**📘 作业 3：Lexicon**

**截止时间：** 2025年10月17日（星期五）晚上11:59  
**课程：** CSCI 60（Krehbiel 教授）

**🧩 概述 (Overview)**

本次作业要求你实现一个名为 **Lexicon** 的类。它和之前的 **Bag** 数据结构类似，但主要区别在于：  
Lexicon 存储的是一组“**不重复的单词**（set of distinct words）”，  
而 Bag 存储的是“可能重复的整数列表（list of possibly repeated integers）”。

像上次一样，完整的接口文件 **lexicon.h** 已经提供；  
你只需要实现对应的 **lexicon.cpp** 文件。

本次任务包括：

1. 修改构造函数与修改函数（mutator），使其反映 Lexicon 的“集合”特性；
2. 重载若干运算符来实现集合运算。

你只需上传 **lexicon.cpp** 文件。  
编译时使用提供的 main.cpp 进行单元测试（unit testing）：

g++ main.cpp lexicon.cpp -std=c++11

**🧱 接口说明 (The interface)**

接口文件 **lexicon.h** 已完全写好，不可修改。  
其中包含每个函数的前置条件（pre-condition）和后置条件（post-condition），  
有部分简单函数已经 inline 定义，剩余部分你要在 **lexicon.cpp** 中实现。

Lexicon 的 [] 运算符反映了单词的存储结构。  
集合操作的三个二元运算符将分别代表：

* 并集 union (|)
* 交集 intersection (&)
* 对称差 symmetric difference (^)

同时还要实现六个比较运算符（==、!=、<=、<、>=、>），  
用来判断一个 Lexicon 是否是另一个的子集、真子集等。

**⚙️ 实现要求 (Implementation)**

你要实现接口中所有函数。建议按照以下阶段完成：

**阶段 1：构造函数**

0参构造函数已为你实现。  
但你需要补充**1参构造函数**（从文件中读取单词）。  
要求：

* 从输入文件读取最多 CAPACITY 个不重复单词；
* 存储顺序与文件中出现顺序一致；
* 忽略重复单词；
* 使用 == 来判断单词是否相同（区分大小写！"Hello", "hello", "HELLO" 视为不同单词）。

举例：  
若 hello.txt 文件包含：

hello

hello

world

则构造 Lexicon lex("hello.txt");  
会产生一个对象：

lex.data\_ = ["hello", "world"];

lex.size\_ = 2;

**阶段 2：布尔成员函数**

实现以下函数：

* contains(word)
* insert(word)
* remove(word)  
  这些函数需要符合 lexicon.h 文件中的前后条件要求。

**阶段 3：集合运算符**

实现三个**成员函数**重载：

* lex1 | lex2：并集（出现在任一 Lexicon 中的单词）
* lex1 & lex2：交集（出现在两个 Lexicon 中的单词）
* lex1 ^ lex2：对称差（仅出现在其中一个 Lexicon 中的单词）

这些运算符应返回一个新的 Lexicon 对象。

**阶段 4：比较运算符**

实现六个**非成员函数**重载：

* lex1 == lex2：若两个 Lexicon 含有完全相同的单词（顺序无关）
* lex1 != lex2：若两个 Lexicon 至少有一个单词不同
* lex1 <= lex2：若 lex1 是 lex2 的子集
* lex1 < lex2：若 lex1 是 lex2 的真子集
* lex1 >= lex2：若 lex1 是 lex2 的超集
* lex1 > lex2：若 lex1 是 lex2 的真超集

这些函数只能通过已公开的接口访问数据（如 contains）。

**🧪 测试程序 (The test program)**

使用提供的 **main.cpp** 来测试你的实现。  
建议边写边测试单个函数（unit test），不要一次性跑所有测试。  
<< 运算符已帮你实现，可以用来打印 Lexicon 内容。

当一切完成后，用提供的测试文件 foundingDocs 和 annaKarenina 运行。  
输出应如下所示（部分展示）：

The US Constitution contains 1680 distinct words.

The Declaration of Independence contains 632 distinct words.

The two documents have 202 words in common.

There are 1908 words in their symmetric difference.

...

Anna Karenina is 350188 words (with repetition).

The number of distinct words reaches our class's capacity.

**✅ 总结**

**你需要提交的文件：**

* lexicon.cpp

**需掌握的知识点：**

* 类的实现与构造函数
* 集合操作（并、交、差）
* 运算符重载（成员与非成员）
* 文件输入输出与字符串处理
* 测试与调试（unit test）