본 매뉴얼은 SocSciBot3 (http://socscibot.wlv.ac.uk/) 한글버젼

김동일 (영남대 '뉴미디어와 사회' 연구실, http://cafe.naver.com/newmas)

박한우 (영남대학교 언론정보학과, http://www.hanpark.net)

SocSciBot 3

Link crawler for the social sciences (new version online Dec 10th, 2004)

SocsciBot 는 연구 목적으로 만들어진 웹 사이트 crawler 이다.

지원 프로그램 SocSciBot Tools 와 Cyclist 와 함께, 사이트에서 링크 분석을 실행하거나 사이트의 수집, 혹은 사이트 수집에 관련된 서치 엔진을 작동시키는데 사용될 수 있다.

이 프로그램은 또한 어떻게 링크 분석과 서치 엔진 작업이 이루어지는지에 대해 설명하기 위한 강의에 사용 되어 질 수 있다.

SocSciBot 는 다음 조건에 따라 이용될 수 있는 프로그램이다. 2004 년 10 월에 나온 새로운 버전은 주로 웹사이트 owner 의 crawling 에 대한 불만을 해소하기 위한 것이다.

SocSciBot and associated software: Conditions of use

- 소프트웨어는 비상업적 목적으로 사용된다. 우리는 또한 이것의 사용에 의해 나타나는 어떤 손해에 대해 책임을 지지 않는다. 그리고 다운 로드된 프로그램의 작동에 의해 발생한 다른 프로그램이나 데이터의 손실에 대해 책임지지 않는다.

- 사용자는 프로그램에서 요구되는 정확한 email 주소를 기입한다. 그리고 사용자의 웹 crawling 에 대해 웹 마스터가 불만을 나타낼 경우에 대비해서 crawling 이 되는 기간 동안 email 을 체크한다. SocSciBot 은 자동적으로 웹 마스터에게 사용자가 crawling 하고 있고 crawling 을 중지하기 위해 웹 마스터가 사용자에게 말할 수 있도록 email 을 하는 option 을 가진다.

- 사용자는 사용하고 있는 대역폭의 여유가 없는 기관의 웹사이트를 crawl 하기 위해 SocSciBot3 을 사용하지 않는다(e.g. 후진국들).

- 사용자는 웹 서버들을 반복해서 crawling 함으로 인해 웹 서버를 오버로드 하지 않아야 하다. e.g. 매일.

- 사용자는 SocSciBot3 의 copy 가 가끔 경고 없이 연결이 끊어질 수 있는 것을 받아들여야 한다. 예를 들면 만약에 SocSciBot3 의 사용에 웹 마스트로부터 어떤 불만이 있는 경우에.

- 사용자는 SocSciBot3 의 사용이 가끔 둔해질 것이라는 것을 받아들여야 한다. 이것은 비윤리적 방식으로 사용되지 않도록 안전하게 하기 위한 것이거나 불만의 원인을 확인하기

위한 것이다. 비윤리적인 사용의 경우를 제외하고, 이 정보는 제 3 자에게 나타나지 않을 것이다.

프로그램은 여기에서 tools 를 처리하는 것을 포함해서 무료로 이용할 수 있다.

데이터를 저장하기에 적당한 폴더를 찾아서 압축을 풀어라.

프로그램이 자동적으로 SocSciBot 의 결과들을 처리하기 때문에 다른 프로그램들을 실행하기 전에(before starting the other programs) SocSciBot 로 데이터를 수집하라.

SocSciBot and associated software: Downloads and instructions

사용자가 사용의 조건에 동의할 때만 프로그램을 다운로드 하라.

- (위의) 사용의 조건들을 읽었고, 그 조건들을 받아들이면, 지금 *SocSciBot, SocSciBot Tools and Cyclist* 를 다운 받아라. 프로그램을 사용하기 위해 새로운 폴더에 파일의 압축을 풀어라. 처음 시작하기 전에 소프트웨어의 몇몇 중요한 부분들의 첫 시작을 설명하는 TUTORIAL 1 을 읽어라. 만약에 SocSciBot 의 이전 버전과 겹친다면, SocSciBot 의 새로운 버전을 실행하기 전에 SocSciBot.ini 파일을 삭제하라.

Tutorials and extra information

- Tutorial 1: Introduction to *SocSciBot, SocSciBot Tools* and *Cyclist* 이곳에서 즉시 SocSciBot 을 다운로드 하는 것에 대한 과정을 숙지하라. 이것은 사용자가 소프트웨어의 핵심 특징과 친숙해 지도록 만들 것이고 그것의 실행 능력에 대한 간단한 증명을 보여 줄 것이다.
- Tutorial 2: Mini link analysis research project case study 소규모 링크 분석 연구 프로젝트를 위해 SocSciBot 과 SocSciBot Tools 를 사용하는 것은 링크 분석 연구에 이용 가능한 주요 특징들을 소개하기 위한 것이다.
- Tutorial 3: Summary of how to use SocSciBot for a link analysis research project.
- Corpus Linguistics Tutorial: Using SocSciBot and Cyclist for text analysis/basic corpus linguistics. *Cyclist* 는 사용자의 다운로드 된 사이트의 텍스트를 위한 concordancer/search engine 인터페이스이다. 이것은 SocSciBot 파일서만 모순이 일어나지 않는다.
- A Frequently Asked Questions list for SocSciBot Tools. 자주하는 질문들은 업데이트되어있다. 이 FAQ 는 또한 어떤 메뉴 항목이 어떤 일에 사용되어야만 하는지를 지정해주는 것을 돕는다.

- SocSciBot3 brief instructions and manual 이 항목은 설명서를 완전하게 이해한 사람들과 crawler 에 대해 더 많은 정보를 얻고자 하는 사람들에게 어떤 그 외의 도움을 주도록 만들어졌다.
- *SocSciBot3 Tools* 매뉴얼은 2005 년 1 월에 완성되었다. 이 매뉴얼은 실행하기에 더욱 쉬워지고 모든 특징들을 어떻게 사용하는지를 설명하도록 시도하는 것을 설명서를 통해 나타낸다.

A major upgrade of SocSciBot3 Tools has been completed at the end of July and a minor upgrade in December 1, 2004.

기술적 지원이 제공 되지 않은 것을 유념하라.

프로그램은 윈도우 95 그리고 그 이상의 사양에서 작동한다. 그리고 5,000 페이지까지의 사이트를 crawl 할 것이고, 제한된 속도를 가진다. 만약 사용자가 더 많은 페이지나 더 빠르게 crawl 하기를 원한다면 email 을 보내 달라. 예를 들어, 우리는 부유한 나라들의 대학 웹 사이트의 더 빠르고 더 많은 crawl 을 허가한다. 데이터베이스 구조와 crawler 를 설명하는 항목이 cybermetrics database site 에 링크되어있다.

프로그램에 의해 보고된 많은 수의 타이틀 바(bar)와 제공된 요약 파일을 무시하라. 이것들은 테스트 목적으로 있는 것이다. 신용 있는 정보는 링크 데이터 파일과 텍스트 데이터 파일에 있다. 그러나 사용자가 이 정보를 얻기 위해 cybermetrics 프로그램들을 사용할 필요가 있을 수는 있다.

SocSciBot 는 그 자체로 사용될 수 있고 근간 서적인 link analysis book 와 연계하여 사용될 수 있다.

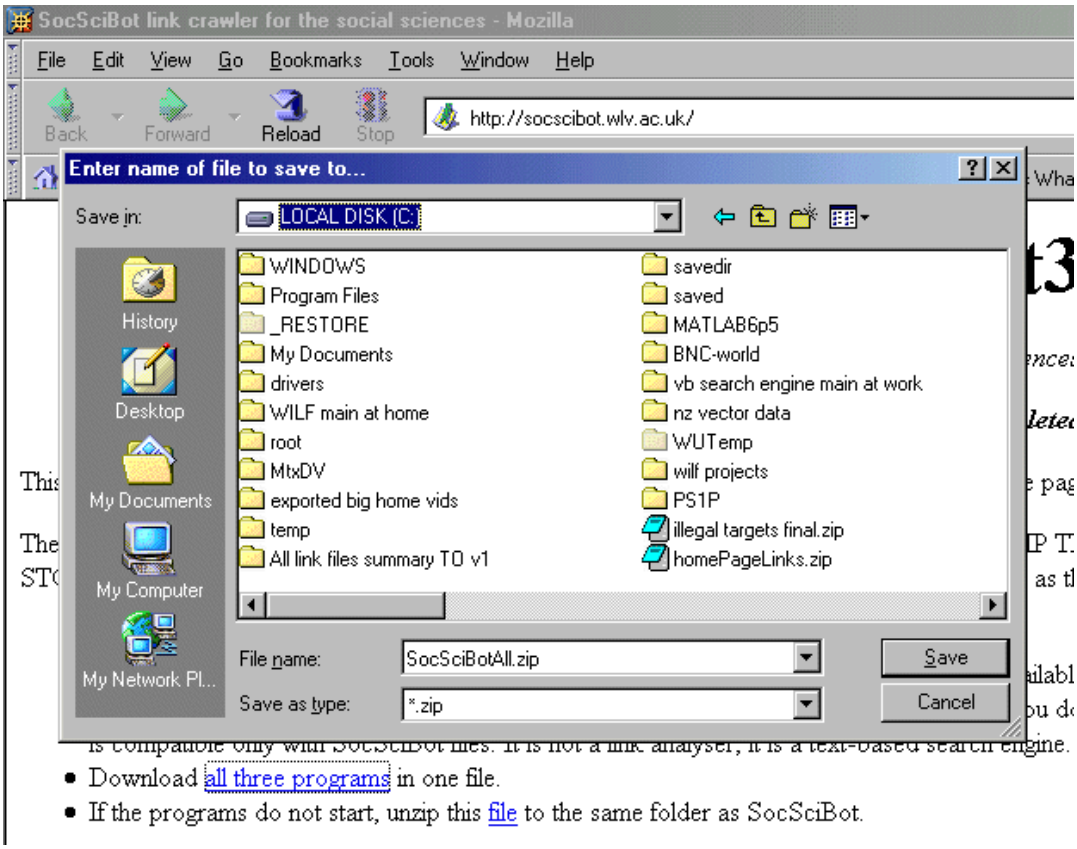
Tutorial 1: Introduction to *SocSciBot*, *SocSciBot Tools* and *Cyclist*

Overview

이 설명서 소개는 링크 데이터를 분석하기 위한 초기 crawling 에서 매우 소규모 SocSciBot 프로젝트의 모든 단계를 통해 이루어진다. 이 프로젝트를 실행하는 것은 SocSciBot 가 무엇을 할 수 있는지를 배우는 가장 쉬운 단계이다.

Step 1: Installing SocSciBot, SocSciBot Tools and Cyclist

1. SocSciBot 웹 사이트 http://socscibot.wlv.ac.uk/ 에 가서 사용자가 사용의 조건에 동의 할 때만 프로그램을 다운로드하기 위한 링크에 따르라. 사용자 컴퓨터에 의해 진행될 때, 데이터를 저장 공간의 충분한 양을 가지고 있는 곳에 프로그램을 저장하도록 지정하라. 이것은 보통 사용자의 컴퓨터 하드 드라이브가 될 것이다. 예, C: 드라이브.



2. 다음에, 사용자가 프로그램을 저장하도록 지정한 장소에 SocSciBotAll.zip 파일의 압축을 풀어라. 이것은 SocSciBot, SocSciBot Tools, 그리고 Cyclist 프로그램을 포함한 여러 가지 새로운 프로그램들을 만들 것이다.

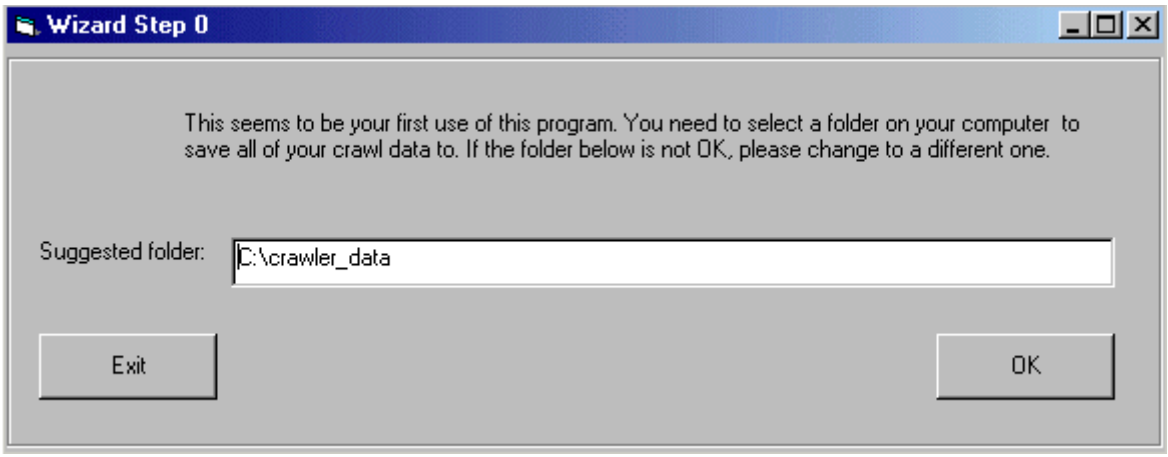
Step 2: Installing Pajek

만약 사용자가 SocSciBot 데이터와 함께 네트워크 다이어그램을 제공받기를 원한다면, 사용자는 Pajek 을 인스톨하도록 추천한다. 사용자는 처음에 SocSciBot 을 시작하기 전에 이것을 시행하도록 요구된다. 왜냐하면 SocSciBot 은 프로그램이 시작될 때, Pajek 을 찾는다. 그리고 Pajek 을 SocSciBot 을 먼저 실행한 후에 인스톨한다면 Pajek 을 찾지 못할 것이다.

1. Pajek 홈 페이지 http://vlado.fmf.uni-lj.si/pub/networks/pajek/에 가서 Pajek 의 최신 버전을 다운로드 해서 인스톨하라.

Step 3: Crawling a first site with SocSciBot

1. 사용자의 컴퓨터에 압축을 푼 폴더에서 SocSciBot 이나 SocSciBot.exe 둘 중 하나를 불러서 파일에서 더블 클릭하면 SocSciBot 이 작동한다. 이것은 아래의 것과 유사한 다이아로그 박스를 제공해야만 한다.

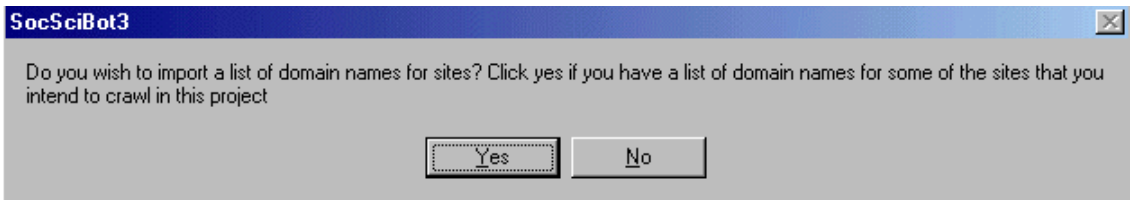


2. 데이터를 저장하기 위해 SocSciBot 에 의해 선택된 폴더를 확인하는 것은 OK 를 클릭하면 받아들여진다. 이것은 사용자가 crawl 하는 어떤 사이트의 웹 마스터들에게 email 을 보내도록 사용될 것이다. 이것은 윤리적인 수행임인 동시에 만약에 웹 마스터가 사용자가 그들의 사이트를 crawling 하는 것에 대해 불만족을 나타내는지 않은지에 대한 문제를 확인하는데 사용자에게 도움을 준다. 웹 마스터들은 사용자의 보스나 네트워크 매니저에게 email 을 하는 대신에 직접적으로 당신에게 email 을 보낼 수 있을 것이다. 사용자는 또한 crawl 의 목적을 설명하는 email 을 포함하는 메시지를 등록할 수 있다. 사용자는 프로젝트에 대해 부가적인 정보에 관련된 페이지의 URL 을 포함하기를 원할 수 있다. 또한, Microsoft Excel 과 Pajek 의 위치에 대한 어떤 질문들에 대해 응답하기를 원할 수 있다.

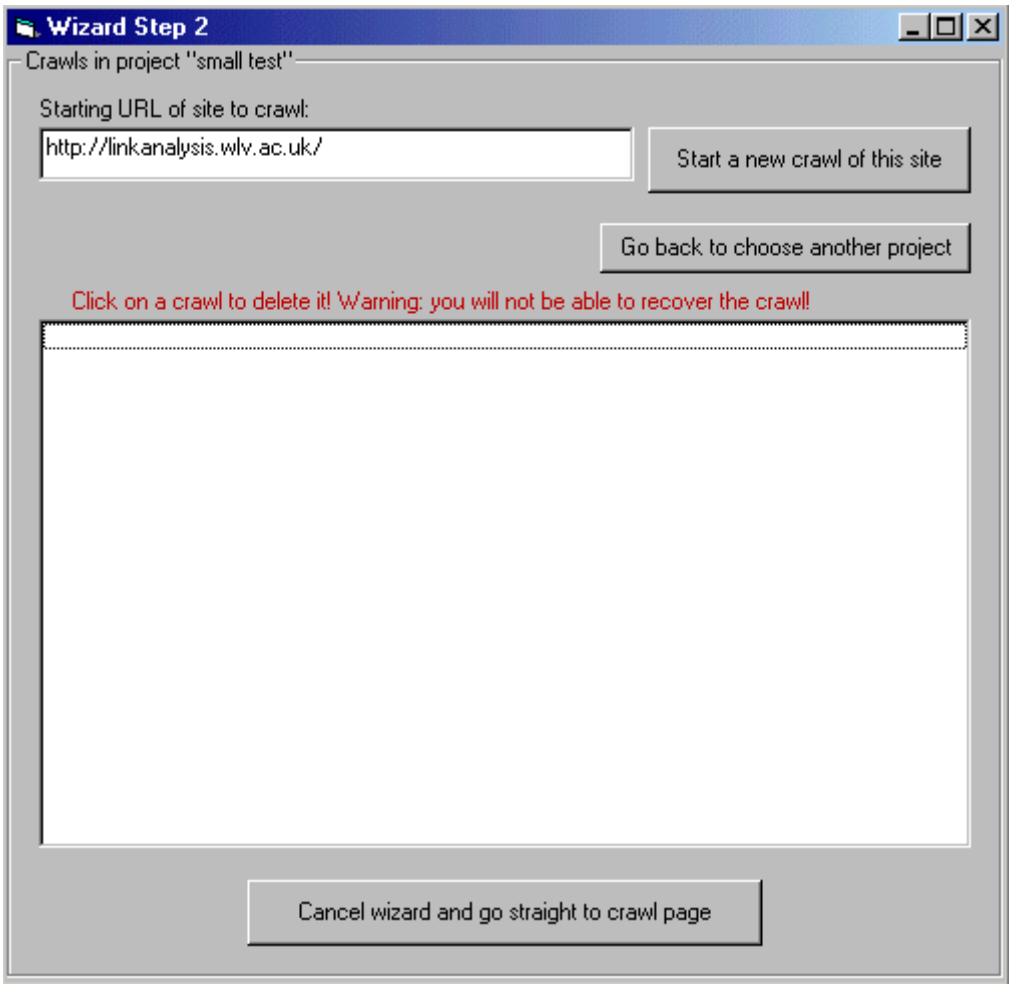
3. 다음 다이아로그 박스, Wizard Step 1 의 아래에 프로젝트의 이름으로 *small test* 를 입력하라. 그 다음에 *start new project* 버튼을 클릭하라. 모든 crawl 들은 프로젝트와 함께 그룹화 된다. 이것은 사용자가 개별적으로 분석된 crawl 그룹의 다른 이름을 가지도록 허용한다.

4. 다음에 질문에 *No* 를 클릭하라. 이것은 사용자가 전문가 수준이 되기 전에는 거의 필요로 하지 않는 확장된 데이터 클리닝(cleaning) 장치이다.

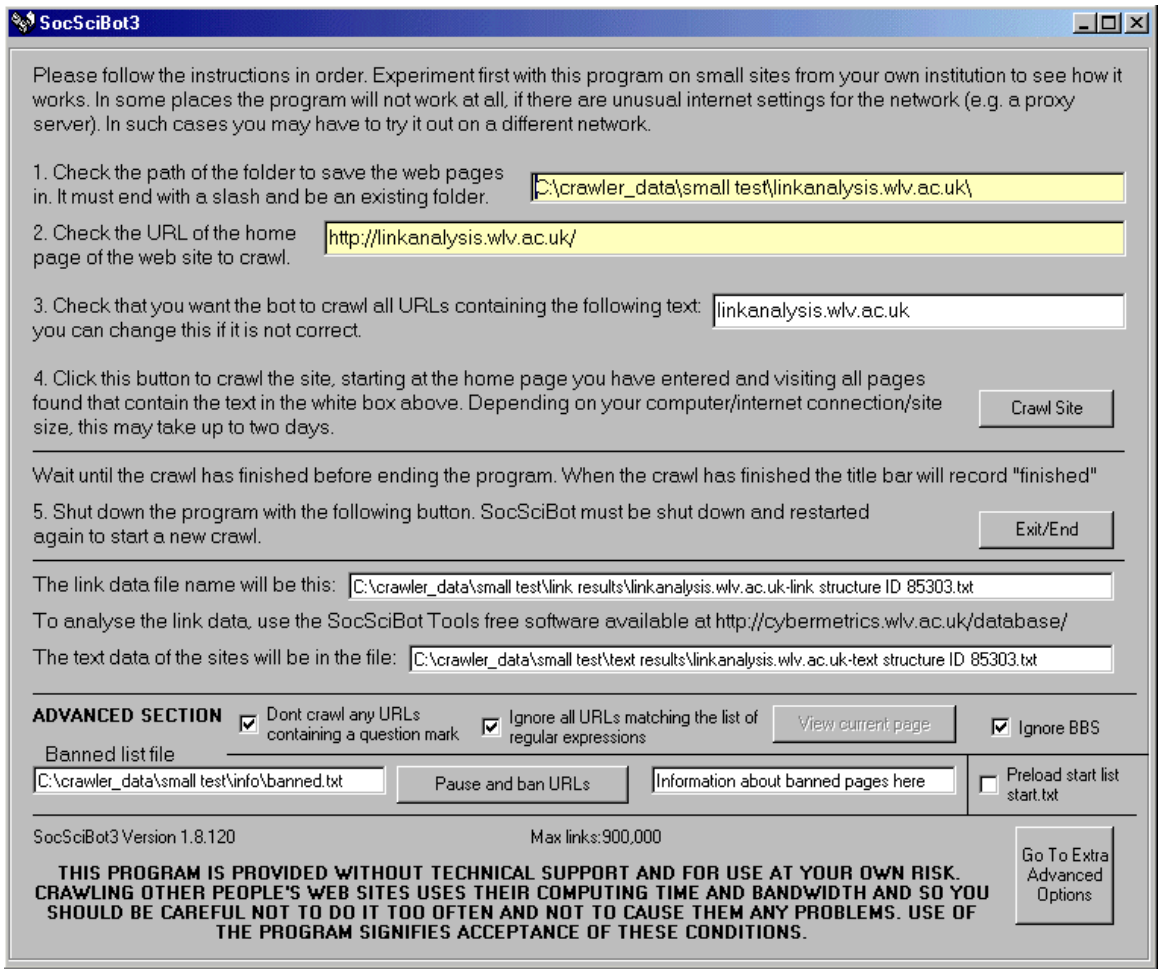
5. 다음에 두 번째 질문에 *No* 를 클릭하라. 이것은 사용자가 전문가 수준이 되기 전에는 거의 필요로 하지 않는 또 다른 확장된 데이터 정소 장치이다.



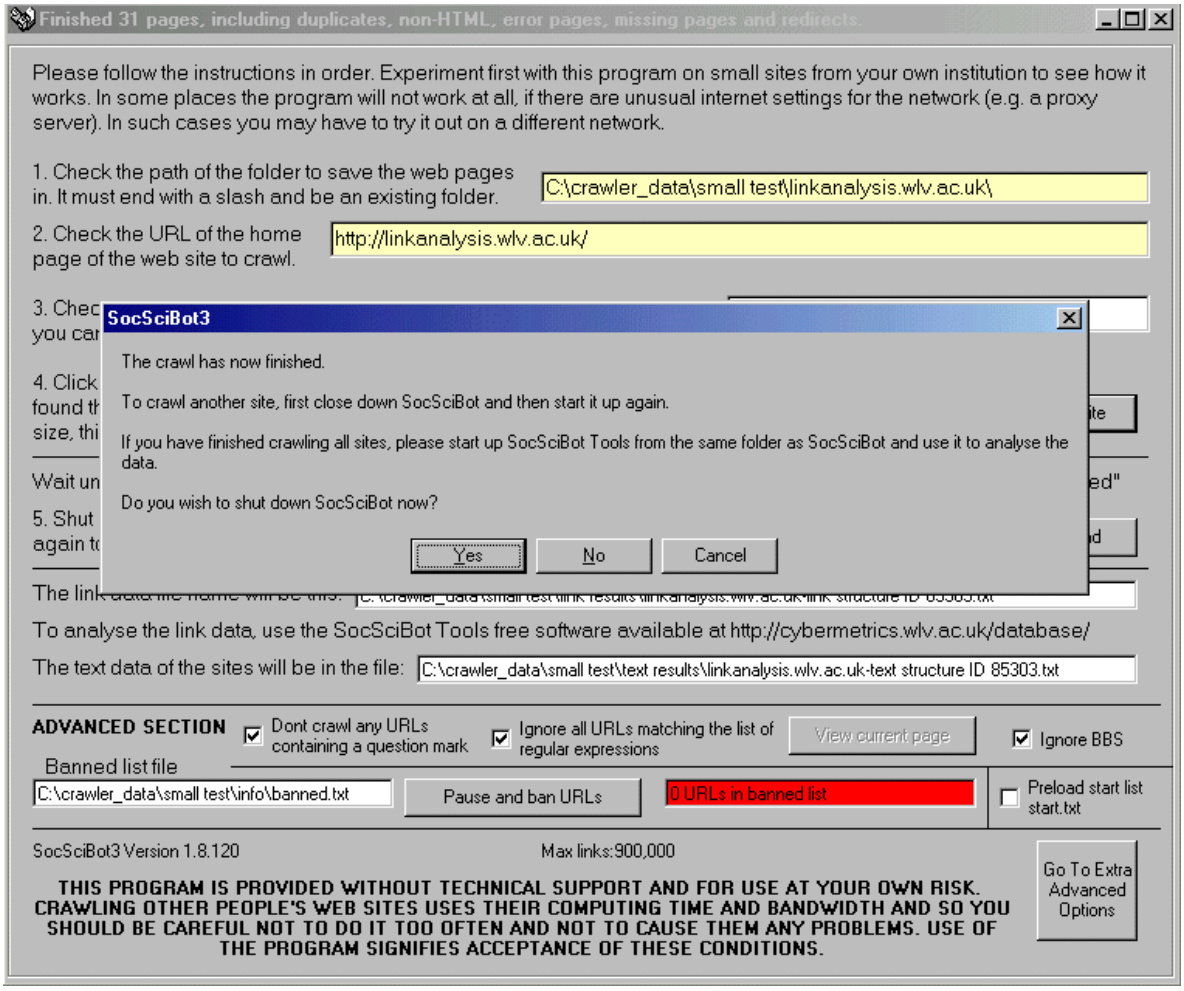
6. wizard step 2 다이아로그 박스에서, crawl 을 위해 사이트의 URL 을 시작하는 곳에 http://linkanalysis.wlv.ac.uk/를 입력하라. 그리고 *Start a new crawl of this site* 를 클릭하라.



7. Crawl 은 작동할 준비가 된다. *Crawl Site* 버튼을 클릭하라. 30 분이나 그 이상 지난 후에 crawl 은 끝난다.



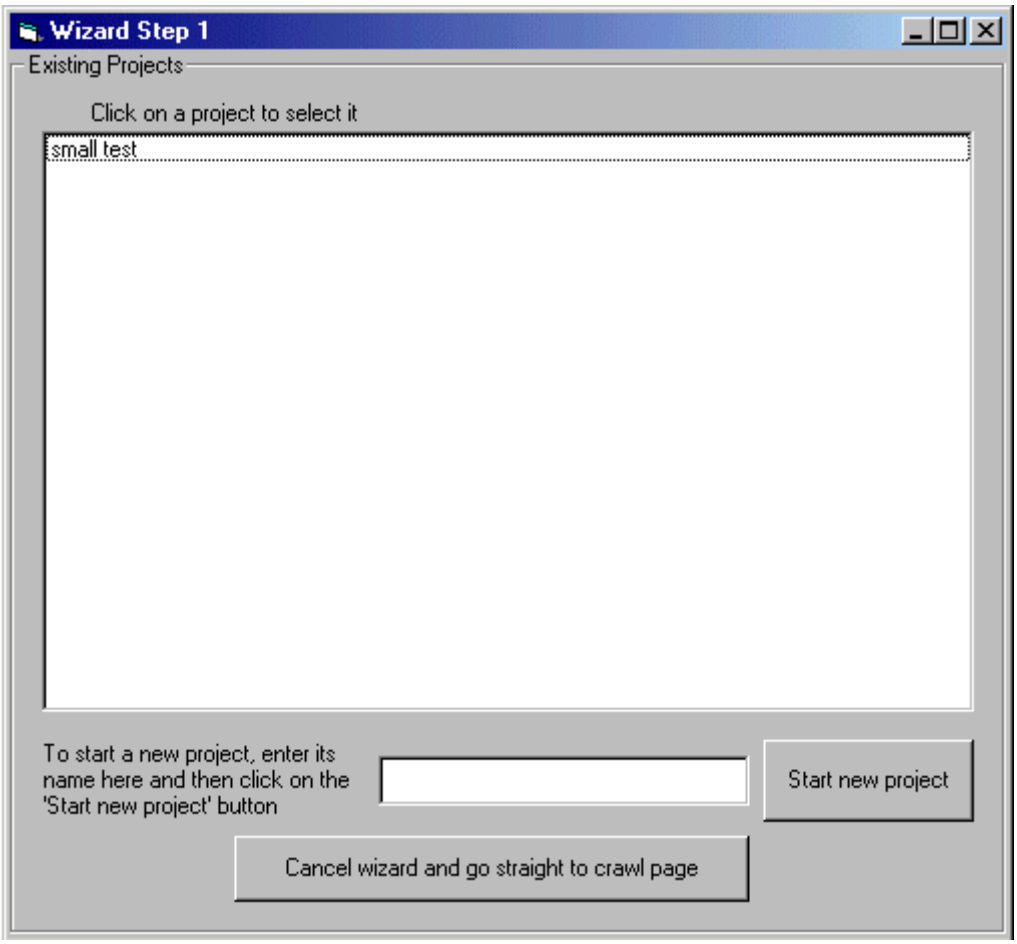
사용자는 crawl 하는 동안 윗부분의 타이틀 바에서 crawl 에 대한 정보를 읽을 수 있다. 그리고 또한 마지막 부분에서도 읽을 수 있다.



8. Crawl 이 완료 되었을 때 SocSciBot 을 끝내기 위해 *Yes* 를 클릭하라. 사용자는 이제 http://linkanalysis.wlv.ac.uk 사이트의 모든 페이지를 crawl 했다. 어떤 흥미로운 간단한 분석을 시행하기 전에, 다음 단계에서 2 개 이상의 사이트를 crawl 할 것이다.

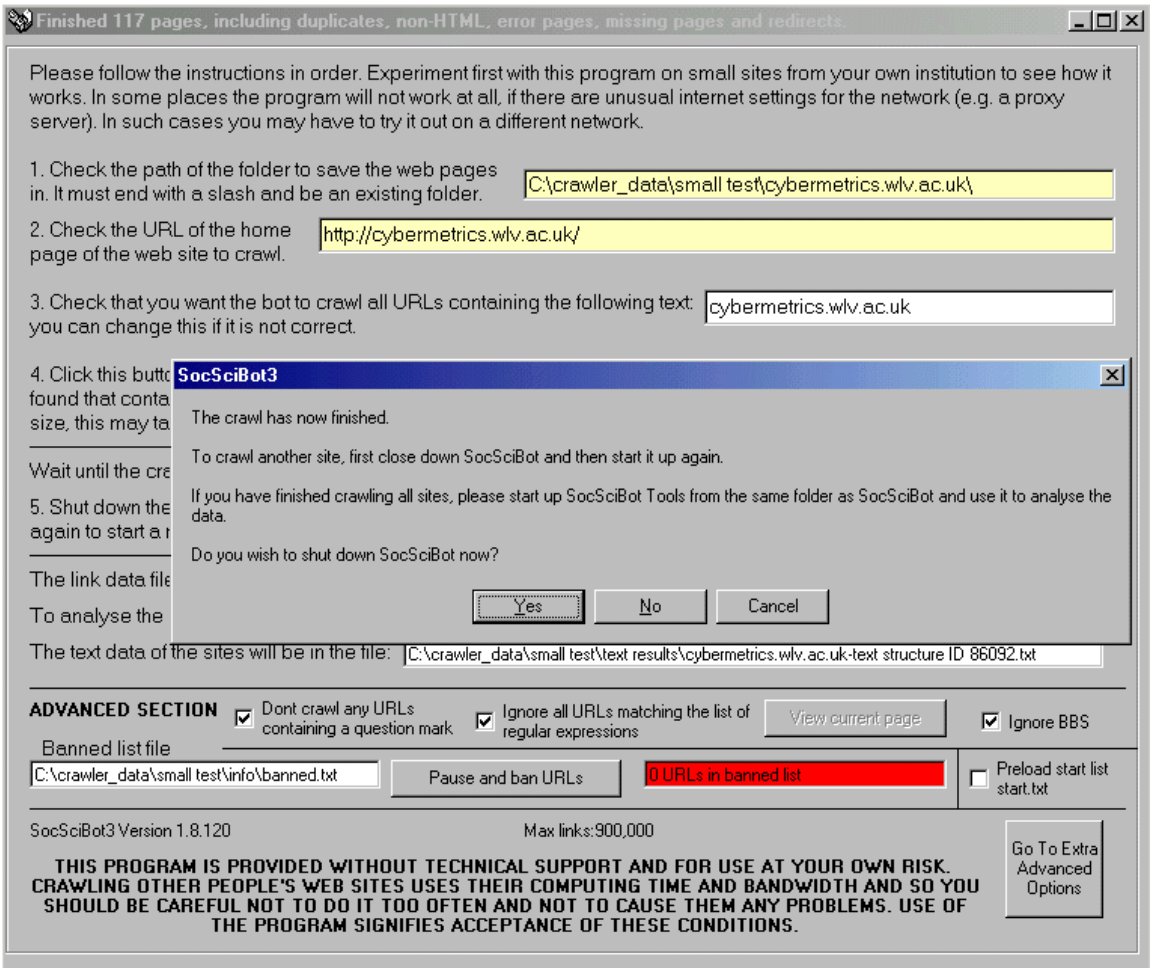
Step 4: Crawling two more sites with SocSciBot

1. 사용자의 컴퓨터에 있는 압축이 해제된 폴더의 SocSciBot 이나 SocSciBot.exe 파일을 더블 클릭해서 SocSciBot 를 다시 시작한다. 이것은 다른 crawl 을 추가하기 위해 이 프로젝트를 선택하는 *small test* 를 Wizard step 1. Click 을 통해 곧바로 사용자에게 지시되어야 한다.



2. Crawl 을 위해 두 번째 사이트의 URL 을 http://cybermetrics.wlv.ac.uk/로 입력하라. 그리고 *Start a new crawl of this site* 를 클릭하라.

3. 다음 화면에서 Crawl 사이트 버튼을 클릭하라. 그리고 crawler 가 완료 될 때까지 기다려라.



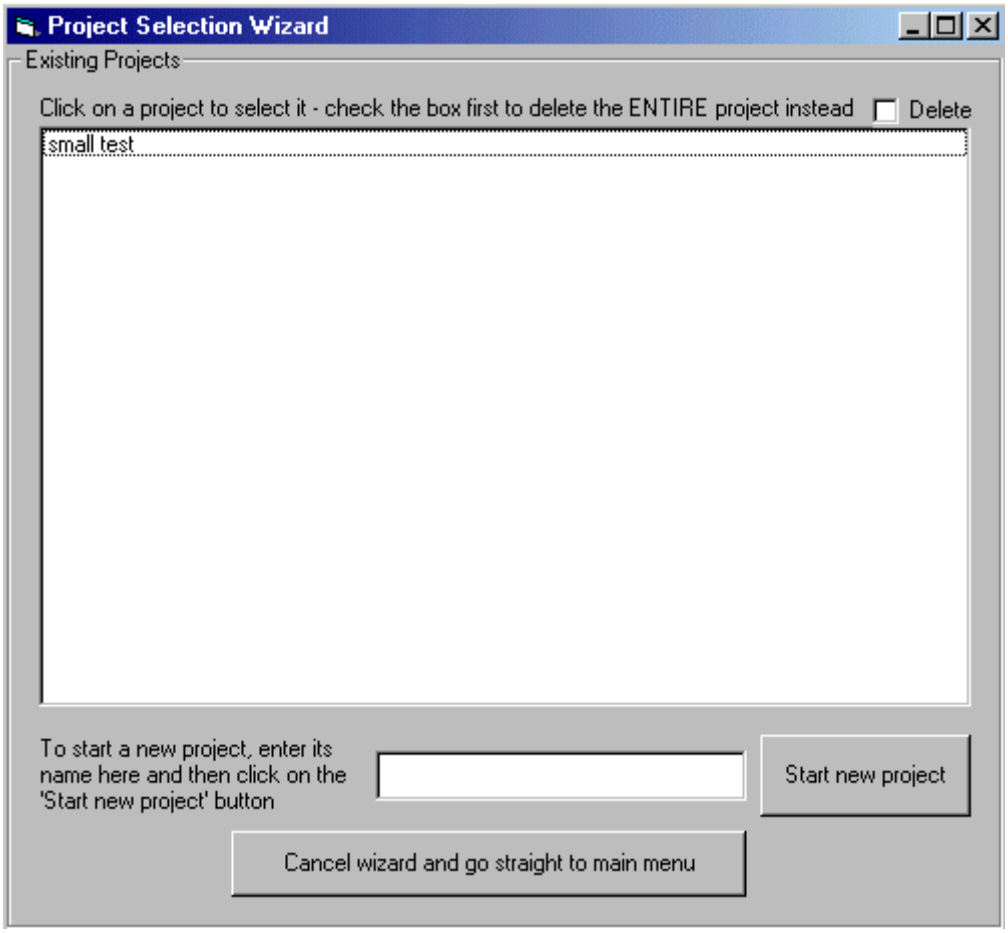
4. Crawl 을 마치기 위해 Yes 를 클릭하라.

5. URL http://socscibot.wlv.ac.uk/로 1~4 단계를 반복하라.

6. 사용자는 이제 성공적으로 3 개의 웹사이트를 crawl 한다. 그리고 그 사이트들을 분석하는 것을 이해한다.

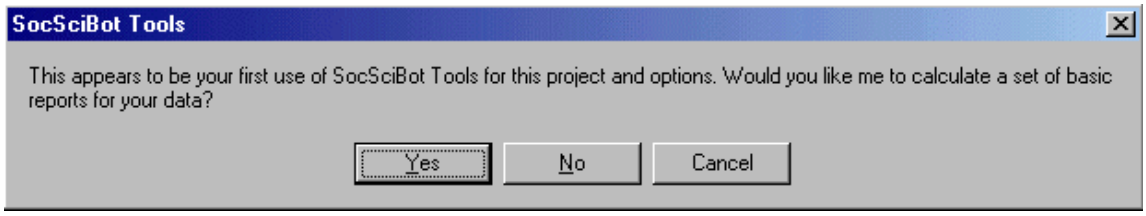
Step 5: Viewing basic reports about the project of three sites with SocSciBot Tools

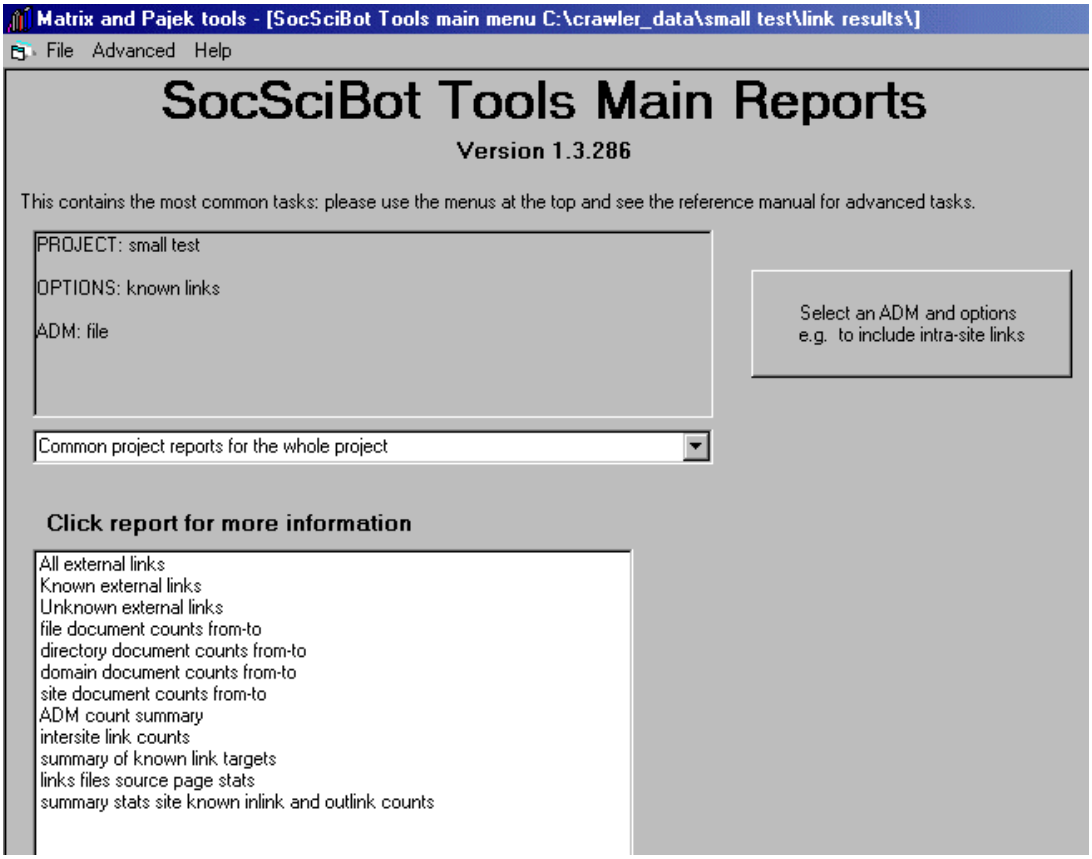
1. 사용자의 컴퓨터에 압축을 푼 폴더에서 SocSciBot Tools 나 SocSciBot Tool.exe 을 더블 클릭해서 SocSciBot Tool 을 시작하라. 이것은 분석을 위해 이 프로젝트를 선택하는데 *small test* 에서 Wizard step 1. Click 을 통해 직접적으로 사용자에게 지시된다.



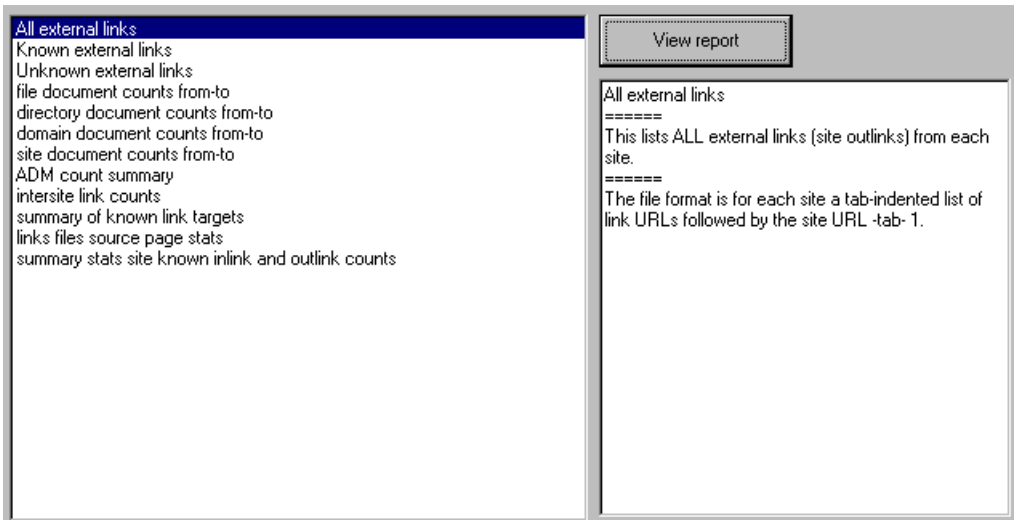
2. 다음의 다이아로그 박스에서 *Use this project* 를 선택하라.

3. 사용자가 일련의 기본적 보고들을 보기를 원하든지 아니든지 상관없이 다음 질문에 Yes 라고 응답하라.

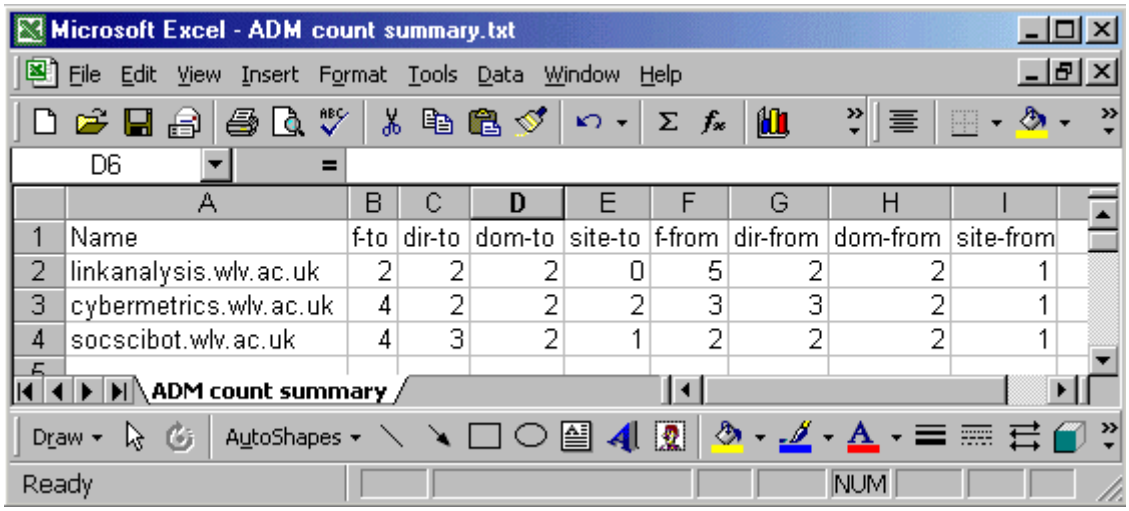




4. 다음 몇 초 후에 보고서들은 계산될 것이다. 그리고 사용자는 화면의 중앙에 메뉴를 따라 내려가면서 그 보고서들을 볼 수 있을 것이다. *All external links*(리스트의 가장 윗부분)를 클릭하라. 더 많은 정보가 화면의 오른쪽에 표시될 것이다. 그리고 크롤 대상이 된 각각의 사이트에 포함된 외부 URL 목표 페이지들의 리스트를 보려면 View report 를 클릭하라(아웃링크 사이트). 모든 보고서들을 동일하게 시행하라. 그리고 그 보고서들이 포함하는 것을 나타나도록 시행하라. 전체 URL 은 정상적으로 주어지지 않는다는 것을 주지하라. 이니셜 http://와 www 은 공간에 저장할 때 잘린다. 만약에 사용자가 컴퓨터에 Excel 을 가지고 있다면 사용자는 때때로 몇몇의 더 많은 버튼들을 작동해야 할 것이다. 그 버튼들은 사용자가 Excel 에서 보고서들을 볼 수 있도록 해줄 것이다. 이 보고서들은 대부분 링크 분석 조사들에 요구되는 링크 정보들을 포함한다.



5. 주요 보고서는 *ADM(Alternative Document Model) count summary* 이다. 이것을 클릭하고 사용자가 Excel 을 가지고 있다면 Excel 에서 View 를 클릭하라(Excel 이 없다면 view report 버튼을 클릭하라). 이것은 프로젝트에서 모든 사이트들에서부터 각각의 외부 사이트에 설정한 링크의 개수를 센다. 이 수치들은 4 개 ADM 의 각각을 보고하는 것이다. 대다수 사람들은 ADM 파일만 필요로 할 것이다(i.e. 표준 링크 수치). 이것은 페이지 인링크 칸(column)(프로그램의 구식 버전에서 f-to)이고, 페이지 아웃링크 칸(프로그램의 구식버전에서 f-from)이다. 예를 들면, linkanalysis.wlv.ac.uk 열(row)을 위한 두 개의 칼럼을 읽으면, 거기에 다른 두 개의 사이트로부터 linkanalysis.wlv.ac.uk 로 두 개의 링크가 있다. 그러나 linkanalysis.wlv.ac.uk 로부터 다른 두 개의 사이트로 5 개의 링크가 있다.

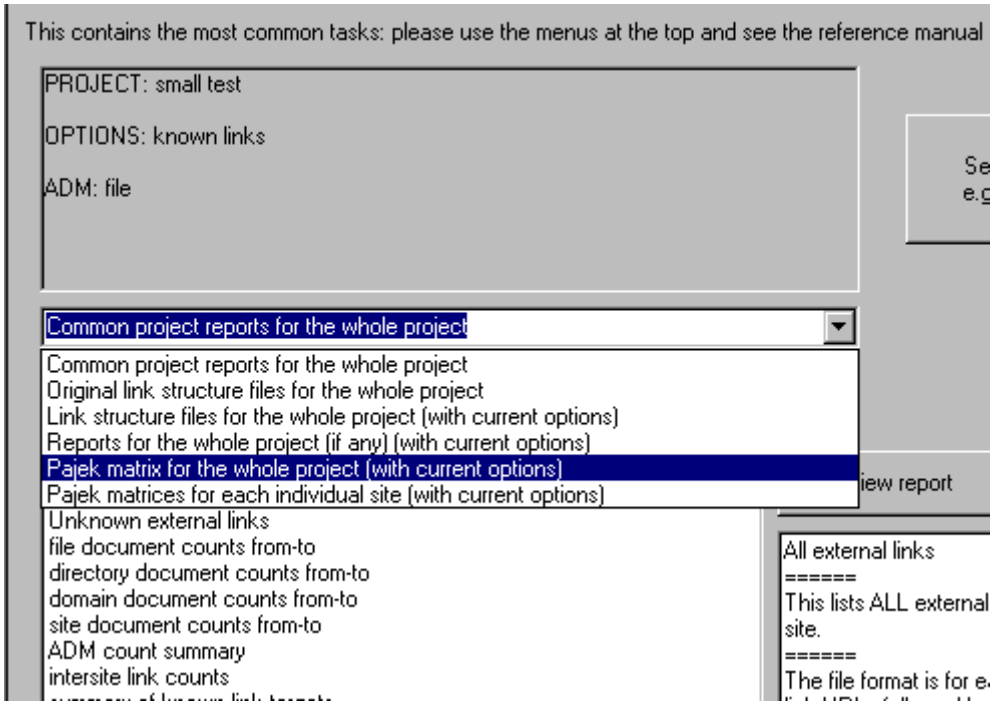


	A	B	C	D	E	F	G	H	I
1	Name	f-to	dir-to	dom-to	site-to	f-from	dir-from	dom-from	site-from
2	linkanalysis.wlv.ac.uk	2	2	2	0	5	2	2	1
3	cybermetrics.wlv.ac.uk	4	2	2	2	3	3	2	1
4	socscibot.wlv.ac.uk	4	3	2	1	2	2	2	1

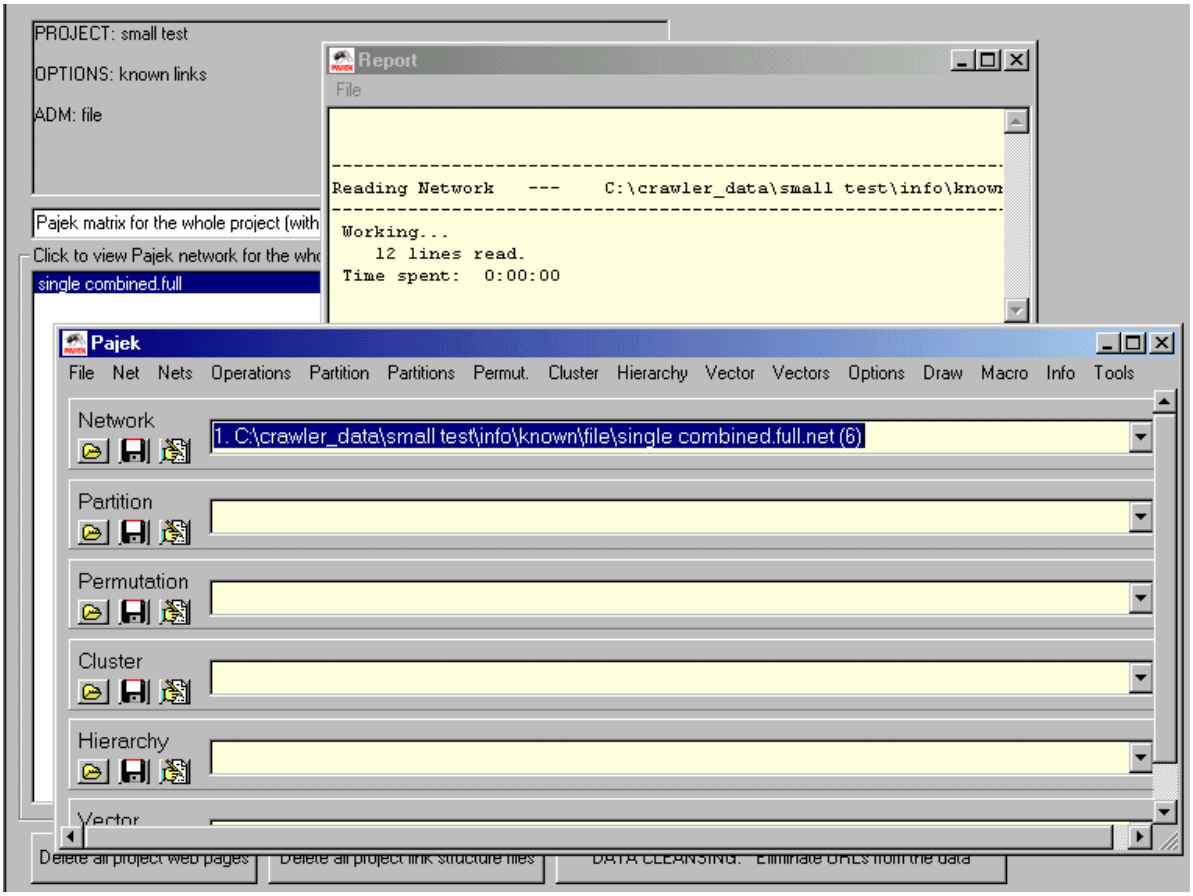
Step 6: Viewing a network diagram with Pajek

사용자가 시스템에 Pajek 을 인스톨했으면, 사용자는 Pajek 이 만들어내는 네트워크 다이아그램을 볼 수 있다.

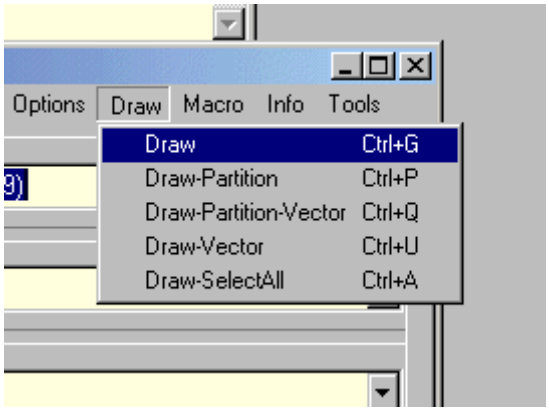
1. *Pajek matrix for the whole project (with current options)* 옵션을 선택하기 위해 화면의 중앙에 박스를 따라 내려가는 것을 사용하라. 만약에 사용자가 이 파일을 계산하기(calculate)를 원하면, *Yes* 를 클릭하라.



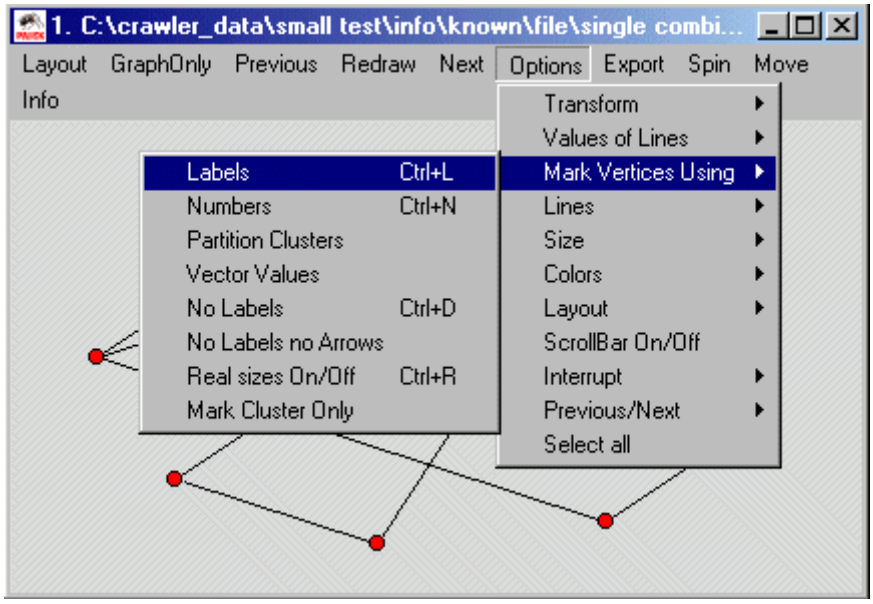
2. Pajek 에서 single.combined.full 을 클릭하라.

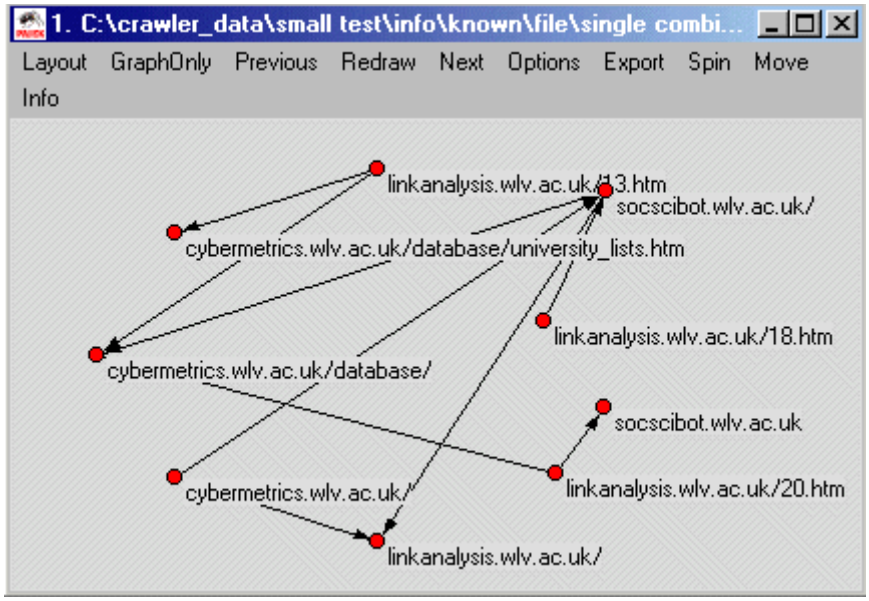


3. 네트워크 데이터는 이제 Pajek 으로 저장된다. 네트워크를 보려면 Pajek 에서 *Draw* 메뉴에서 *Draw* 을 선택하라.

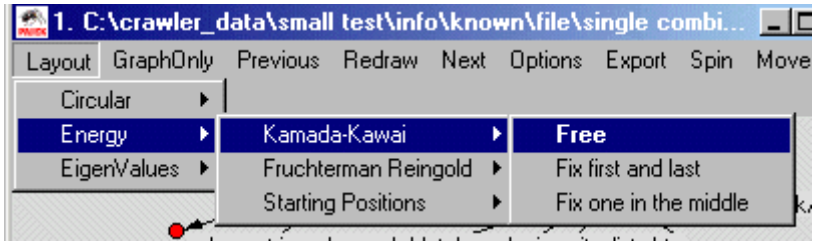


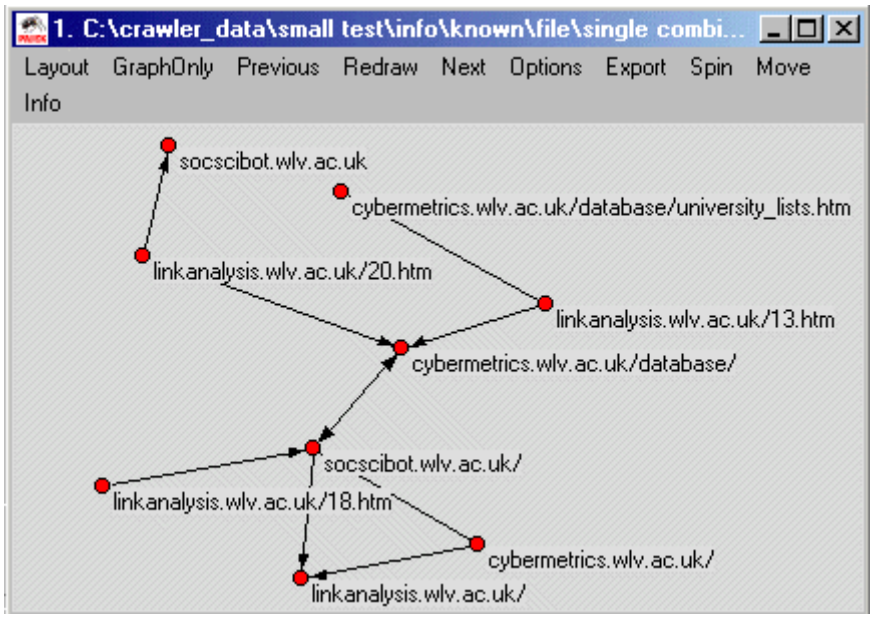
4. 네트워크에 labels 라는 것이 없으면(사이트 도메인 이름), Options 메뉴를 선택하라. 그리고 *Mark Vertices using* 에서 *Labels* 를 선택하라(혹은 Control+L 을 해도 된다). 이것은 internal site 링크를 제외한, inter-site 링크의 네트워크를 보여준다.





5. 네트워크 다이어그램의 개선된 레이아웃을 가지기 위해선, *Layout, Energy, Kamada-kawai, Free* 의 Kamada-Kawai 포지션닝 알고리듬을 선택하고 결과를 관찰하라.

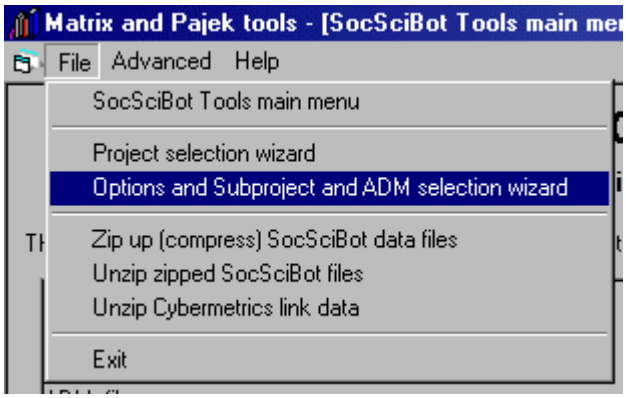


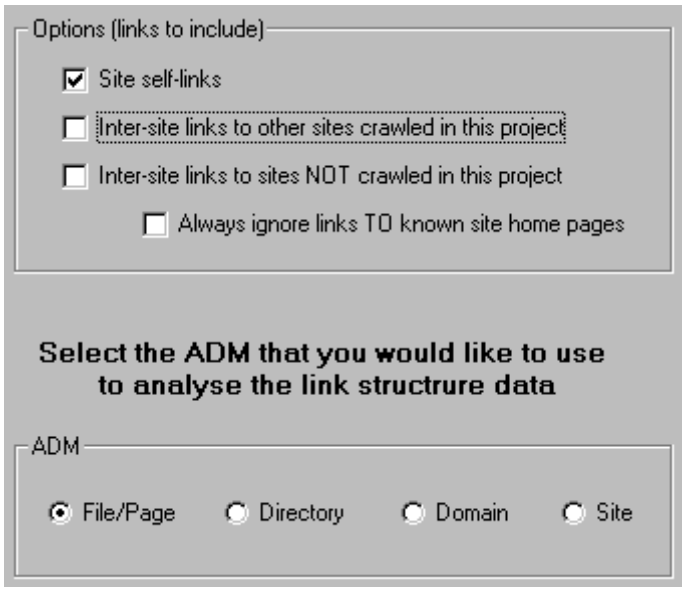


Step 7: Viewing site diagrams with Pajek

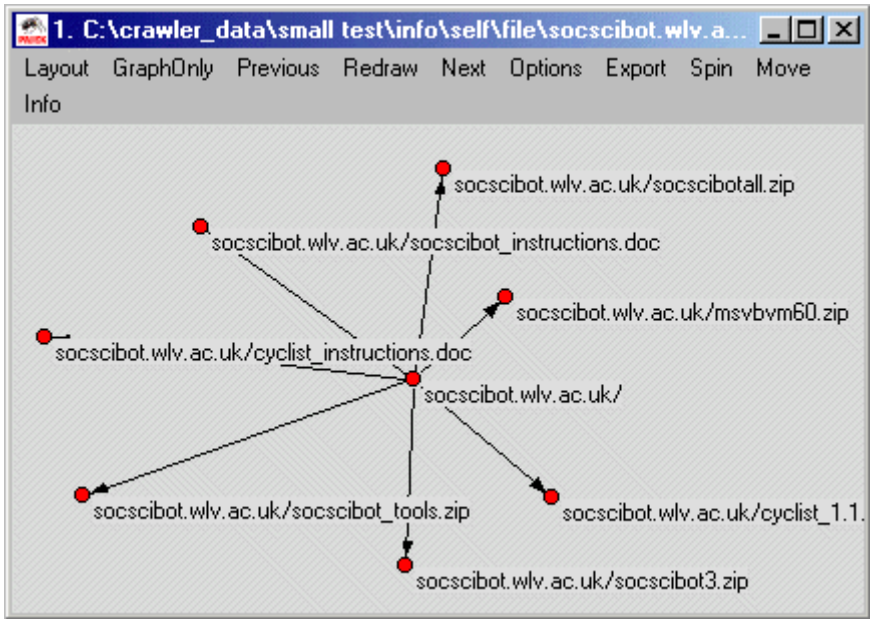
1. 만약에 사용자가 inter-site 연결보다 오히려 각각의 개개 사이트의 다이어그램을 보고 싶다면, 이것 또한 가능하다. 사용자가 메뉴를 따라 내려가서 *Pajek matrices for each individual site (with current options)*를 선택하면, 사용자는 SocSciBot

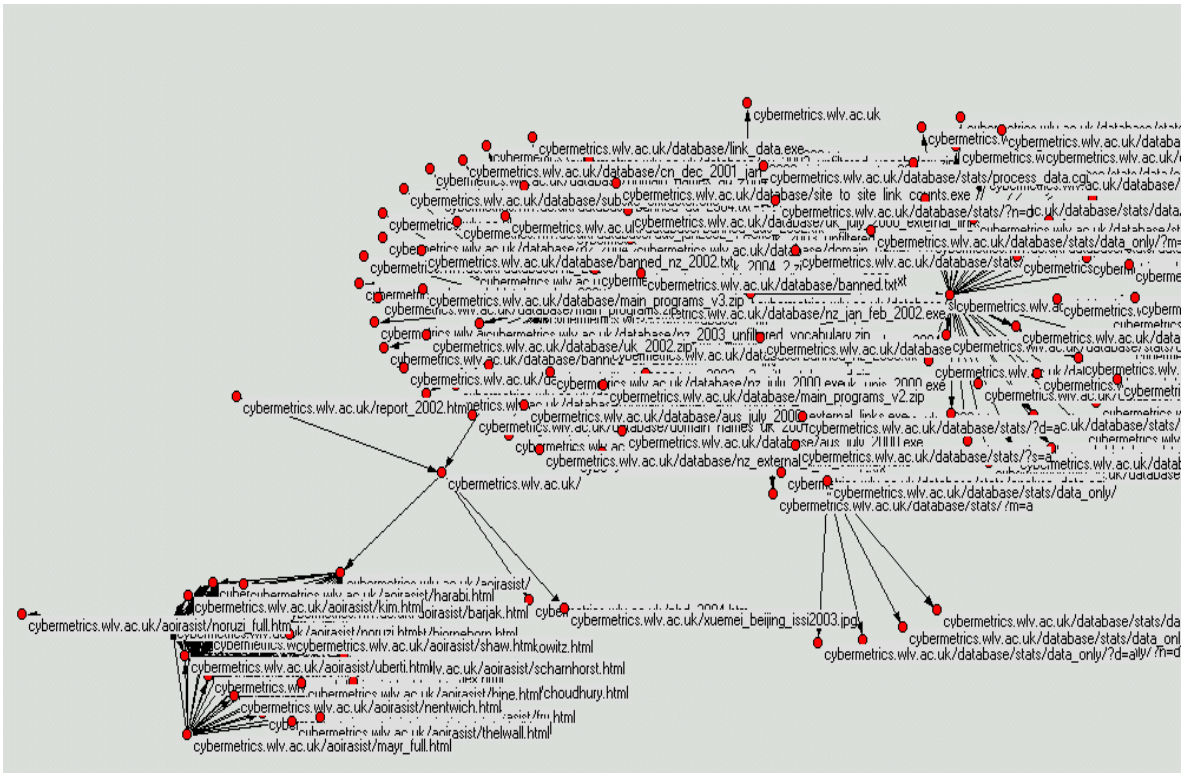
Tools 의 디폴트가 모든 internal site 링크를 무시하기 때문에, 이것을 얻지 못할 것이다. 사용자가 원하는 것이 무엇인지 말할 수 있는가? 사용자는 원하는 internal site 링크와 만약에 사용자가 웹사이트의 internal 구조의 다이어그램을 원한다면 링크의 어떤 다른 타입도 안 된다는 것을 SocSciBot Tools 에게 말할 필요가 있다. SocSciBot Tool 에서 이것을 실행하기 위해선, *File* 메뉴에서 *Options and Subproject and ADM selection wizard* 선택하고, site self-links 옵션을 선택하라.





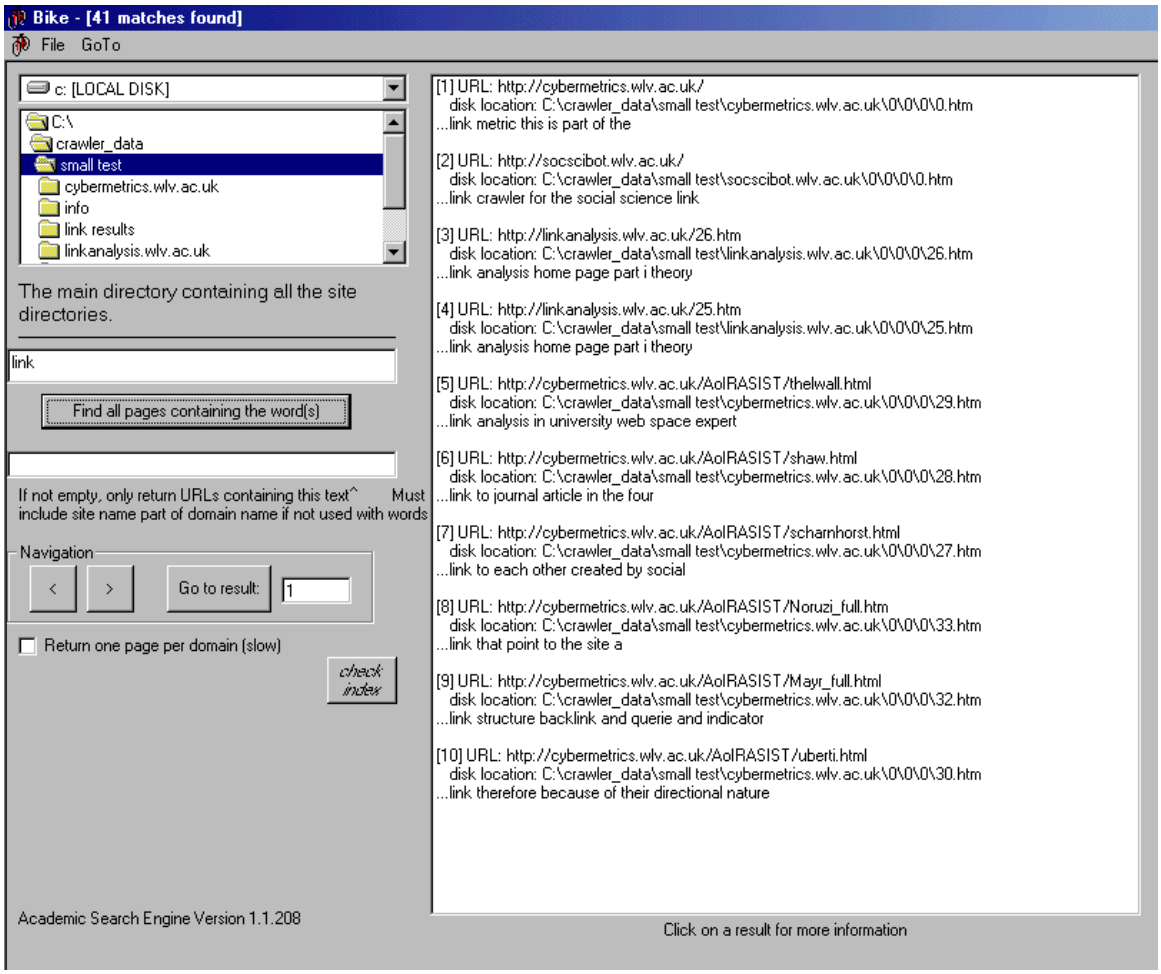
2. 이제 아래 메뉴에서 *Pajek matrices for each individual site (with current options)* 을 선택하고 클릭해서 파일을 보라. 사용자는 Pajek 에서 개개의 사이트 네트워크를 얻을 수 있다. 아래는 Kamada-Kawai 알고리듬으로 다시 그려진 두 개의 네트워크이다. 두 번째 것은 해석하기 어려울 정도로 많은 선들이 있다.





Step 8: Using Cyclist as a Search Engine

1. 사용자의 컴퓨터에 압축을 해제한 폴더에서 Cyclist 나 Cyclist.exe 둘 중 하나의 파일을 불러서 더블 클릭하여 Cyclist 를 실행하라. Cyclist 는 링크 분석 프로그램이 아닌 텍스트 서치 엔진이다.
2. 질문에 응답하고 20 초나 그 이후에 계산이 된 후에 사용자는 표준 서치 엔진 타입 인터페이스를 얻을 것이다. "link" 그리고 나서 사용자에게 주어진 정보를 보려면 오른쪽 사이드에서 결과들을 클릭하는 것처럼 선택할 일상적인 단어를 찾아보라.



End

이 설명서의 단계는 규모에 상관없이 작거나 큰 프로젝트에 동등하게 적용된다. 유일한 차이점은 데이터를 처리하기 위해서 site crawl 과 SocSciBot Tools 그리고 Cyclist 를 위한 상당한 시간이 걸릴 수도 있는 것이 대규모 프로젝트라는 것이다.

실제 프로젝트에서 데이터를 수집할 때 가능하면, 어떤 잘못된 결과가 나타날 수 있는 경우를 대비해서 Cyclist 나 SocSciBot Tools 를 실행하기 전에 사용자의 데이터를 백업하라.

Tutorial 2: Mini Link Analysis Research Project Case Study

Overview

이 설명서 소개는 아주 작은 가상 링크 분석 리서치 프로젝트의 단계에 사용된다. 이 프로젝트는 표준 리서치 프로젝트를 위해 SocSciBot 와 SocSciBot Tools 를 어떻게 사용하는지를 배우기에 가장 쉬운 방법을 사용자에게 제공해 주도록 고안되었다. Tutorial 3 은 특히 만약에 사용자의 리서치 프로젝트가 상당히 다른 목표/정보를 요구하는 것일 때, 사용자만의 프로젝트에서 도움을 줄 리서치 프로젝트에 더 일반적인 충고를 제공할 것이다.

Setting up a new project and crawling the sites

프로젝트는 웹사이트의 작은 수를 crawling 하는 것과 분석을 포함한다. 이것들은 첫 프로젝트(Tutorial 1)에서 crawl 된 웹사이트와 유사할 것이다. 왜냐하면 웹사이트들은, 저자가 알기로는, 사라지지 않을 사이트를 필요로 하기 때문이다(because they need to be sites that I know will not disappear).

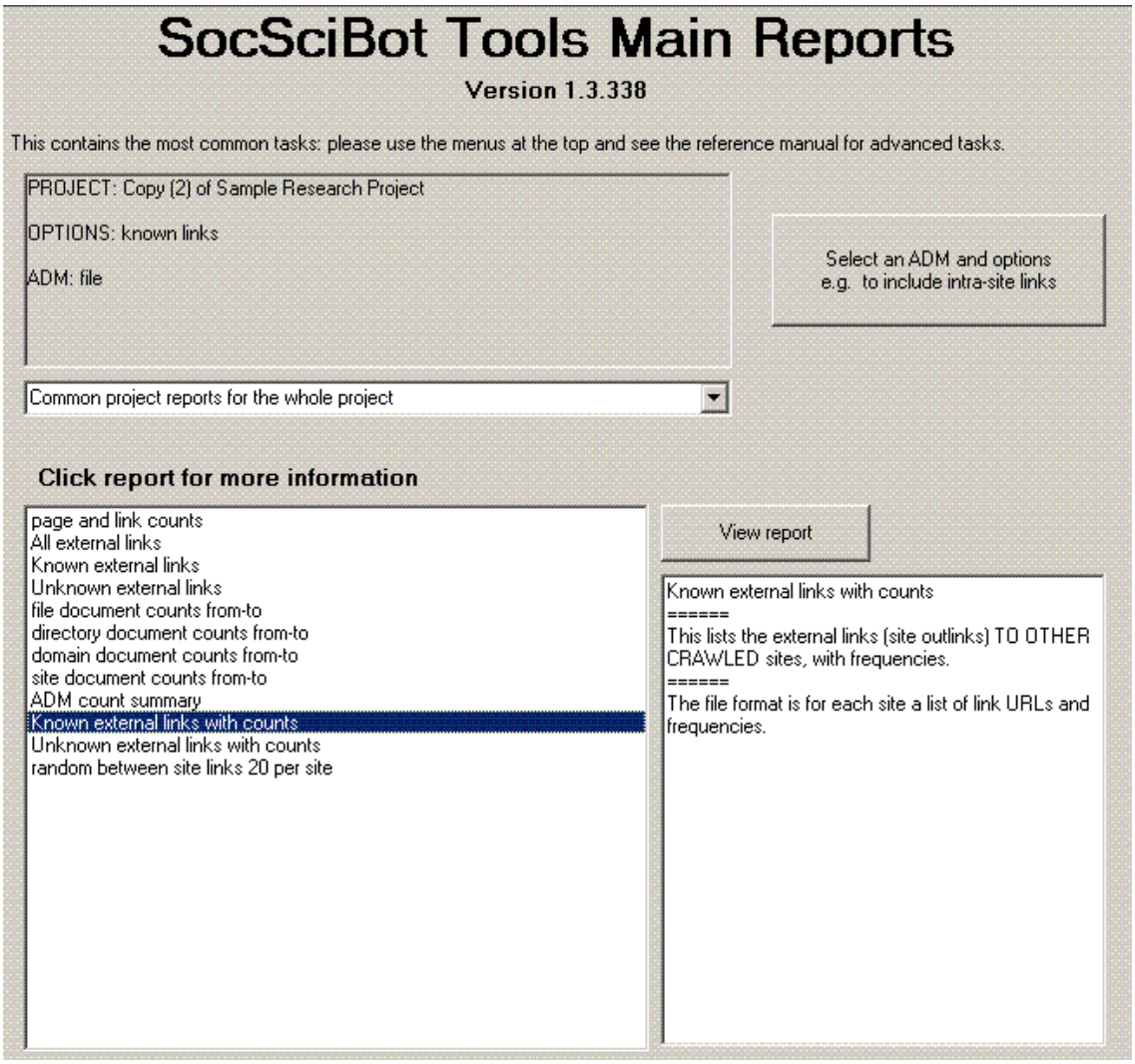
사이트는 다음과 같다. 아래 새로운 프로젝트에서 *Sample Research Project* 와 crawl 사이트를 불러서 새로운 프로젝트를 설치하라. 새로운 프로젝트를 설치하기 위해, 사용자가 SocSciBot 를 시작할 때, 첫 화면에, 존재하는 프로젝트 이름을 클릭하는 대신에, 아래에 박스 안에 이름을 기입하고 *Create New Project* 버튼을 클릭하라.

- http://cba.scit.wlv.ac.uk/
- http://cybermetrics.wlv.ac.uk/
- http://linkanalysis.wlv.ac.uk/
- http://www.scit.wlv.ac.uk/~cm1993/ [note the symbol before cm1993]

Data cleansing

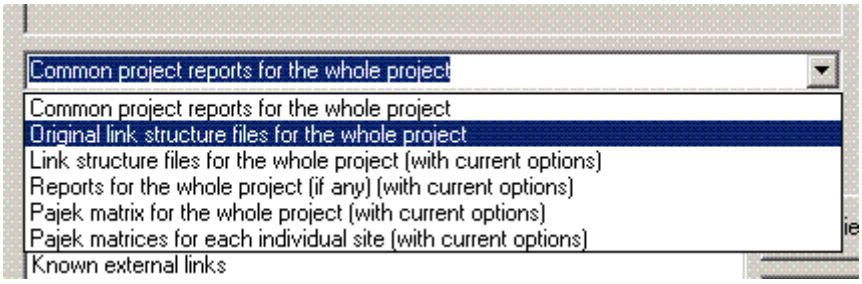
사용자가 위의 모든 사이트들을 crawl 했을 때, SocSciBot Tools 와 새로운 프로젝트, *Sample Research Project* 를 선택하라. 처음에, 데이터 분석의 가장 시간을 많이 소비하는 단계는 데이터 셋에서 변칙들(anomalies)을 확인하고 제거하는 것이다. 이상적으로, 다운로드된 각각의 페이지는 사용자의 리서치 프로젝트를 위한 항목과 적합한 것인지 확인하는 과정을 거칠 필요가 있다. 예를 들면, 많은 리서치 프로젝트들은 사이트에서 모든 페이지들이 사이트의 소유권자에 의해 만들어진 것이고, 다른 사람들의 웹 페이지들의 복사본이 아니어야 한다는 것을 요구한다. 이것은 만약에 많은 링크들과 함께 어떤 사람의 웹사이트의 매우 많은 복사본이 있다면 결과의 커다란 차이를 만들 수 있다. 이것의 예로서, Linux Gazette 웹사이트는 많은 세계 웹사이트 서버들에 많은 컴퓨터 소프트웨어 문서로서 복사되어 있다. 이것들의 연산 복사본들은 전형적으로 그들의 공식적 홈페이지에 링크되어 있다. 그리고 이 링크들은 어떤 링크 분석을 망칠 수 있다. 문제들은 만약 한 출처로부터

많은 링크가 있다면 발생하기 때문에, 복사된 페이지를 찾는 가장 좋은 방법은 가장 확률이 높은 목표 페이지를 확인하면서 시작하는 것이다. SocSciBot Tools 메인 보고서 메뉴에서, 데이터 셋에 가장 확률이 높은 목표 페이지의 두 가지 보고서가 있다: 데이터 셋에서 사이트들 사이에 모든 링크 리스트인 *Known external links with counts* 와 데이터 셋의 외부 사이트 링크의 리스트인 *Unknown external links with counts* 이다. 클릭해서 그리고 view report button 을 클릭해서 이들 두 개를 보라.

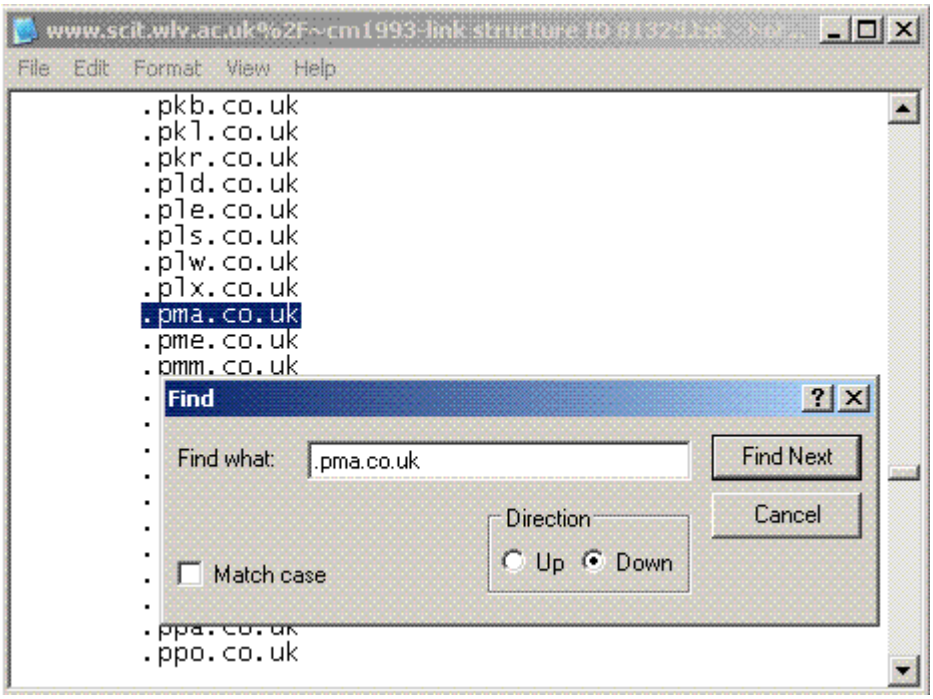


두 개의 리스트에서 사용자는 페이지가 매우 높게 목표된 페이지가 없다는 것을 알게 된다. 그러나 리스트의 하나를 조사해보면 사용자는 .co.uk 사이트들에 링크 되어있는 이상하게 긴 리스트를 보게 될 것이다. 이것은 변칙이 될 수 있는가? 이것을 알아내는 것은 중요하다. 왜냐하면 만약에 분석이 이 링크들을 포함한다면 거기엔 수많은 링크들이 포함되어 있기 때문이다. 결정을 위해, 사용자는 페이지들을 방문할 필요가 있을 것이다. 그리고 왜 이 링크들이 만들어졌는지 알아 볼 필요가 있다. 페이지들을 찾는 것은 시간 낭비이다: 사용자는 그것들을 찾기 위해 수동으로 링크 구조 파일들을 조사해야만 할 것이다.

.pma.co.uk 라 불리는 이상한 링크 URL 들 중 하나를 선택하라. 링크 구조 파일을 보기 위해, 화면 중앙에 아래로 내려가는 메뉴에서 전체 프로젝트를 위한 원본 링크 구조 파일을 선택하라.



차례로 각각의 파일을 클릭하라. 노트패드 윈도우는 사이트의 링크 구조를 포함해서 보여줄 것이다. 파일은 각각의 페이지의 URL 을 포함한다(http://나 http://www 없이). 그리고 각각의 링크 URL(출처가 있는 페이지의 URL)을 포함한다. 사용자는 이 파일들 중에서 하나로 URL .pma.co.uk 를 찾아야 한다(팁: URL 서치를 위한 노트패드의 Find 장치(Edit | Find)를 사용하라.



.pma.co.uk 에서 스크롤을 내리면, 사용자는 페이지 .scit.wlv.ac.uk/~cm1993/penn/p.htm 에서 유래한 URL 을 찾을 것 이다(i.e. http://www.scit.wlv.ac.uk/~cm1993/penn/p.htm). 만약에 사용자가 이 페이지를 방문하면, 사용자는 Penn UK Business Directory 로 알려진 아주 큰 사이트의 부분을 보게 될 것이다. 이 사이트는 단순히 UK .co.uk 웹 사이트의 거대한 리스트가 있다(구식 리서치 프로젝트의 부분). 사용자는 페이지들의 이 세트가 사용자의 데이터 세트에 포함되지 않아야 한다는 것을 결정하도록 가정해야 한다. 왜냐하면, 다른 위치에서 사이트의 복사본이 있기 때문이다. 전체 미니 사이트가 SocSciBot Tools 의 데이터 정화 피처를 사용되지 않을 수 있다: 아래와 같이 금지된 항목.

화면의 아래에서 DATA CLEANSING 버튼을 클릭하라. 새 파일이 나타날 것이다. 이것을 위해 우리는 아래에 다음 두 줄을 추가해야만 한다.

[scit.wlv.ac.uk]

http://www.scit.wlv.ac.uk/~cm1993/penn

사각의 괄호에서 첫 번째 줄은 단지 도메인 네임만 등록된 노트에서 제외될 페이지들의 사이트 도메인 네임을 확인한다. 그리고 처음 www.은 잘려져야만 한다. 두 번째 줄은 http://www.scit.wlv.ac.uk/~cm1993/penn 과 함께 시작하는 모든 URL 을 제외하기 위한 설명서이다. 그래서 전체 Penn UK 사이트는 제거 될 것이다.

이제 파일을 저장하고 Yes 버튼을 클릭하라. Penn UK 페이지들은 데이터 세트에서 제거 될 것이다. 그리고 기초 통계들은 재계산될 것이다. 페이지들이 정말로 링크 구조를 보여주면서 진행되고 카운트 파일과 함께 알려지지 않은 외부 링크에 의해 진행되어 왔는지 체크하라.

Interpreting the link counts

학술적 커뮤니케이션을 나타내는 것과 같은 링크 카운트에 대해 결론을 맺기 위해서, 사용자는 사용자의 가설을 명확히 하기 위한 몇몇의 단계를 수행할 필요가 있다. Link Analysis: An Information Science Approach 책을 보거나 이 복합 이슈의 토론을 위한 저널 기사인 Interpreting social science link analysis research: A theoretical framework 을 보라. 그러나 발생될 수 있는 한 표준 단계는 일부 일반적인 추정들이 링크 카운트의 적당한 해석에 대해 만들어 질 수 있기 위해서 링크의 랜덤 샘플을 분류하는 것이다. SocSciBot Tools 는 이 목적으로 정확하게 링크의 랜덤 샘플을 제공한다. 파일: 리포트에서 사이트 당 랜덤링크 20 개 그리고/혹은 선택박스 아래로 선택된 전체 프로젝트 보고서(현재 옵션과 함께) 섹션을 위한 랜덤 링크.

이 파일에는 각각의 사이트 내에서 랜덤으로 배열되어 crawl 된 사이트 당 20 개의 링크가 있다. 선행 조사에서와 같이, 우리는 모두 40 개 링크를 방문하기를 원할 것이다(i.e. 링크 원본 페이지와 링크 타겟 페이지). 그리고 일련의 정당한 분류 항목에 따라서 링크 창조를 위한 이유를 분류해야 할 것이다. 예를 들면, 매우 간단한 주제들은 (a) 학술적 목적과 관련될 수 있고, (b) 학술적 활동과는 관련이 없을 것이다. 그러나 이런 간단한 예와는 대조적으로 사용자는 각각의 카테고리를 위한 세부 항목을 부여할 필요가 있다. 4 개의 사이트와 우리는 40 링크로 정했기 때문에, 우리는 사이트 당 랜덤 링크 20 개 파일에서 처음 10 개 랜덤 링크를 얻는다. 그리고 그것들을 분류한다.

Obtaining the important link information for your research question

이것은 쉬운 부분이다: 리포트 영역에서 가능한 다른 리포트를 추출해서 탐구할 수 있는 리서치 의문의 종류에 대해서 생각해 보라.

Tutorial 3: Summary of how to use SocSciBot for a link analysis research project

Overview

이 설명서는 링크 데이터의 마지막 분석에 처음 crawl 에서 링크 분석 리서치 프로젝트의 주요 단계들을 요약한다. 이 설명서는 또한 사용자의 리서치 프로젝트를 위한 SocSciBot 과 SocSciBot Tools 를 사용하기 위해 사용자에게 가능한 주요 정보를 제공하도록 고안되었다. 이 부분을 시작하기 전에 설명서 1 과 2 를 훑어보라.

Setting up project goals

첫 단계는 crawl 할 웹사이트를 결정하고 이들 사이트들의 홈페이지의 URL 을 수집하는 것이다. 이 단계는 사용자의 리서치 의문에 전체적으로 의존한다. 그러나 사용자는 사용자가 crawl 하기 위한 실제적인 것이 있는지 없는지 알기 위해서 crawl 하고자 하는 각각의 사이트에 얼마나 많은 페이지들이 있는지 측정하는데 search engine advanced searches 을 사용하기를 바랄 것이다. 각각의 crawl 은 만약 사이트가 100 페이지이상을 포함한다면 잠시 시간이 걸릴 것이다. 그래서 사용자의 crawl 을 완료하기 위한 충분한 예상 시간을 세워두길 바란다.

전체 크기의 연구 전에 작은 규모의 선행 연구를 시행해보는 것은 좋은 생각이다. 이 선행연구의 목적은 전체 연구가 사용자에게 사용자의 리서치를 위해 필요로 하는 정보를 어떻게 제공할 가능성이 있는지 접근하기 위한 것이다.

Setting up a new SocSciBot project and crawling the sites

초기 SocSciBot 시작 화면에서 사용자의 리서치에 적당한 이름을 만들어서 새 프로젝트를 시작하라. 현존하는 프로젝트에 crawl 을 추가하지 말라. 왜냐하면 이것은 링크 분석 결과를 난잡하게 만들 것이기 때문이다. 새 프로젝트를 시작하기 위해, SocSciBot 을 시작할 때, 첫 번째 화면에서 현존는 프로젝트 이름을 클릭하는 대신에 아래 박스에 이름을 기입하고 *Create New Project* 이름을 클릭하라. 일단 사용자가 새 프로젝트를 선택했다면, 사용자는 두 번째 SocSciBot 화면에서 홈페이지의 URL 을 기입하면서 사이트를 crawl 시작할 준비가 되었다. 사용자의 데이터 세트에서 각각의 새 사이트를 위해 사용자는 이전 crawl 의 끝에서 SocSciBot 을 닫을 필요가 있다. 그런 다음 다시 시작해야 한다. 만약 사용자가 사이트를 crawl 하는데 어렵이 있다면 Tutorial 1 에서 지시사항을 참고하라.

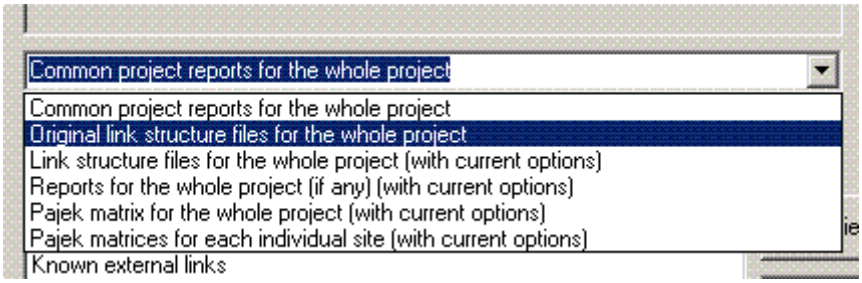
Data cleansing

"처음 그리고 데이터 분석의 가장 많은 시간을 소비하는 것은 데이터 세트에서 변칙을 확인하고 제거하라. 이상적으로 각각의 다운된 페이지들은 사용자의 리서치 프로젝트를 위한 항목에 맞는지 확인할 필요가 있다"는 것을 상기하라(Tutorial 2). 페이지의 큰 규모의 수집을 위한 추천 방법은 다음과 같다.

1. 페이지의 포함 혹은 제외되는 일련의 항목을 만들어라. 그리고
2. 원하지 않는 원본에서 유래한 것이 있는지 없는지 확인하기 위해 가장 높이 타겟된 페이지의 링크를 조사하라.

이것은 모든 원하지 않은 페이지들을 제거할 것이다. 그러나 링크의 가장 영향력이 있는 원천들을 제거 할 것이다. SocSciBot 메인 리포트 메뉴의 데이터 세트에서 가장 높이 타겟된 페이지들은 표준 리포트에 있다: *Known external links with counts*, (데이터 세트에서 사이트들 사이에 모든 링크), 그리고 *Unknown external links with counts*, (데이터 세트 밖의 사이트의 링크). 가장 일반 링크 목표를 확인하기 위해 이것들을 사용하라.

링크 타겟들을 위한 일부 원본 페이지를 찾는 것은 더 어려운 것이다. 사용자는 화면의 중앙에서 메뉴를 내려 보면 *Original link structure files for the whole project* 를 선택하면 찾을 수 있는 적당한 advanced search engine 링크 서치를 만드는 것을 시도할 수 있다. 혹은 프로젝트를 위한 각각의 링크 구조 파일에서 링크 타겟들을 위한 서치를 할 수 있다. 사용자는 URL 을 위한 서치를 위해 *Notepads Find* 장치(Edit | Find)를 사용할 수 있다는 것을 기억하라.



Collections of pages can be excluded using SocSciBot Tools's data cleansing feature: the Banned List. Before using this feature, which may take a while for SocSciBot Tools to process, follow two steps:

페이지의 수집은 SocSciBot Tools 의 데이터 정화 작업을 사용함으로써 지워질 수 있다: 금지된 리스트. 진행을 위한 SocSciBot Tools 에서 시간이 걸릴 수 있는 이 장치를 사용하기 전에 다음 두 단계를 시행하라.

1. 잘못될 수 있는 것을 대비해서 사용자의 데이터 세트의 백업 카피 본을 만들어라(e.g. zip 파일에서 프로젝트 폴더를 압축하라).
2. 금지된 리스트 피처를 시작하기 전에 제외하기를 원하는 모든 원본 URL 의 리스트를 수집하라. 이것은 사용자만이 금지된 리스트 피쳐를 사용해야만 한다는 것을 의미한다.

화면 아래에 DATA CLEANSING 버튼을 클릭하라. 새 파일이 나타날 것이다. 사용자는 사용자의 crawl 에서 페이지들이 제외 되어야만 하는 것을 SocSciBot Tools 에게 전달하는 이 파일의 끝에 다음 라인을 추가해야만 한다. 아래는 설명되는 예이다.

[wlv.ac.uk]
http://www.scit.wlv.ac.uk/~cm1993/penn/
http://www.wlv.ac.uk/linuxgazette/

사각의 괄호 안에 첫 번째 라인은 페이지가 단지 도메인 네임을 기입되어야 하는 노트에서 제외될 사이트의 도메인 네임을 확인한다. 그리고 어떤 초기 www. 은 잘려진다. 다음 사각의 괄호까지 이것 아래의 모든 라인들은 도메인 네임에서 (wlv.ac.uk)를 포함하는 사이트의 crawl 에서 텍스트(i.e. 위의 예에서 http://www.scit.wlv.ac.uk/~cm1993/penn/ 혹은 http://www.wlv.ac.uk/linuxgazette/)와 함께 모든 페이지 시작을 제외하는 지시로서 SocSciBot Tools 에 의해 해석된다. 위의 예에서, 모든 penn 과 linuxgazette 디렉토리 그리고 이것들의 어떤 서브디렉토리는 제거될 것이다.

Known external links with counts 파일과 *Unknown external links with counts* 파일, 페이지들이 링크 구조 파일을 보여주는 것이 잘 진행되는지 확인하라.

Interpreting the link counts

"학술적인 커뮤니케이션을 나타내는 것과 같은 링크 카운트에 대한 추론을 만들기 위해서, 사용자는 사용자의 가설을 확인할 몇몇의 단계를 거칠 필요가 있다는 것을 상기하라(Tutorial 2). Link Analysis: An Information Science Approach 책을 보거나 복합 이슈의 토론을 위한 저널 기사인 Interpreting social science link analysis research: A theoretical framework 을 보라. 그러나 일어날 수 있는 한 표준단계는 일부 일반적 추론들이 링크 카운트의 적당한 해석에 대해 만들어질 수 있는 링크의 랜덤 샘플을 분류하는 것이다. SocSciBot Tools 는 정확하게 이 목적을 위해 링크의 랜덤 샘플을 제공한다. 파일은: 선택 박스 아래로 선택되는 *Reports and/or Random links for the whole project reports (with current options)* 섹션에서 *random links 20 per site* 이다."

사용자는 링크를 위해 적당한 분류 계획을 만들 필요가 있다. 그리고 링크의 샘플을 분류해야 한다. 예를 들면, 사용자는 선행 연구를 위해 40 개를 분류하기를 원할 수 있고 전체 연구를 위해선 160 개를 원할 수 있다. 사용자는 각각의 사용자 카테고리에서 세부 항목을 부여할 필요가 있다. 추천되는 접근은 링크가 원래 계획과 조화되지 않는다는 것을 확인 했을 때, 몇 개의 링크를 방문하고 필요하면 분류계획을 확장할 수 있는 직관에 의한 분류 계획으로 시작하는 것이다.

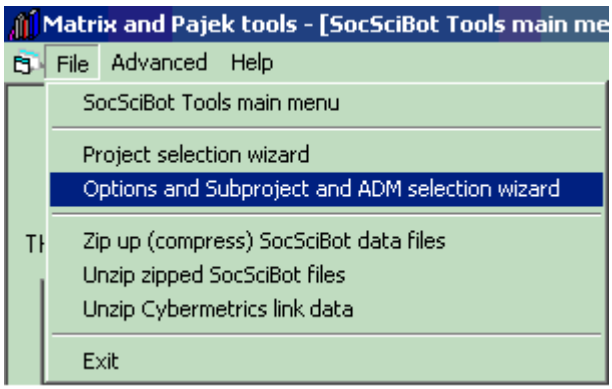
Obtaining the important link information for your research question

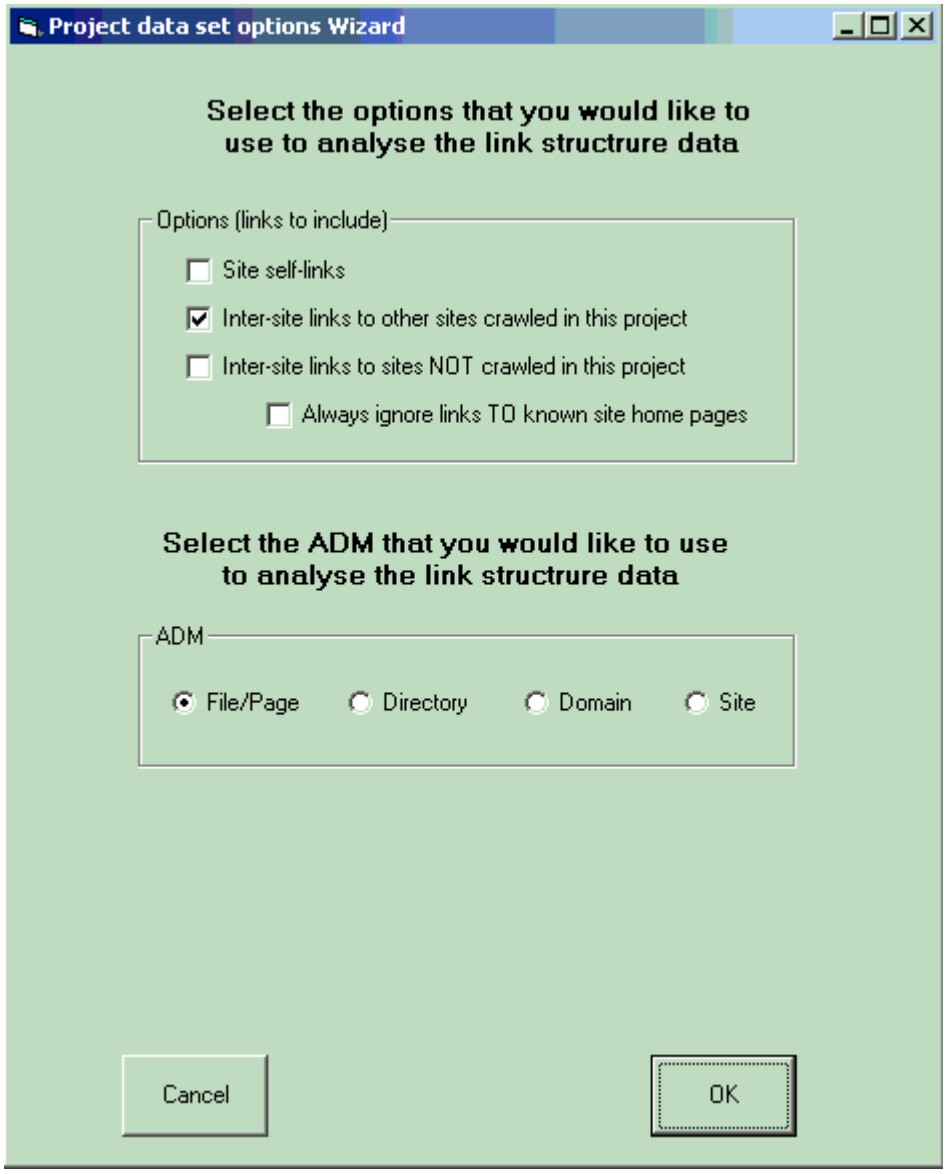
사용자가 원하는 링크 데이터를 획득하기 위해 치러질 단계들은 주로 사용자가 관심 있어하는 링크의 유형에 달려있다. 만약에 사용자가 사용자가 crawl 해왔던 사이트들

사이의 링크에 관심이 있다면, 사용자는 SocSciBot Tools 에 의해 자동적으로 제공된 리포트를 사용 할 수 있다. 만약에 사용자가 링크의 다른 타입에 관심이 있다면 사용자는 새로운 리포트를 요구하기 전에 사용자가 관심 있어 하는 형태를 등록할 필요가 있을 것이다. 링크의 일부 형태들은 SocSciBot Tools 옵션에서 자동적으로 조달될 것이다. 다음 리스트는 상황을 요약한다.

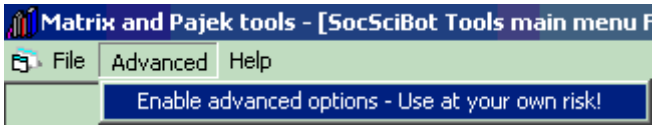
1. 사이트들 사이에 링크를 crawl 하라. 이것은 SocSciBot Tools 를 위한 디폴트 세팅이다: 제공된 리포트를 사용하라.
2. 다음에 나오는 한 개 혹은 그 이상의 데이터 세트에서 모든 링크는 subproject selection wizard 를 사용하면서 선택될 수 있다.
    1. 각각의 사이트 내에서 링크들은 crawl 되었다. (i.e. 사이트 self-links)
    2. crawl 된 세트의 외부 모든 사이트들의 링크
    3. 홈페이지에서 링크들은 없다.
    4. 사이트들 사이에 링크 혹은 디렉토리 사이에 링크들 혹은 도메인과 같은 특별 카운트 방법(대체 문서 모델:ADM)을 사용하면서 카운트된 링크
3. 특별 도메인 링크들(e.g. .edu 도메인의 모든 링크 나 .co.uk 도메인의 모든 링크)은 SocSciBot Tools 의 확장된 피처를 사용함으로써 선택 될 수 있다. 그러나 이것들에서의 리포터들은 자동적으로 계산되어지지 않을 것이다: 사용자는 SocSciBot 의 확장된 피처를 배워야만 할 것이다. 대체적으로, 만약 너무 많은 페이지들이 crawl 되지 않았다면 사용자는 링크의 형태를 제거하거나 사용자가 관심이 없는 페이지를 제거하기 위해 직접 수작업으로 링크 구조 파일들을 편집할 수 있을 것이다.

위의 2 를 위해 보라. subproject selection wizard 는 아래에서 보여주는 것과 같이 file 메뉴에서 찾을 수 있다.





3 에서 확장된 옵션은 Advanced 메뉴를 사용해서 선택된다. 이것은 문서화 되지 않은 많은 새롭고 강력한 메뉴 선택을 제공 할 것이다. 이것들을 사용하는 것은 어려울 것이다. 그래서 만약 절대적으로 필요할 때만 이것을 사용하기를 바란다.



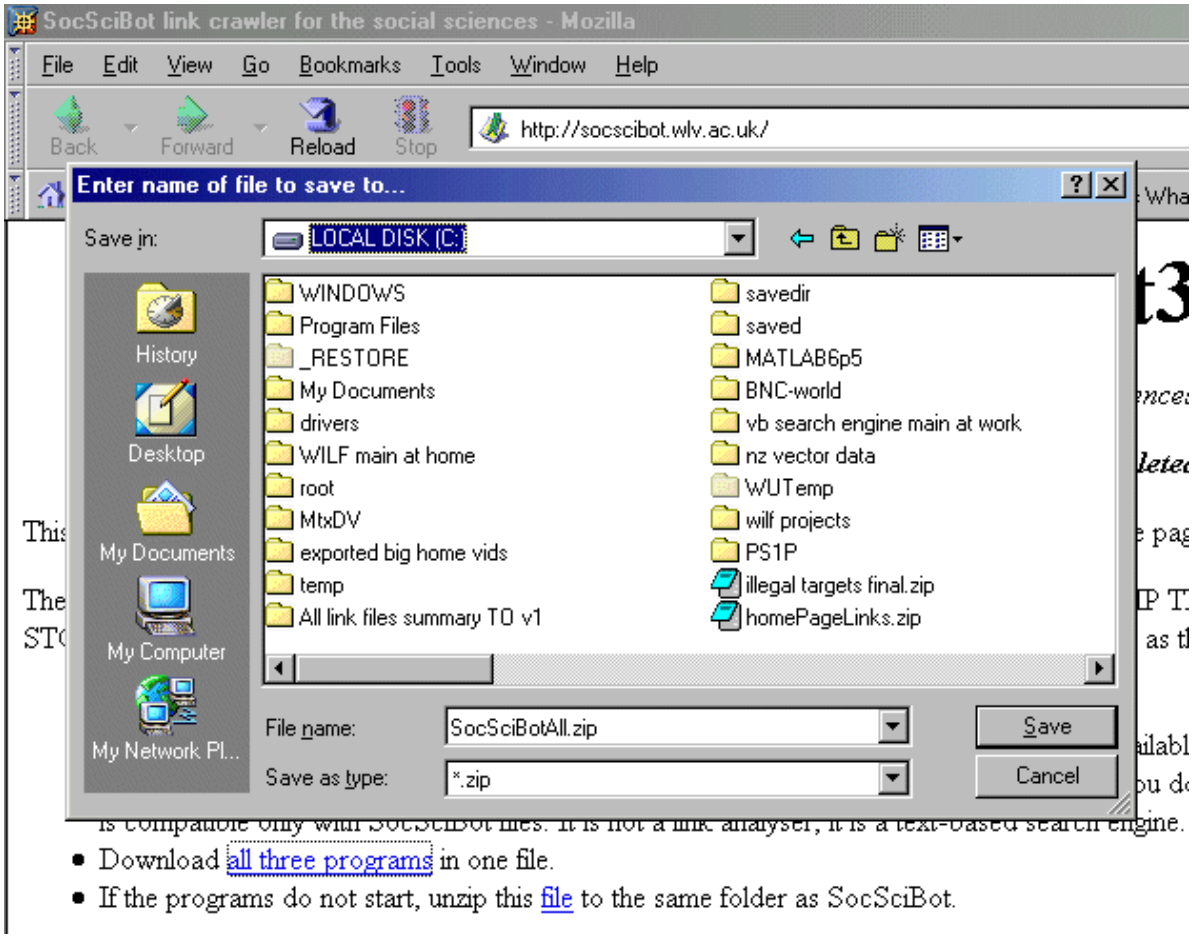
Corpus Linguistics Tutorial: Using SocSciBot and Cyclist for Text Analysis/Basic Corpus Linguistics

Overview

Cyclist 는 SocSciBot 에 의해 다운된 웹사이트를 위한 단어 빈도 통계와 컨코던스/서치 엔진 인터페이스를 제공하는 프로그램이다. 이 설명서는 일련의 웹사이트에서 단어 빈도 어휘를 얻기 위한 필요한 단계들을 훑어본다. 처음 몇 단계들은 Tutorial 1 과 유사하다. 그래서 만약에 사용자가 이것을 완수했다면, 사용자는 Step 4 로 바로 가서 볼 수 있다.

Step 1: Installing SocSciBot, SocSciBot Tools and Cyclist

1. SocSciBot 웹 사이트 http://socscibot.wlv.ac.uk/ 에 가서 사용자가 사용의 조건에 동의 할 때만 프로그램을 다운로드하기 위한 링크에 따르라. 사용자 컴퓨터에 의해 진행될 때, 데이터를 저장 공간의 충분한 양을 가지고 있는 곳에 프로그램을 저장하도록 지정하라. 이것은 보통 사용자의 컴퓨터 하드 드라이브가 될 것이다. 예, C: 드라이브.



2. 다음에, 사용자가 프로그램을 저장하도록 지정한 장소에 SocSciBotAll.zip 파일의 압축을 풀어라. 이것은 SocSciBot, SocSciBot Tools, 그리고 Cyclist 프로그램을 포함한 여러 가지 새로운 프로그램들을 만들 것이다.

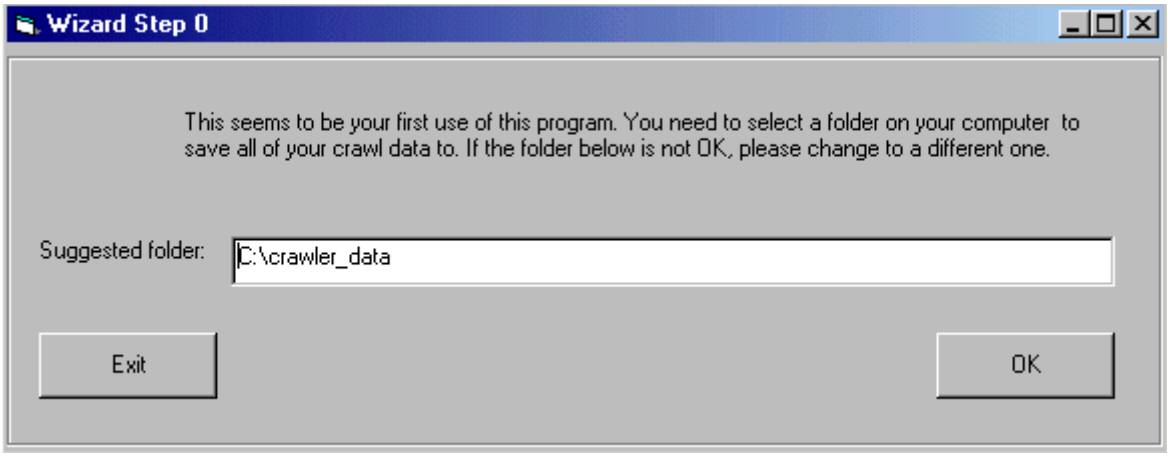
Step 2: Installing Pajek

만약 사용자가 SocSciBot 데이터와 함께 네트워크 다이어그램을 제공받기를 원한다면, 사용자는 Pajek 을 인스톨하도록 추천한다. 사용자는 처음에 SocSciBot 을 시작하기 전에 이것을 시행하도록 요구된다. 왜냐하면 SocSciBot 은 프로그램이 시작될 때, Pajek 을 찾는다. 그리고 Pajek 을 SocSciBot 을 먼저 실행한 후에 인스톨한다면 Pajek 을 찾지 못할 것이다.

1. Pajek 홈 페이지 http://vlado.fmf.uni-lj.si/pub/networks/pajek/에 가서 Pajek 의 최신 버전을 다운로드 해서 인스톨하라.

Step 3: Crawling a first site with SocSciBot

1. 사용자의 컴퓨터에 압축을 푼 폴더에서 SocSciBot 이나 SocSciBot.exe 둘 중 하나를 불러서 파일에서 더블 클릭하면 SocSciBot 이 작동한다. 이것은 아래의 것과 유사한 다이아로그 박스를 제공해야만 한다.

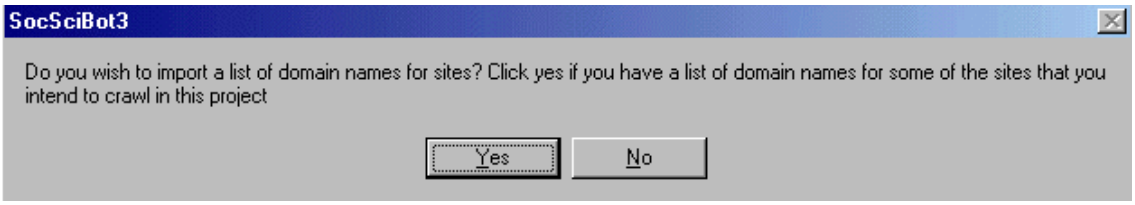


2. 데이터를 저장하기 위해 SocSciBot 에 의해 선택된 폴더를 확인하는 것은 OK 를 클릭하면 받아들여진다. 이것은 사용자가 crawl 하는 어떤 사이트의 웹 마스터들에게 email 을 보내도록 사용될 것이다. 이것은 윤리적인 수행임인 동시에 만약에 웹 마스터가 사용자가 그들의 사이트를 crawling 하는 것에 대해 불만족을 나타내는지 않은지에 대한 문제를 확인하는데 사용자에게 도움을 준다. 웹 마스터들은 사용자의 보스나 네트워크 매니저에게 email 을 하는 대신에 직접적으로 당신에게 email 을 보낼 수 있을 것이다. 사용자는 또한 crawl 의 목적을 설명하는 email 을 포함하는 메시지를 등록할 수 있다. 사용자는 프로젝트에 대해 부가적인 정보에 관련된 페이지의 URL 을 포함하기를 원할 수 있다. 또한, Microsoft Excel 과 Pajek 의 위치에 대한 어떤 질문들에 대해 응답하기를 원할 수 있다.

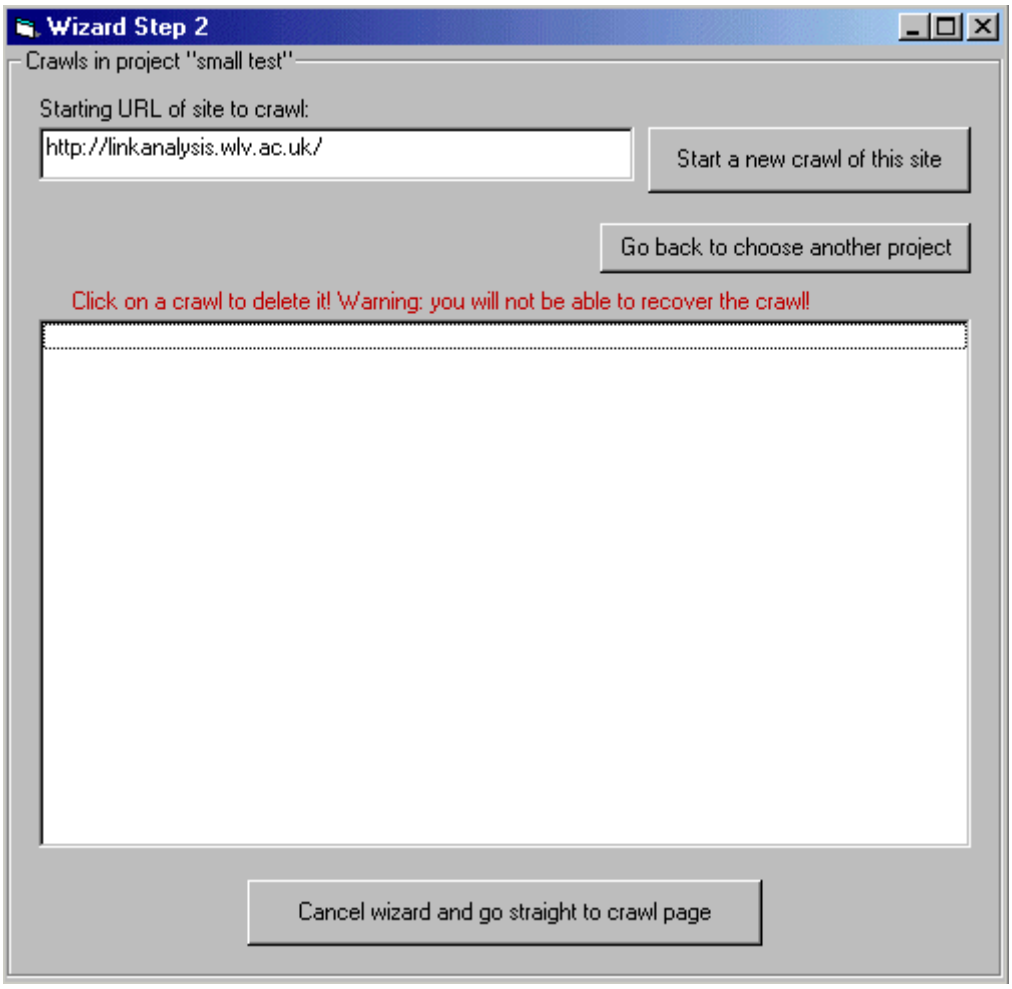
3. 다음 다이아로그 박스, Wizard Step 1 의 아래에 프로젝트의 이름으로 *small test* 를 입력하라. 그 다음에 *start new project* 버튼을 클릭하라. 모든 crawl 들은 프로젝트와 함께 그룹화 된다. 이것은 사용자가 개별적으로 분석된 crawl 그룹의 다른 이름을 가지도록 허용한다.

4. 다음에 질문에 *No* 를 클릭하라. 이것은 사용자가 전문가 수준이 되기 전에는 거의 필요로 하지 않는 확장된 데이터 클리닝(cleaning) 장치이다.

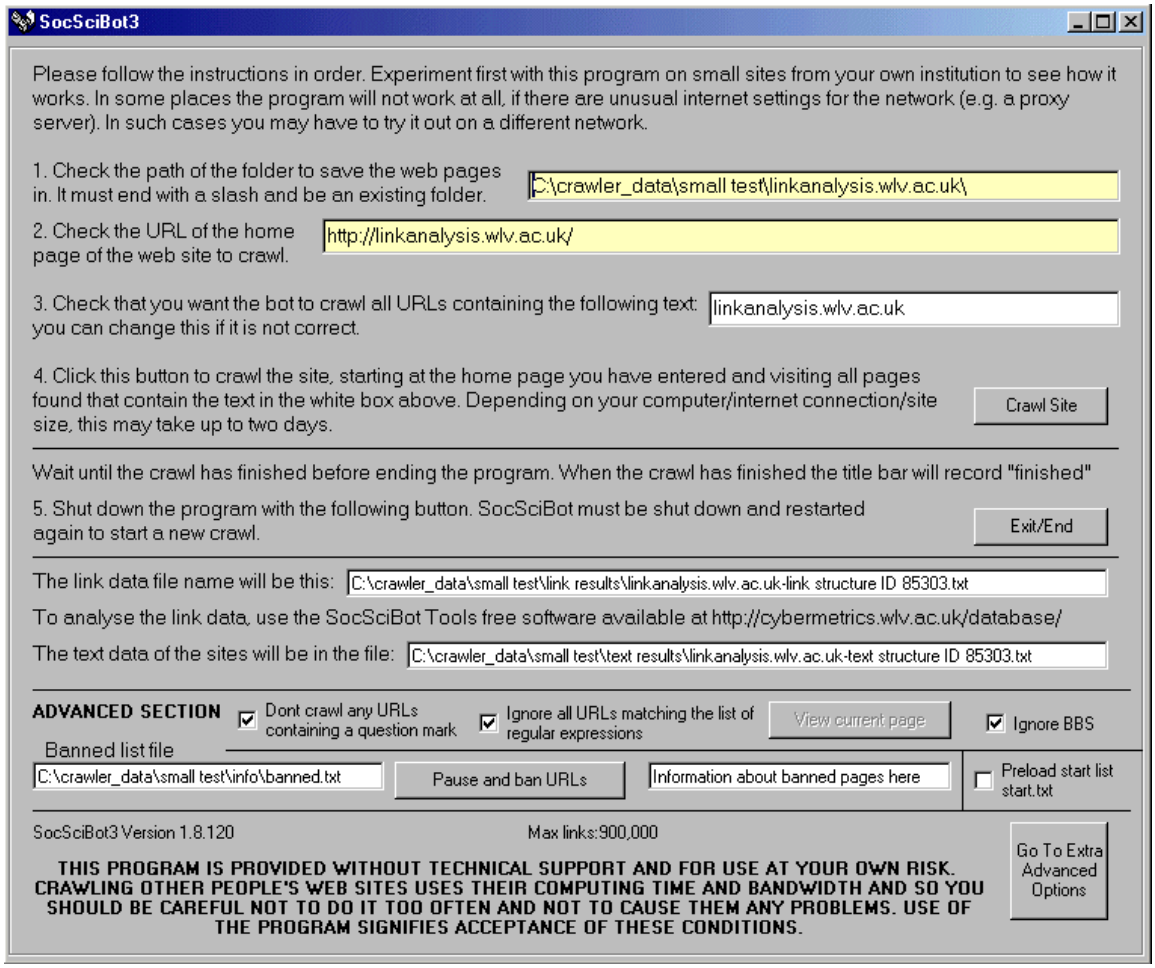
5. 다음에 두 번째 질문에 *No* 를 클릭하라. 이것은 사용자가 전문가 수준이 되기 전에는 거의 필요로 하지 않는 또 다른 확장된 데이터 클리닝(cleaning) 장치이다.



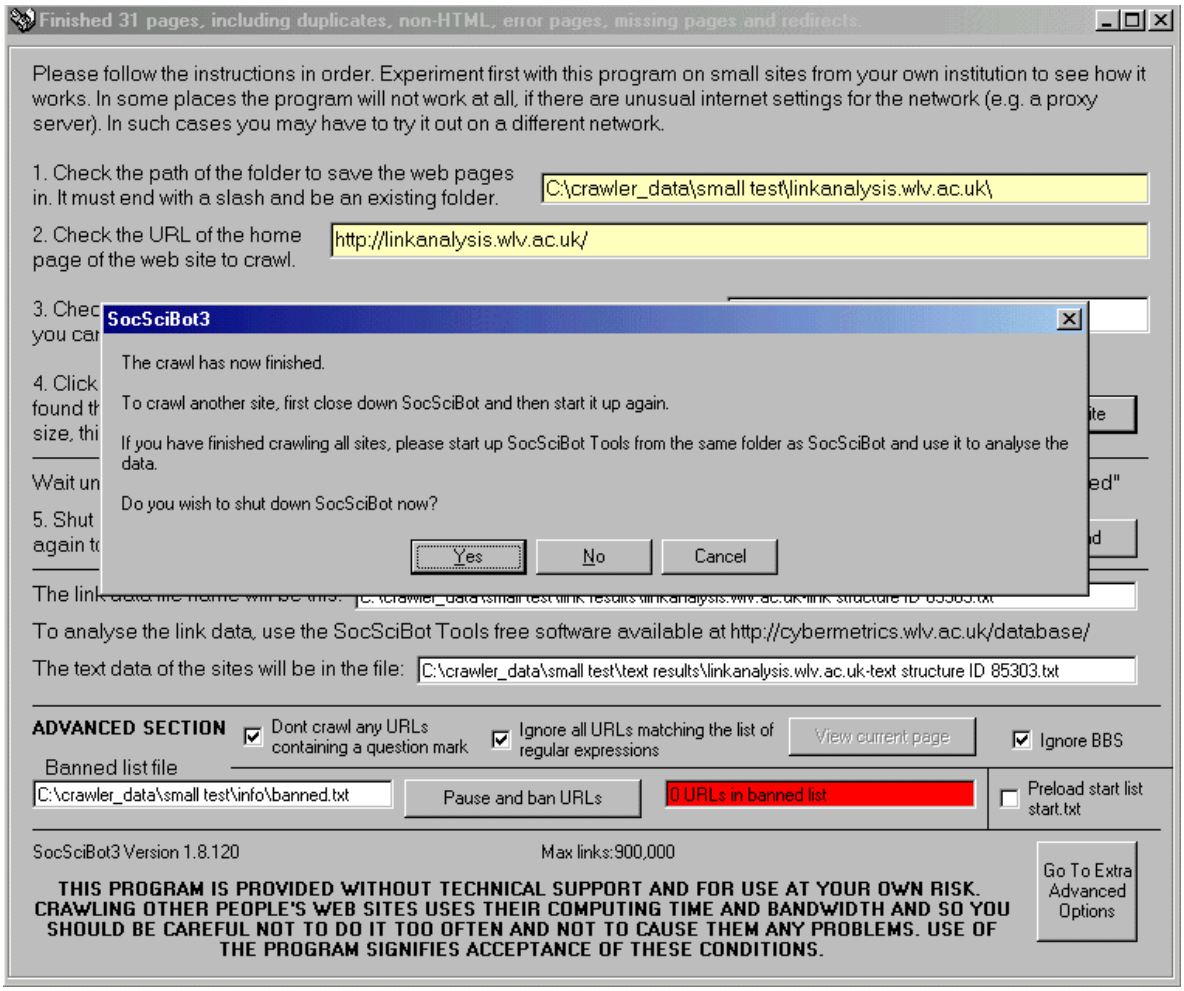
6. wizard step 2 다이아로그 박스에서, crawl 을 위해 사이트의 URL 을 시작하는 곳에 http://linkanalysis.wlv.ac.uk/를 입력하라. 그리고 *Start a new crawl of this site* 를 클릭하라.



7. Crawl 은 작동할 준비가 된다. *Crawl Site* 버튼을 클릭하라. 30 분이나 그 이상 지난 후에 crawl 은 끝난다.



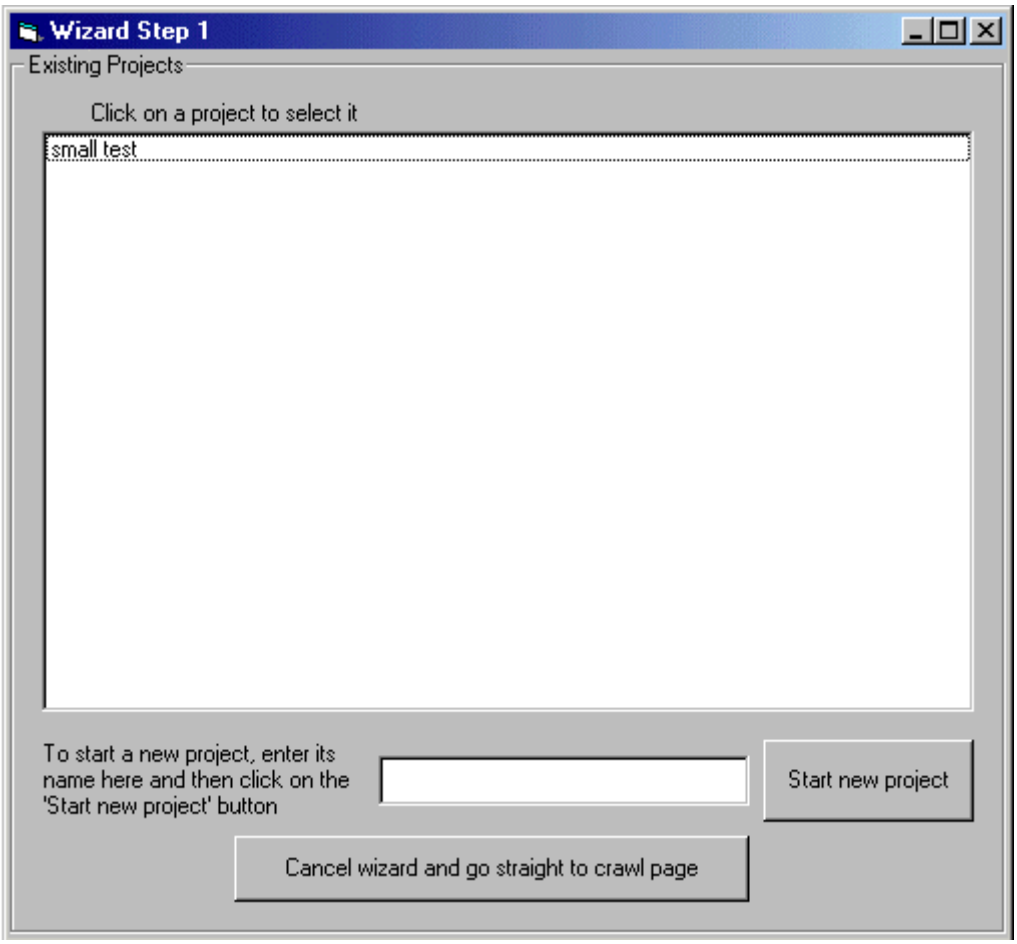
사용자는 crawl 하는 동안 윗부분의 타이틀 바에서 crawl 에 대한 정보를 읽을 수 있다. 그리고 또한 마지막 부분에서도 읽을 수 있다.



8. Crawl 이 완료 되었을 때 SocSciBot 을 끝내기 위해 *Yes* 를 클릭하라. 사용자는 이제 http://linkanalysis.wlv.ac.uk 사이트의 모든 페이지를 crawl 했다. 어떤 흥미로운 간단한 분석을 시행하기 전에, 다음 단계는 2 개 이상의 사이트를 crawl 할 것이다.

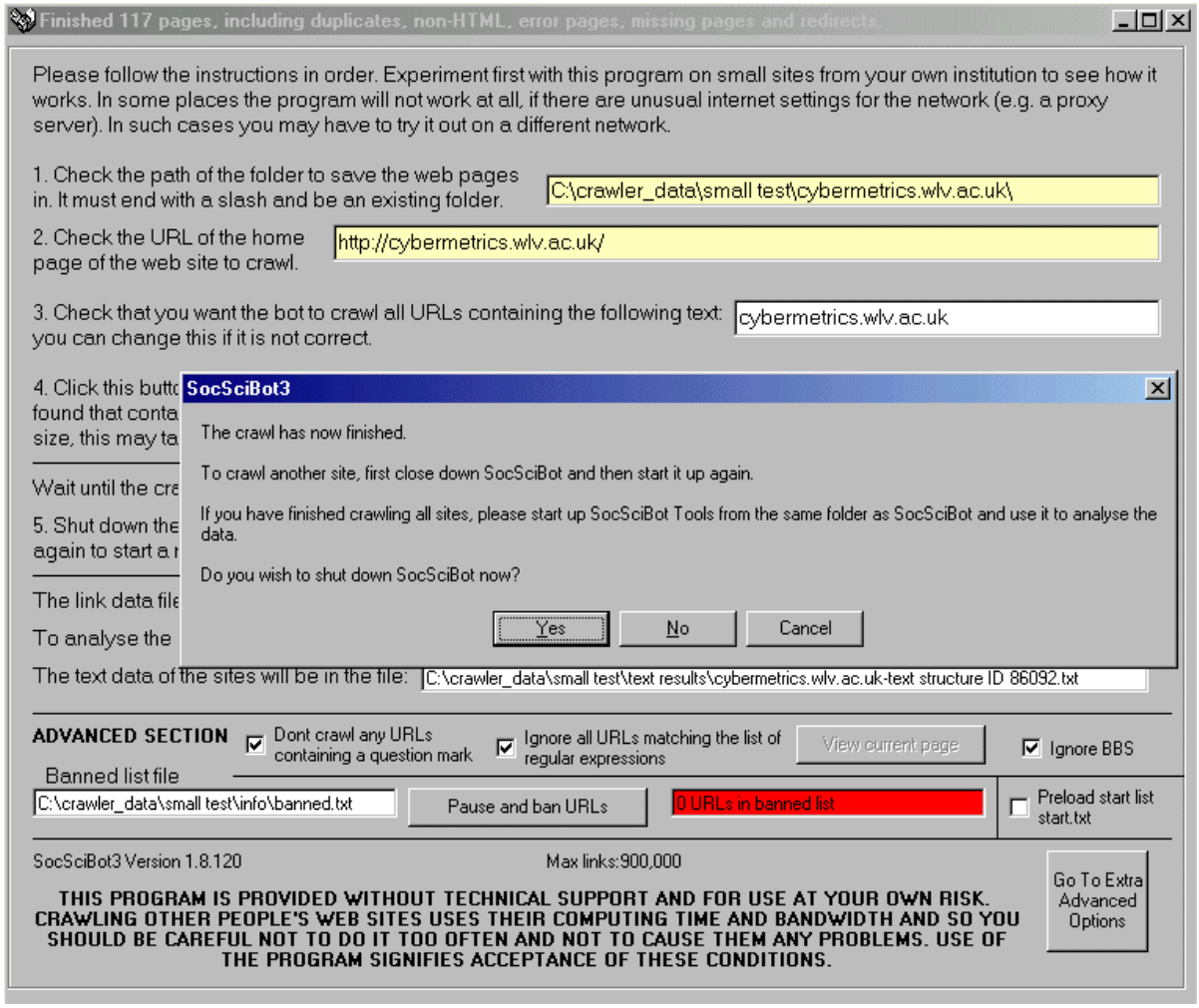
Step 3: Crawling two more sites with SocSciBot

1. 사용자의 컴퓨터에 있는 압축이 해제된 폴더의 SocSciBot 이나 SocSciBot.exe 파일을 더블 클릭해서 SocSciBot 을 다시 시작한다. 이것은 다른 crawl 을 추가하기 위해 이 프로젝트를 선택하는 *small test* 를 Wizard step 1. Click 을 통해 곧바로 사용자에게 지시되어야 한다.



2. Crawl 을 위해 두 번째 사이트의 URL 을 http://cybermetrics.wlv.ac.uk/로 입력하라. 그리고 *Start a new crawl of this site* 를 클릭하라.

3. 다음 화면에서 Crawl 사이트 버튼을 클릭하라. 그리고 crawler 가 완료 될 때까지 기다려라.



4. Crawl 을 마치기 위해 Yes 를 클릭하라.

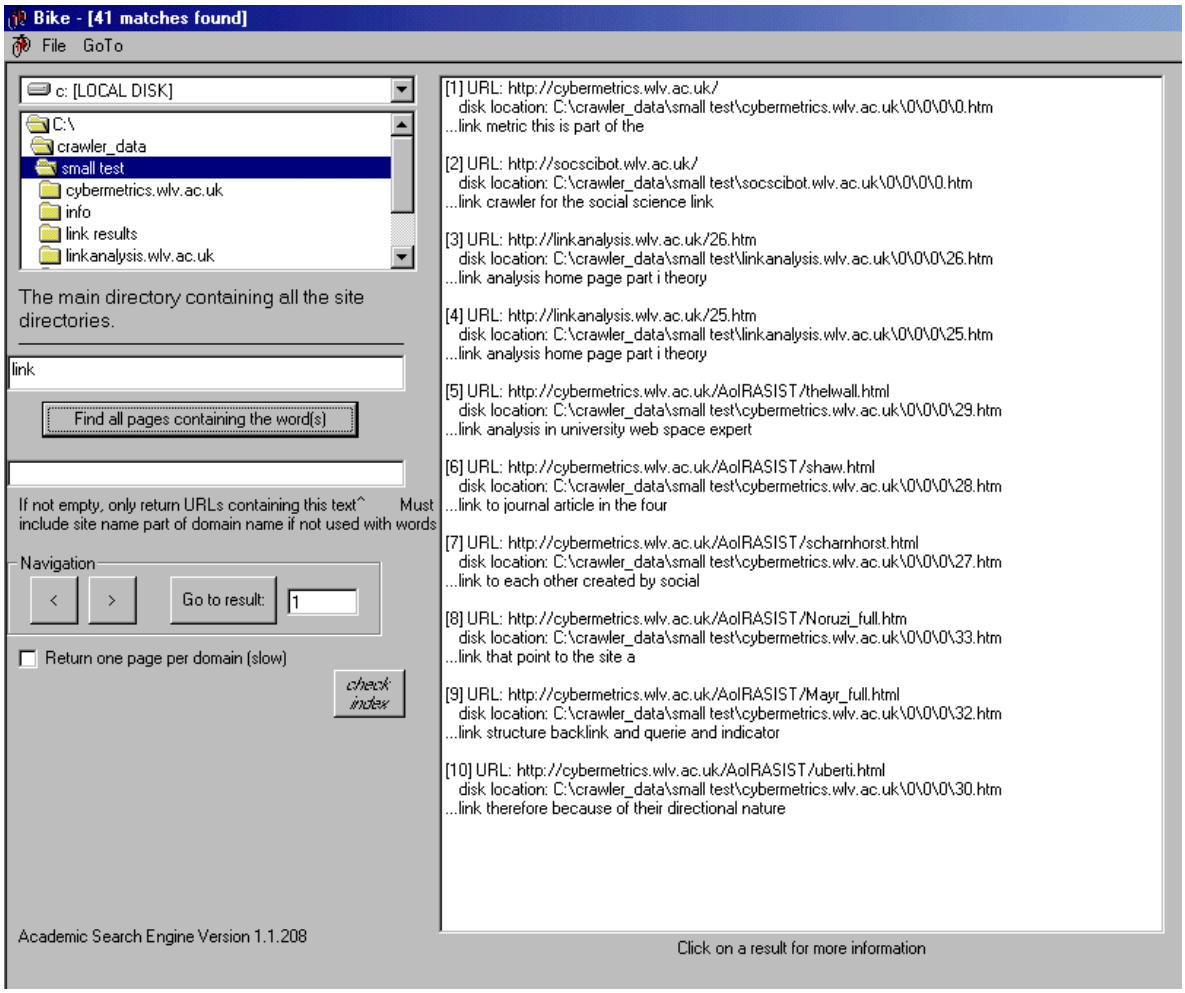
5. URL http://socscibot.wlv.ac.uk/로 1~4 단계를 반복하라.

6. 사용자는 이제 성공적으로 3 개의 웹사이트를 crawl 한다. 그리고 그 사이트들을 분석하는 것을 이해한다.

Step 4: Using Cyclist as a concordancer

1. 사용자의 컴퓨터에 압축이 해제된 것의 폴더에서 Cyclist 나 Cyclist.exe 둘 중 하나를 불러서 파일을 더블 클릭하여 Cyclist 를 시작하라. Cyclist 는 컨코던스/서치 엔진의 기능을 도움을 준다. 그리고 corpus 단어 빈번 분석자의 기능을 보좌한다.

2. 질문에 응답하라. 그리고 계산이 20 초나 그이상정도 지난 후에 사용자는 표준 컨코던스/서치 엔진 타입 인터페이스를 얻을 것이다. “링크”와 같은 일상적인 단어를 찾도록 시도하라. 아래 예에서, 몇몇의 부가 정보와 함께, 프로젝트에 41 페이지들은 단어 링크(혹은 링크들)를 포함하고 첫 10 개는 리스트가 된다.



3. 다른 결과 페이지들을 위한 네비게이션이 가능하다. 다음 혹은 이전 페이지로 가도록 화살표 버튼을 사용하라. 혹은 다른 페이지에 스킵한 결과의 넘버를 타이프하라.

4. 각각의 페이지에 대한 더 많은 정보는 화면의 오른쪽 옆의 결과를 클릭하면 찾을 수 있다.

- 결과의 *Top* 라인을 클릭하는 것은 인터넷 Explorer 에 원래 페이지를 저장할 것이다(사용자가 아마도 흥미가 있지 않은, top 에 페이지에 대한 일부 추가 정보와 함께).
- 결과의 *second* 라인을 클릭하는 것은 Notepad 안으로 원본 페이지의 HTML 원천 코드를 저장할 것이다(사용자가 아마도 흥미 없는, top 에서 페이지에 대한 일부 추가 정보와 함께).
- 결과의 *third and last* 라인을 클릭하는 것은 페이지에서 일부 부가적인 단어들을 보여 줄 것이다.

Step 5: Obtaining the corpus statistics

Cyclist 는, 다음과 같이, 여러 가지 단어 빈번 통계를 계산한다.

- 프로젝트에서 어떤 웹사이트의 어떤 웹 페이지에서 생긴 단어의 전체 단어 리스트를 위해, 메뉴 아이템 *Info / Save Word Frequency Summary of Whole Project* 를 선택하라.

사용자는 그런 다음 각각의 단어의 발생 수와 함께 파일 이름을 정할 것이고 전체 어휘는 파일에 저장될 것이다.

- 프로젝트에서 각 사이트에 웹 페이지에서 발생한 전체 단어의 리스트를 위해, 메뉴 아이템 *Info / Save Word Frequency Summary of Each Individual Domain* 을 선택하라. 사용자는 파일이름을 정할 것이고 각 사이트의 전체 어휘는 각 단어의 발생 수와 함께 분리된 파일에 저장될 것이다(사용자가 기입한 파일이름에 기초한 이름들). 각 라인에서 사용자가 아마 필요로 하지 않은 첫 번째 번호는 단어의 ID 이다.
- 프로젝트에서 단어의 전체 수, 혹은 유일한 단어의 전체 수를 위해, *Info / Word Count For all Sites* 를 선택하라. 각 분리된 사이트에서 전체 단어의 수를 노트하는 것은 위와 같이 *Info / Save Word Frequency Summary of Each Individual Domain* 에 의해 창조된 파일에 주어진다.

작은 사이트를 위해, 엑셀과 같은 스프레드 쉬트는 어휘를 분류하는데 유용하다.

Other information and options

Setting up your own project and crawling the sites

초기 SocSciBot 시작 화면에서 사용자의 리서치에 적당한 이름이 주어진 새 프로젝트를 시작하라. 현존하는 프로젝트에 crawl 을 추가하지 말라. 왜냐하면 이것은 링크 분석 결과를 난잡하게 만들기 때문이다. 새 프로젝트를 시작하기 위해서, 사용자가 SocSciBot 를 시작했을 때, 첫 번째 화면에서, 현존하는 프로젝트 이름을 클릭하는 대신에, 아래 박스에서 이름을 기입하라. 그리고 Create New Project 버튼을 클릭하라. 일단 사용자가 새 프로젝트를 선택 했다면, 사용자는 두 번째 SocSciBot 화면에서 홈페이지 URL 을 기입하면서 사이트를 crawl 을 시작할 준비를 하라. 사용자의 데이터 세트에서 각각의 새 사이트를 위해, 사용자는 이전 crawl 의 마지막에 SocSciBot 을 닫을 필요가 있다. 그런 다음 다시 시작하라. 만약에 사용자가 사용자의 사이트들을 crawling 하는데 어려움이 있다면 Tutorial 1 에서 지시사항을 참고하라.

Excluding unwanted pages from the sites

사용자는 사이트의 crawl 이 너무 많은 단어를 찾을 수 있다는 것을 알 수 있다. 예를 들면, crawl 은 사이트의 네덜란드 언어 버전과 사이트의 English 언어 버전 모두 다 포함했을지도 모른다. 그러나 사용자는 단지 English 언어 버전만을 원한다. 사용자는 금지된 리스트 피처를 사용하여 crawling 그리고/혹은 어휘로부터 제외된 페이지를 얻을 수 있다. 이것은 *pattern* 에 기초해서 crawl 에서 URL 을 제외한다. 예를 들면, 만약에 웹 사이트가 English 와 네덜란드 부분을 가지고 있다면 http://www.wlv.ac.uk/du/의 시작 URL 을 가진 모든 네덜란드 페이지들과 함께 그리고 나서 등록한 http://www.wlv.ac.uk/du/는 모든 네덜란드 페이지들을 단번에 제외하도록 할 것이다.

- 페이지들은 Pause and Ban URLs 버튼을 클릭하면 SocSciBot crawl 전이나 하는 동안에 제외 될 수 있다. 이것이 정답을 얻는데 교활한 것이기 때문에 면밀하게 지시사항을 탐독하라.
- 페이지들은 SocSciBot crawl 후에 제외 될 수 있다. 그러나 SocSciBot Tools 프로그램을 사용할 때만 그렇다. crawl 후에 페이지들을 제외하기 위해 SocSciBot Tools(SocSciBot 과 Cyclist 로서 동일한 zip 파일에 있는)를 시작하고 사용자가 사용자의 프르젝트를 선택한 후에 화면의 아래에 Data Cleansing 버튼을 클릭하라. 어떤 데이터 정화를 하기 전에, 오류가 날 수 있는 것을 대비해서, 사용자의 crawl 데이터를 백업하는 것은 좋은 생각이다(e.g. 프로젝트 디렉토리의 zip 파일 복사본을 가져와서). 처음에 작은 사이트를 시도해 보는 것도 좋은 생각이다.

Selecting different options for indexing the sites

만약 사용자가 사이트에서 다른 형태의 단어 stemming 을 필요로 한다면, Porter Algorithm 을 포함하는 일부 이용 가능한 옵션이 있다. 다른 단어 stemming 옵션을 위해 새로운 어휘를 창조하기 위해선 사용자가 사이트를 재색인할 필요가 있다. 이것을 실행하기 위해, Cyclist 에서 *GoTo* 메뉴에서 *Make Index* 를 선택하라. 사용자가 원하는 옵션을 선택하고 *1. Make Index* 버튼을 클릭하라. 기존의 어휘에 덮어 쓰여서 새 어휘는 만들어 질 것이다. 이 과정이 완료 될 때, *GoTo* 메뉴에서 *Search Engine Interface* 를 선택하라.