Inverted File (역파일) 생성

㈜아이브릭스

프로그래밍 테스트

1. 배경

I. 정보검색

정보검색이란 문서 내에 있는 내용, 문서의 메타데이터, 데이터베이스 등에서 정보를 찾는 것을 말한다.

- **데이터 검색**: 구조가 있는 데이터(structured data)의 집합에 대하여 정확한 질의어(query)를 사용하여 조건에 맞는 결과 집합을 얻어내는 것으로, 일반적인 데이터베이스에서 SQL과 같은 문법에 의해 쿼리조건에 맞는 결과를 찾아내는 것을 말한다.
- **문서 검색**: 구조가 없는 데이터(unstructured data)에 대하여 자유로운 형식의 모호성 있는 질의어를 사용하여 관련된 결과 목록을 얻는 것으로, 대표적인 예가 자연어질의에 대하여 관련 있는 순으로 결과 목록을 보여 주는 구글, 네이버 등의 검색엔진을 들 수 있다.

이러한 정보검색시스템은 데이터 수집, 색인, 랭킹, 표현, 사용자 피드백이라는 다섯 가지 요소로 구성되어 있으며, 다섯 가지 요소 중 색인은 아래와 같이 정의된다.

색인이란 문서 데이터에 대하여, 각 단어 별 문서리스트를 생성한 것을 의미하며, 흔히 ***역문서리스트(inverted list)라는 용어로도 표현***된다. 한편 색인방식에 따라 데이터집합을 한꺼번에 색인하는 일괄색인(batch indexing)과 점증색인(incremental indexing)으로 구분될 수 있는데, 뉴스검색은 대표적으로 점증색인을 적용하는 분야이다. 정보검색을 위한 색인과정에서 중요한 것은 주어진 문서에서 색인어를 추출하는 과정인데, 언어적 특성과 상관없이 적용될 수 있는 n-gram 방식과, 자연언어처리의 형태소분석을 통한 방식이 존재한다.

2. 문제 설명

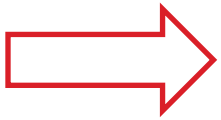
I. Inverted File (역파일) 생성

이전 페이지에서 설명한 역문서 리스트는 물리적인 파일로 존재하게 되며, 검색기에서는 이 파일을 이용하여 빠른 검색을 지원할 수 있게 된다. 이러한 물리적인 파일을 역파일(Inverted File)이라고 부른다.

역파일의 생성 과정은 아래와 같으며,
본 문제는 입력에 대한 역파일을 생성하는 것이다.

Doc	Text
1	It is what it is
2	What? What is is is it
3	It is a banana. It!

역파일



Word	Doc	Frequency
a	3	1
banana	3	1
is	2	3
	1	2
	3	1
it	1	2
	3	2
	2	1
what	1	1
	2	2

3. 프로그램 실행

프로그램은 아래와 같이 두개의 인자를 받아 실행할 수 있어야 한다.

Problem <input file> <output file>

Ex) Problem input output

4. 입력

입력 파일의 라인은 하나의 문서라고 가정한다. 하나의 문서에는 문서ID와 내용이 기록되어 있다.
문서 ID는 숫자로만 이루어져 있으며, 문서ID를 제외한 내용의 길이는 1024 byte를 초과하지 않는다.

입력 예)
1 It is what it is
2 What? What is is is it
3 It is a banana. It!

5. 출력

출력 파일은 역파일 구조이다. 하나의 라인에는 하나의 단어만 기록하며 단어와 관련된 문서ID 및 그 빈도수를 기록해야 한다.
단어는 아래의 출력 예와 같이 오름차순으로 정렬되어 있어야 한다.

또한, 단어와 관련된 문서ID 리스트는 아래 기준으로 정렬되어야 한다.

1) 단어 빈도수의 내림차순
2) 단어 빈도수가 같은 경우에는 문서ID의 오름차순

입력 예에 대한 출력은 아래와 같다.

출력 형식)
[단어] [문서ID] [빈도수] [문서ID] [빈도수]

출력 예)
a 3 1
banana 3 1
is 2 3 1 2 3 1
it 1 2 3 2 2 1
what 2 2 1 1

6. 제약/주의사항

단어는 모두 소문자로 변환하여 빈도수를 계산해야 한다.

단어는 연속으로 이루어진 숫자와 영문자로 구성된다고 가정한다. (특수문자는 제거한다.)

- ✓ I'm a boy ▷ i, m, a, boy
- ✓ inverted index(inverted file) ▷ inverted, index, inverted, file
- ✓ 32-bit OS ▷ 32, bit, os

입력되는 특수문자는 키보드로 입력이 가능한 ASCII 범위로 제한된다.

출력파일에서 단어는 알파벳 순으로 정렬되어 있어야 하며, 문서ID 리스트는 빈도수의 내림차순으로 정렬되어 있어야 한다.

"빈도수가 같은 경우에는 문서ID의 오름차순이여야 한다."

프로그래밍 언어로는 Java, Node.js 중 선택하여 진행할 수 있다.

최종 제출 문서는 big 입력 파일을 수행하여 나온 결과 파일(output.big) 과 실행 파일, 그리고 소스 압축파일이다.

※ 테스트에 제공되는 파일은 본 설명문서와 small 입력파일, small 결과파일, big 입력파일 총 4개의 파일로 구성됩니다.