

## 1-1) 각 텍스트 파일들로부터 색인어 추출

```
aa.c
1  #include <stdio.h>
2  #include <string.h>
3
4  int main()
5  {
6      FILE *fp = fopen("aaa.txt", "w"); //파일 쓰기
7      FILE *read = fopen("fnames.txt", "r"); //파일 읽기
8      char buffer[100]; //한줄 읽을때마다 임시저장용 버퍼
9
10     while(1)
11     {
12         fgets(buffer, sizeof(buffer), read); //기사제목 한줄을 읽어옴
13         buffer[strcspn(buffer, "\n")] = 0;
14         //fgets사용시 뒤에 개행문자가 붙기 때문에 제거해줌
15         if(feof(read)) break; //마지막라인 2번 읽는 것 방지
16         fputs("index2018.exe ", fp);
17         fputs(buffer, fp);
18         fputs(" index-", fp);
19         fputs(buffer, fp);
20         fputs("\n", fp);
21         //index2018.exe 기사제목 index-기사제목 포맷으로 파일에 쓰기
22     }
23
24     fclose(fp);
25     fclose(read);
26     return 0;
27 }
```

위 코드를 통해 배치파일 생성 후 실행하여

색인어만 추출 된 623개의 파일 생성

- index-IT-20HPnotebook.txt
- index-IT-64bitPC.txt
- index-IT-2007newWebIR.txt
- index-IT-alanco.txt
- index-IT-amazon.txt
- index-IT-analogTV.txt
- index-IT-asianux.txt
- index-IT-auction.txt
- index-IT-baidu.txt
- index-IT-bido.txt
- index-IT-billJobs.txt
- index-IT-bioChip.txt
- index-IT-blackberry.txt
- index-IT-blasterWorm.txt
- index-IT-blogMusic.txt
- index-IT-bluetooth.txt
- index-IT-hot.txt

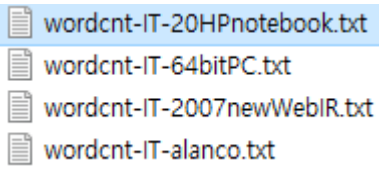
index-IT-20HPnotebook.txt - 메모장				
파일(F)	편집(E)	서식(O)	보기(V)	도움말(H)
한국HP	한국	HP	내년	20인치
20	인치	노트북	첫선	
데스크톱PC		데스크	톱	PC
대체	수	20인치급	20	인
치급	대형	노트북	내년	첫선
보	게임마니아	게임	마니아	겨냥
400만원	400	만원	초고성능	초고
성능	데스크톱PC		데스크	톱
PC	데스크톱PC도		도	등장
한국HP	한국	HP	가정	사무실
노트북PC	노트북	PC	사용	수요층
증가	증가함	데스크톱PC		데스크
톱	PC	대체	수	20인치
20	인치	19인치급	19	인
치급	대형	노트북PC	노트북	PC
내년초	출시	7일	7	일
HP	노트북	LCD	패널	사이즈
12.1인치	12.1	인치	최대	20인치
20	인치	다양	다양화	윈도비스
타	윈도비	스타	태블릿PC	태블릿
PC	라인업	강화	계획	
12.1인치	12.1	인치	이하	PDA
아이팩	중점	두	네트워크	모빌리티
모빌	리티	기능	강화	내비게이
션	기능	등	멀티미디어	기능
추가	계획	울트라모바일		울트라

## 1-2) 색인어 추출 결과 저장

```
aa.c 1-2bat.c
1  #include <stdio.h>
2  #include <string.h>
3
4  int main()
5  {
6      FILE *fp = fopen("bbb.txt", "w"); //파일
7      FILE *read = fopen("fnames.txt", "r"); //
8      char buffer[100]; //한줄 읽을때마다 임시저장
9
10     while(1)
11     {
12         fgets(buffer, sizeof(buffer), read); //
13         buffer[strcspn(buffer, "\n")] = 0;
14         //fgets사용시 뒤에 개행문자가 붙기 때문에
15         if(feof(read)) break; //마지막라인 2번 읽
16         fputs("wordcount.exe -new ", fp);
17         fputs(buffer, fp);
18         fputs(" wordcnt-", fp);
19         fputs(buffer, fp);
20         fputs("\n", fp);
21         //index2018.exe 기사제목 index-기사제목
22     }
23
24     fclose(fp);
25     fclose(read);
26     return 0;
27 }
```

1-1과 동일한 코드로 배치파일 생성 후 색인어와 빈도가 나타난 623개의 파일 생성

(페이지 맞춤을 위해 조금만 캡처함)



- wordcnt-IT-20HPnotebook.txt
- wordcnt-IT-64bitPC.txt
- wordcnt-IT-2007newWebIR.txt
- wordcnt-IT-alanco.txt

```

#include <stdio.h>
#include <string.h>

int main()
{
    FILE *fnames = fopen("fnames.txt", "r"); //fnames.txt 파일 읽기모드로 열기
    char buffer[50]; //파일이름을 받아올 임시용

    char word[9] = "wordcnt-"; //wordcnt-"filename" 형식의 파일을 열기위해
    int index = 1; //DID 부여용

    while(1)
    {
        char filename[100]; //DID를 적을 파일의 이름을 만들기 위한 문자열
        fgets(buffer, sizeof(buffer), fnames); //기사제목을 한줄 읽어옴
        buffer[strcspn(buffer, "\n")] = 0; //fgets 함수의 개행문자제거
        if(feof(fnames)) break; //마지막라인 2번 읽는것 방지
        strcat(filename, word); //filename = wordcnt-
        strcat(filename, buffer); //filename = wordcnt-ITnews...

        FILE *fp = fopen(filename, "r+w"); //filename으로 문서열기
        fseek(fp, 0, SEEK_SET); //파일포인터를 파일의 처음으로 옮김
        fprintf(fp, "DID : %d\n", index); index++; //DID 번호삽입
        fclose(fp);
        filename[0] = '\0'; //다음 파일을 열기위해 filename 초기화
    }

    fclose(fnames);
    return 0;
}

```

DID : 6

1 12월 1 14:15:57 1 1996년 1 1월부터  
 1 2002/09/24 1 2002/09/25 2 2006년 2 2006년  
 까지 12006년으로 2 2007년 1 2007년까지 1 2007년  
 부터 1300만대 1 31일까지 1 3분의 1 85%에  
 2 : 1 <한세희기자 1 FCC로 2 FCC의 1  
 J 1 LA타임스가 1 TV 1 TV는 1 W 1  
 hahn@etnews.co.kr> 1 '방송깃발' 1 "이 1 "정부와 1  
 "정치적 2 ○ 1 가격 1가전업체에 1 가전업체 1  
 가진 1강요한다는 1 강화했다 1 같은 1같이 1개발하도  
 록 1 거의 1거쳐 1것 3 것으로 1 것은 2것을  
 1 것이라는 1 것이었다면 1 경매액은 1 경매에  
 1 경제적 1 계속할 1 골자로 1 공동  
 1 공화당의 1 관련 1관리에 1 관한 1구입해야  
 1 권역의 1 규모의 2 규정하고 2 그러나  
 1 그치고 1 기기에 1 기다릴 3 기술을  
 1 기존 1기존의 2 나섰다 1 내용들을 1  
 내용의 1 논쟁이 1 높은 1당사자 1 대규모  
 1 대부분이 1 대역은 1 대하 1대해 1더 1

### 1-3) 전체 파일들에 대한 색인어들의 문서빈도 계산

```
#include <stdio.h>
#include <string.h>

int main()
{
    FILE *fp = fopen("ccc.txt", "w"); //파일 쓰기
    FILE *read = fopen("fnames.txt", "r"); //파일 읽기
    char buffer[100]; //한줄 읽을때마다 임시저장용 버퍼
    while(1)
    {
        fgets(buffer, sizeof(buffer), read); //기사제목 한줄을 읽어옴
        buffer[strcspn(buffer, "\n")] = 0;
        //fgets사용시 뒤에 개행문자가 붙기 때문에 제거해줌
        if(feof(read)) break; //마지막라인 2번 읽는 것 방지
        fputs("wordcount.exe -new -uniq wordcnt-",fp); //파일이름 형식
        fputs(buffer, fp);
        fputs(" uniq-", fp);
        fputs(buffer, fp);
        fputs("\n", fp);
        //index2018.exe 기사제목 index-기사제목 포맷으로 파일에 쓰기
    }

    fclose(fp);
    fclose(read);
    return 0;
}
```

동일한 코드로 배치파일 생성후 623개의 파일들에 대해 unique 색인어 목록 추출 후

```
C:\Users\Wabcd\Downloads\KLT2010-TestVersion-20180806\WEXE\uniq-ITnews<>>copy *.txt all.txt
uniq -f 2005-2007-ITnews.txt all.txt
```

cmd에서 copy 명령어 사용 후 all.txt 만들어 wordcount 실행한 결과

1	→	1	"1000고지	1	"1분	1	"2라운드	1	"3세대
1			"3차원	1			"6개월내	1	"DRM
1			"HAM의	1			"IT투자	1	"OH
1			"PC	1	"SW	1	"SW·HW·칩까지	1	"SW는
1			"SW도	1			"에코봇 II'이라는	1	"e메일주소
"m비즈	1		"▶▷신혼방◀◀	1			"가격이	1	"개인화
1			"검색	1			"검색순위	1	"검색엔진
서비스와	1		"공개된	2			"구글	1	"구글로부터
"구글은	1		"구글처럼	1			"국어정보화	1	"나
1			"난	1			"눈에	1	"늘어가는
"노인	1		"타고	1			"타고	1	"타고

#### 1-4) TID : term 출력과 <TID, DF> 테이블 구성

<div>table5.txt - 메모장</div> <div>파일(F) 편집(E) 서식(O)</div> <div>1 : "1분 2 : "3세대 3 : "4세대 4 : "DRM 5 : "IPv6 6 : "OH 7 : "SW 8 : "SW는 9 : "'에코봇 II'이라는 10 : "m비즈 11 : "가격이 12 : "검색 13 : "검색엔진 14 : "공개된 15 : "구글로부터 16 : "구글처럼 17 : "나 18 : "난 19 : "눈에 20 : "늘어도 21 : "타고"</div>	<div>1-4.c - 메모장</div> <div>파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)</div> <pre>#include &lt;stdio.h&gt; #include &lt;string.h&gt;  int main() {     FILE *fp = fopen("output.txt", "r");     FILE *output = fopen("table5.txt", "w");     int TID = 1;     int table[40000];      while(!feof(fp))     {         char term[100];         int DF;         fscanf(fp, "%d%s", &amp;DF, &amp;term);         table[TID] = DF;         fprintf(output, "%d : %s\n", TID, term);         TID++;     }      fclose(fp);     fclose(output);     return 0; }</pre>
--	---

좌측의 코드로 우측의 파일을 만들고 table 구성