한국어 메시지의 내용분석을 위한 KrKwic 소프트웨어의 소개

박 한 우

영남대 언론정보학과 교수

hanpark@ynu.ac.kr

http://www.hanpark.net

한국교육과정평가원 특강

2006년 11월 7일

***Loet Leydesdorff와 협동작업임**

http://www.leydesdorff.net

Journal of the Korean Data Analysis Society, Vol. 6, No. 5, October 2004, pp. 1377-1387

한국어의 내용분석을 위한 KrKwic 프로그램의 이해와 적용*: Daum.net에서 제공된 지역혁신에 관한 뉴스를 대상으로

박한우[1] Loet Leydesdorff[2]

요 약

본 논문은 영어권에서 개발된 PullText 소프트웨어를 한국어로 작성된 메시지의 내용분석을 위하여 변형한 KrKwic 프로그램을 소개한다. 논문의 목적은 KrKwic의 사용을 원하는 학생, 일반인, 연구자의 이해를 돕고자 실제 사례를 통하여 분석절차를 구체적으로 제시하는 것이다. 논문은 크게 KrKwic 프로그램이 기반하고 있는 이론적 모델인 사회 네트워크적 내용분석에 대한 개괄과 Daum.net에서 제공된 지역혁신에 관한 뉴스를 이용하여 수행한 실제 분석사례로 구성되어 있다.

주요용어 : PullText, KrKwic, 한국어 분석 프로그램, 내용분석, 사회 네트워크 분석.

1. 서론

내용분석이란 커뮤니케이션 메시지의 의미 혹은 핵심 아이디어를 조사하는 연구방법이다 (Krippendorff, 1980). 내용분석의 대상은 문자로부터 영상까지 다양하지만 일반적으로 문자로 작성된 메시지를 대상으로 한다. 내용분석은 전통적으로 연구자가 직접 문서를 읽고, 코딩하고, 분석해

KrKwic?

- **KrKwic**

- Korean Key Words In Context

- 네델란드 암스테르담 대학교의 Loet Leydesdorff 교수가 개발한 Full Text 소프트웨어를 한국어 분석을 위해 변형

KrKwic의 이론적 배경

- 커뮤니케이션 메시지의 의미는 어디에?
  - 자주 사용되는 단어에
  - 단어 간 관계망: 매스 + 커뮤니케이션
  - Semantic Network Analysis

전통적 내용분석의 문제점

- **기존 방법의 한계점 (Danowski, 1993)**

  - 연구자가 임의로 만든 분석항목에 너무 의존적
  - 개념적으로 조잡하고
  - 노동 비용 등이 비교적 많이 들며
  - 외적(external) 타당성이 제한되어 있으며
  - 연구자의 성향에 영향을 받는 이데올로기적

컴퓨터 내용분석의 등장

- **컴퓨터 소프트웨어를 이용한 내용분석**

- 컴퓨터 내용분석 프로그램의 목록
  http://academic.csuohio.edu/kneuendorf/content/cpuca/ccap.htm
- 그러나, 대부분이 영어권를 위한 것
- 한국어를 위한 KLiwc 등이 있으나 어떤 (문법적 또는 심리적) 범주체계에 따라 단어들을 분류하거나 한국어의 전산처리 방식에 초점을 둠

KrKwic 소프트웨어의 구성

- **크게 3개의 하위 소프트웨어로 구성**

- KrKwic: 단어 빈도 분석을 통해 핵심어, 주요 이미지, 중요 이슈를 파악함

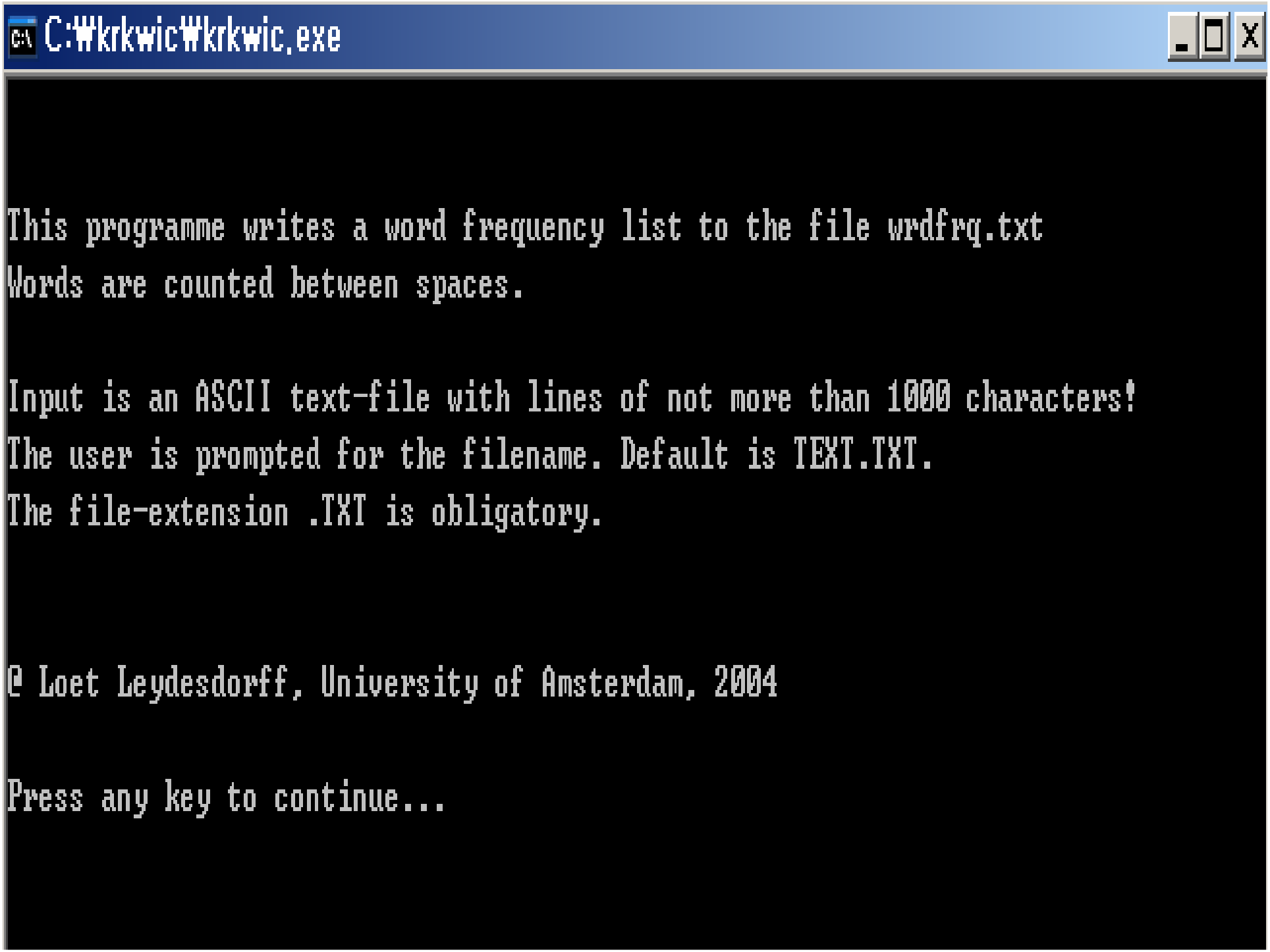
- KrTitle: 논문, 웹사이트, 기사, 특허, 법조문 등의 제목과 요약문 또는 주관식 응답, 드라마나 영화의 대사, 조직 목표, 광고 카피, 일상 대화 같이 비교적 짧은 메시지

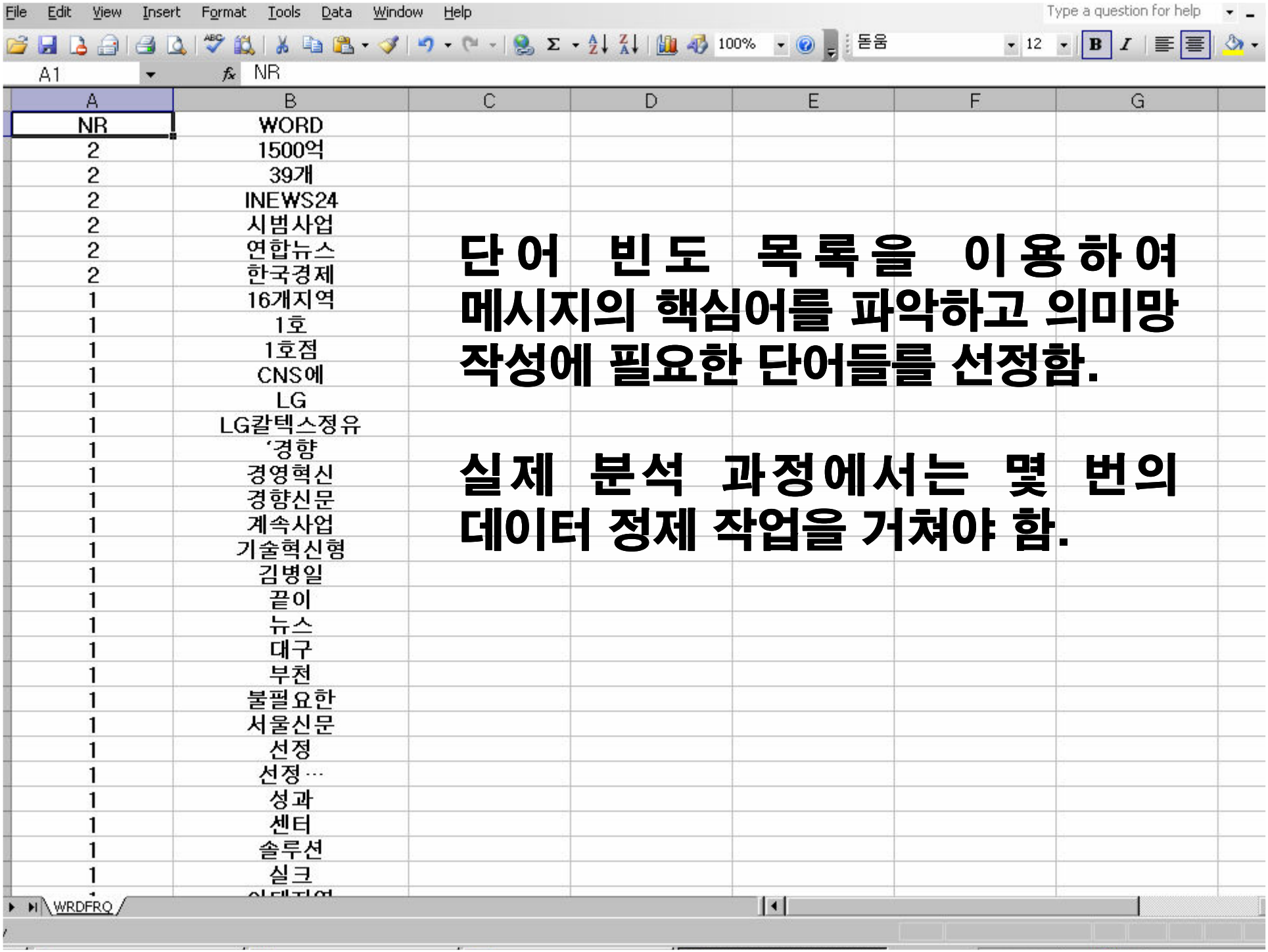
- KrText: KrTitle로 처리하기에 분량이 비교적 많은 메시지를 독립적인 파일로서 취급하여 분석

title_text - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

기술혁신 중소기업 육성
16개지역 특성화 1500억 지원
LG칼텍스정유 파업 끝이 안보인다
지역 특성화 선정… 대구 - 애견, 진주 - 실크
부천 '경향 뉴스 & 커피' 1호점 탄생
지역혁신 특성화 39개 시범사업 선정··3년간 1500억 투입
[화제] 하나로텔레콤, '주니어보드'..경영혁신 성과 '짭짤'
김병일 장관 "불필요한 계속사업 중단"
지역혁신특성화 시범사업 39개 선정
LG CNS에 아태지역 1호 어바이어 솔루션 센터 오픈





NR	WORD
2	1500억
2	39개
2	INEWS24
2	시범사업
2	연합뉴스
2	한국경제
1	16개지역
1	1호
1	1호점
1	CNS에
1	LG
1	LG칼텍스정유
1	'경향
1	경영혁신
1	경향신문
1	계속사업
1	기술혁신형
1	김병일
1	끝이
1	뉴스
1	대구
1	부천
1	불필요한
1	서울신문
1	선정
1	선정…
1	성과
1	센터
1	솔루션
1	실크

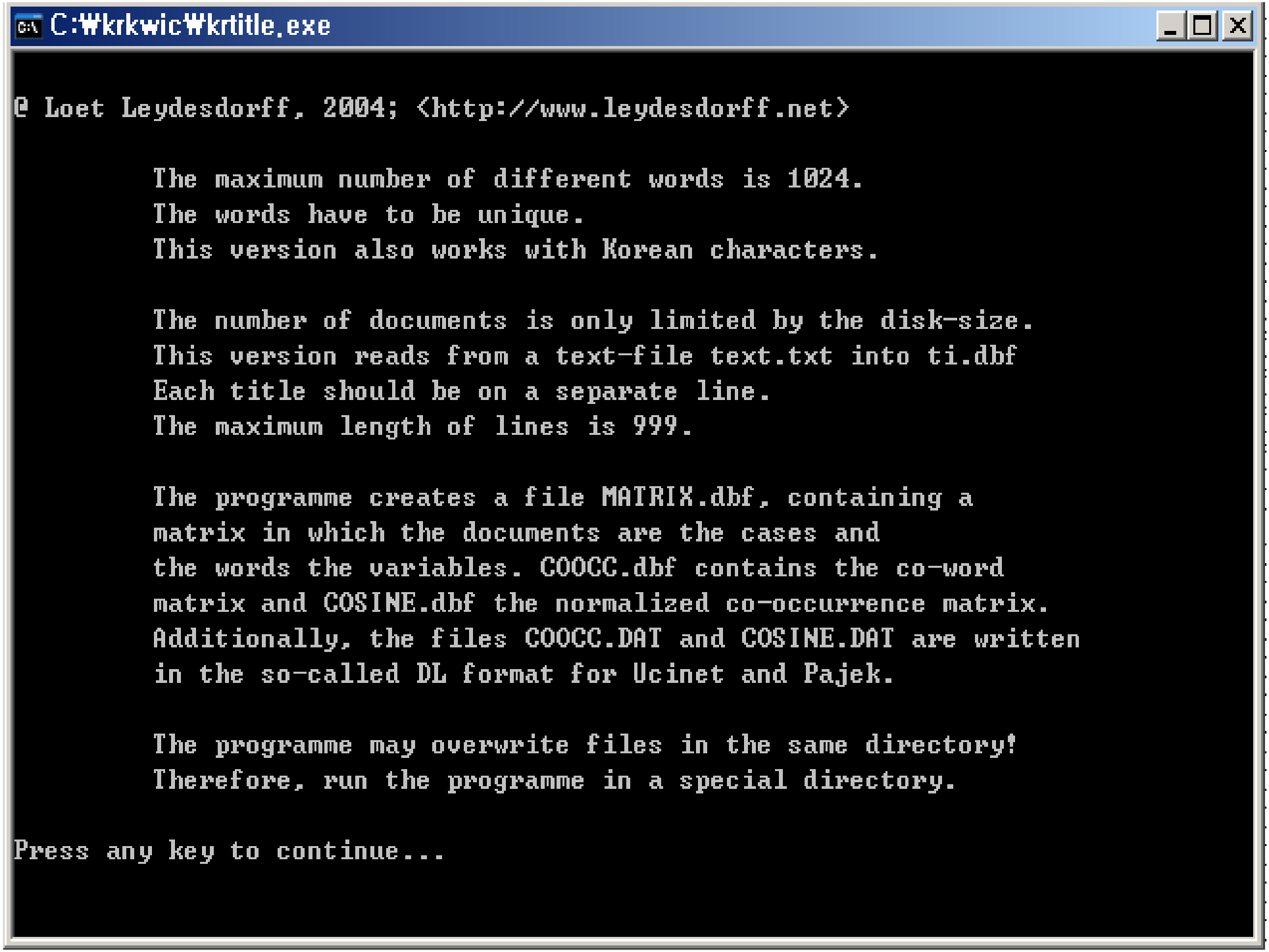
단어 빈도 목록을 이용하여 메시지의 핵심어를 파악하고 의미망 작성에 필요한 단어들를 선정함.

실제 분석 과정에서는 몇 번의 데이터 정제 작업을 거쳐야 함.

title_words.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

16개지역
기술혁신
선정
성과
지역
지역혁신
특임
특성화



KrTitle 아웃풋

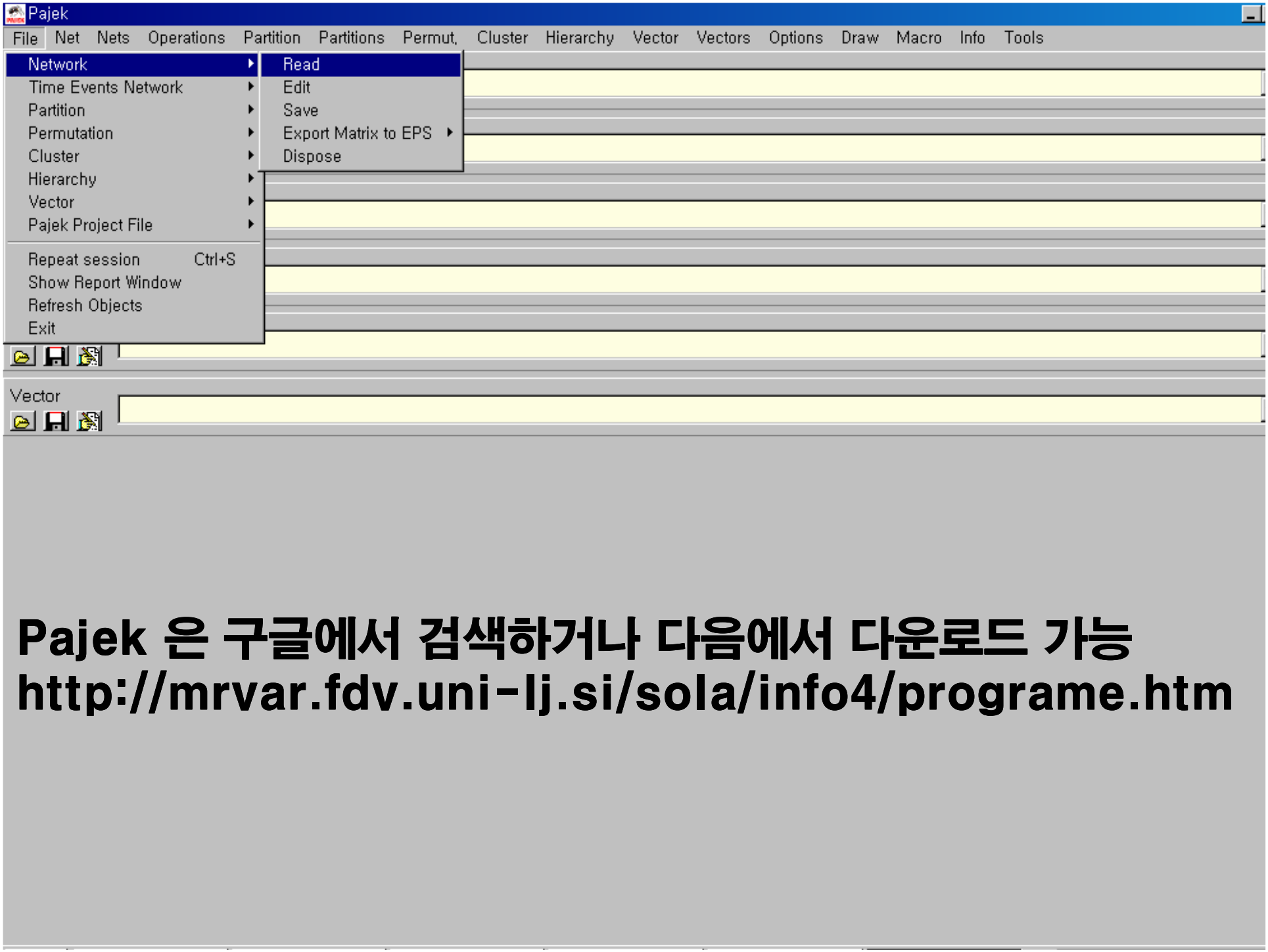
- 크게 3개의 결과 파일이 생성됨

- matrix.dbf: 메시지(사례) * 단어(변인) 행렬로 각 칸의 값은 단어가 메시지에서 출현한 빈도

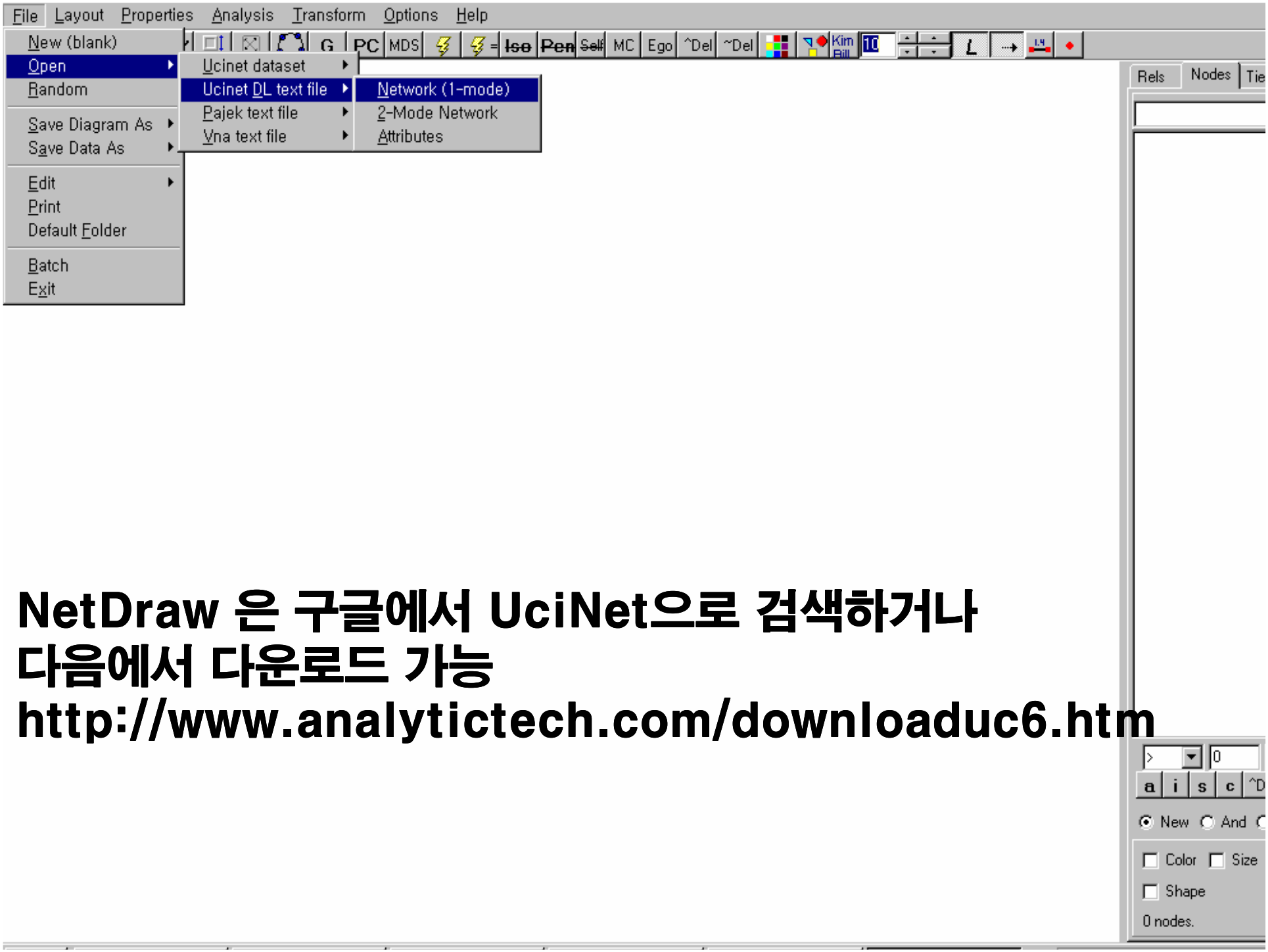
- coocc.dat와 coocc.dbf: 단어 * 단어 공출현빈도 행렬로 각 칸의 값은 단어들이 메시지에서 동시에 출현한 빈도

- cosine.dat와 cosine.dbf: 단어 * 단어 코사인 행렬로 각 칸의 값은 단어 간 거리

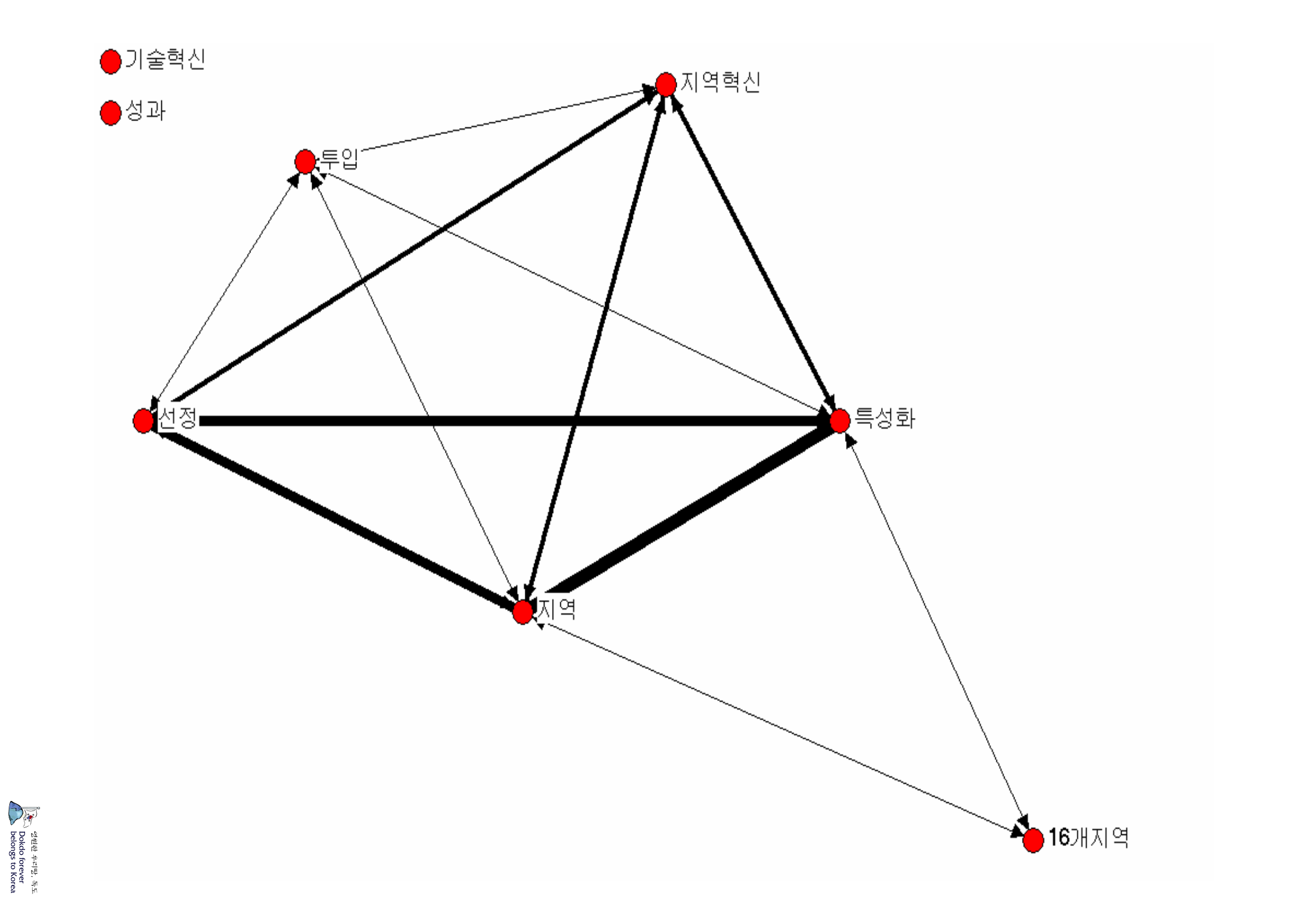
단어들	개지역	기술혁신	선정	성과	지역	지역혁신	투입	특성화
개지역								
기술혁신								
선정								
성과								
지역								
지역혁신								
투입								
특성								

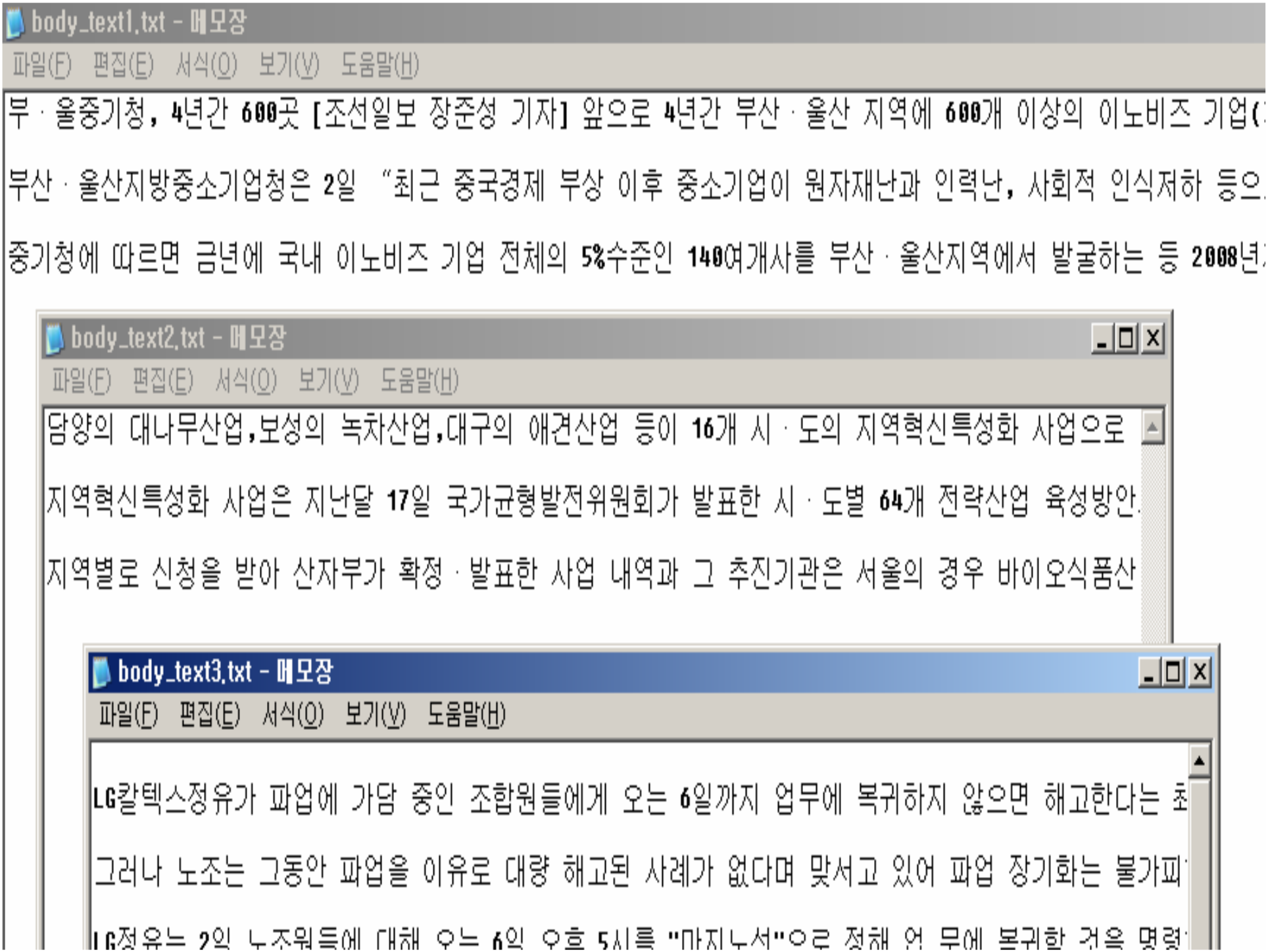


Pajek 은 구글에서 검색하거나 다음에서 다운로드 가능
http://mrvar.fdv.uni-lj.si/sola/info4/programe.htm



NetDraw 은 구글에서 UciNet으로 검색하거나
다음에서 다운로드 가능
http://www.analytictech.com/downloaduc6.htm





대권예비주자의 신문기사 네트워크 분석과 홍보전략:
조선일보와 한겨레신문에 나타난 고건과 박근혜 관련 기사를 중심으로

남인용
부경대 신문방송학과 교수

박한우
영남대 언론정보학과 교수

배애진
영남대 언론정보학과
<뉴미디어와 사회> 연구실 연구원

Work in progress

본 논문은 2006년 11월4일(토) 한국광고문화회관에서 개최되는
한국광고학회(http://www.koads.or.kr)의 2006년 가을학술대회에서 발표됨

<표-?> 5.31 지방선거 전후 조선일보 고건 관련기사 제목에 나타난 주요어

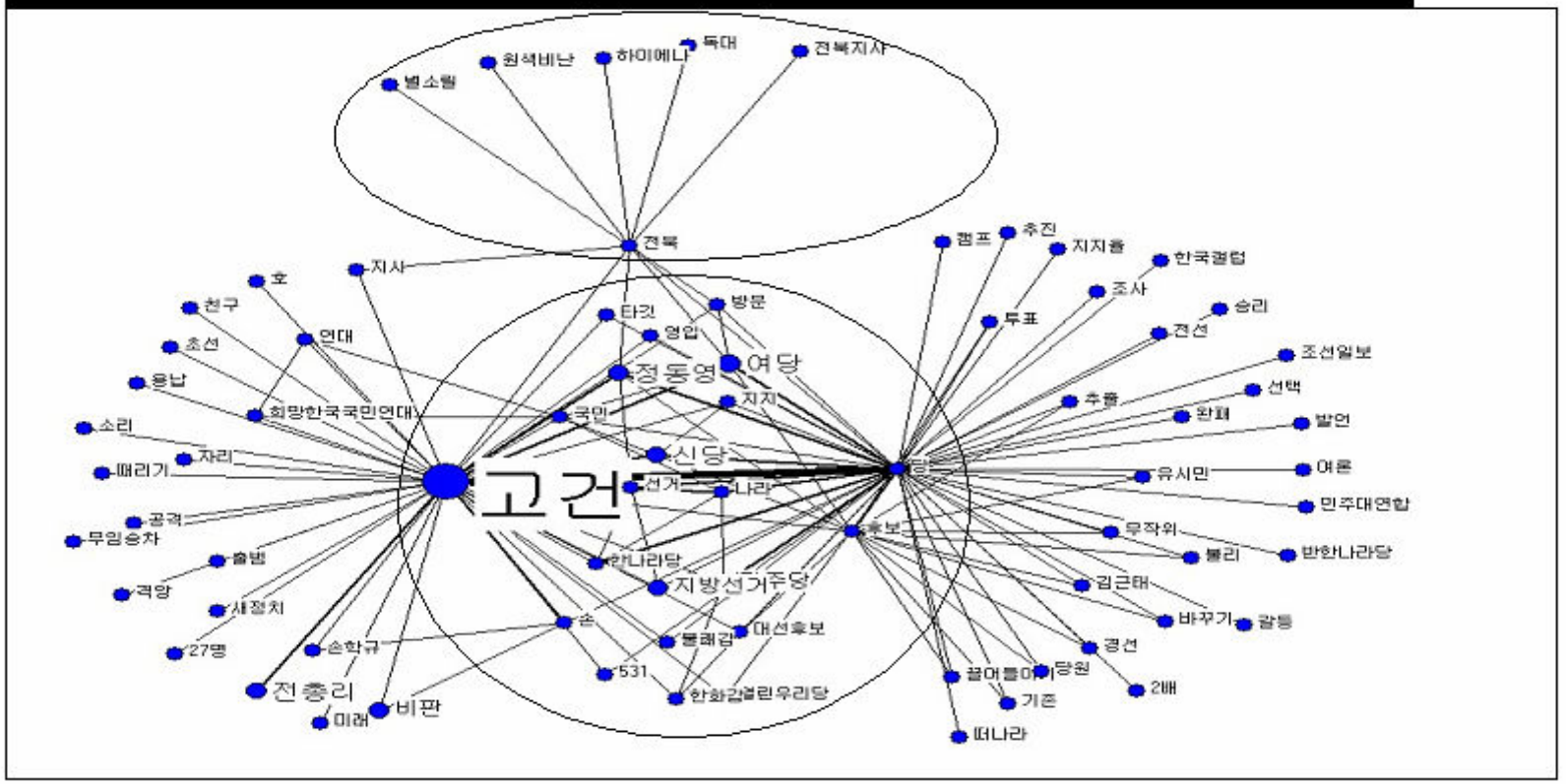
2006년 3월 1일~5월 31일		2006년 6월 1일~8월 31일	
사용된 단어	빈도수	사용된 단어	빈도수
고건	13	고건	17
전총리	5	신당	5
정동영	4	여당	3
여당	3	531	2
지방선거	3	국민	2
2일, 27명, 90분, 격앙, 경제, 고장, 공격, 교양강좌, 국정의달인, 독대, 동문서답, 때리기, 무산, 무임승차, 문화, 미래, 민주당, 민주대연합, 반한나라당, 방문, 별소릴, 비상, 비판, 서울시장, 선거, 소리, 손학규, 역할, 연대, 열린우리당, 용납, 원색비난, 유명, 인기, 인사, 자리, 자문그룹, 전략, 전북, 전북지사, 전선, 정치시스템, 주장, 지사, 짱, 초선, 추진, 출동, 출범, 충북대, 친구, 하이에나, 행정, 환경, 회동	1	대선후보	2
		민주당	2
		비판	2
		새정치	2
		한나라당	2
		한화갑	2
		희망한국국민연대	2
		106명, 2배, 47.3%, 65.4%, 갈등, 결성, 경선, 공식, 국세청, 기존, 김근태, 깃발, 까닭, 끌어들이기, 나라, 남동공단, 당, 당원, 돛단배, 떠나라, 뜻, 명단, 무작위, 미래, 바꾸기, 바람, 박근혜, 반발, 발기인, 발언, 발표, 밝혀, 방문, 불리, 불쾌감, 비전, 빗댄, 살림, 선거, 선택, 손, 손짓, 손학규, 승리, 시류, 신계륜, 쏠림, 아마추어정부, 압승, 여론, 열린우리당, 영입, 완패, 유시민, 이념, 이명박, 인터뷰, 일자, 정동영, 정책, 정파, 조사, 조선일보, 조직, 중심, 지지, 지지율, 철회, 초월, 추출, 출범, 캠프, 타깃, 투표, 필요, 한국갤럽, 현상, 현정부, 호, 호남, 후보	1

<표-?> 한겨레신문 고건 관련기사 제목에 나타난 주요어

2006년 3월 1일 ~ 2006년 5월 31일		2006년 6월 1일 ~ 2006년 8월 31일	
단어	빈도수	단어	빈도수
고건	9	고건	15
시장	2	531	4
연대	2	후폭풍	4
전총리	2	구상	2
정동영	2	대선행보	2
12일, 가속페달, 강현욱지사, 개혁세력, 기획, 날선, 당, 당신, 대꾸없는, 대답없는, 독자엔진, 때리기, 만나나, 물갈이, 물건너간, 민주개혁세력연합론, 민주당, 복귀, 선거, 선긋자, 손잡고, 수혈, 시나리오, 신호탄, 여당, 왜, 의장, 의혹, 이래도저래도, 이럴바엔, 이명박, 장착, 점심, 지도부, 지방선거, 초선, 치른, 탈당설, 통합, 특혜분양, 한독단지, 함께, 확실히, 회동	1	시동	2
		중도통합기구	2
		출범	2
		희망한국국민연대	2
		106명, 연대, 4년전, 거사, 계속, 고장난정치, 고치자, 고통, 공개, 국민, 굳건, 내품으로, 닮은꼴, 몰두, 민생경제, 민주당, 박근혜, 박범신, 받을까, 발기인, 보유세, 본격적, 서로, 선두, 성토, 시간고민, 신경전, 알쏭달쏭화법, 어렵다, 여론, 연말, 열린우리당, 올려, 움직인다, 은근한, 이명박, 전총리, 정계개편, 정몽준, 정세현, 주춤, 지금, 지지율, 책임, 추격, 추진, 퇴로없이, 판단, 편가르기, 한나라당, 함께, 현정부, 희망연대, 힘	1
총 기사 수	9	총 기사 수	15
사용된 총 단어 수	49	사용된 총 단어 수	63
총 단어의 출현 빈도 수	61	총 단어의 출현 빈도 수	89

<그림-?>은 조선일보의 기사제목에 나타난 고건의 의미(semantic) 네트워크 지도이다. 중앙에는 '고건', '정동영', '여당', '신당', '지방선거'를 중심으로 여러 단어들이 촘촘히(densely) 연결된 클러스터가 있다. 중앙 클러스터는 지방선거 정국에서 고건과 제휴하려는 여러 정치 행위자들과 이러한 움직임을 묘사하는 단어들로 구성되어 있다. 지도의 윗부분에는 고건의 부정적 이미지와 연관된 단어들이 모여 있다. 좌측에는 삼각형의 형태로 '희망한국국민연대', '국민', '연대'가 하나의 클러스터를 만들고 있으며, 그 아래에는 '전총리'가 '고건'과 독자적으로 연결되어 있다.

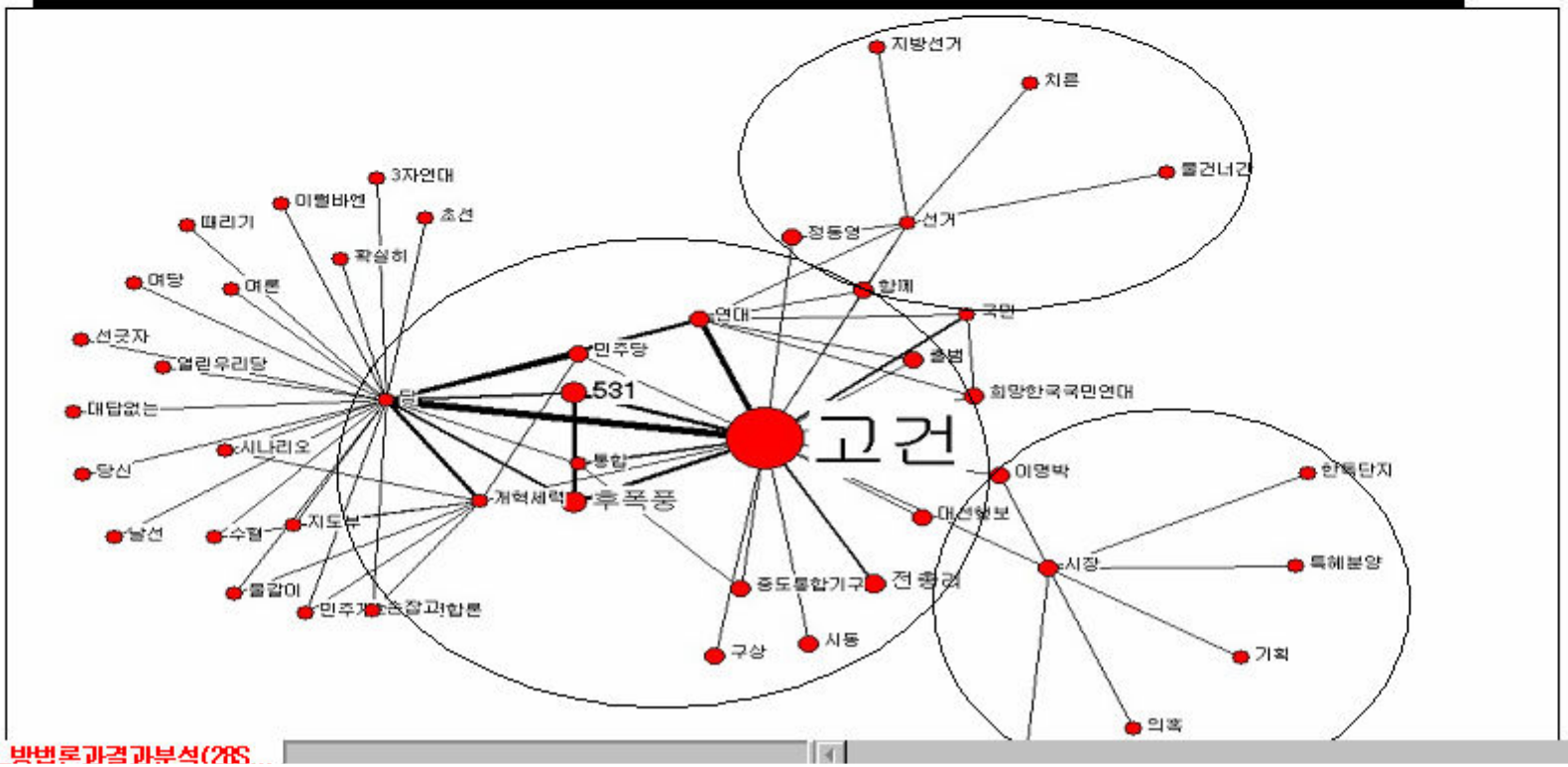
<그림-?> 조선일보의 고건 관련 기사 제목에 나타난 주요어의 네트워크 지도



* 다른 단어와의 관계 강도가 2이상인(greater than or equal to) 경우만 표시함

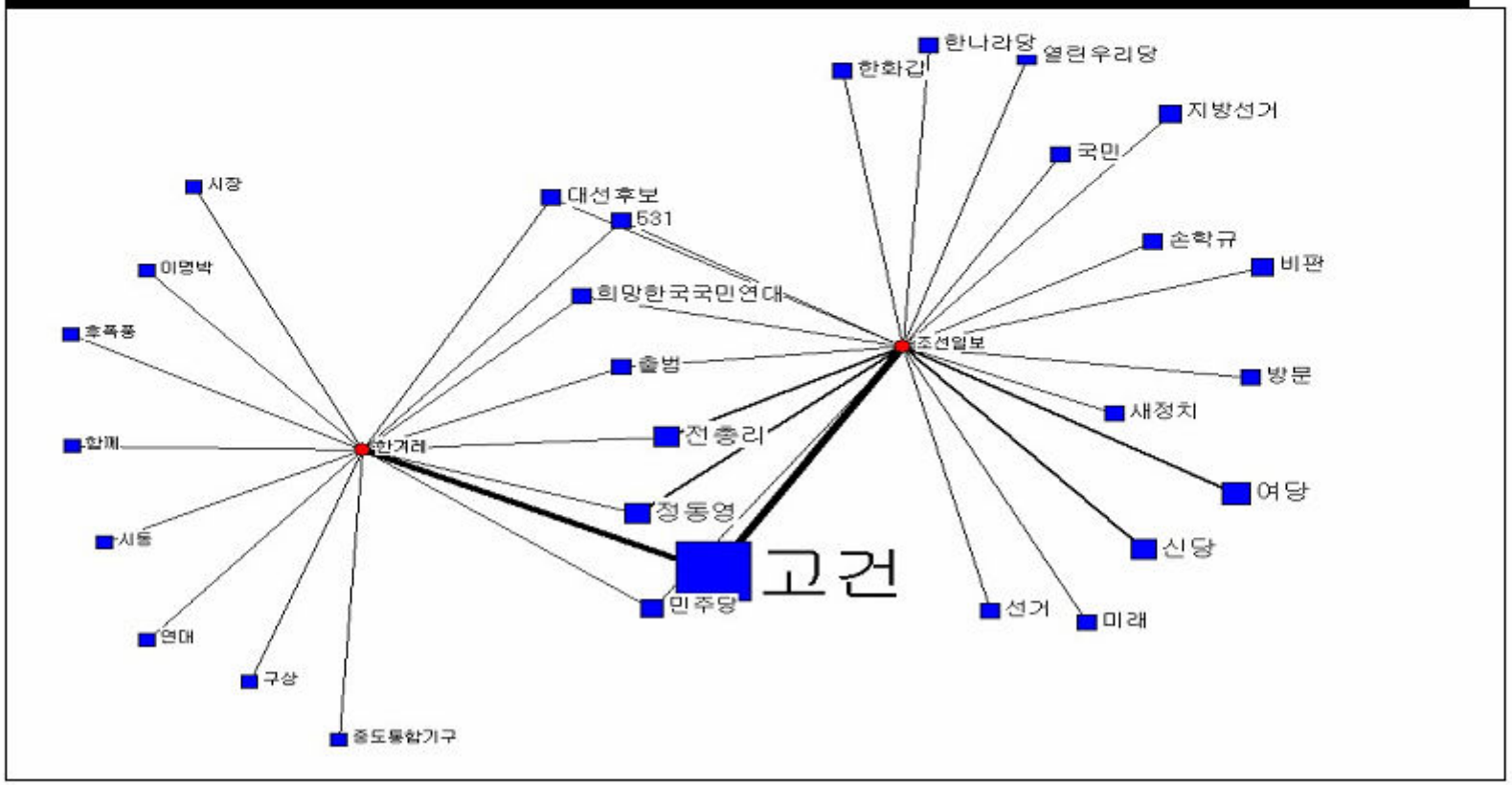
한겨레신문의 네트워크 지도를 보면 1개의 주요한(major) 클러스터(cluster)와 3개의 작은(minor) 클러스터가 있다. 첫째, '고건'이라는 단어는 '당', '연대', '후폭풍', '531', '민주당', '개혁세력', '희망한국연대', '국민', '전총리', '대선행보' 등과 함께 중앙에서 핵심(prominent) 클러스터를 형성하고 있다. 이 클러스터를 보면, 한겨레신문은 조선일보가 정동영을 비롯한 기존 정치 세력과의 연대를 강조하는 것과 달리 지방선거의 전후 정국에서 고건의 잠재적(potential) 역할에 초점을 맞추는 것으로 보인다. 둘째, 오른쪽 위에는 지방 선거 국면에서 고건이 열린 우리당과 관련된 단어들이 모여 있다. 셋째, 우측 아래에는 서울 시장 시절의 행적에 대한 평가와 관련된 단어들이 모여 있다. 재미있게도 서울 시장을 논의하는 클러스터는 이명박과 연결되어 있다.

<그림-?> 한겨레신문의 고건 관련 기사 제목에 나타난 주요어의 네트워크 지도



들이 수직으로 배열되어 있다. 오른쪽에는 조선일보가 사용한 상징어의 목록이 왼쪽에는 한겨레신문의 주요어가 위치해 있다. 네트워크 지도를 보면, 두 신문 모두와 연결된 단어들은 '고건', '전총리', '정동영', '희망한국국민연대', '민주당', '531', '대선후보' 등이 있다. 이 단어들을 살펴보면, 두 신문 모두 고건 전총리가 대선후보와 관련하여 기존 정당과의 연합을 시도하거나 국민연대를 출범하는 것을 강조하고 있음을 알 수 있다. 조선일보에만 연결된 단어들은 '한나라당', '열린우리당', '한화갑', '신당' 등이 있으며, 한겨레신문에만 연결된 단어들은 '중도통합기구', '이명박', '연대' 등이 있다.

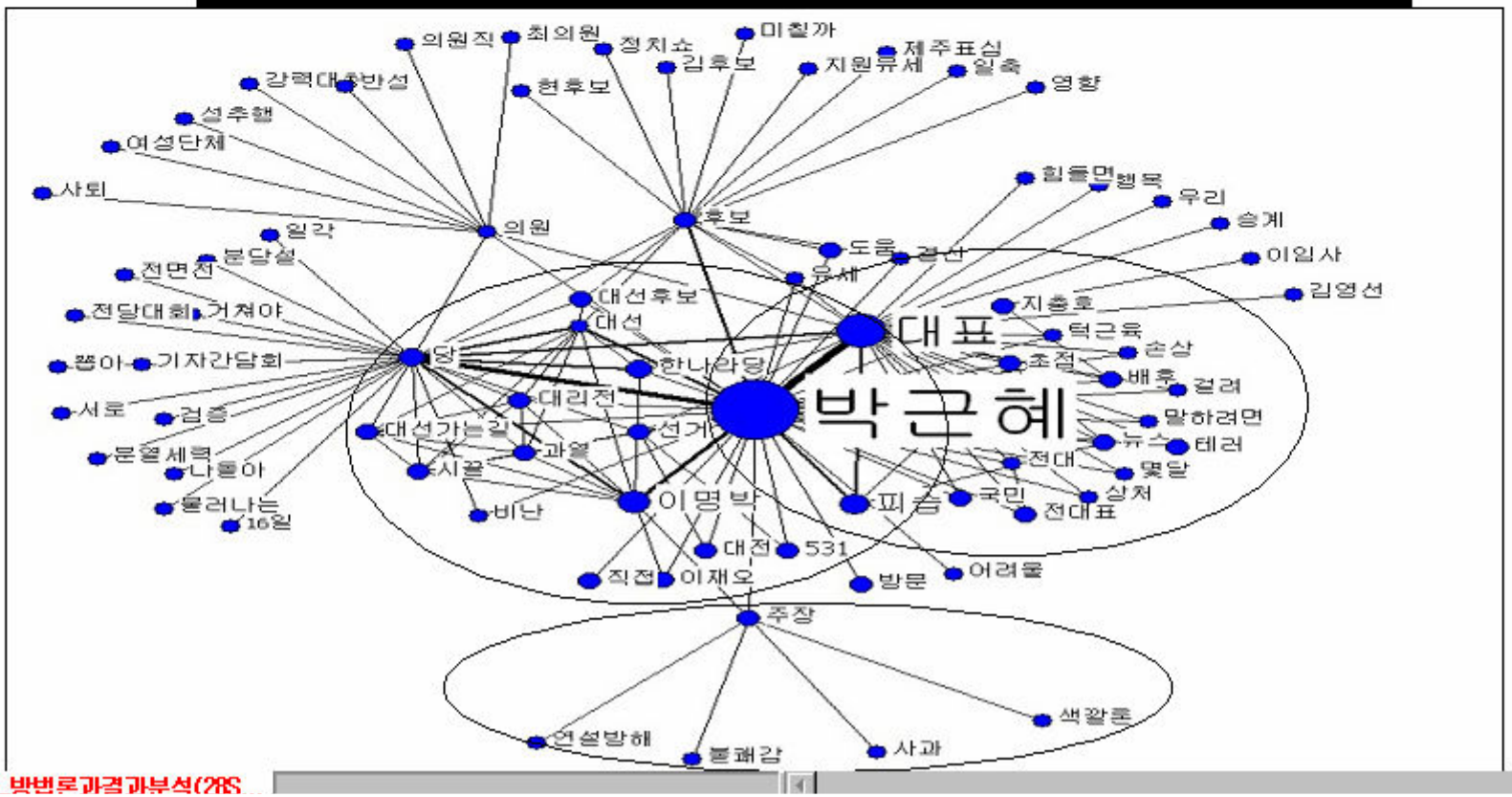
<그림-?> 조선과 한겨레의 고건 관련 기사 제목에 나타난 주요어의 네트워크 지도



* 각 신문에서 출현빈도가 2회 이상인 단어들만 표시함

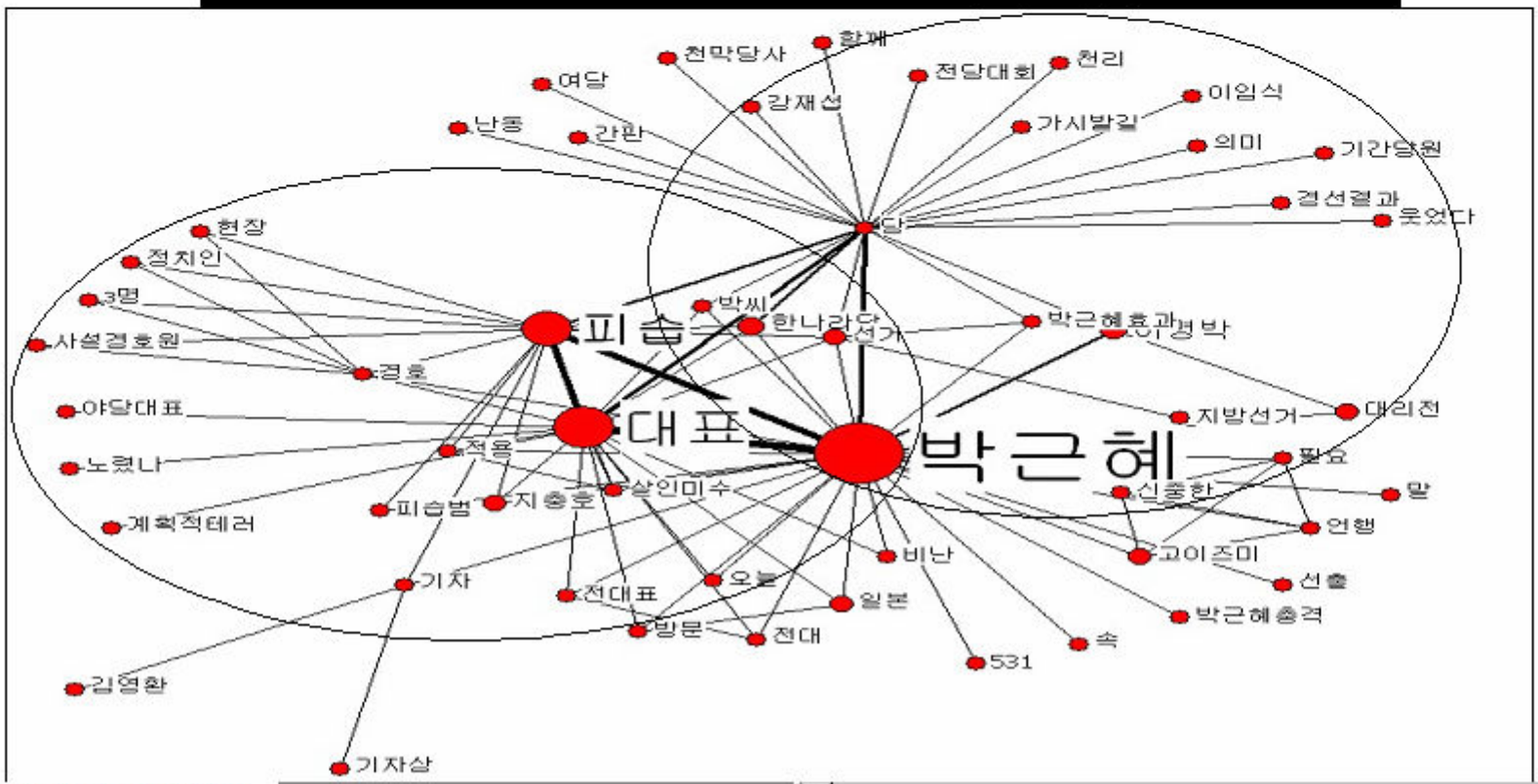
있다. 첫 번째 클러스터는 박근혜, 대표, 이명박, 당, 한나라당, 대리전, 이재오가 중심이 되어 형성되고 있다. 이 클러스터를 보면, 한나라당 전당대회에서 박근혜와 이명박의 대결구도가 나타나고 있음이 보인다. 다음으로 우측 중앙을 보면, '박근혜'와 '대표'라는 단어를 축(axis)으로 하여 '피습'이라는 단어가 연결되면서 독자적인 그룹을 형성되고 있다. 이 두 번째 클러스터는 박근혜 대표 피습사건과 관련된 단어들이 촘촘히 모여있다. 핵심 클러스터들 아래에 있는 작은 클러스터는 한나라당 전당대회로 인해 박근혜와 이명박이 서로 불편한 감정을 가지고 있음을 나타내는 단어들이 모여 있다. 한편 성추행 사건으로 이슈가 되었던 '최의원'과 관련된 단어들이 좌측 위에서 발견되었다.

<그림-?> 조선일보의 박근혜 관련 기사 제목에 나타난 주요어의 네트워크 지도



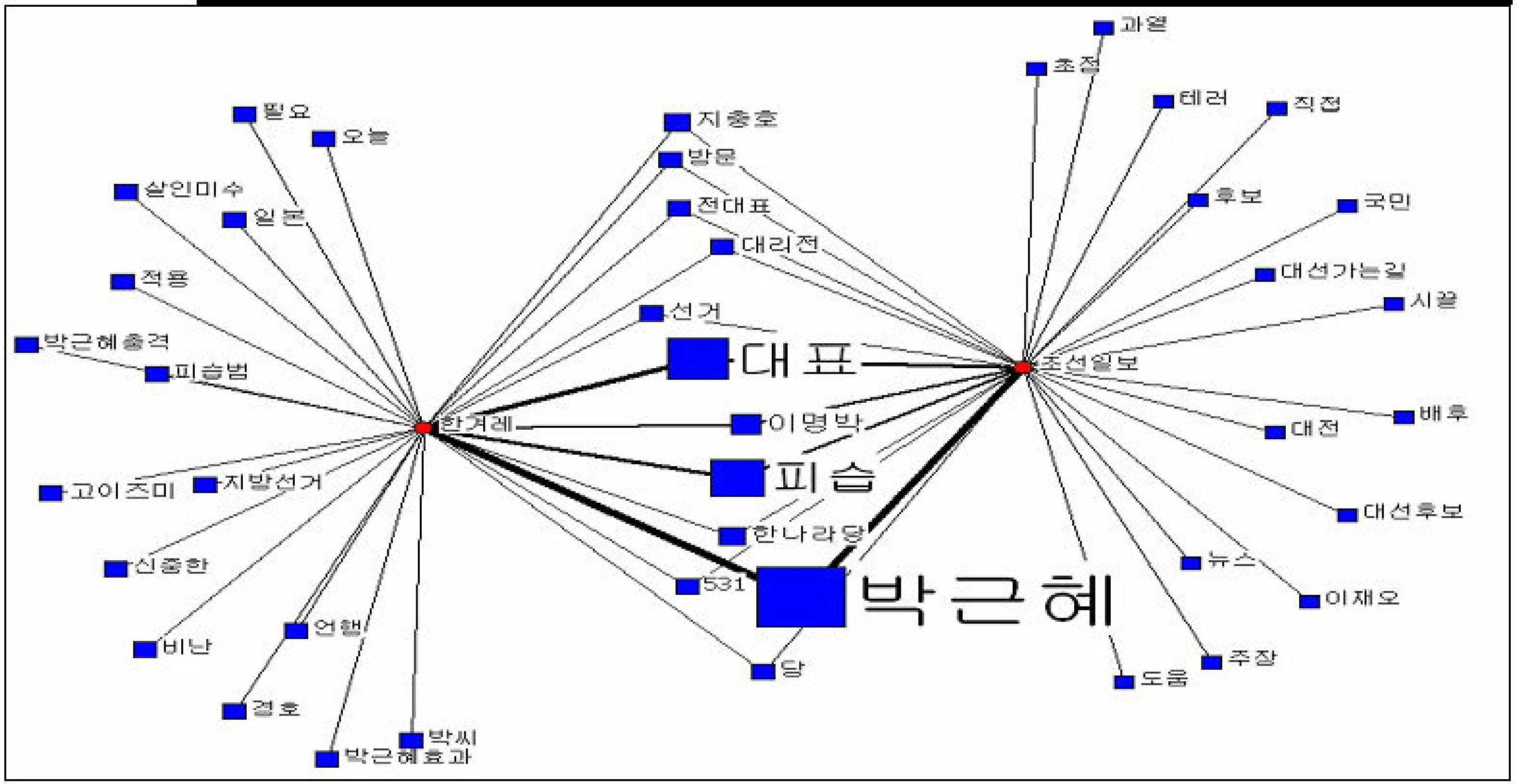
시들이 나타나 있다. 재미있게도 조선일보의 네트워크 지도와 비교해서, 한겨레신문에서는 '피습'과 관련된 상징어가 빈번히 등장하면서 박 대표의 피습사건이 두드러지게 나타나고 있다. 이것은 '피습'이라는 단어의 동심원 크기와 '박근혜'와 '피습'을 연결하는 선의 굵기에서 확인할 수 있다. 두 번째 클러스터는 '박근혜', '선거', '한나라당', '박근혜 효과', '당', '이명박', '대리전' 등의 단어들을 중심으로 형성되어 있다. 클러스터의 위쪽에서는 한나라당 전당대회와 관련한 단어들이 모여 있는데, '강재섭', '박근혜', '웃었다' 등의 단어에서 전당대회 결과가 박근혜에게 유리하게 작용된 것을 짐작할 수 있다. 그리고 삼각형의 형태로 '박근혜', '이명박', '대리전'이 클러스터의 밑 부분에서 서로 연결되어 있다.

<그림-?> 한겨레의 박근혜 관련 기사 제목에 나타난 주요어의 네트워크 지도



<그림-?>을 보면, 중앙 부분에는 조선일보와 한겨레신문이 공통적으로 사용하는 단어들이 수직으로 배열되어 있다. 오른쪽에는 조선일보가 기사제목에서 2회 이상 사용한 단어들이, 왼쪽에는 한겨레신문의 주요어가 나열돼 있다. 네트워크 지도를 보면, 두 신문 모두와 연결된 단어들은 '박근혜', '대표', '피습', '이명박', '한나라당', '531', '대리전', '지충호' 등이 있다. 이 단어들을 살펴보면, 두 신문 모두 박근혜 대표 피습 사건, 531 지방선거, 한나라당 전당대회를 핵심 이슈로 하고 있음을 알 수 있다. 조선일보에만 연결된 단어들은 '테러', '대전', '배후', '대선후보' 등이 있으며, 한겨레신문에만 연결된 단어들은 '고이즈미', '박근혜효과', '박근혜충격', '피습범', '지방선거' 등이 있다.

<그림-?> 조선과 한겨레의 박근혜 관련 기사 제목에 나타난 주요어의 네트워크 지도



* 각 신문에서 출현빈도가 2회 이상인 단어들만 표시함

감 사 합 니 다!!!

Interesting & growing area
Promising approach

- **영남대학교 <뉴미디어와 사회> 연구실**
  - * 홈페이지 : http://www.hanpark.net
  - * 카페: http://cafe.naver.com/newmas.cafe
  - * E-mail : hanpark@ynu.ac.kr
    parkhanwoo@hotmail.com

Many thanks to my assistants!