|  |
| --- |
| به نام خداوند جان و خرد  پروژه – فاز ۱  نام درس : مبانی بازیابی اطلاعات  استاد درس: دکتر زهرا زجاجی  حل تمرین‌ها: محمدرضا توکلیان  مهلت تحویل: 03/04/1401  سامانه تحویل: lms.ui.ac.ir |

در این پروژه شما به پیاده‌سازی عملی مطالبی که در کلاس، فصل ساخت ایندکس‌ها آموخته‌اید میپردازید.

در اولین فاز این پروژه شما باید برای مجموعه داده‌ای که به داده شده است، ایندکس بسازید.

طبق مطالبی که در کلاس درس گفته شده است، در این فاز شما باید با استفاده از روش BSBI[[1]](#footnote-1) ایندکس مورد نظر را ساخته و تمام اسنادی که داده شده است را وارد ایندکس نمایید. برنامه‌ی نهایی شما باید بتواند یک کوئری را از کاربر گرفته و اسناد مربوط به آن کوئری را بازیابی کند. در ادامه به جزئیات پروژه و چگونگی پیادهسازی و ارزیابی پروژه می‌پردازیم.

در بخش ۴.۲ کتاب آمده است که برای برای بهینه سازی ساخت ایندکس اولین گام استفاده از termID به جای term است که در واقع باید از اعداد به جای رشته استفاده شود. پس در اولین مرحله شما باید کلاس IdMap را پیاده سازی کنید. این کلاس باید بتواند یک string را به مقدار عددی تبدیل کند.

class IdMap:  
 *"""Helper class to store a mapping from strings to ids."""* def *\_\_init\_\_*(*self*):  
 *self*.str\_to\_id = {}  
 *self*.id\_to\_str = []  
  
 def *\_\_len\_\_*(*self*):  
 *"""Return number of terms stored in the IdMap"""* return *len*(*self*.id\_to\_str)  
  
 def \_get\_str(*self*, *i*):  
 *"""Returns the string corresponding to a given id (`i`)."""  
 ### Begin your code  
  
 ### End your code* def \_get\_id(*self*, *s*):  
 *"""Returns the id corresponding to a string (`s`).   
 If `s` is not in the IdMap yet, then assigns a new id and returns the new id.  
 """  
 ### Begin your code  
  
 ### End your code* def *\_\_getitem\_\_*(*self*, *key*):  
 *"""If `key` is a integer, use \_get\_str;   
 If `key` is a string, use \_get\_id;"""* if *type*(*key*) is *int*:  
 return *self*.\_get\_str(*key*)  
 elif *type*(*key*) is *str*:  
 return *self*.\_get\_id(*key*)  
 else:  
 raise *TypeError*

در گام بعدی برای بهینه سازی ذخیره سازی posting list ها شما باید کلاس UncompressedPostings را پیاده‌سازی کنید. این کلاس باید بتوان posting list ها به صورت bytearray در دیسک ذخیره کند.

import array  
class UncompressedPostings:  
  
 @staticmethod  
 def encode(*postings\_list*):  
 *"""Encodes postings\_list into a stream of bytes  
  
 Parameters  
 ----------  
 postings\_list: List[int]  
 List of docIDs (postings)  
  
 Returns  
 -------  
 bytes  
 bytearray representing integers in the postings\_list  
 """* return array.array('L', *postings\_list*).tobytes()  
  
 @staticmethod  
 def decode(*encoded\_postings\_list*):  
 *"""Decodes postings\_list from a stream of bytes  
  
 Parameters  
 ----------  
 encoded\_postings\_list: bytes  
 bytearray representing encoded postings list as output by encode  
 function  
  
 Returns  
 -------  
 List[int]  
 Decoded list of docIDs from encoded\_postings\_list  
 """* decoded\_postings\_list = array.array('L')  
 decoded\_postings\_list.frombytes(*encoded\_postings\_list*)  
 return decoded\_postings\_list.tolist()

# to test to your implementation:

x = UncompressedPostings.encode([1,2,3])  
*print*(x)  
*print*(UncompressedPostings.decode(x))

اصلی‌ترین کلاسی که در این پروژه شما پیاده سازی میکنید کلاس InvertedIndex و کلاس BSBIIndex است. توضیحات هر متد این کلاس‌ها در قسمت کامنت کدها آمده است. در فایل main.py اجرای نهایی پروژه شما آمده است. در این فایل یک instance از روی کلاس BSBIIndex ساخته می‌شود و با فراخوانی متد retrieve و ورودی یک کوئری اسنادی که در مرحله ساخت ایندکس ساخته اید را بازیابی کند.

* لطفا ساختار توابع را تغییر ندهید. اضافه کردن متدها و ماژول های اضافه مجاز است.
* به خوانایی و تمیزی کد نمره اختصاص داده خواهد شد

1. Blocked sort-based indexing [↑](#footnote-ref-1)