챕터 소스 데이터 분석 – text analytics 2

오혜린/김지솔

1.‘nltk’폴더에 들어가‘nltk.corpus’파일 안에 있는 ‘gutenberg’라는 파일을 열어 그 텍스트 파일을 사용한다.

2. ‘alice’의 값을 gutenberg 파일 중 ‘carroll-alice.txt’로 선택한 후‘sample\_text’의 값은 ‘We will ~’ 이하에 있는 문장으로 설정하면 된다.

3. print len(alice) : ‘alice.txt’에 있는 글자 수를 구해 print.

4. print alice[0:100] : ‘alice.txt’에 있는 텍스트 중, 처음부터 100번째 글자까지 출력한다는 표시이다.

5.‘default\_st’의 값을 ‘nltk.sent\_tokenize’로 설정하고‘alice\_sentences’의 값은 default\_st의 값 안에 있는 텍스트 ‘alice’로 설정한다.‘sample\_sentences’의 값은 default\_st의 값 안에 있는 텍스트 ‘sample\_text’로 설정한다.

6.‘Total sentences in sample\_text에서 ’를 프린트하고 ‘sample\_text’의 글자 수를 출력한다.

7.‘Sample text sentences : -‘를 프린트하고 위의 ‘sample\_text’를 줄에 맞춰 정렬해서 출력

8.‘Total sentences in alice는‘와 alice.txt의 글자 수를 출력한다는 뜻이다.

9.‘First 5 sentences in alice는‘를 출력하고 alice.txt. 에서 첫 5문장을 출력하라는 듯이다.

10.‘nltk.corpus’에서 ‘europar1\_raw’를 연다

11.‘german\_text’의 값을 ‘german’파일에 있는 ‘ep-00-01-17.de’로 설정

12.‘german\_text’에서 100번째 글자까지 출력한다.

13.‘german\_text’로 설정하고 언어는 ‘german’으로 한다.

14.‘german\_tokenizer’의 값은 ‘nltk.data.load’에서 가져온다.‘german\_sentences’는 위의 tokenizer에서 ‘german\_text’를 기본값으로 가진다.

15.‘german\_tokenizer’의 변수형을 출력 class ‘german\_tokenizer’을 출력

16.‘german\_sentences’의 문장 5줄을 출력한다.

17.‘punkt\_st’의 값을 ‘nltk.tokenize.PunktSentenceTokenizer()’로 설정하고

위에 입력한 ‘sample\_sentences’의 값은 ‘punkt\_st.tokenize(sample\_text)로 설정

18.‘SENTENCE\_TOKENS\_PATTERN’을 전체 모든 기호 및 문자로 설정

19. sentence의 값을 "The brown fox wasn't that quick and he couldn't win the race" 로 정해준다.

20. default\_wt의 값을 nltk.word\_tokenize로 정해주고 words의 값을 default\_wt(sentence)로 정해준다 따라서 word를 print했을 때 단어별로 문장을 나눠서 출력해준다.

e.g. ['The', 'brown', 'fox', 'was', "n't", 'that', 'quick', 'and', 'he', 'could', "n't", 'win', 'the', 'race']

21. word는 treebanck\_wt.tokenize(sentence)로 정해준 후 print 했을 경우 문장하나하나를 구문 분석하여 코퍼스로 구성해 단어별로 문장을 나눠서 출력해준다.

e.g. ['The', 'brown', 'fox', 'was', "n't", 'that', 'quick', 'and', 'he', 'could', "n't", 'win', 'the', 'race']

The rest was so difficult to analyzed.

.

.

.