CS-320 Project Report

Programming Language Detection

Abstract: The idea is to develop a program in python that will recognise the programming language of given code like C/C++, Java, HTML.

Submitted by

16010108 Anupriya Sinha
16010123 Haripriya B

5^{th} Semester :: August-November 2018 session
Bachelor of Technology in Computer Science and Engineering

Under the guidance of:
Mr. Himangshu Sarma

Department of Computer Science and Engineering
Indian Institute of Information Technology Manipur
Mantripukhri, Imphal, India - 795002

Chapter 1

Introduction and Architecture

1.1 Introduction to problem statement

The idea is to develop a program in python that will recognise the programming language of given code like C/C++, Java, HTML. The approach used is regular expression to search/match the basic syntax of programming language.

1.2 Architecture

Following softwares and platforms has been used:

- OS: Ubuntu 18.04
- Jupyter Notebook
- Python 3.6

The Jupyter Notebook is an open-source IDE that allows users to share and create the documents that contain live code, narrative text, etc. The purpose of using Jupyter notebook in this project is the convenience it provide to run the code.

Python language has been used because of it is simple, easy to understand and code.

Chapter 2

Implementation

The developed program takes text file(.txt extension) input and reads the whole file content as a single string. After that different regular expression has been used to find the class of programming language to which the given code belongs.

2.1 Algorithm

1. Read the given file as a single string
2. if string satisfies the regex for java
   print it is java file
3. else if string satisfies the regex for HTML
   print it is HTML file
4. else if string satisfies the regex for C/C++
   and if string contains cout/cin
   print it is C++ file
   else if it contains printf/scanf
   print it is C file

   else
   print it is C/C++ file

2.2 Regular expression used

- Regular expression for Java

  (**default**| **public**| **private**| **protected**)?[\s]**class**[\s][A-Z]+[\w]*{[\s\w;
  /*=+-]***public**[\s]+**static**[\s]***void**[\s]***main**[\s]*[(]**String**[\s]+[\w]+[[]?
  []]?[)][\s\w]*{

  A java class can be declared with an access modifier like default, private, public, protected; so that is taken as an optional checking in the regular expression. A class name should contain capital letter at the starting, which is also checked. The main function defined as "public static void main" is matched with spacing in between. The arguement type is checked to be String followed by any identifier.

- Regular expression for HTML

```
[\s\w;/*<>=!+-]*<[\s]*html[\s\w;/*=!+-]*>[\s\w]*<[\s]*title[\s\w;/
*=!+-]*>[\s\w]*<[\s]*/title[\s]*>[\s\w;/*<>=!+-]*<[\s]*body[\s\w;/
*=!+-]*>[\s\w;/*<>.=!+-]*<[\s]*/body[\s]*>[\s\w;/*<>.=!+-]*<[\
s]*/html[\s]*>[\s]*
```

  A proper HTML file should contain the opening and closing tag of html along with the opening and closing tags of title and body which is checked in the regular expression. In between these tags there can be any other tags like head, footer, paragraph,etc.

- Regular expression for C/C++

```
(int |void |char |float |double )[\s]*main[(][\w]*[)]
```

  A C/C++ program's main function can have int or float or double or char or void as return types which are checked in the regular expression. With only main function a program can be of C family. In order to classify as C, existance of printf and scanf is checked and to classify as C++, existance of cin and cout is checked.

```
([\s]*#[\s]*include[\s]*[<\"][\w]*.h[\">][\s]*){2,}
```

  The inclusion of header files in C/C++ program is matched with the include keyword and .h extension.

Chapter 3

Conclusion

A program in python has been developed that recognizes few programming languages like C/C++, Java, HTML.

3.1 Future Scope

- This project can be extended to recognize many more number of programming languages.
- Machine learning and Deep learning approachs can be included further which may give better result.
- Web application can be developed for online prediction of programming languages for given code snippet.

3.2 Limitations

- Comment containing code snippet will also be recognized as programming file if it matches the regular expression.
- A text file containing code snippet will also be recognized as programming file if it matches the regular expression.
- Only standard syntax is checked, besides this any error in remaining code is not considered.
- Standard syntax are only considered.