Homework #2
406.424 인터넷응용
2016년도 봄학기

제출기한: **2016년 4월 16일 23:59까지**

1. (50점) 아래 3가지 조건들을 만족하도록 Galago 토큰화 함수들을 수정하시오.
(어떤 원리로 토큰화 함수가 수정되었는지 각 소문제마다 설명과 함께 MS-Word 문서 파일 형태로 작성하고, TokenizerExample.java 파일을 함께 제출하시오.)

1-1. (20점) 따옴표로 시작해서 따옴표로 끝나는 단어는 따옴표만 없애시오. 그리고 단어 도중에 따옴표가 나오는 경우 따옴표를 포함한 뒤의 글자들을 모두 삭제하시오.
- 예시: 'hello' --> hello, imlab's --> imlab, 'hello'world' --> hello

1-2. (10점) ".com"으로 끝나는 단어는 토큰화되지 않도록 수정하시오.
- 예시: naver.com --> naver.com

1-3. (20점) 마침표(.)로 연결된 단어에서, 마침표 앞, 뒤, 및 사이에 있는 글자가 모두 1개일 경우 마침표를 삭제하고, 2개 이상일 경우 토큰화되지 않도록 수정하시오.
- 예시: i.b.m --> ibm, ieee.803.99 --> ieee.803.99, 127.0.0.1 --> 127.0.0.1

2. (50점) 강의 자료를 참고하여 bible.txt 파일을 2가지 토큰화 함수를 바꾸어가며 사용하여 Zipf's law를 확인하고자 한다.
(결과에 대한 설명을 MS-Word 문서 파일로 작성하고, ZipfsLaw.java 파일을 함께 제출하시오.)

2-1. (20점) Galago 기본 토큰화 함수(TagTokenizer.class)들을 통해 단어들의 출현 빈도와, 그런 빈도를 보이는 단어들의 수를 측정하여 **<이런 횟수를 보이는 단어들의 수, 단어가 나온 횟수>** 의 형태로 나열하고, 이를 바탕으로 Zipf 분포를 그리시오. 그리고 Zipf's law(r*f=k)에서의 k값들을 구하고 주어진 bible.txt가 Zipf's law를 따르는지 설명하시오.

2-2. (20점) 1번 문제에서 수정한 토큰화 함수(TokenizerExample.class)들을 사용하여 (2-1)의 과정을 반복하시오.

2-3. (10점) (2-1)의 결과와 (2-2)의 결과가 어떻게 다른지 서로 비교하시오.

ETL 사이트에 업로드된 예시 코드를 응용하시오. 과제는 MS-Word나 PDF의 파일 형태로, Java 코드 결과물과 함께 압축하여 ETL 사이트에 제출하시오.