Language identification

by

Andreas Socratous

Final Year project submitted in partial fulfilment of the requirements for the Degree of

Bachelor of Science in Computer Science

Department of Computer Science

School of Sciences

University of Nicosia

May, 2016

***Language Identification***

***By***

***Andreas Socratous***

This Final Year Project has been accepted in partial fulfilment of the requirements for the Degree of

Bachelor of Science in Computer Science

Project Advisor:

/ /

(Name) (Signature) (Date)

Examiner:

/ /

(Name) (Signature) (Date)

Examiner:

/ /

(Name) (Signature) (Date)

Department of Computer Science

School of Sciences

University of Nicosia

**Abstract**

The purpose of this project is to develop a program that is using N-Grams, in order to identify the language of the given file.

In summary, the program compiles with several embedded files, in order to develop a profile for each language, in the current case, is English, French and Spanish. Then, the program will try to compare the input profile with each of the pre-made profiles, calculate the distance between the two compared profiles and find the smallest distance. The language profile which have the smallest distance, means that the probability of this language to be the input file is the maximum.

In order to create the profiles, there was needed to split each file into tokens, and from tokens to N-Grams. The N-Grams take from 1 to the length of the token and spit it into sub-strings. The data structure is used is at first Linked Lists for test and then Hash Table, for better runtime. The programming language used is C#, because, there are a lot of helpful ready-made functions to use and an “improved” Hash Table, called the Dictionary.

**Acknowledgements**

I would like to thank my parents for their encouragement and financial support during my studies. Also my project advisor for his critical questions, which forced me to read more, think critically and write with better clarity and for teaching me that we learn by doing. Special thanks also go to my friend John Smith, who read my project report and helped me with the grammar and style. I would also like to express my gratitude to my Manager in work, Mr Yiannis Ioannou and my co – workers in Computer Center in University of Nicosia, which helped me a lot to the understanding and solving the project.

**Table of Contents**

**Chapter 1:**

**1.1 Identifying the Problem**

The problem in this project was to identify the language using N-Grams. In order to do so, Dr Athena Stassopoulou, my advisor, suggested some papers for reading to understand the differences between the different solutions that problem has. Some solutions were the Naive Bayes approach and the N-Grams approach. For this project, the method used was N-Grams.  
  
After reading the (enter text here) paper, I decided to use the N-Grams method, because in my opinion, it was more programmatically efficient and practical to be implemented.   
  
The idea behind N-Grams with a text, for example “beer” is that the text will be broken into: (consider where ‘\_’ is the space character)

1. Bi – Grams : \_b, be, ee, er, r\_
2. Tri – Grams : \_\_b, \_be, bee, eer, er\_, r\_\_
3. Quad – Grams : beer

That means that the N-Gram algorithm should be applied from number 1 to the length n of the current token. This will be done with every token.  
  
Now, let’s talk about the tokens. A token is basically every single world individually in all the text documents. Every text document will be spitted into tokens and stored in a data structure. In this case, Linked Lists are used.

* 1. **Decision of the Programming language**

Based on my current knowledge on programming languages, I decided to use C# for the implementation of this project. This programming language provides many flexibilities both with GUI (Graphical User Interface) and with libraries. It has many build-in ready-made functions that will be very helpful for the implementation of this project.

* 1. **Designing the interface**