search_engine Report

Python programming and practice

223479 Minchae Choi

1. Introduction

A. Project Purpose and Background

To review and practice what I learned in 'Python programming and practice' class.

B. Goal

Developing a search engine that returns similarity scores when given a query from user.

2. Requirements

A. User requirements

Should return sentences that is similar to the user's query

B. Functional Requirements

- Preprocess sentences within the search target and store them in a list.
- Receive an input English string (query) from the user and preprocess it.
- Calculate the similarity between the query and sentences within the search target. (Similarity is based on the count of the same "word.")
- Rank the sentences based on similarity.
- Output the top 10 ranked sentences to the user from the ranked sentences.

3. Design and Implementation

A. Implementation Details

- Preprocess sentences within the search target and store them in a list.

```
def preprocess(sentence):
    preprocessed_sentence = sentence.strip().split(" ") # make tokens by splitting
    return preprocessed_sentence

def indexing(file_name):
    file_tokens_pairs = []
    lines = open(file_name, "r", encoding="utf8").readlines() # read the file and add line by line in lines
    for line in lines:
        tokens = preprocess(line)
        file_tokens_pairs.append(tokens) # add tokens in file_tokens_pairs
    return file_tokens_pairs
```

(1) input : file

(2) output : file_tokens_pairs (2D array of tokens for each word in sentences)

(3) explanation : Get lines list from readlines(), and then make tokens list by using preprocess function line by line. In preprocess function, trimming the sentence with strip(), and then create the tokens list with split(" "), which divides the string based on spaces.

- Receive an input English string (query) from the user and preprocess it.

```
# 2. Input the query
query = input("영어 쿼리를 입력하세요.")
preprocessed_query = preprocess(query)
```

(1) input : query from user

(2) output : preprocessed_query

(3) explanation : trim and split the input query by preprocess function

- Calculate the similarity between the query and sentences within the search target.

```
def calc_similarity(preprocessed_query, preprocessed_sentences):
    score_dict = {}
    for i in range(len(preprocessed_sentences)):

        # exception for case sensitivity
        sentence = preprocessed_sentences[i]
        query_str = ' '.join(preprocessed_query).lower()
        sentence_str = ' '.join(sentence).lower()
        preprocessed_query = set(preprocess(query_str))
        preprocessed_sentence = preprocess(sentence_str)

        # Calculate the score of similarity
        file_token_set = set(preprocessed_sentence)
        all_tokens = preprocessed_query | file_token_set
        same_tokens = preprocessed_query & file_token_set
        similarity = len(same_tokens) / len(all_tokens)
        score_dict[i] = similarity
    return score_dict
```

(1) Input : preprocessed_query, preprocessed_sentences

(2) Output : score_dict

(3) Explanation : To ignore case sensitivity, we need to convert both the query and sentence to lowercase. However, the lower() function cannot be applied to a list type, so we must use the join function to convert the list into a string. After the conversion, the string must be turned into a set to calculate similarity. Similarity is calculated by taking the intersection of the query and the file, and dividing its length by the union's length of the query and the file.

- Rank the sentences based on similarity.

```
# 5. Print the result
if sorted_score_list[0][1] == 0.0:
    print("There is no similar sentence.")
else:
    print("rank", "Index", "score", "sentence", sep = "₩t")
    rank = 1
    for i, score  in sorted_score_list:
        print(rank, i, score, ' '.join(file_tokens_pairs[i]), sep = "₩t")
        if rank == 10:
            break
        rank = rank + 1
```

(1) Input: sorted_score_list (A sorted list of key values from score_dict)

(2) Output: None / Print the sentences and scores with the top 10 similarity scores. If there are no similar sentences, print "There is no similar sentence."

(3) Explanation: If the score of the first item in the sorted list is 0, it is determined that there are no similar sentences, and a message is displayed. If that's not the case, then from 1st place to 10th place, display the rank, index of the sentence, and sentence.

4. Testing

A. Test Resutls for Each Functionality

- Preprocess sentences within the search target and store them in a list.

```
["You'll", 'be', 'picking', 'fruit', 'and', 'generally', 'helping', 'us', 'do', 'all', 'the', 'usual', 'farm', 'work.']
['In', 'the', 'Middle', 'Ages,', 'cities', 'were', 'not', 'very', 'clean,', 'and', 'the', 'streets', 'were', 'filled', 'with', 'garbag
e.']
['For', 'the', 'moment', 'they', 'may', 'yet', 'be', 'hiding', 'behind', 'their', 'apron', 'strings,', 'but', 'sooner', 'or', 'later',
'their', 'society', 'will', 'catch', 'up', 'with', 'the', 'progressive', 'world.']
['Do', 'you', 'know', 'what', 'the', 'cow', 'answered?"', 'said', 'the', 'minister.']
['Poland', 'and', 'Italy', 'may', 'seem', 'like', 'very', 'different', 'countries.']
['Mr.', 'Smith', 'and', 'I', 'stayed', 'the', 'whole', 'day', 'in', 'Oxford.']
['The', 'sight', 'of', 'a', 'red', 'traffic', 'signal', 'gave', 'him', 'an', 'idea.']
['So', 'they', 'used', 'pumpkins', 'instead.']
['2.', 'a', 'particular', 'occasion', 'of', 'state', 'of', 'affairs:', 'They', 'might', 'not', 'offer', 'me', 'much', 'money.']
["I'm", 'especially', 'interested', 'in', 'learning,', 'horse-riding', 'skills,', 'so', 'I', 'hope', "you'll", 'include', 'informatio
n', 'about', 'this.']
['Instead,', 'the', 'devil', 'gave', 'him', 'a', 'single', 'candle', 'to', 'light', 'his', 'way', 'through', 'the', 'darkness.']
['It', 'shines', 'over', 'the', 'sea.']
['He,', 'too,', 'was', 'arrested,', 'and', 'a', 'bomb', 'was', 'thrown', 'at', 'his', 'house.']
['It', 'seems', 'that', 'the', 'high', 'temperature', 'and', 'pressure', 'on', 'the', 'star', 'made', 'its', 'carbon', 'surface', 'tur
```

- Receive an input English string (query) from the user and preprocess it.

```
영어 쿼리를 입력하세요.hello my name is minchae
['hello', 'my', 'name', 'is', 'minchae']
```

- Calculate the similarity between the query and sentences within the search target.

```
영어 쿼리를 입력하세요.hello my name is minchae
{0: 0.0, 1: 0.0, 2: 0.0, 3: 0.0, 4: 0.0, 5: 0.0, 6: 0.0, 7: 0.0, 8: 0.0, 9: 0.0, 10: 0.0, 11: 0.0, 12: 0.0, 13: 0.0, 14: 0.0, 15: 0.0,
16: 0.0, 17: 0.09090909090909091, 18: 0.0, 19: 0.0, 20: 0.07692307692307693, 21: 0.041666666666666664, 22: 0.0, 23: 0.0588235294117647
05, 24: 0.05555555555555555, 25: 0.0, 26: 0.0, 27: 0.0625, 28: 0.0, 29: 0.0, 30: 0.0, 31: 0.1, 32: 0.0, 33: 0.0, 34: 0.0, 35: 0.0, 36:
0.0, 37: 0.0, 38: 0.0, 39: 0.0, 40: 0.0, 41: 0.0, 42: 0.0, 43: 0.0, 44: 0.0, 45: 0.1111111111111111, 46: 0.0, 47: 0.0, 48: 0.0, 49: 0.
0, 50: 0.08333333333333333, 51: 0.08333333333333333, 52: 0.0, 53: 0.0, 54: 0.0, 55: 0.05263157894736842, 56: 0.0, 57: 0.0, 58: 0.0, 5
9: 0.0, 60: 0.0, 61: 0.0, 62: 0.0, 63: 0.0, 64: 0.0, 65: 0.05263157894736842, 66: 0.0, 67: 0.0, 68: 0.0, 69: 0.03571428571428571, 70:
0.06666666666666667, 71: 0.0, 72: 0.0, 73: 0.07692307692307693, 74: 0.0, 75: 0.06666666666666667, 76: 0.0, 77: 0.08333333333333333, 7
8: 0.07692307692307693, 79: 0.0, 80: 0.047619047619047616, 81: 0.0, 82: 0.06666666666666667, 83: 0.0, 84: 0.0, 85: 0.0, 86: 0.0, 87:
0.0, 88: 0.0, 89: 0.0, 90: 0.0, 91: 0.0, 92: 0.0, 93: 0.0, 94: 0.0, 95: 0.0, 96: 0.05, 97: 0.0, 98: 0.041666666666666664, 99: 0.0, 10
0: 0.0, 101: 0.0, 102: 0.0, 103: 0.0, 104: 0.0, 105: 0.0, 106: 0.05, 107: 0.1111111111111111, 108: 0.0, 109: 0.0, 110: 0.0, 111: 0.071
42857142857142, 112: 0.07142857142857142, 113: 0.0, 114: 0.0, 115: 0.0, 116: 0.0, 117: 0.0, 118: 0.0, 119: 0.06666666666666667, 120:
0.07142857142857142, 121: 0.0, 122: 0.0, 123: 0.0, 124: 0.0, 125: 0.0, 126: 0.0, 127: 0.0, 128: 0.038461538461538464, 129: 0.05, 130:
0.0, 131: 0.0, 132: 0.0, 133: 0.0, 134: 0.0, 135: 0.0, 136: 0.0, 137: 0.0, 138: 0.09090909090909091, 139: 0.0, 140: 0.0, 141: 0.0, 14
2: 0.0, 143: 0.0, 144: 0.0, 145: 0.04, 146: 0.0, 147: 0.0, 148: 0.0, 149: 0.0, 150: 0.0, 151: 0.0, 152: 0.0, 153: 0.0, 154: 0.0, 155:
0.0, 156: 0.0, 157: 0.0, 158: 0.0, 159: 0.0, 160: 0.0, 161: 0.0, 162: 0.0, 163: 0.05263157894736842, 164: 0.0, 165: 0.0, 166: 0.0, 16
7: 0.0, 168: 0.0, 169: 0.0, 170: 0.0, 171: 0.0, 172: 0.0, 173: 0.09090909090909091, 174: 0.043478260869565216, 175: 0.0, 176: 0.0, 17
```

- Rank the sentences based on similarity.

```
영어 쿼리를 입력하세요.hello my name is minchae
rank    Index   score   sentence
1       679     0.5     My name is Mike.
2       526     0.2857142857142857      Bob is my brother.
3       538     0.2857142857142857      My hobby is traveling.
4       453     0.25    My mother is sketching them.
5       241     0.2222222222222222      My father is running with So-ra.
6       336     0.2222222222222222      My family is at the park.
7       212     0.2     My sister Betty is waiting for me.
8       505     0.18181818181818182     My little sister Annie is five years old.
9       610     0.15384615384615385     I would raise my voice and yell, "LUNCH IS READY!"
10      190     0.14285714285714285     It is Sunday.
11      314     0.14285714285714285     This is Washington.
12      710     0.14285714285714285     Travel is exciting.
13      45      0.11111111111111111     This method is called *acupuncture.
14      107     0.11111111111111111     But this is very interesting.
15      293     0.11111111111111111     B : When is it?
16      519     0.11111111111111111     Taking pictures is his job.
17      597     0.11111111111111111     The earth is in danger.
```

- Output the top 10 ranked sentences to the user from the ranked sentences.

```
영어 쿼리를 입력하세요.hello my name is minchae
rank    Index   score   sentence
1       679     0.5     My name is Mike.
2       526     0.2857142857142857      Bob is my brother.
3       538     0.2857142857142857      My hobby is traveling.
4       453     0.25    My mother is sketching them.
5       241     0.2222222222222222      My father is running with So-ra.
6       336     0.2222222222222222      My family is at the park.
7       212     0.2     My sister Betty is waiting for me.
8       505     0.18181818181818182     My little sister Annie is five years old.
9       610     0.15384615384615385     I would raise my voice and yell, "LUNCH IS READY!"
10      190     0.14285714285714285     It is Sunday.
```

B. Final Test Screenshot

```
영어 쿼리를 입력하세요.hello my name is minchae
rank    Index   score   sentence
1       679     0.5     My name is Mike.
2       526     0.2857142857142857      Bob is my brother.
3       538     0.2857142857142857      My hobby is traveling.
4       453     0.25    My mother is sketching them.
5       241     0.2222222222222222      My father is running with So-ra.
6       336     0.2222222222222222      My family is at the park.
7       212     0.2     My sister Betty is waiting for me.
8       505     0.18181818181818182     My little sister Annie is five years old.
9       610     0.15384615384615385     I would raise my voice and yell, "LUNCH IS READY!"
10      190     0.14285714285714285     It is Sunday.
```

```
영어 쿼리를 입력하세요.Hello
There is no similar sentence.
```

5. Resutls and conclusion

A. Result

developed a basic search engine

B. Conclusion

could learn about the overview of search engines