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: ۱۴۰۰/۰۴/۰۵

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] 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
```

[1] Blocked sort-based indexing

```
        ### 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.