고급프로그래밍(B반) 특별과제

주어진 문자열에서 n개의 연속된 부분 문자열을 “n-gram”이라고 한다.

예를 들면, “ALIGAME”이라는 문자열이 주어졌을 때, 이 문자열의 5-gram들은 ALIGA, LIGAM, IGAME이 된다.

특히 1-gram, 2-gram, 3-gram은 unigram, bigram, trigram 이라고 부른다.

이제, 주어진 문단을 입력 받아서, 가장 빈도가 높은 unigram, bigram, trigram을 찾는 프로그램을 작성하시오.

문단 속에는 영문자들로 만들어진 문자열들이 여러 개 나열되어 있으며, 빈칸이나 쉼표, 마침표 등의 문자들은 문자열로 인정하지 않는다.

문단 내에 여러 연속 문자열들에서 가장 빈도가 높은 한 개의 영문자, 연속된 2개짜리 문자열, 3개짜리 문자열을 찾는 것이다.

만약, 가장 빈도가 높은 부분 문자열이 여러 개인 경우, 사전적 순서에서 가장 먼저 나오는 것으로 한다.

[ Input File : 주어진 파일 - **ngramCheck.txt** 또는 **ngramData.txt**]

파일 속에는 여러 개의 문단이 (10개 미만으로 제한할 것) 있고, 하나의 문단은 여러 개의 라인으로 구성된다.

각 문단 첫 줄에는 읽어 들일 라인 수(한자리 숫자로 가정함)가 있고 아래로 그 만큼의 라인이 문단을 구성한다

각각의 라인은 20자를 넘지 않도록 제공되며, 최소한 한 개의 영문자가 존재한다

문자열을 구성하는 영문자 외에 이들 간의 구분자는 space, comma, period, end-of-line(CR, LF)로만 되어 있다.

문단 첫 줄에 **0** (zero)가 있으면, 입력 파일의 끝을 의미한다

[ Output File : 출력 파일명 - **ngramOut.txt** ]

각 문단명 (예를 들면, “Paragragh #1:”)을 먼저 출력한 다음, 문단에서 빈도가 가장 높은 unigram, bigram, trigram을 차례로 출력한다.

다음 문단과는 한 줄을 띄어서 구분한다.

**Sample input (ngramCheck.txt)**

|  |
| --- |
| **1**  z z. z,z. z z. z,z.  0  go go go  go go go  ali ali ali  **3**  a a. a,a.  bc bc  abcd abcd abcd  **0** |

**Sample output (ngramOut.txt)**

|  |
| --- |
| Paragraph #1  unigram: z  bigram: go  trigram: ali  Paragraph #2  unigram: a  bigram: bc  trigram: abc |

※ 주의사항

1) 위 Sample과 같이 ngramCheck.txt 파일로 먼저 프로그램 결과를 확인해 보고,

과제 제출 시에는 ngramData.txt를 입력 파일로 한 결과를 제출바랍니다.

(ngramCheck.txt와 ngramData.txt 파일은 이메일로 모두 전달드리겠습니다.)

2) 문자열이 “AAAAAA”일 때, bigram의 개수는? (“AA” 3개? or “AA” 5개?)

▶ 본 과제에서는 5개의 “AA”라고 판정한다.

|  |
| --- |
| * 과제 제출 안내  1. 평가 배점 : **20점** 2. **제출마감 : 5월31일(일) 23:59까지 LMS 제출** 3. 제출할 내역 (별도 보고서 파일 작성 불필요)    * + - * 이클립스 프로젝트 폴더를 압축하여 제출   (소스파일, 입력파일(ngramData.txt), 출력파일(ngram.Out.txt) 등)   * + - * + LMS 제출 화면에서 Delay coupon 사용 개수 기재 |