데이터 과학

H01: Word Count

Kookmin University

과제 개요

- 주어진 문서(위키피디아 문장 데이터셋)에서 가장 많이 등장하는 단어 1000개가 각각 몇 번 등장했는지 조사하고, 지프의 법칙을 따르는지 알아보기
- 위키피디아 문장 데이터셋:
  - https://www.kaggle.com/mikeortman/wikipedia-sentences

세부내용

문제 1. 문서를 입력받아 자주 등장하는 단어 1000개를 자주 등장하는 순서대로 출력하는 python 프로그램 작성하기

- 단어의 조건
  - 각 단어에는 영어 알파벳과 숫자만 존재한다고 가정
  - 모든 알파벳은 소문자로 변경
  - 예) 길동 is one of Tom’s best friends.
    → [‘is’, ‘one’, ‘of’, ‘tom’, ‘s’, ‘best’, ‘friends’]
- 입출력
  - 문서의 내용은 stdin으로 입력됨
  - 결과를 stdout으로 출력
  - 결과는 많이 등장한 단어부터 순서대로 출력
    - 각 줄에는 단어와 등장횟수를 tab으로 구분하여 출력

세부내용

문제 2. matplotlib을 활용하여 각 단어의 출현 횟수 순위(x축)와 각 단어의 출현 횟수 (y축)를 log scale에서 출력하고, 지프의 법칙을 따르는지 확인하기

- 지프의 법칙
  - 상수 c와 s에 대해, 단어의 출현 횟수 순위 k와 단어의 출현 횟수 n 사이에 다음의 관계가 있을 경우 지프의 법칙을 따른다고 함

$$n = ck^{-s}$$

  - 주어진 데이터에 대해서 상수 c와 s의 값을 대략적으로 구해보기

세부내용

문제 2. matplotlib을 활용하여 각 단어의 출현 횟수 순위(x축)와 각 단어의 출현 횟수 (y축)를 log scale에서 출력하고, 지프의 법칙을 따르는지 확인하기

- 입출력
  - 문제 1의 결과가 stdin으로 입력됨
  - plt.savefig()를 통해 파일로 결과 그래프 출력
    - 파일명: ‘h01_<이름>_<학번>_plot.png’
  - 대략적으로 계산한 상수 c와 s를 stdout으로 출력

제출

- 다음 세 파일을 압축하여 제출
  - **압축 파일명:** h01_<이름>_<학번>.zip
  - **파일1:** 문제1 파이썬 스크립트 파일 제출
    - 파일명: h01_<이름>_<학번>_wc.py
  - **파일2:** 문제2 파이썬 스크립트 파일 제출
    - 파일명: h01_<이름>_<학번>_plot.py
  - **파일3:** 문제2 결과 그래프 파일 제출
    - 파일명: h01_<이름>_<학번>_plot.png

Questions?