AnsweringPhraseQueries

September 7, 2023

[13]:
```
import re
import os
from collections import defaultdict
from progressbar import progressbar
```

[109]:
```
DATA_PATH = "f:/Datasets/bbc"
```

[110]:
```
def tokenize_text(text: str, min_len: int=3, to_lower=True)->list:
    pattern = re.compile(r"[A-Za-z0-9]+[-|']{0,1}[A-Za-z0-9]+")
    if to_lower == True:
        return [w.lower() for w in re.findall(pattern, text) if len(w)>=min_len]
    else:
        return [w for w in re.findall(pattern, text) if len(w)>=min_len]
```

[111]:
```
doc_id = 0
inverted_index = defaultdict(list)
all_texts= []
for foldername in progressbar(os.listdir(DATA_PATH)):
    full_folder_path = os.path.join(DATA_PATH, foldername)
    if os.path.isdir(full_folder_path):
        filenames = os.listdir(full_folder_path)
        for filename in filenames:
            full_file_path = os.path.join(full_folder_path, filename)
            with open(full_file_path) as fp:
                text = fp.read()
                all_texts.append(text)
                tokens = tokenize_text(text)
                bigrams = [tokens[i] + " " + tokens[i+1] for i in␣
 ↪range(len(tokens)-1)]
                for token in tokens:
                    if doc_id not in inverted_index[token]:
                        inverted_index[token].append(doc_id)
                for token in bigrams:
                    if doc_id not in inverted_index[token]:
                        inverted_index[token].append(doc_id)
                # break
            doc_id += 1
    # break
```

```
print(str(len(all_texts)) + " documents indexed in " + str(len(foldername)) + "␣
 ↪folders")
```

```
100% (6 of 6) |#########################| Elapsed Time: 0:00:13 Time:  0:00:13
```

```
2225 documents indexed in 4 folders
```

[103]:
```
def intersect(p1: list, p2: list)->list:
    results = []
    i = 0
    j = 0
    while i < len(p1) and j < len(p2):
        if p1[i] == p2[j]:
            results.append(p1[i])
            i += 1
            j += 1
        elif p1[i] < p2[j]:
            i += 1
        else:
            j += 1
    return results
```

[118]:
```
def extended_intersect(query_terms: list)->list:
    query_terms = sorted(query_terms, key=lambda x:len(inverted_index[x]))
    # print(query_terms)
    results = inverted_index[query_terms[0]]
    for i in range(1, len(query_terms)-1):
        results = intersect(results, inverted_index[query_terms[i]])
    return results
```

[119]:
```
def parse_phrase_query(query: str)->list:
    query_tokens = tokenize_text(query)
    # print(query_tokens)
    if len(query_tokens) > 2:
        return [query_tokens[i] + " " + query_tokens[i+1] for i in␣
 ↪range(len(query_tokens)-1)]
    elif len(query_tokens) == 2:
        return [query_tokens[0] + " " + query_tokens[1]]
    else:
        return [query]
```

[121]:
```
%%timeit
extended_intersect(parse_phrase_query("world business report"))
```

```
19.1 µs ± 1.14 µs per loop (mean ± std. dev. of 7 runs, 10,000 loops each)
```

[115]:
```
all_texts[1005]
```

[115]:
```
'UK heading wrong way - Howard\n\nTony Blair has had the chance to tackle the
problems facing Britain and has failed, Michael Howard has said.\n\n"Britain is
heading in the wrong direction", the Conservative leader said in his New Year
message. Mr Blair\'s government was a "bossy, interfering government that takes
decisions that should be made by individuals," he added. But Labour\'s campaign
spokesman Fraser Kemp responded: "Britain is working, don\'t let the Tories
wreck it again". Mr Howard also paid tribute to the nation\'s character for its
generous response to the Asian quake disaster. The catastrophe was overshadowing
the hopes for the future at this usually positive time of the year, Mr Howard
said.\n\n"We watched the scenes of destruction with a sense of disbelief. The
scale, the speed, the ferocity of what happened on Boxing Day is difficult to
grasp. "Yet Britain\'s response has shone a light on our nation\'s character.
The last week has shown that the warm, caring heart of Britain beats as strong
as ever." He went on to reflect on the values that "most Britons hold dear".
Looking ahead to the coming general election, he pledged to "turn these beliefs
into reality" and set out the choices he says are facing Britain. "How much tax
do people want to pay? Who will give taxpayers value for money, the clean
hospitals and good, disciplined schools they want? "Who can be trusted to get a
grip on the disorder on our streets and the chaos in our immigration
system?"\n\nMr Blair has failed to tackle these problems, he claimed, saying he
has the "wrong solution" to them.\n\n"The result is big government and higher
taxes eroding incentives, undermining enterprise and denying people choice.
"Worst of all, it is a government that has wasted people\'s money and failed to
tackle the problems families face today." The Tories, he said, can cut crime and
improve public services without asking people to pay more taxes. "We can have
progress without losing what makes Britain great - its tolerance, the respect
for the rule of law, the ability of everyone to fulfil their potential. "We
simply need to change direction. The election will give Britain the chance to
change." This is the record Mr Blair will have to defend in the coming months,
he said, urging voters to hold him to account.\n\nBut Labour spokesman Mr Kemp
said: "It would be more appropriate for this message to come out on 1 April, not
1 January." "Let us never forget that when Michael Howard was in government
Britain suffered mass unemployment, 15% interest rates, record home
repossessions, and the introduction of the poll tax. "With Labour Britain is
working. Rather than alluding to false promises Michael Howard should be
starting 2005 with an apology to the British people for the misery that the
government, of which he was a member, inflicted upon the country.\n'
```

[ ]: