
heartwood	0.00292569	0.00064061
addition	0.00175541	0.00064061
aerial	0.01872440	0.01089045
flowers	0.00526624	0.00064061
…	…	…

กรณีศึกษาที่ 2 ฟีเจอร์ที่ประกอบด้วย 3 กลุ่มพร้อมกับ Class ของประโยคที่เป็นความสัมพันธ์แบบสตัฟฟ์เมื่อ Class ={“yes”, “no”} ฟีเจอร์กลุ่มแรกคือคำหรือสมาชิก (Element) ของเซต C กลุ่มที่ 2 คือคำหรือสมาชิก ของเซต S และกลุ่มที่ 3 คือ คำหรือสมาชิกต่างๆใน เซต A โดยจะใช้ทุกคำทั้งหมดที่อยู่ระหว่างแนวคิดสารเคมีกับแนวคิดของพืช ซึ่งได้แสดงตัวอย่างของความน่าจะเป็นของ a_i (เมื่อ i=1,2,..,nและnคือจำนวนคำที่อยู่ระหว่างคำที่เป็นแนวคิดของสารเคมีและคำที่เป็นแนวคิดของพืช) ดังตารางที่3.14ถึงตารางที่3.23สำหรับ n=10

ตารางที่ 3.14 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_1	Class = ‘YES’	Class = ‘NO’
heartwood	0.00128866	0.00079177
isolated	0.50193299	0.06650831
extract	0.04317010	0.04275534
addition	0.00193299	0.00079177
obtained	0.00193299	0.00395883
…	…	…

ตารางที่ 3.15 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.10159480	0.00999286
obtained	0.00177200	0.00356888
heartwood	0.00590667	0.00071378
leaves	0.01476669	0.00499643
parts	0.02303603	0.01213419
…	…	…

ตารางที่ 3.16 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.03943662	0.00470746
aerial	0.02140845	0.01143241
flowers	0.00507042	0.00067249
leaves	0.01464789	0.00067249
roots	0.01746479	0.00134499
…	…	…

ตารางที่ 3.17 แสดงความน่าจะเป็นของคำที่ 4 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_4	Class = ‘YES’	Class = ‘NO’
aerial	0.01302225	0.00128783
heartwood	0.00325556	0.00064392
fruits	0.00868150	0.00128783
leaves	0.01302225	0.00064392
identified	0.00217037	0.00064392
…	…	…

ตารางที่ 3.18 แสดงความน่าจะเป็นของคำที่ 5 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_5	Class = ‘YES’	Class = ‘NO’
strain	0.00053591	0.00189873
aerial	0.01339764	0.00063291
growth	0.00160772	0.00126582
stems	0.00643087	0.00063291
colors	0.00053591	0.00126582
…	…	…

ตารางที่ 3.19 แสดงความน่าจะเป็นของคำที่ 6 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_6	Class = ‘YES’	Class = ‘NO’
inhibit	0.00163755	0.00064851
investigation	0.00054585	0.00129702
leaves	0.00545852	0.00064851
regions	0.00054585	0.00129702
stems	0.00109170	0.00064851
…	…	…

ตารางที่ 3.20 แสดงความน่าจะเป็นของคำที่ 7 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_7	Class = ‘YES’	Class = ‘NO’
potential	0.00167038	0.00066181
inhibit	0.00167038	0.00066181
agents	0.00055679	0.00529451
leaves	0.00111359	0.00066181
resulted	0.00111359	0.00132363
…	…	…

ตารางที่ 3.21 แสดงความน่าจะเป็นของคำที่ 8 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_8	Class = ‘YES’	Class = ‘NO’
potential	0.00172018	0.00068540
stems	0.00114679	0.00068540
natural	0.00057339	0.00274160
trunk	0.00057339	0.00137080
seeds	0.00516055	0.00068540
…	…	…

ตารางที่ 3.22 แสดงความน่าจะเป็นของคำที่ 9 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_9	Class = ‘YES’	Class = ‘NO’
bacterial	0.00058617	0.00140647
combined	0.00058617	0.00140647
fractionated	0.00058617	0.00210970
leaves	0.00234467	0.00070323
stems	0.00117233	0.00070323
…	…	…

ตารางที่ 3.23 แสดงความน่าจะเป็นของคำที่ 10 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO”เมื่อ n = 10 ในกรณีศึกษาที่ 2

a_{10}	Class = ‘YES’	Class = ‘NO’
assay	0.00059809	0.00144404
stems	0.00299043	0.00072202
fractionated	0.00059809	0.00216606
seeds	0.00179426	0.00072202
inhibition	0.00059809	0.00216606
…	…	…

กรณีศึกษาที่ 3 ฟีเจอร์ที่ประกอบด้วย 1กลุ่มพร้อมกับ Class ของประโยคที่เป็นความสัมพันธ์แบบสตัฟฟ์เมื่อ Class ={“yes”, “no”} ฟีเจอร์กลุ่มนี้คือ คำหรือสมาชิกต่างๆใน เซต A โดยจะใช้คำทั้งหมดเหล่านี้ a_1, a_2, .., a_r ที่มีความถี่มากที่สุด ภายใต้ขนาดกรอบหน้าต่าง r ของคำต่างๆระหว่างแนวคิดสารเคมีกับแนวคิดของพืช ซึ่งได้แสดงตัวอย่างของความน่าจะเป็นของ a_i (เมื่อ i = 1, 2, .., r)ภายใต้ขนาดกรอบหน้าต่าง r ดังตารางที่ 3.24 ถึงตารางที่ 3.28 สำหรับ r = 5 ตารางที่ 3.29 ถึงตารางที่ 3.32 สำหรับ r = 4 ตารางที่ 3.33 ถึงตารางที่ 3.35 สำหรับ r = 3

ตารางที่ 3.24 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 5 ในกรณีศึกษาที่ 3

a_1	Class = ‘YES’	Class = ‘NO’
heartwood	0.00136893	0.00073314
isolated	0.45995893	0.07331378
extract	0.05065024	0.04325513
addition	0.00205339	0.00073314
obtained	0.00068446	0.00366569
…	…	…

ตารางที่ 3.25 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 5 ในกรณีศึกษาที่ 3

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.09699625	0.00801068
roots	0.02252816	0.00667557
parts	0.02440551	0.01134846
leaves	0.01564456	0.00734312
seeds	0.00750939	0.00333778
…	…	…

ตารางที่ 3.26 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 5 ในกรณีศึกษาที่ 3

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.04045211	0.00441640
aerial	0.02201071	0.01072555
parts	0.01963117	0.00126183
leaves	0.01011303	0.00126183
roots	0.0184414	0.00126183
…	…	…

ตารางที่ 3.27 แสดงความน่าจะเป็นของคำที่ 4 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 5 ในกรณีศึกษาที่ 3

a_4	Class = ‘YES’	Class = ‘NO’
aerial	0.01600000	0.00120919
parts	0.01428571	0.00060459
fruits	0.00914286	0.00120919
leaves	0.01371429	0.00060459
stems	0.00400000	0.00060459
…	…	…

ตารางที่ 3.28 แสดงความน่าจะเป็นของคำที่ 5 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 5 ในกรณีศึกษาที่ 3

a_5	Class = ‘YES’	Class = ‘NO’
parts	0.01634724	0.00059524
aerial	0.01409245	0.00059524
leaves	0.00958286	0.00059524
stems	0.00676437	0.00059524
bark	0.00507328	0.00059524
…	…	…

ตารางที่ 3.29 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 4 ในกรณีศึกษาที่ 3

a_1	Class = ‘YES’	Class = ‘NO’
leaves	0.00986842	0.02148887
isolated	0.49736842	0.07828089
extract	0.05263158	0.02916347
investigated	0.00065789	0.00460476
roots	0.00921053	0.00613968
…	…	…

ตารางที่ 3.30 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 4 ในกรณีศึกษาที่ 3

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.10935601	0.00981767
bark	0.00546780	0.00280505
parts	0.02308627	0.00981767
leaves	0.01761847	0.00631136
stems	0.00303767	0.00210379
…	…	…

ตารางที่ 3.31 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 4 ในกรณีศึกษาที่ 3

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.02537486	0.00329381
aerial	0.02133795	0.00922266
obtained	0.00288351	0.00329381
leaves	0.01384083	0.00131752
roots	0.01557093	0.00131752
…	…	…

ตารางที่ 3.32 แสดงความน่าจะเป็นของคำที่ 4 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 4 ในกรณีศึกษาที่ 3

a_4	Class = ‘YES’	Class = ‘NO’
aerial	0.01551247	0.00062933
parts	0.01385042	0.00062933
roots	0.00387812	0.00692259
leaves	0.01218837	0.00062933
stems	0.00387812	0.00062933
…	…	…

ตารางที่ 3.33 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 3 ในกรณีศึกษาที่ 3

a_1	Class = ‘YES’	Class = ‘NO’
heartwood	0.00135962	0.00074516
isolated	0.49558124	0.06333830
extract	0.05234534	0.04470939
addition	0.00203943	0.00074516
obtained	0.00203943	0.00372578
…	…	…

ตารางที่ 3.34 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ r = 3 ในกรณีศึกษาที่ 3

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.10627719	0.00405954
stems	0.00310752	0.00202977
pigment	0.00186451	0.00067659
leaves	0.01429459	0.00744249
fruit	0.00435053	0.00067659
…	…	…

ตารางที่ 3.35 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO” เมื่อ r = 3 ในกรณีศึกษาที่ 3

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.03775811	0.00191449
heartwood	0.00294985	0.00063816
addition	0.00176991	0.00063816
aerial	0.02005900	0.01084876
flowers	0.00530973	0.00063816
…	…	…

กรณีศึกษาที่ 4 ฟีเจอร์ที่ประกอบด้วย 1 กลุ่มพร้อมกับ Class ของประโยคที่เป็นความสัมพันธ์แบบสตัฟฟ์เมื่อ Class ={“yes”, “no”} ฟีเจอร์กลุ่มนี้คือ คำหรือสมาชิกต่างๆใน เซต A โดยจะใช้ทุกคำทั้งหมดที่อยู่ระหว่างแนวคิดสารเคมีกับแนวคิดของพืช ซึ่งได้แสดงตัวอย่างของความน่าจะเป็นของ a_i (เมื่อ i = 1, 2, .., n และ n คือจำนวนคำที่อยู่ระหว่างคำที่เป็นแนวคิดของสารเคมี และคำที่เป็นแนวคิดของพืช) ดังตารางที่ 3.36 ถึงตารางที่ 3.45 สำหรับ n = 10

ตารางที่ 3.36 แสดงความน่าจะเป็นของคำที่ 1 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืช ในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_1	Class = ‘YES’	Class = ‘NO’
heartwood	0.00134771	0.00074906
isolated	0.51347709	0.08164794
extract	0.05256065	0.04344569
addition	0.00202156	0.00074906
obtained	0.00202156	0.00374532
…	…	…

ตารางที่ 3.37 แสดงความน่าจะเป็นของคำที่ 2 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_2	Class = ‘YES’	Class = ‘NO’
extract	0.10864198	0.00885559
obtained	0.00123457	0.00885559
heartwood	0.00246914	0.00068120
leaves	0.01666667	0.00749319
parts	0.02160494	0.01158038
…	…	…

ตารางที่ 3.38 แสดงความน่าจะเป็นของคำที่ 3 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_3	Class = ‘YES’	Class = ‘NO’
extract	0.04110393	0.00450161
aerial	0.01996477	0.01093248
flowers	0.00528479	0.00064309
leaves	0.01409278	0.00128617
roots	0.01761597	0.00064309
…	…	…

ตารางที่ 3.39 แสดงความน่าจะเป็นของคำที่ 4 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_4	Class = ‘YES’	Class = ‘NO’
aerial	0.01351351	0.00122850
heartwood	0.00337838	0.00061425
fruits	0.00394144	0.00122850
leaves	0.01295045	0.00061425
identified	0.00225225	0.00061425
…	…	…

ตารางที่ 3.40 แสดงความน่าจะเป็นของคำที่ 5 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_5	Class = ‘YES’	Class = ‘NO’
strain	0.00055772	0.00182149
aerial	0.00836587	0.00060716
growth	0.00167317	0.00121433
stems	0.00669269	0.00060716
colors	0.00055772	0.00121433
…	…	…

ตารางที่ 3.41 แสดงความน่าจะเป็นของคำที่ 6 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_6	Class = ‘YES’	Class = ‘NO’
inhibit	0.00171429	0.0018797
investigation	0.00057143	0.00125313
leaves	0.00514286	0.00062657
regions	0.00057143	0.00125313
stems	0.00285714	0.00062657
…	…	…

ตารางที่ 3.42 แสดงความน่าจะเป็นของคำที่ 7 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_7	Class = ‘YES’	Class = ‘NO’
potential	0.00175336	0.00063939
inhibit	0.00175336	0.00063939
agents	0.00058445	0.00511509
leaves	0.00116891	0.00063939
resulted	0.00058445	0.00255754
…	…	…

ตารางที่ 3.43 แสดงความน่าจะเป็นของคำที่ 8 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_8	Class = ‘YES’	Class = ‘NO’
potential	0.00180941	0.00066269
stems	0.00120627	0.00066269
natural	0.00060314	0.00265076
trunk	0.00060314	0.00132538
seeds	0.00542823	0.00066269
…	…	…

ตารางที่ 3.44 แสดงความน่าจะเป็นของคำที่ 9 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_9	Class = ‘YES’	Class = ‘NO’
bacterial	0.00061690	0.00135777
combined	0.00061690	0.00135777
fractionated	0.00061690	0.00203666
leaves	0.00246761	0.00067889
stems	0.00123381	0.00067889
…	…	…

ตารางที่ 3.45 แสดงความน่าจะเป็นของคำที่ 10 ที่อยู่ระหว่างแนวคิดของสารเคมีและแนวคิดของพืชในประโยคที่เป็น class “YES/NO” เมื่อ n = 10 ในกรณีศึกษาที่ 4

a_{10}	Class = ‘YES’	Class = ‘NO’
assay	0.00063052	0.00139373
stems	0.00315259	0.00069686
fractionated	0.00063052	0.00209059
seeds	0.00189155	0.00069686
inhibition	0.00063052	0.00209059
…	…	…

4. การสกัดความสัมพันธ์แบบสตัฟฟ์ (Stuff Relation Extraction)

ขั้นตอนการสกัดความสัมพันธ์แบบ สตัฟฟ์ เป็นการค้นหาและสกัดประโยคที่มีความสัมพันธ์แบบ สตัฟฟ์จากคลังข้อมูลสำหรับใช้ทดสอบ ด้วยสมการ (4) ร่วมกับค่าความน่าจะเป็นของฟีเจอร์ต่างๆตามกรณีศึกษา1-4แสดงในอัลกอริทึมการสกัดความสัมพันธ์แบบสตัฟฟ์ (ภาพที่ 3.4) และอัลกอริทึม การสกัดฟีเจอร์จากประโยค (ภาพที่ 3.5)โดย R คือความสัมพันธ์แบบสตัฟฟ์

L is a list of sentence.

i is a index of sentence list.

W is a set of word in each sentence.

C is natural-product-compounds concept set.

S is the natural-sources concept set.

A is a set of the high frequency words existing between $\langle c_1, c_2, .., c_i \}$ and $\langle s_1, s_2, .., s_j \rangle$

cf is a array of natural-product-compound occurrence features from $\langle c_1, c_2, .., c_i \rangle$

sf is a array of natural-source occurrence features from $\langle s_1, s_2, .., s_j \}$

af is a array of the high-frequency-word occurrence features existing between $\langle c_1, c_2, .., c_i \rangle$ and $\langle s_1, s_2, .., s_j \rangle$ varying to window sizes r (r = 3, 4, 5) and n (where n is the total number of words existing between the natural-product-compound concept words and the natural-source concept words)

STUFF_RELATION_EXTRACTION

1. { i ← 0, R ← ϕ
2. Array [max1] cf, Array [max2] sf, Array [max3] af
3. initialize each element of cf[], sf[], and af[] with " "
4. while (i ≤ length[L]) do
5. { ExtFea(cf[], sf[], af[])
6. Case 1 where max3 = r
7. StuffRelationForNaturalProduct = $\arg\max_{class \in Class} P(class) \prod_{num=1}^{max1} P(c_{num}) \prod_{num=1}^{max2} P(s_{num}) \prod_{num=1}^{max3} P(a_{num})$
8. Case 2 where max3 = n
9. StuffRelationForNaturalProduct = $\arg\max_{class \in Class} P(class) \prod_{num=1}^{max1} P(c_{num}) \prod_{num=1}^{max2} P(s_{num}) \prod_{num=1}^{max3} P(a_{num})$
10. Case 3 where max3 = r
11. StuffRelationForNaturalProduct = $\arg\max_{class \in Class} P(class) \prod_{num=1}^{max3} P(a_{num})$
12. Case 4 where max3 = n
13. StuffRelationForNaturalProduct = $\arg\max_{class \in Class} P(class) \prod_{num=1}^{max3} P(a_{num})$
14. if (StuffRelationForNaturalProduct == "yes") then
15. R = R ∪ {i} ;
16. }
17. return R
18. }

ภาพที่ 3.4 แสดงอัลกอริทึมของการสกัดความสัมพันธ์แบบสตัฟฟ์

```
ExtFeat(Array [max1] cf,  Array [max2] sf, Array [max3] af)
1.    { j ⟵ 0; k ⟵ 0
2.    initialize each element of cf[], sf[], and af[] with " "
3.      while(FindPositionOfS(w_j)) do
4.            j++;
5.      sf[]=FindElementsOfSfromPlantSourceBase
6.      while(FindPositionOfC (w_k)) do
7.            k++;
8.      cf[]=FindElementOfCfromChemNet
9.      Array [max3]  temp
10.     if( j > k ) then
11.       for(m ⟵ k to j)
12.           temp[m] = w_m
13.        Next m
14.     else if( k > j ) then
15.           for(m ⟵ j to k)
16.               temp[m] = w_m
17.           Next m
18.     Array featureA[] = SortingWord(temp[])
19.     af[]= featureA[]
20.   }
```

ภาพที่ 3.5 แสดงอัลกอริทึมของการสกัดฟีเจอร์สำหรับความสัมพันธ์แบบสตัฟฟ์

การวัด การวัดประสิทธิภาพของระบบ การสกัดความสัมพันธ์แบบสตัฟฟ์ จะอ้างอิงความถูกต้องจากเอกสารที่ผ่านการกำ กับจากผู้เชี่ยวชาญ ซึ่งการวัดประสิทธิภาพ จะวัดโดยใช้ค่าความถูกต้อง (precision) ค่าความระลึก (recall) และค่า F-measure ซึ่งทั้ง 3 ค่า สามารถคำ นวณได้ดังนี้

$$\text{Precision(P)} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \quad (7)$$

$$\text{Recall(R)} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNagative}} \quad (8)$$

$$\text{F} - \text{measure} = \frac{(\beta^2+1)\text{PR}}{\beta^2\text{P}+\text{R}} \quad (9)$$

โดย βคือค่าพารามิเตอร์ที่แสดงสัดส่วนความสำคัญระหว่างค่าความถูกต้องและค่าความระลึก โดยทั่วไป βจะมีค่าเท่ากับ 1

3.2 เครื่องมือที่ใช้

ในการวิจัยครั้งนี้ใช้เครื่องมือที่ประกอบไปด้วยส่วนของฮาร์ดแวร์ (Hardware) และซอฟต์แวร์ (Software) สรุปได้ดังต่อไปนี้

3.2.1ฮาร์ดแวร์(Hardware)

1.เครื่องแมคบุ๊กโปร(Macbook Pro) โดยมีรายละเอียดดังนี้

Model: Early 2011

CPU: 2.3GHz (2410M) Intel Core i5 พร้อม 3MB on-chip L3cache

RAM: 4GB 1333 MHz DDR3

VGA: Intel HD Graphics 3000 พร้อม 384 MB DDR3 SDRAM shared พร้อม main memory

Hard disk: 320 GB

Monitor: LED Display 13.3" @Ressolution 1,280 × 800

2.เครื่องคอมพิวเตอร์ตั้งโต๊ะ (Desktop Computer) โดยมีรายละเอียดดังนี้

CPU:3.40 GHz Intel Core i7-2600

RAM: 4GB

VGA: AMD Radeon HD 6450

Hard disk: 1TB

Monitor: Lenovo LED Display 21" @Ressolution 1,600 × 900

3.เครื่องบริการ (Server Computer) โดยมีรายละเอียดดังนี้

CPU:2.66GHzIntel Core i7-920พร้อม 8MB on-chip L3cache

RAM: 12GB 1333 MHz DDR3

VGA: AMD Radeon HD 6450

Hard disk: 2x500 GB

3.3.2ซอฟต์แวร์(Software)

1.ซอฟต์แวร์ระบบปฏิบัติการ (Operating System Software)โดยมีรายละเอียดดังนี้

Microsoft® Windows® 7 Enterprise

Mac OS X Lion 10.7.5

Linux: CentOS

2.ซอฟต์แวร์ปฏิบัติการประยุกต์ (Application Software) โดยมีรายละเอียดดังนี้

Adobe®DreamWeaver® CS5

Apache Web Server

PHP: Hypertext Preprocessor

MySQL RDBMS

Eclipse Classic (INDIGO)