**Project #1**

M1505.001600 정보모델링기법과 응용

2018년도 봄학기

제출기한: **2018년 4월 16일 23:55까지**

1. (40점) 팀별로 주어진 url의 CNN 영어 신문 기사를 crawling하여 다음과 같이 분석하시오.

   1-1. (20점) BeautifulSoup를 사용해 신문 기사의 제목, 저자(여러 명일 경우 모두), 날짜, 영상 제목, 기사 본문 내용을 출력하시오.

   1-2. (5점) 1-1의 결과 얻어낸 기사 본문을 단어 단위로 tokenize하시오.

   1-3. (15점) 1-2의 결과 tokenize된 단어들을 POS_Tagging하고 각 품사의 등장 빈도를 count하여 품사를 빈도 순으로 나열하시오.

| 조 | 배정 url |
|---|---|
| 1 | http://money.cnn.com/2018/04/02/investing/bahrain-oil-field-discovery/index.html |
| 2 | http://money.cnn.com/2018/04/02/investing/spotify-investor-ipo-lead-edge-capital/index.html |
| 3 | http://money.cnn.com/2018/04/02/investing/trump-amazon-trade-war-wall-street/index.html |
| 4 | http://money.cnn.com/2018/04/02/news/economy/china-united-states-november-trade-deals/index.html |
| 5 | http://money.cnn.com/2018/04/02/news/economy/drug-prices-medicare/index.html |
| 6 | http://money.cnn.com/2018/04/02/news/epa-emissions-rules/index.html |
| 7 | http://money.cnn.com/2018/04/02/technology/alibaba-eleme-deal-food-delivery/index.html |
| 8 | http://money.cnn.com/2018/04/02/technology/business/intel-apple-mac/index.html |
| 9 | http://money.cnn.com/2018/04/02/technology/business/spotify-ipo/index.html |
| 10 | http://money.cnn.com/2018/04/02/technology/mark-zuckerberg-tim-cook-facebook-apple/index.html |

2. (30점) 아래 3가지 조건들을 만족하도록 word_tokenizer 함수를 작성하시오.

   2-1. (12점) 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

   2-2. (6점) ".com"으로 끝나는 단어는 토큰화되지 않도록 하시오.

   - 예시: naver.com --> naver.com

   2-3. (12점) 마침표(.)로 연결된 단어에서, 마침표 앞, 뒤, 및 사이에 있는 글자가 모두 1개일 경우 마침표를 삭제하고, 0개 혹은 2개 이상일 경우 토큰화되지 않도록 하시오.

   - 예시: i.b.m --> ibm, ieee.803.99 --> ieee.803.99, 127.0.0.1 --> 127.0.0.1

3. (30점) Zipf의 법칙을 확인할 수 있는 python 코드를 작성하시오.

   3-1. (20점) 주어진 텍스트(bible.txt)로부터 각 단어의 등장 빈도를 count하여 Zipf의 법칙이 성립하는지 확인할 수 있는 python 코드를 작성하시오

   3-2. (10점) 주어진 텍스트 외에 Zipf의 법칙을 만족하는 텍스트를 찾아(출처 자유) Zipf의 법칙을 확인하시오.

   - 텍스트 파일을 읽어 들일 경우 텍스트 파일도 함께 제출

***eTL에 업로드된 예시 코드를 응용하시오. 발표자료 겸 보고서는 PPT나 PDF의 파일 형태로, python 코드 결과물과 함께 압축하여 eTL 사이트에 제출하시오.***