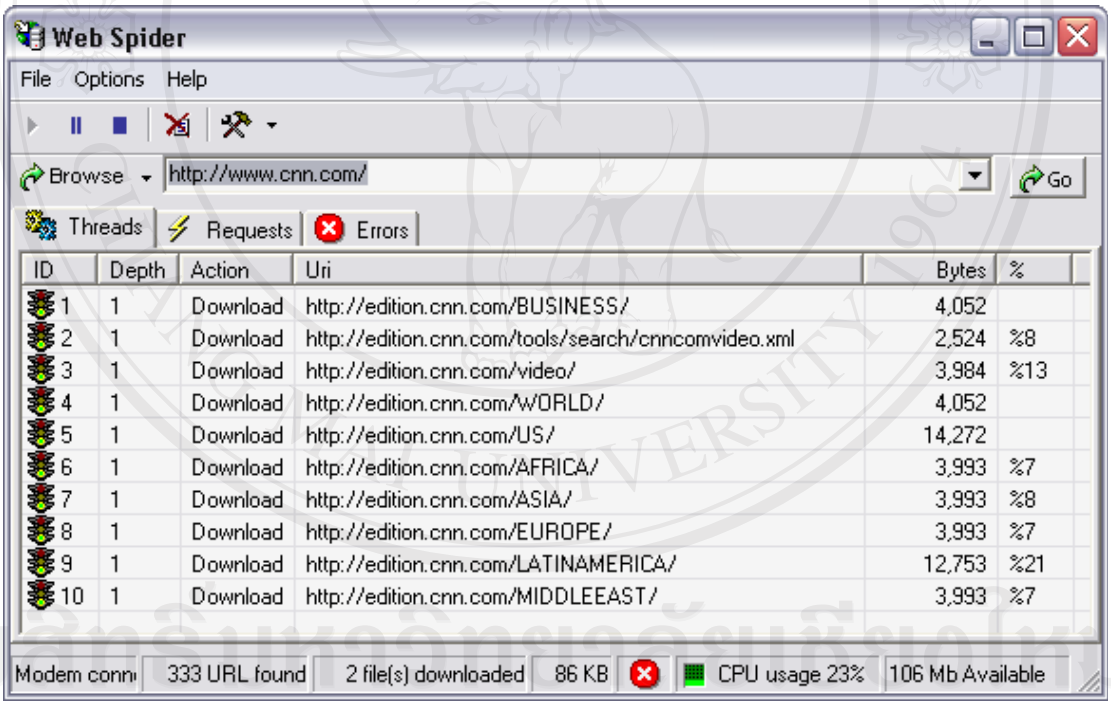
ภาคผนวก ก

โปรแกรมเว็บสไปเดอร์

เว็บสไปเดอร์ หรือเรียกว่าเว็บ ครอว์เลอร์ เป็นโปรแกรมที่เข้าเยี่ยมเวิลด์ไวด์เว็บโดยดำเนินตามแบบแผนอัตโนมัติ เว็บ ครอว์เลอร์จะเก็บหน้าสำเนาของเว็บเพจที่ทำการเยี่ยมชมทั้งหมดภายหลังการประมวลผล เว็บ ครอว์เลอร์สามารถใช้สำหรับงานซ่อมบำรุงอัตโนมัติบนเว็บไซต์เช่น การตรวจสอบการเชื่อมโยงหรือตรวจสอบโค้ดเอชทีเอ็มเอล นอกจากนี้โปรแกรมเว็บ ครอว์เลอร์สามารถใช้ในการรวบรวมประเภทเฉพาะข้อมูลจากเว็บเพจ เช่น การรวบรวมอีเมล ฯลฯ

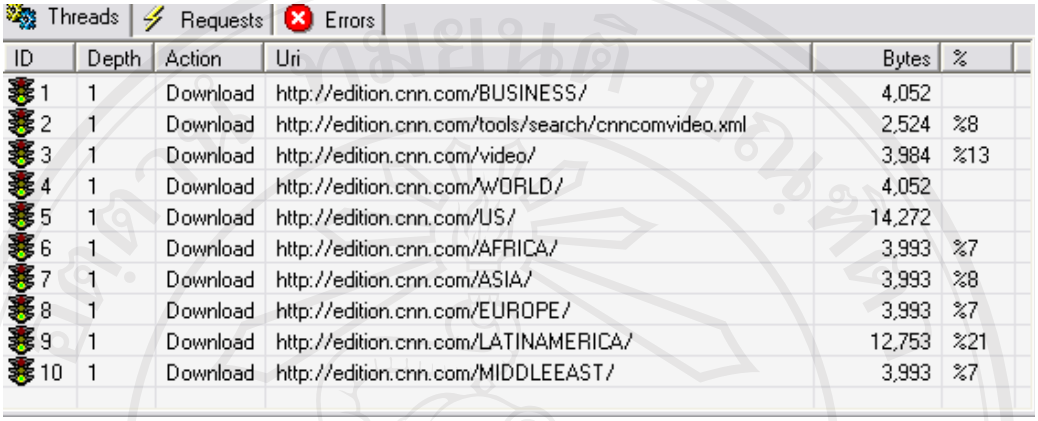


รูป 1 เว็บสไปเดอร์

ซอฟต์แวร์เว็บ สไปเดอร์มี 3 มุมมองที่สามารถปฏิบัติตามขั้นตอนการรวบรวมข้อมูลตรวจสอบรายละเอียดและดูข้อผิดพลาดในการรวบรวมข้อมูล

ภาพของเทรด (Threads view)

ในภาพของเทรดแสดงถึงแต่ละเส้นทางของการเข้าคิวของ URLs และการเชื่อมต่อของยูอาร์แอลเพื่อทำการเก็บหน้าสำเนาของเว็บเพจดังแสดงในรูป

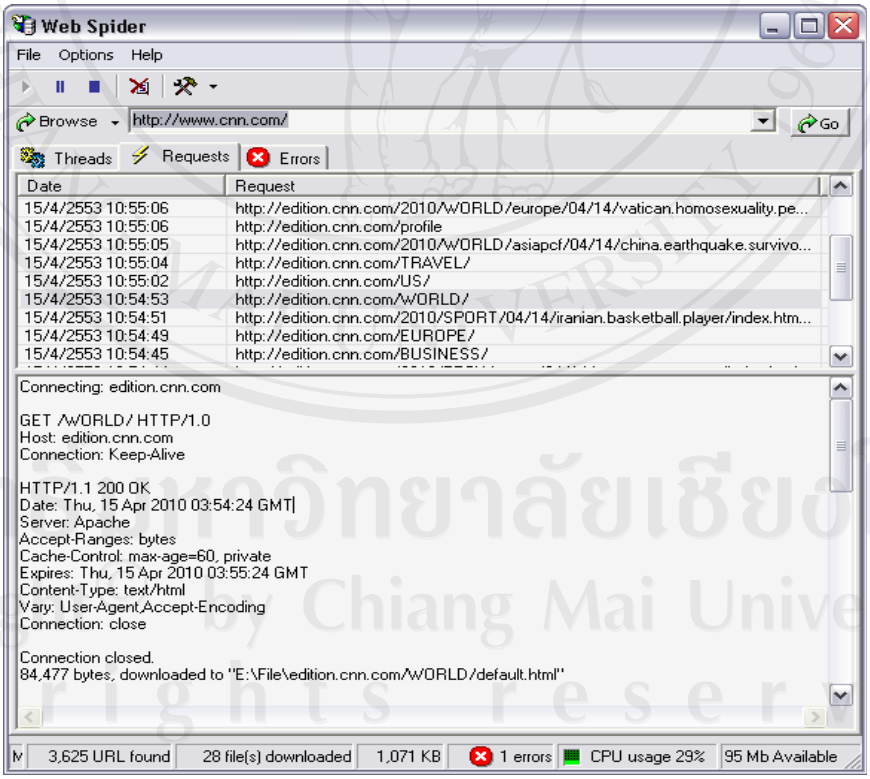


ID	Depth	Action	Uri	Bytes	%
1	1	Download	http://edition.cnn.com/BUSINESS/	4,052	
2	1	Download	http://edition.cnn.com/tools/search/cnncomvideo.xml	2,524	%8
3	1	Download	http://edition.cnn.com/video/	3,984	%13
4	1	Download	http://edition.cnn.com/WORLD/	4,052	
5	1	Download	http://edition.cnn.com/US/	14,272	
6	1	Download	http://edition.cnn.com/AFRICA/	3,993	%7
7	1	Download	http://edition.cnn.com/ASIA/	3,993	%8
8	1	Download	http://edition.cnn.com/EUROPE/	3,993	%7
9	1	Download	http://edition.cnn.com/LATINAMERICA/	12,753	%21
10	1	Download	http://edition.cnn.com/MIDDLEEAST/	3,993	%7

รูป 2 ของจำนวนเทรด

ภาพการร้องขอ (Requests view)

ภาพแสดงดูรายการล่าสุดของหน้าดาวน์โหลดในภาพของเทรด ดังต่อไปนี้



รูป 3 การร้องขอ

มุมมองนี้ช่วยให้คุณสามารถดูการตอบสนองแต่ละส่วนของหัวคำขอ

```
GET /WORLD/ HTTP/1.0
Host: edition.cnn.com
Connection: Keep-Alive
```

คุณสามารถดูการตอบสนองแต่ละส่วนของหัวคำขอ

```
HTTP/1.1 200 OK
Date: Thu, 15 Apr 2010 03:54:24 GMT
Server: Apache
Accept-Ranges: bytes
Cache-Control: max-age=60, private
Expires: Thu, 15 Apr 2010 03:55:24 GMT
Content-Type: text/html
Vary: User-Agent,Accept-Encoding
Connection: close

Connection closed.
84,477 bytes, downloaded to "E:\File\edition.cnn.com/WORLD/default.html"
```

และรายการที่พบ URL ที่สามารถใช้ได้ในหน้าดาวน์โหลด

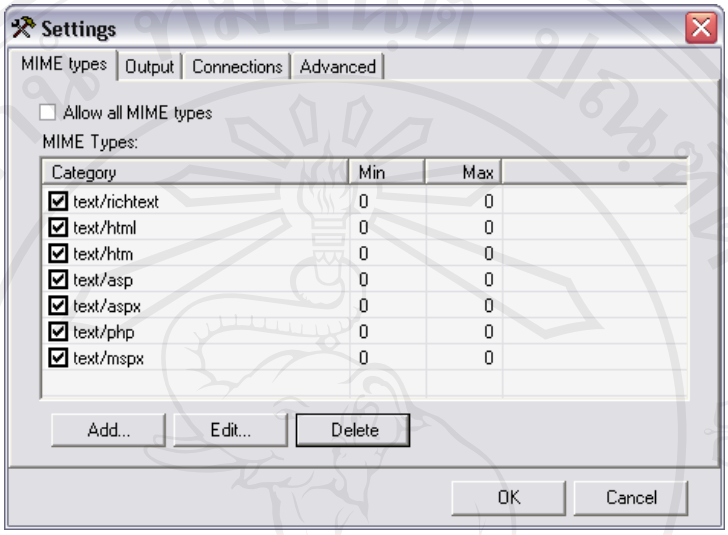
```
Parsing page ...
Found: 289 ref(s)
http://edition.cnn.com/2010/SPORT/tennis/04/14/history.of.tennis.federer.henryIII/index.html
http://edition.cnn.com/2010/SPORT/04/14/iranian.basketball.player/index.html
http://edition.cnn.com/2010/LIVING/04/13/russian.adoption.families/index.html
http://edition.cnn.com/2010/WORLD/meast/04/14/jordan.israel/index.html
http://edition.cnn.com/2010/WORLD/americas/04/14/costa.rica.trial/index.html
http://edition.cnn.com/2010/WORLD/africa/04/14/nigeria.jonathan/index.html
http://edition.cnn.com/2010/WORLD/europe/04/14/medvedev.iran.russia/index.html
http://edition.cnn.com/linkto/afghanistan.blogs.cnn.html
http://edition.cnn.com/.element/ssi/auto/3.0/sect/WORLD/gallery_content_0.html
```

การตั้งค่าเว็บครอว์เลอร์(Crawler Settings)

การตั้งค่าเว็บครอว์เลอร์ไม่ค่อยซับซ้อนจะมีการเลือกตัวเลือกคล้ายๆ กับเว็บครอว์เลอร์ในท้องตลาดทั่วไป รวมถึงการตั้งค่าเช่นการสนับสนุนชนิดของ MIME (Multipurpose Internet Mail Extensions), ดาวน์โหลดโฟลเดอร์, จำนวนกระทู้ทำงานและอื่นๆ

ประเภทของ MIME (MIME types)

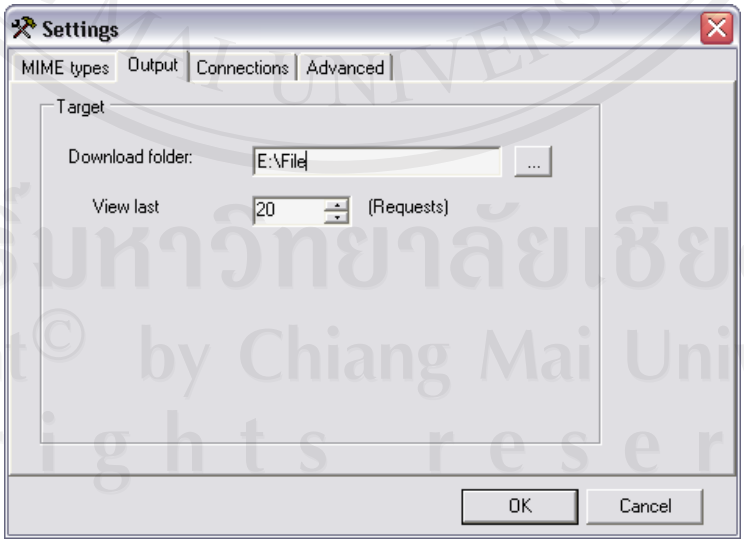
ประเภท MIME ประเภทที่ได้รับการสนับสนุนให้ดาวน์โหลดโดยเว็บครอว์เลอร์และรวมถึงรูปแบบเริ่มต้นที่จะใช้ ผู้ใช้สามารถเพิ่มแก้ไขและลบประเภท ผู้ใช้สามารถเลือกให้ชนิด MIME ทั้งหมดในรูปดังต่อไปนี้



รูป 4 ประเภทของ MIME

แหล่งเก็บหน้าสำเนาของเว็บเพจ (Output)

การตั้งค่าแหล่งเก็บหน้าสำเนาของเว็บเพจ รวมโฟลเดอร์ดาวน์โหลดและแสดงจำนวนคำขอให้ในมุมมองการตรวจสอบรายละเอียดคำขอ

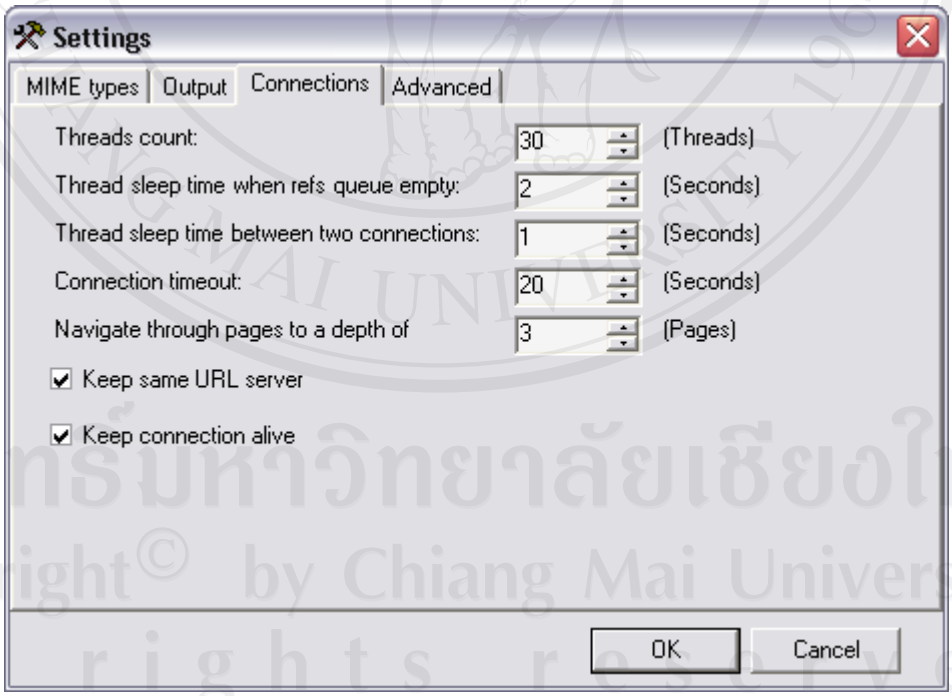


รูป 5 แหล่งเก็บหน้าสำเนาของเว็บเพจ

การเชื่อมต่อ (Connections)

มีการตั้งค่าการเชื่อมต่อ:

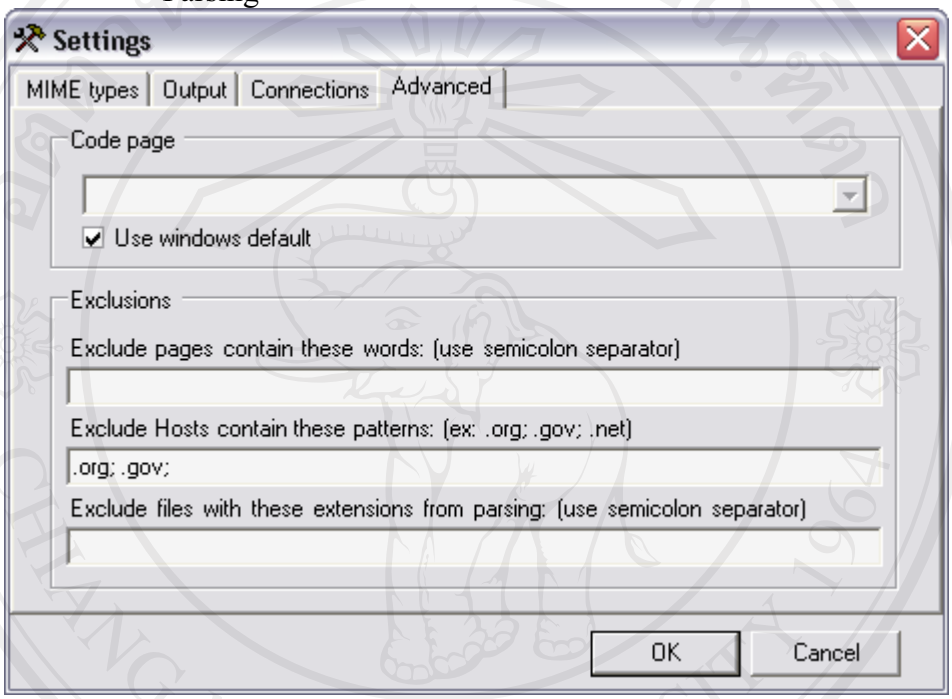
- Thread count : จำนวนของการทำงานของภาพของเทรดในเว็บครอว์เลอร์
- Thread sleep time when the refs queue is empty: เวลาในแต่ละเทรดพักเมื่อ refs ในคิวว่างลง
- Thread sleep time between two connections: เวลาในแต่ละเทรดพักหลังจากการจัดการทั้งหมดของคำร้อง ซึ่งเป็นค่าที่มีความสำคัญมากในการป้องกันในการโหลดจำนวนมากของโปรแกรม
- Connection timeout: หมายถึงการส่งและรับ Timeout ทุกซ็อกเก็ตของครอว์เลอร์
- Navigate through pages to a depth of : แสดงความลึกของนำทางในการรวบรวมข้อมูล
- Keep same URL server : จำกัดการรวบรวมข้อมูลยูอาร์แอลเดียวกันของต้นฉบับยูอาร์แอล
- Keep connection alive: เก็บการเชื่อมต่อซ็อกเก็ตสำหรับการร้องขอภายหลังเพื่อการหลีกเลี่ยงการเชื่อมต่ออีกครั้ง



รูป 6 การเชื่อมต่อ

การตั้งค่าขั้นสูง (Advanced)

- Code page สำหรับการเข้ารหัส Text page ที่ดาวน์โหลด
- รายการของหน้าเว็บที่ผู้ใช้กำหนดรายการ โดยจำกัดเว็บที่มีคำที่ไม่ต้องการ
- รายชื่อโฮสต์ที่ผู้ใช้กำหนดรายชื่อโฮสต์ โดยจำกัดนามสกุลที่ไม้ต้องการ
- รายการข้อมูลที่ผู้ใช้กำหนดรายชื่อไฟล์ โดยจำกัดชื่อไฟล์ที่ไม้ต้องการหลังทำการ Parsing



รูป 7 การตั้งค่าขั้นสูง

จุดที่น่าสนใจ (Points of Interest)

1. การจัดการเทรด (Thread management)

จำนวนของเทรดในครอว์เลอร์ผู้ใช้สามารถกำหนดผ่านการตั้งค่าได้ โดยค่าเริ่มต้นจะกำหนดไว้ที่ 10 เทรด แต่สามารถเปลี่ยนจากแท็บการตั้งค่า “การเชื่อมต่อ” ในโค้ดครอว์เลอร์สามารถแก้ไขได้จากคุณสมบัติของ ThreadCount จากในโค้ดต่อไปนี้

```
// number of running threads
private int nThreadCount;
private int ThreadCount
{
   get   {   return nThreadCount;   }
   set
   {
      Monitor.Enter(this.listViewThreads);
```

```
try
{
  for(int nIndex = 0; nIndex < value; nIndex ++)
  {
    // check if thread not created or not suspended
    if(threadsRun[nIndex] == null ||
      threadsRun[nIndex].ThreadState != ThreadState.Suspended)
    {
      // create new thread
      threadsRun[nIndex] = new Thread(new
          ThreadStart(ThreadRunFunction));
      // set thread name equal to its index
      threadsRun[nIndex].Name = nIndex.ToString();
      // start thread working function
      threadsRun[nIndex].Start();
      // check if thread dosn't added to the view
      if(nIndex == this.listViewThreads.Items.Count)
      {
        // add a new line in the view for the new thread
        ListViewItem item =
          this.listViewThreads.Items.Add(
          (nIndex+1).ToString(), 0);
        string[] subItems = { "", "", "", "0", "0%" };
        item.SubItems.AddRange(subItems);
      }
    }
    // check if the thread is suspended
    else if(threadsRun[nIndex].ThreadState ==
          ThreadState.Suspended)
    {
      // get thread item from the list
      ListViewItem item = this.listViewThreads.Items[nIndex];
      item.ImageIndex = 1;
      item.SubItems[2].Text = "Resume";
      // resume the thread
      threadsRun[nIndex].Resume();
    }
  }
  // change thread value
```

```
            nThreadCount = value;
        }
        catch(Exception)
        {
        }
        Monitor.Exit(this.listViewThreads);
    }
}
```

ถ้า ThreadCode เพิ่มขึ้นโดยผู้ใช้ โค้ดสร้างเทรดใหม่หรือหยุดการระงับเทรดชั่วคราวจะทำงาน ระบบจะมีกระบวนการทำงานพิเศษเพื่อระงับเทรดชั่วคราว แต่ละการทำงานของเทรจะต้องมีจำนวนเทรดเท่ากับ index ใน array ด้วย ถ้าหากจำนวนค่าเทรดมากกว่า ThreadCount ก็ยังคงทำงานและย้ายการทำงานไปสำหรับโหมดหยุดพักชั่วคราว

2. ความลึกของครอว์เลอร์ (Crawling depth)

ความลึกนี้ช่วยให้เว็บครอว์เลอร์เดินทางไปตามลำดับการเชื่อมต่อของเว็บเพจ ซึ่งแต่ละยูอาร์แอลจะมีความลึกเท่ากับความลึกของยูอาร์แอลเริ่มต้นบวกหนึ่ง โดยความลึกระดับ 0 ถูกกำหนดจากผู้ใช้ในตอนแรก ส่วนยูอาร์แอลระดับความลึกอื่นจากเว็บเพจอื่นๆ จะทำการถูกจัดความลึกและทำอย่างนี้จนจบของคิวในยูอาร์แอลซึ่งหมายถึงกระบวนการที่เรียกว่า "First in First Out" และทุกๆ เทรดสามารถเพิ่มคิวได้ทุกเวลา ตามโค้ดดังต่อไปนี้

```
void EnqueueUri(MyUri uri)
{
    Monitor.Enter(queueURLS);
    try
    {
        queueURLS.Enqueue(uri);
    }
    catch(Exception)
    {
    }
    Monitor.Exit(queueURLS);
}
```

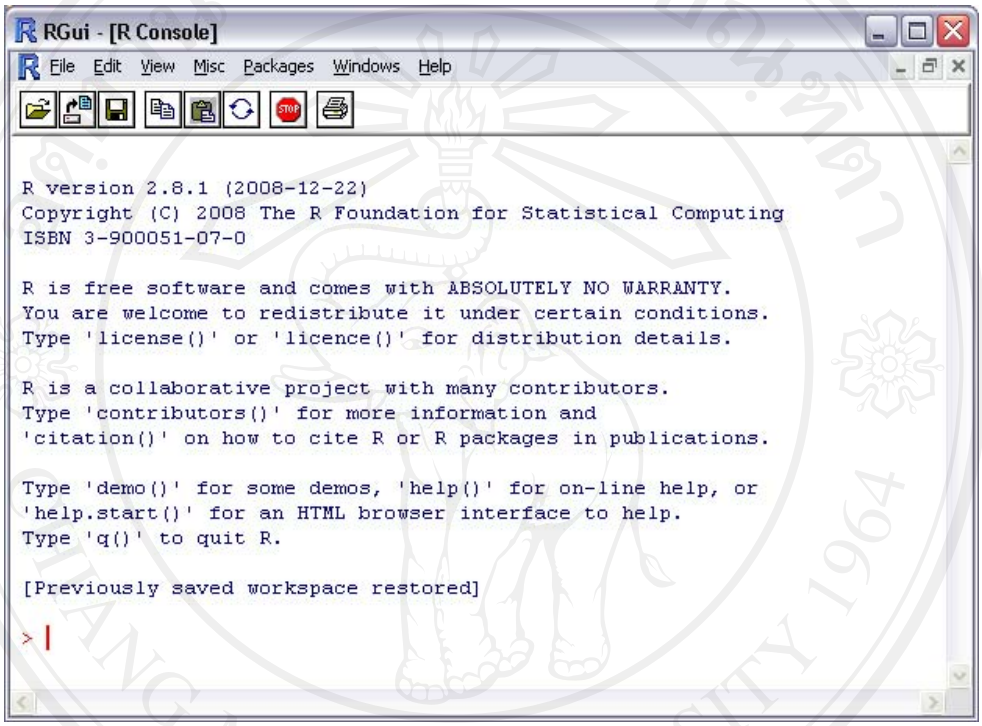
และแต่ละเทรดสามารถเรียกยูอาร์แอลแรกในคิวที่จะร้องขอดังแสดงในโค้ดต่อไปนี้

```
MyUri DequeueUri()
{
    Monitor.Enter(queueURLS);
    MyUri uri = null;
    try
    {
        uri = (MyUri)queueURLS.Dequeue();
    }
    catch(Exception)
```

```
    {
    }
    Monitor.Exit(queueURLS);
    return uri;
}
```

ภาคผนวก ข

โปรแกรมภาษาอาร์



คำสั่งที่ใช้ในการทดลองในโปรแกรมภาษาอาร์มีดังนี้

1) เตรียมเอกสารเพื่อสร้างคำสำคัญเอกสารของกลุ่ม โดยใช้คำสั่ง

```
#นำเข้าแพกเก็จทีเอ็ม (Term Mining)
>library(tm)

#เลือกเอกสารที่นามสกุล .txt เท่านั้น
> txt <- system.file("texts", "txt", package = "tm")


#กำหนดที่อยู่ของเอกสารและอ่านเอกสารที่เป็นภาษาอังกฤษเท่านั้น
> (ovid <- Corpus(DirSource(txt), readerControl = list(reader = readPlain,
+ language = "en_US")))
```

2) สร้างคำสำคัญเอกสารของแต่ละกลุ่มทดสอบโดยใช้หลักการทำดัชนีเอกสาร โดยใช้คำสั่ง

```
#ทำการกำจัดเครื่องหมายต่างๆ และเปลี่ยนอักษรตัวพิมพ์ใหญ่ให้เป็นตัวพิมพ์เล็ก
> reuters <- tmMap(reuters, stripWhitespace)

#ทำการให้คำตัวเล็กทั้งหมด
> reuters <- tmMap(reuters, tmTolower)

#ทำการกำจัดคำศัพท์ที่ไม่มีผลกระทบต่อความหมายโดยทั่วไป
> reuters <- tmMap(reuters, removeWords, stopwords("english"))

#ตัดคำเพื่อหารากศัพท์
> tmMap(reuters, stemDoc)

#ทำการสร้างคำสำคัญจากชุดสำหรับเรียนรู้
>tdm <- TermDocMatrix(reuters)


#ทำการสร้างพจนานุกรมของคำสำคัญ-เอกสาร
>dictionary <- Dictionary(tdm)
```

3) สร้างเมตริกซ์ความถี่ของคำสำคัญ-เอกสาร ของแต่ละกลุ่มทดสอบโดยใช้คำสั่ง

```
#ทำการสร้างเมตริกซ์ความถี่ของคำสำคัญ-เอกสาร
> tdmD <- TermDocMatrix(reuters, list(dictionary = dictionary))

#ทำการค่าน้ำหนักของคำสำคัญ-เอกสาร
>TFIDF<-weightTfIdf(tdmD)
```

4) สร้างแบบจำลองของแต่ละกลุ่มทดสอบโดยใช้หลักการซัพพอร์ตเวกเตอร์แมชชีนโดยใช้คำสั่ง

```
#นำเข้าแพกเก็จอี1071และแพกเก็จคลาส
> library(e1071)
> library(class)

#ทำการตัดเมตริกซ์เพื่อเลือกข้อมูลและคลาสที่ต้องการจัดกลุ่ม
>x <- subset(Dmodel, select = -Type)
>y <- Type

#ทำการสร้างโมเดลโดยใช้ซัพพอร์ตเวกเตอร์แมชชีนในแพกเก็จอี1071
>model <- svm(x, y)
```

5) ทำการทำนายของแต่ละกลุ่มทดสอบ โดยใช้ข้อมูลทดสอบของแต่ละกลุ่มทดสอบเข้ากระบวนการทำดัชนีเอกสารและสร้างเมตริกซ์ความถี่ของคำสำคัญ -เอกสาร เพื่อทำนายตามกลุ่มทดสอบโดยใช้คำสั่ง

```
# ทำการทำนายตามโมเดลที่ได้สร้างในข้อที่ 4
pred <- predict(model, x)
# หรือใช้คำสั่งนี้
pred <- fitted(model)
```

ประวัติผู้เขียน

ชื่อ-สกุล นายเจษฎา เทโวขัติ

วัน เดือน ปี เกิด 13 กันยายน 2526

ประวัติการศึกษา สำเร็จการศึกษามัธยมศึกษาตอนปลาย โรงเรียนดอยสะเก็ดวิทยาคม ปีการศึกษา 2544

สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยเชียงใหม่ ปีการศึกษา 2548

ประสบการณ์ 2549 โปรแกรมเมอร์บริษัท ไอโซแคร์ ซิสเตมส์ จำกัด (ISOCARE SYSTEM CO.,LTD)

2549-ปัจจุบัน เจ้าหน้าที่ดูแลระบบสารสนเทศเพื่อการจัดการและระบบเชื่อมต่อของบริษัทอเมริกันอินเตอร์แนชชั่นแนลแอสชัวรันส์ จำกัด (ประกันวินาศภัย)