605.744: Information Retrieval
Programming Assignment #2: Inverted Files

Sabbir Ahmed

September 18, 2022

1 Introduction

This paper describes the enhancements and features added to the Information Retrieval program started in Assignment 1. Modifications include improvement in performance and efficiency in normalizing text and generating statistics from the pre-generated corpus and addition of binary inverted files.

2 Technical Background

All of the source code is in Python 3.10. The program is split into several modules and follows an object oriented structure. The following is the directory structure of the source code:

The source code for all of the files are attached in Appendix A.

The total number of non-empty lines of code for the program doubled to just under 400. However, with an average execution time to process the sample files reduced to 37 seconds.

2.1 Existing Classes

2.1.1 Driver

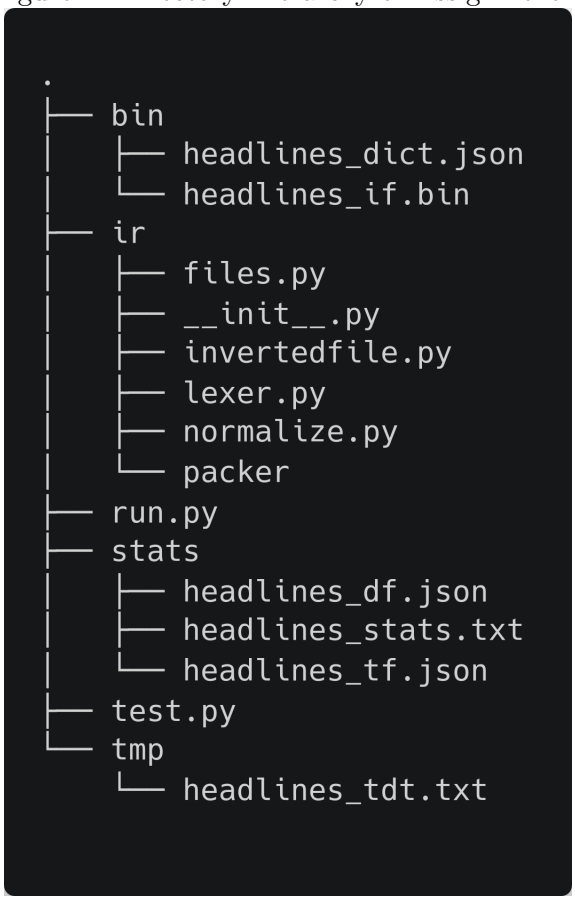
The driver script for the program is split between `run.py` and `test.py`. The former script, while maintaining the responsibility for the same list of tasks as the previous iteration, also generates the binary inverted files and dictionaries for the sample corpus and saves it to disk. The latter script loads the inverted files and dictionaries and looks up the test terms as specified in the prompt.

2.1.2 `normalize.Normalizer`

The `Normalizer` class received the following modifications:

1. expansion of word contractions was replaced by removing all of the stopwords. The stopwords are a mixture of tokens from `nltk.corpus.stopwords.words("english")` and the contractions hash table keys. The larger list of "STOPWORDS" is represented as a Python set() object for efficient lookups.
2. all of the normalizing methods accept generators as input to improve performance on memory.

Figure 1: Directory Hierarchy of Assignment 2



2.1.3 `lexer.Lexer`

The `Lexer` class received the following modifications:

1. the Counter object `tf_in_doc` was added to maintain a term-frequency of the current document. This addition allows for the new method `term_doc_tf(self, doc_id: str)` that generates tuples of `(term-string, doc-ID, term-document-frequency)`. These tuples are saved on disk in files named "sample_tdt.txt" for further sorting and processing to generate the inverted files.

2.2 New Classes

2.2.1 `invertedfile.InvertedFile`

The `InvertedFile` class is responsible for building the binary inverted files and dictionaries. The class ingests the "sample_tdt.txt" files, sorts (in memory) the tuples as detailed in the prompt, and converts the values into 4-byte integer formats. The class also provides a method to look up tokens in the generated inverted files and dictionaries.

2.2.2 `Packer`

The `Packer` class is responsible for encoding and decoding fixed-format binary data.

2.2.3 `files` Classes

The `IO` class was split into 2 further classes, `Formatter` and `DataFile`. These classes provide support for file utility functions in the program, including reading and writing to plain files, JSON files, and binary files, formatting statistics outputs, generating file paths, etc.

3 Statistics and Observations

With the modifications made to the text normalization process, the statistics generated from the input file have changed. The new top 10 more frequent tokens now include "market", "announc", and "report", which are terms that expected in headlines.

In terms of the inverted files and dictionaries, the space they occupy on disk combined are significantly lower than the original document.

File	Size (in bytes)	Description
headlines.txt	39381610	Input corpus file
headlines_dict.json	4275427	Generated dictionary JSON file
headlines_if.bin	27774312	Inverted binary file

Table 1: Sizes of Files Computed Through the `stat` Command on a Debian Based Linux

4 Testing

The outputs are attached in Appendix B.

A Source Code

Code Listing 1: ./ir/files.py

```
import json
from pathlib import Path
import re
from typing import Any

from .lexer import Lexer
from .normalize import Normalizer


class IO:
    @staticmethod
    def read(filename: str) -> str:

        with open(f"{filename}.txt") as fp:
            return fp.read()

    @staticmethod
    def dump(filename: str, data: str) -> None:

        with open(f"{filename}.txt", "w") as fp:
            fp.write(data)
        print(f"Dumped to '{filename}.txt'")
```

```
    @staticmethod
    def read_json(filename: str) -> Any:

        with open(f"{filename}.json") as fp:
            return json.loads(fp.read())

    @staticmethod
    def dump_json(filename: str, data: Any) -> None:

        with open(f"{filename}.json", "w") as fp:
            json.dump(data, fp)
        print(f"Dumped json to '{filename}.json'")

    @staticmethod
    def read_bin(filename: str) -> bytes:

        with open(f"{filename}.bin", "rb") as fp:
            return fp.read()

    @staticmethod
    def dump_bin(filename: str, data: bytes) -> None:

        with open(f"{filename}.bin", "wb") as fp:
            fp.write(data)
        print(f"Dumped binary to '{filename}.bin'")


class DataFile:
    def __init__(self, filename: Path) -> None:

        self.filename = filename
        self.num_docs: int = 0

        self.df_file_name: str = f"stats/{filename.stem}_df"
        self.tf_file_name: str = f"stats/{filename.stem}_tf"
        self.stats_file_name: str = f"stats/{filename.stem}_stats"

        self.tdt_file_name: str = f"tmp/{filename.stem}_tdt"

        self.inv_file_name: str = f"bin/{filename.stem}_if"
        self.dict_name: str = f"bin/{filename.stem}_dict"

    def ingest(
        self,
        prep: Normalizer,
        lex: Lexer,
        term_doc_tf: list[tuple[str, str, int]],
    ) -> None:

        doc_id: str = ""
        doc_id_re = re.compile(r"\d+")
        line_num: int = 0

        with open(self.filename) as fp:
            for line in fp:
                match line_num % 4:

                    # line containing DocID
                    case 0:
```

```
                        doc_id = next(doc_id_re.finditer(line)).group()
                        self.num_docs += 1

                    # line containing document
                    case 1:

                        # normalize document through the preprocessing pipeline
                        prep.set_document(line)
                        prep.process()

                        # add processed tokens to the lexer
                        lex.add(prep.get_tokens())

                        # save records of term-DocID-tf
                        term_doc_tf.extend(lex.term_doc_tf(doc_id))

                    # empty lines
                    case _:
                        pass

                line_num += 1

        print("Processed", self.num_docs, "documents.")


class Formatter:

    hr: str = "------------------------------\n"

    table_header: str = f"{'Word':<12} | {'TF':<6} | {'DF':<6}\n{hr}"

    @staticmethod
    def __format_tf_df(term: str, tf: int, df: int) -> str:

        return f"{term:<12} | {tf:<6} | {df:<6}\n"

    @staticmethod
    def format_stats(lex: Lexer, num_docs: int = 0) -> str:

        contents: str = ""

        contents += f"{Formatter.hr}"
        contents += f"{num_docs} documents.\n"

        contents += f"{Formatter.hr}"
        contents += f"Collections size: {lex.get_collection_size()}\n"
        contents += f"Vocabulary size: {lex.get_vocab_size()}\n"
        contents += f"\n{Formatter.hr}"

        contents += "Top 100 most frequent words:\n"
        contents += Formatter.table_header
        for term in lex.get_top_n_tf_df(100):
            contents += Formatter.__format_tf_df(*term)

        contents += f"\n{Formatter.hr}"
        contents += "500th word:\n"
        contents += Formatter.table_header
        contents += Formatter.__format_tf_df(*lex.get_nth_freq_term(500))
```

```
        contents += f"\n{Formatter.hr}"
        contents += "1000th word:\n"
        contents += Formatter.table_header
        contents += Formatter.__format_tf_df(*lex.get_nth_freq_term(1000))

        contents += f"\n{Formatter.hr}"
        contents += "5000th word:\n"
        contents += Formatter.table_header
        contents += Formatter.__format_tf_df(*lex.get_nth_freq_term(5000))

        contents += f"\n{Formatter.hr}"
        single_occs: int = lex.get_single_occs()
        contents += "Number of words that occur in exactly one document:\n"
        contents += f"{single_occs} ({round(single_occs / lex.get_vocab_size() *
                                              100, 2)}%)\n"

        return contents

    @staticmethod
    def format_term_doc_tf(term_doc_tf: list[tuple[str, str, int]]) -> str:

        contents: str = ""
        for line in term_doc_tf:
            contents += " ".join(str(i) for i in line) + "\n"

        return contents
```

Code Listing 2: ./ir/normalize.py

```
import re
from typing import Iterator, Generator

import nltk

# fmt: off
STOPWORDS: set[str] = {
    # contractions
    "aren't", "ain't", "can't", "could've", "couldn't", "didn't", "doesn't", "don'
                                                t",
    "hadn't", "hasn't", "haven't", "he'd", "he'll", "he's", "i'd", "i'll",
    "i'm", "i've", "isn't", "it'll", "it'd", "it's", "let's", "mightn't",
    "might've'", "mustn't", "must've'", "shan't", "she'd", "she'll", "she's", "
                                                should've",
    "shouldn't", "that'll", "that's", "there's", "they'd", "they'll", "they're", "
                                                they've",
    "wasn't", "we'd", "we'll", "we're", "we've", "weren't", "what'll", "what're",
    "what's", "what've", "where's", "who'd", "who'll", "who're", "who's", "who've"
                                                ,
    "won't", "wouldn't", "would've", "y'all", "you'd", "you'll", "you're", "you've
                                                ",
    # NLTK stopwords
    "a", "all", "am", "an", "and", "any", "are", "as",
    "at", "be", "because", "been", "being", "but", "by", "can",
    "cannot", "could", "did", "do", "does", "doing", "for", "from",
    "had", "has", "have", "having", "he", "her", "here", "hers",
    "herself", "him", "himself", "his", "how", "i", "if", "in",
    "is", "it", "its", "itself", "just", "let", "may", "me",
    "might", "must", "my", "myself", "need", "no", "nor", "not",
    "now", "o", "of", "off", "on", "once", "only", "or",
```

```
    "our", "ours", "ourselves", "shall", "she", "should", "so", "some",
    "such", "than", "that", "the", "their", "theirs", "them", "themselves",
    "then", "there", "these", "they", "this", "those", "to", "too",
    "very", "was", "we", "were", "what", "when", "where", "which",
    "who", "whom", "why", "will", "with", "would", "you", "your",
    "yours", "yourself", "yourselves",
}
# fmt: on


class Normalizer:
    def __init__(self) -> None:

        self.document: str = ""
        self.tokens: Generator[str, None, None]

        self.ws_re: re.Pattern[str] = re.compile(r"([A-Za-z]+'?[A-Za-z]+)")
        self.snow: nltk.stem.SnowballStemmer = nltk.stem.SnowballStemmer(
            "english"
        )

    def set_document(self, document: str) -> None:

        self.document = document[:-1]

    def __to_lower_case(self, document: str) -> str:

        return document.lower()

    def __split_document(self, document: str) -> Generator[str, None, None]:

        return (x.group(0) for x in self.ws_re.finditer(document))

    def __remove_stopwords(
        self, tokens: Generator[str, None, None]
    ) -> Generator[str, None, None]:
        return (word for word in tokens if word not in STOPWORDS)

    def __stem(self, tokens: Iterator[str]) -> Generator[str, None, None]:

        return (self.snow.stem(token) for token in tokens)

    def get_tokens(self) -> Generator[str, None, None]:

        return self.tokens

    def process(self) -> None:

        # convert the entire document to lower-case
        doc_lc: str = self.__to_lower_case(self.document)

        # split the document on its whitespace
        tokens: Generator[str, None, None] = self.__split_document(doc_lc)

        # remove contractions and stopwords
        no_sw: Generator[str, None, None] = self.__remove_stopwords(tokens)

        # stem tokens
        self.tokens = self.__stem(no_sw)
```

Code Listing 3: ./ir/lexer.py

```
from collections import Counter
from typing import Generator


class Lexer:
    def __init__(self) -> None:

        self.tf: Counter[str] = Counter()
        self.df: Counter[str] = Counter()
        self.tf_in_doc: Counter[str] = Counter()

    def add(self, tokens: Generator[str, None, None]) -> None:

        # create a Counter for the document
        self.tf_in_doc.clear()
        self.tf_in_doc.update(tokens)

        # update the total term-frequency values with the Counter
        self.tf.update(self.tf_in_doc)

        # increment the document-frequency values
        self.df.update(self.tf_in_doc.keys())

    def term_doc_tf(
        self, doc_id: str
    ) -> Generator[tuple[str, str, int], None, None]:

        for term in self.tf_in_doc:
            yield term, doc_id, self.tf_in_doc[term]

    def get_tf(self) -> Counter[str]:

        return self.tf

    def get_df(self) -> Counter[str]:

        return self.df

    def get_collection_size(self) -> int:

        return self.tf.total()

    def get_vocab_size(self) -> int:

        return len(self.tf)

    def get_top_n_tf_df(
        self, n: int
    ) -> Generator[tuple[str, int, int], None, None]:

        top_n_tf = self.tf.most_common(n)
        for tf in top_n_tf:
            term, freq = tf
            yield term, freq, self.df[term]
```

```
    def get_nth_freq_term(self, n: int) -> tuple[str, int, int]:

        term, freq = self.tf.most_common(n)[-1]
        return term, freq, self.df[term]

    def get_single_occs(self) -> int:

        single_occs: int = 0
        for df in self.df.values():
            if df == 1:
                single_occs += 1

        return single_occs
```

Code Listing 4: ./run.py

```
from pathlib import Path

from ir.files import IO, Formatter, DataFile
from ir.normalize import Normalizer
from ir.lexer import Lexer
from ir.invertedfile import InvertedFile
from ir.packer import Packer


def process_document(filename: Path) -> None:

    prep = Normalizer()
    lex = Lexer()
    data = DataFile(filename)

    term_doc_tf: list[tuple[str, str, int]] = []
    data.ingest(prep, lex, term_doc_tf)
    term_doc_tf_str: str = Formatter.format_term_doc_tf(term_doc_tf)
    IO.dump(data.tdt_file_name, term_doc_tf_str)
    IO.dump_json(data.df_file_name, lex.get_df())
    IO.dump_json(data.tf_file_name, lex.get_tf())

    contents: str = Formatter.format_stats(lex, data.num_docs)
    IO.dump(data.stats_file_name, contents)

    invf = InvertedFile()
    invf.vocabulary(lex.get_df(), lex.get_tf())
    invf.ingest(IO.read(data.tdt_file_name))
    inv_file = invf.get_inverted_file_raw()
    if_data = Packer.encode(inv_file)
    IO.dump_bin(data.inv_file_name, if_data)
    IO.dump_json(data.dict_name, invf.get_dictionary())


if __name__ == "__main__":

    process_document(Path(__file__).parent.parent / "data" / "headlines.txt")
```

B Outputs

Code Listing 5: Statistics of 'headlines.txt'

```
500000 documents.
```

```
Collections size: 3518452
Vocabulary size: 111614
```

Top 100 most frequent words:

Word	TF	DF
new	27023	26505
market	14814	12930
after	13025	12977
announc	11090	11065
up	10814	10666
over	10007	9961
say	9875	9828
year	9442	9250
report	8920	8758
day	8578	8414
global	8272	8180
man	8211	8126
open	7930	7875
get	7838	7785
us	7754	7655
more	7682	7577
out	7638	7583
first	7591	7532
world	7268	7160
septemb	7266	7208
win	7226	7155
week	6879	6751
polic	6581	6520
launch	6426	6414
make	6398	6366
show	6138	6058
share	5886	5763
school	5817	5645
about	5741	5686
take	5725	5714
state	5712	5572
servic	5695	5566
back	5611	5538
home	5594	5480
video	5569	5499
top	5569	5505
plan	5515	5452
busi	5489	5329
one	5479	5289
time	5467	5379
industri	5405	5282
game	5248	5097
china	5216	5109

inc	5065	4496
help	4992	4950
call	4953	4925
set	4906	4890
th	4898	4749
citi	4891	4776
group	4883	4792
stock	4777	4708
into	4771	4760
live	4761	4677
review	4718	4691
updat	4714	4692
research	4667	4557
against	4662	4647
two	4647	4584
rate	4643	4551
fall	4635	4613
kill	4614	4597
compani	4448	4389
nation	4418	4361
award	4328	4221
best	4321	4250
high	4281	4225
million	4279	4220
releas	4262	4248
meet	4212	4190
star	4174	4114
car	4139	4070
price	4090	4052
offic	4067	4017
offer	4058	4036
watch	4035	3791
deal	4026	3981
refuge	3979	3936
chang	3906	3879
free	3903	3803
manag	3890	3780
down	3844	3822
big	3832	3749
name	3801	3777
season	3772	3749
use	3763	3741
hous	3751	3706
lead	3727	3720
news	3708	3641
look	3705	3677
hit	3705	3698
fire	3687	3591
sale	3685	3614
technolog	3683	3551
team	3677	3630
secur	3670	3541
charg	3663	3648
confer	3663	3637

footbal	3575	3527
start	3558	3541
work	3557	3502

500th word:

Word	TF	DF
agreement	1314	1310

1000th word:

Word	TF	DF
exhibit	736	724

5000th word:

Word	TF	DF
manipul	92	92

Number of words that occur in exactly one document:
52962 (47.45%)