<빅데이터 최신기술 1차과제>

한글 문장의 유사도 계산

20153167 김현중

공통 어절 개수에 의한 유사도 계산.

```
def uzul(a, b):
    uzul_tokenized_a = a.split() # 문장 1을 어절 단위로 쪼개기
    uzul_size_a=len(uzul_tokenized_a) # 문장 1의 어절 개수
    uzul_tokenized_b = b.split() # 문장 2을 어절 단위로 쪼개기
    uzul_size_b=len(uzul_tokenized_b) # 문장 2의 어절 개수
    uzul_intersection = set(uzul_tokenized_a).intersection(set(uzul_tokenized_b))  # 문장 1과 문장 2에서의 교집합 어절 배열
    uzul_size_intersection = len(uzul_intersection)  # 문장 1과 문장 2의 공통된 어절 개수
    if uzul_size_a <= uzul_size_b:
        uzul_short_doc = uzul_size_a
    else:
        uzul_short_doc = uzul_size_b  ## 문장 1과 문장 2의 어절 개수를 비교하여 짧은 문자의 어절 개수를 short_doc에 저장
    simillarity = (float(uzul_size_intersection) / float(uzul_short_doc))  ## 공통 어절 개수 / 짧은 문장 어절 개수 = 유사도
    return simillarity*100   ##어절 유사도 반환
```

구현 Step

1. uzul이라는 함수에서는 두개의 문장을 어절 단위로 쪼갠다
   (split 함수 이용)

2 . 쪼개진 두개의 어절 배열을 비교하여 교집합 배열을 구한다
   (intersection 이용)

3. 교집합 배열의 크기를 구해준다
(교집합 배열의 크기는 공통된 어절의 개수가 된다)

4. 두개의 어절 배열 크기를 비교하여 크기가 더 작은 배열이 짧은 문장의 배열이 된다.

5. 유사도 식 (공통된 어절 개수 / 짧은 문장의 어절 개수) 을 이용하여 유사도를 반환해준다.

공통 음절 개수에 의한 유사도 계산.

```
def umzul(a, b):
    umzul_tokenized_a = set(a) # 문장 1을 음절 단위로 쪼개기절
    umzul_size_a=len(umzul_tokenized_a) # 문장 1의 음절 개수
    umzul_tokenized_b = set(b) # 문장 2를 음절 단위로 쪼개기
    umzul_size_b=len(umzul_tokenized_b) # 문장 2의 음절 개수
    umzul_intersection = set(umzul_tokenized_a).intersection(set(umzul_tokenized_b)) # 문장 1과 문장 2에서의 교집합 음절 배열
    umzul_size_intersection=len(umzul_intersection) # 문장 1과 문장 2의 공통된 음절 개수
    if umzul_size_a <= umzul_size_b:
        umzul_short_doc = umzul_size_a
    else:
        umzul_short_doc = umzul_size_b  ## 문장 1과 문장 2의 음절 개수를 비교하여 짧은 문자의 음절 개수를 short_doc에 저장
    simillarity = (float(umzul_size_intersection) / float(umzul_short_doc))  ## 공통 음절 개수 / 짧은 문장 음절 개수 = 유사도
    return simillarity*100   ##음절 유사도 반환
```

구현 Step
1. umzul이라는 함수에서는 두개의 문장을 어절 단위로 쪼갠다
(set 함수 이용)

2 . 쪼개진 두개의 음절 배열을 비교하여 교집합 배열을 구한다
(intersection 이용)

3. 교집합 배열의 크기를 구해준다
(교집합 배열의 크기는 공통된 음절의 개수가 된다)

4. 두개의 음절 배열 크기를 비교하여 크기가 더 작은 배열이 짧은 문장의 배열이 된다.

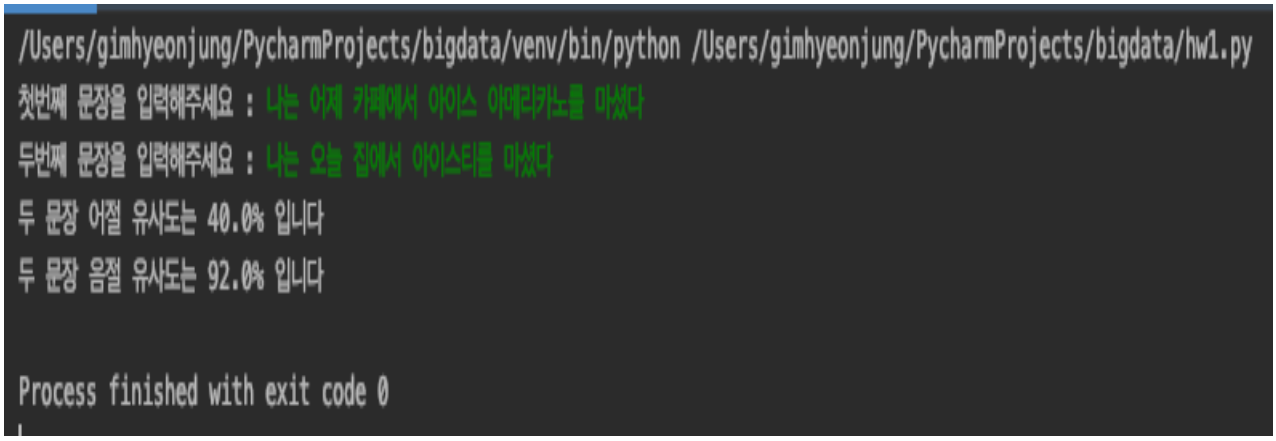
5. 유사도 식 (공통된 음절 개수 / 짧은 문장의 음절 개수) 을 이용하여 유사도를 반환해준다.

메인 함수

```
doc1 = raw_input("첫번째 문장을 입력해주세요 : ")
doc2 = raw_input("두번째 문장을 입력해주세요 : ")
print("두 문장 어절 유사도는 " + str(uzul(doc1,doc2))+"% 입니다")
print("두 문장 음절 유사도는 " + str(umzul(doc1,doc2))+"% 입니다")
```

공통어절, 공통음절 개수에 의한 유사도 함수를 작성하였으므로 두개의 문장을 입력 받아 함수를 호출하여 준다.

출력 결과



최종적으로 두 문장을 입력하게 되면
어절 유사도와 음절 유사도가 나오게 된다.