Homework

두 개의 영어 텍스트 문서의 파일명이 주어졌을 경우, 두 문서의 텍스트를 읽어온 후, 두 문서의 모든 문장(sentence)들을 서로 비교해서 연속된 단어가 **4개 이상** 나오는 모든 연속 된 단어를 찾는 간단한 표절 검사 프로그램을 작성하라. (단, 문장 구별자로는 **마침표(.), 느낌표(!), 물음표(?)**만 사용하고, 단어를 구별하기 위해서는 우선 영어 알파벳을 제외한 **모든 기호를 스페이스로 변환 후** 스페이스를 기준으로 분리하라. - 마찬가지로, 파이썬의 기본 문자열 함수를 제외한 외부 문장 분석 등의 라이브러리 함수 사용 금지)

- 비교 대상인 두 문서의 파일명은 각각 “c:/d1.txt”, “c:/d2.txt”로 Hard 코딩한다. (즉, 사용자로부터 파일명을 입력받지 않고, 코드에 직접 위의 파일명을 기입한다)
- 실행 결과인 공통 연속 문자열이 4개 이상인 경우, 가장 많은 단어가 연속인 경우만 출력한다(공통 문자열의 subset은 출력하지 않는다.) 예를 들어, 다음과 같은 2개의 문장을 비교할 때, 출력 결과물은 상호 부분집합이 아닌 최대의 공통 문자열을 갖는 2개 문자열만 출력하면 된다.

[비교문장]

가) “The Jupyter Notebook is an open-source web application that allows you to create and share documents”

나) “The Jupyter Notebook is an open-source web application, which allows you to create and share documents”

[실행결과]

1. The Jupyter Notebook is an open source web application
2. allows you to create and share documents