به نام یکدانه دردانه

سید عارف طباطبایی 9831040

فاز 1:

زیر بخش 1:

ابتدا با json.load داک‌های خبری را می‌خوانیم و آن‌ها را به صورت tuple های سه تایی شامل title، url و content در می‌آوریم.

در هر content، ابتدا punctuation ها را حذف می‌کنیم. سپس با استفاده از کتابخانه parsivar، متن مربوطه را ابتدا normalize و سپس tokenize می‌کنیم. سپس توکن‌ها را ریشه‌یابی می‌کنیم و در ادامه نیز stopwords را حذف می‌کنیم. برای مثال خواهیم داشت:

به ازای متن ورودی:

```
'!تقدیم به تو . امیدوارم امروز حالت خوب باشه ?سلام خوبی'
```

نتیجه:

```
'سلام', 'خوبی', 'امیدوار', 'امروز', 'حالت', 'خوب', 'باشه', 'تقدیم', 'تو'
```

زیر بخش 2:

ساختمان شاخص مکانی را با استفاده از تابع defaultdict در کتابخانه collections ایجاد می‌کنیم:

```
# Create the inverted index
index = defaultdict(lambda: {'num': 0, 'positions': defaultdict(lambda: [0, []])})
```

شامل num و positions که در اینجا num به تعداد بار استفاده از یک کلمه در تمام doc ها و positions به تعداد بار استفاده از آن کلمه در هر doc و موقعیت مکانی‌های آن‌ها اشاره دارند.

مطابق تکه کد زیر تمام doc ها را زیر و رو کرده و شاخص مکانی را می‌سازیم:

```
for i, doc in preprocessed_data.items():
    if int(i) % 1000 == 0:
        print(f'{int(i)} have been processed.')
    content = doc['content']
    for j, token in enumerate(content):
        index[token]['num'] += 1
        index[token]['positions'][i][0] += 1
        index[token]['positions'][i][1].append(j)
```

در انتها مطابق زیر برای مثال، شاخص مکانی را برای کلمه خبر چاپ می‌کنیم:

```
term = index['خبر']
print(f"Total frequency: {term['num']}")
print("Positions:")
for doc_id, positions in term['positions'].items():
    print(f"  Document {doc_id}: total_number: {positions[0]}, positions: {positions[1]}")
print()
```

نتیجه:

```
Total frequency: 1881
Positions:
  Document 14: total_number: 1, positions: [6]
  Document 19: total_number: 1, positions: [13]
  Document 29: total_number: 1, positions: [420]
  Document 48: total_number: 1, positions: [360]
  Document 56: total_number: 1, positions: [131]
  Document 78: total_number: 2, positions: [12, 25]
  Document 79: total_number: 1, positions: [55]
  Document 135: total_number: 1, positions: [26]
  Document 142: total_number: 1, positions: [838]
  Document 148: total_number: 2, positions: [320, 682]
  Document 150: total_number: 1, positions: [22]
  Document 200: total_number: 1, positions: [31]
  Document 203: total_number: 1, positions: [67]
  Document 206: total_number: 1, positions: [26]
  Document 211: total_number: 2, positions: [18, 61]
  Document 216: total_number: 2, positions: [124, 135]
  Document 218: total_number: 2, positions: [32, 93]
  Document 219: total_number: 1, positions: [80]
  Document 220: total_number: 1, positions: [94]
  Document 222: total_number: 1, positions: [25]
  Document 223: total_number: 2, positions: [62, 86]
  Document 226: total_number: 1, positions: [13]
```

```
  Document 236: total_number: 1, positions: [122]
  Document 238: total_number: 1, positions: [154]
  Document 259: total_number: 1, positions: [64]
  Document 300: total_number: 2, positions: [50, 55]
  Document 304: total_number: 2, positions: [150, 244]
  Document 306: total_number: 1, positions: [163]
  Document 314: total_number: 8, positions: [54, 58, 73, 77, 98, 154, 158, 172]
  Document 319: total_number: 1, positions: [198]
  Document 353: total_number: 1, positions: [33]
  Document 354: total_number: 1, positions: [338]
```

و....