Real-World Engineering Projects

Language Identification Programming Project

Background Lecture

Overview

- Overview of the project
- History and Motivation
- Details of the Project
  - Resources
  - Step 1: Calculate the histogram of letter frequencies
  - Step 2: Analyze example texts
  - Step 3: Identify the language of the text
  - Report
- Example Code
- Teamwork
- Summary
- References

Overview

In this project, you will write a computer program which can identify the language of a written text.

Step 1: Calculate a histogram of letter frequencies.

Step 2: Analyze example texts.

Step 3: Identify the language of the text.

History and Motivation

“There is no need to do more than mention the obvious fact that a multiplicity of languages impedes the cultural interchanges between the people of the earth, and is a serious deterrent to international understanding.” – Warren Weaver [1], 1949

History and Motivation

Information, in print and on the internet, is available in many languages.

- There are ~6900 active languages in the world [2].

- Wikipedia contains articles in 262 languages and only ~23% are in English [3].

- Worldpress contains ~7,000,000 blogs, and only ~36% are in English [4].

Before a text can be translated, the language of the text must be identified.

Details of the Project: Resources

Project Gutenberg:

- http://www.gutenberg.org (or http://www.promo.net/pg)
- Online database dating to 1971
- Over 30,000 books
- Books in 46 languages
- Over 50 books in 13 languages
- Books in plain text
- All books are public domain
- Files contain a license in English

Details of the Project: Resources

Other resources:

- Text editor (or Integrated Development Environment)
- Compiler
- Spreadsheet (for plotting)

Details of the Project: Step 1

Step 1: Write a computer program to calculate the histogram of letter frequencies. The program should read in a text file and output the number of a's, b's, and so on. Both upper and lowercase letters should be included in the letter frequencies.

This process is illustrated below.

Details of the Project: Step 1

Letters 'a' and 'o' are highlighted. The text [6] has 21 'a's and 11 'o's.

“It was in Warwick Castle that I came across the curious stranger whom I am going to talk about. He attracted me by three things: his candid simplicity, his marvelous familiarity with ancient armor, and the restfulness of his company – for he did all the talking.”

Details of the Project: Step 1

Letters 'a' and 'o' are highlighted. The text [6] written in English, Italian, and German.

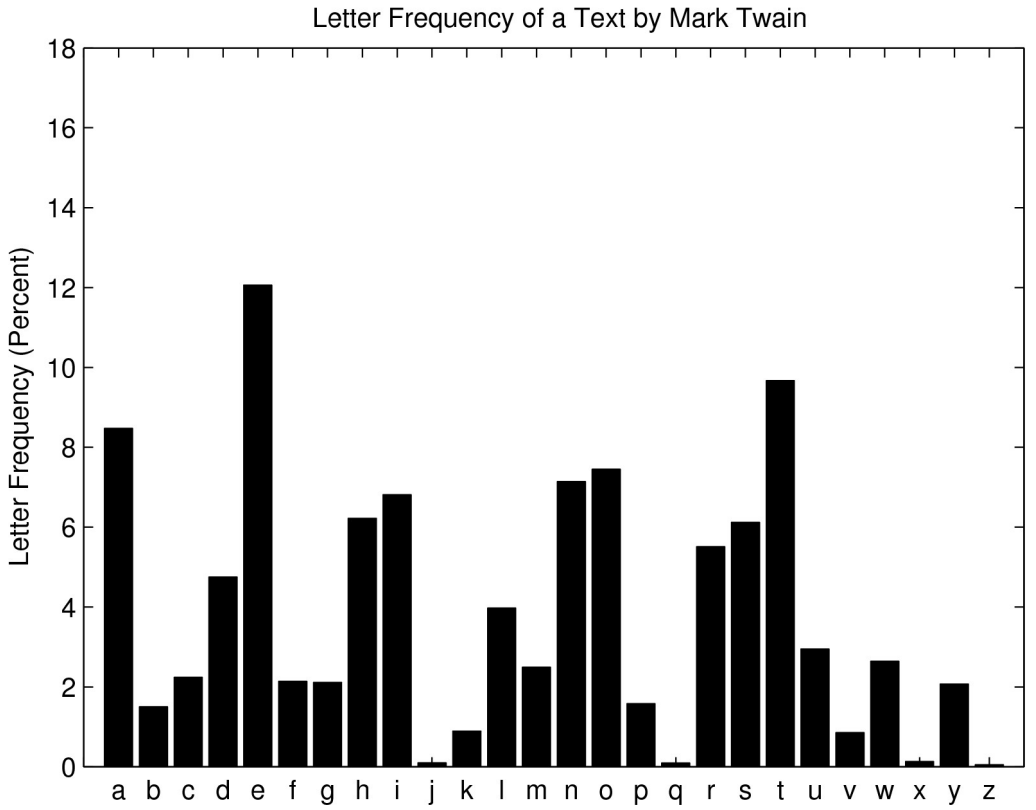
“It was in Warwick Castle that I came across the curious stranger whom I am going to talk about.”

“Era nel castello di Warwick che ho trovato lo Sconosciuto curioso di quale sto andando parlare.”

“Es war im Warwick Schloss, dass ich über auf den neugierigen Fremden zufallig stieß, den ich sprechen werde.”

Details of the Project: Step 1

An example histogram:



Details of the Project: Step 2

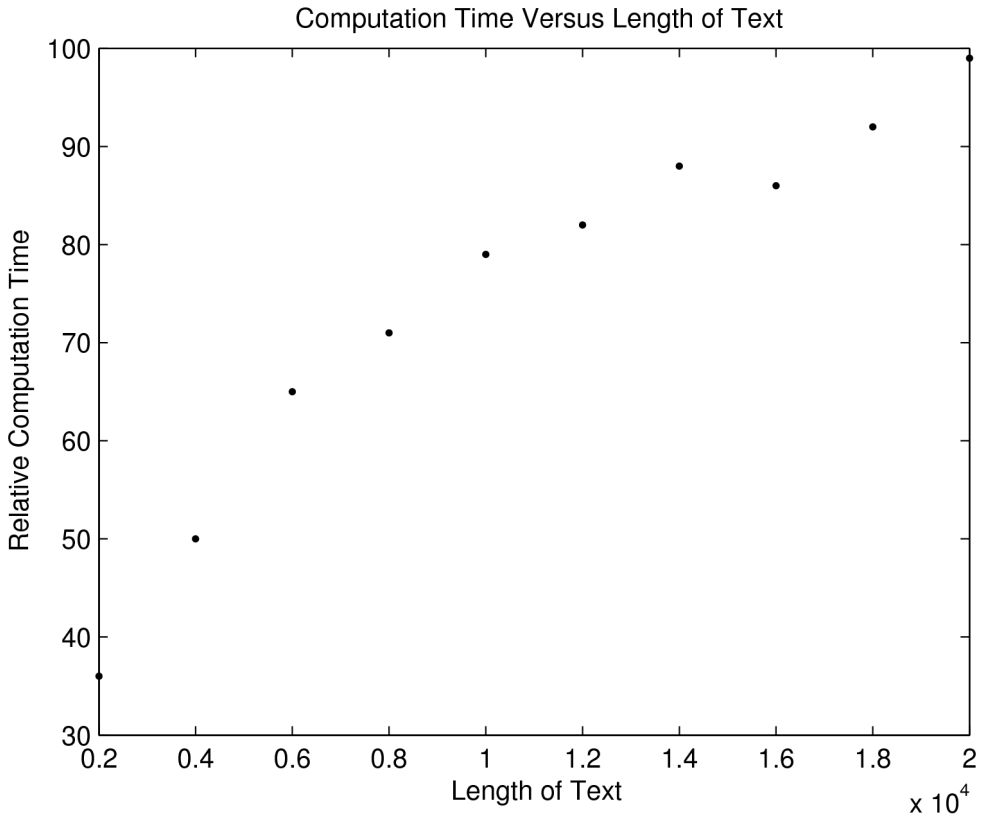
All engineering projects involve trade-offs.
In this step, you will explore the trade-offs in this project.

Pick an example text. Produce a plot of letter frequency versus the length of text analyzed for vowels.

Produce a plot that shows the time for computing the histogram versus the length of text analyzed.

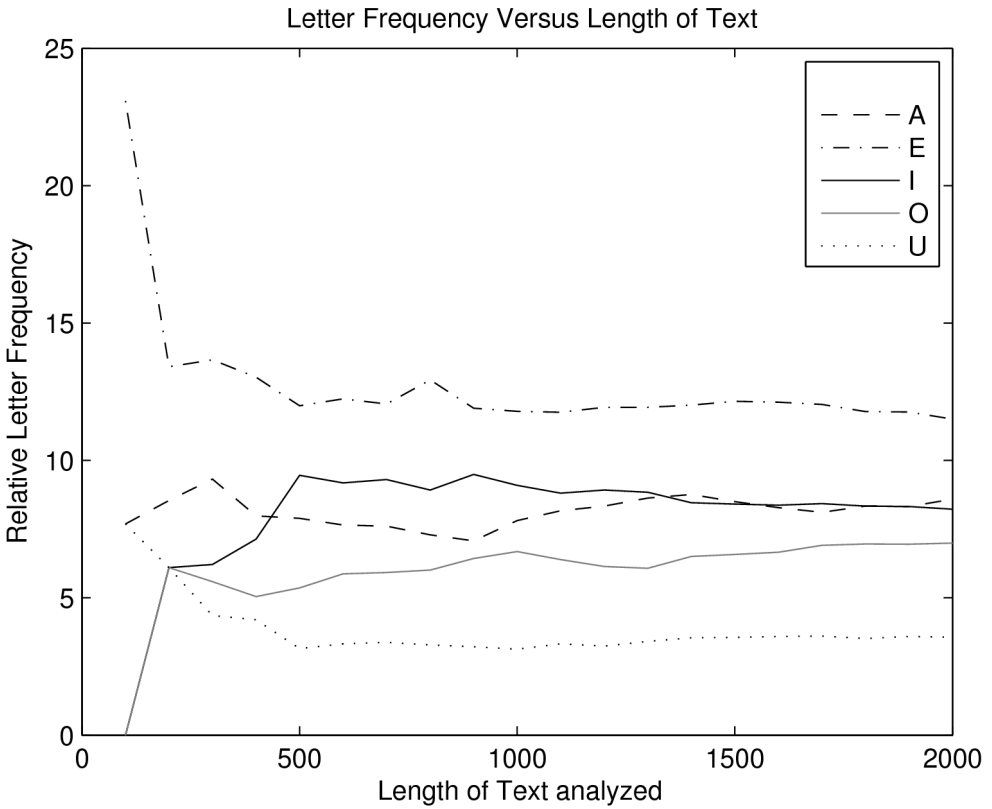
Details of the Project: Step 2

Example figure:



Details of the Project: Step 2

Example figure:



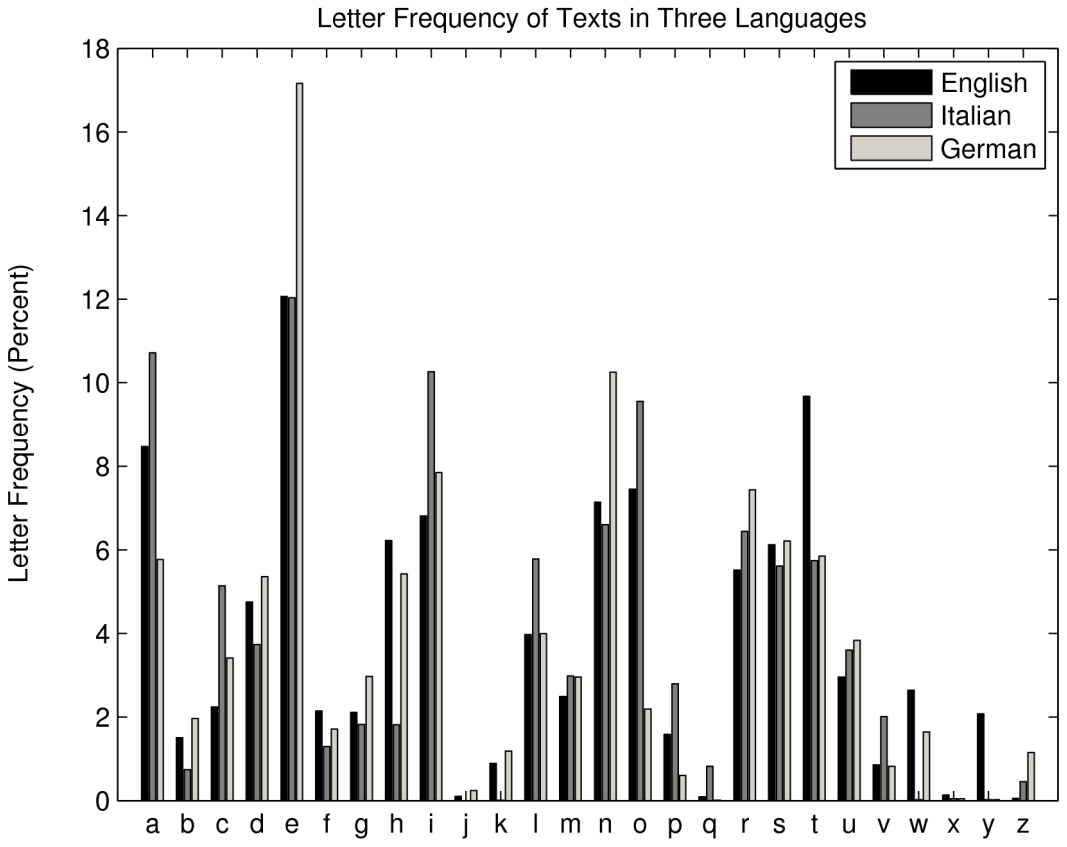
Details of the Project: Step 3

Modify your program so that it can identify the language of the text. More specifically, your program should be able to distinguish between English, German, and Italian.

Each text will have a slightly different histogram, but there will be similarities in the histograms of all English texts, all Italian texts, and all German texts.

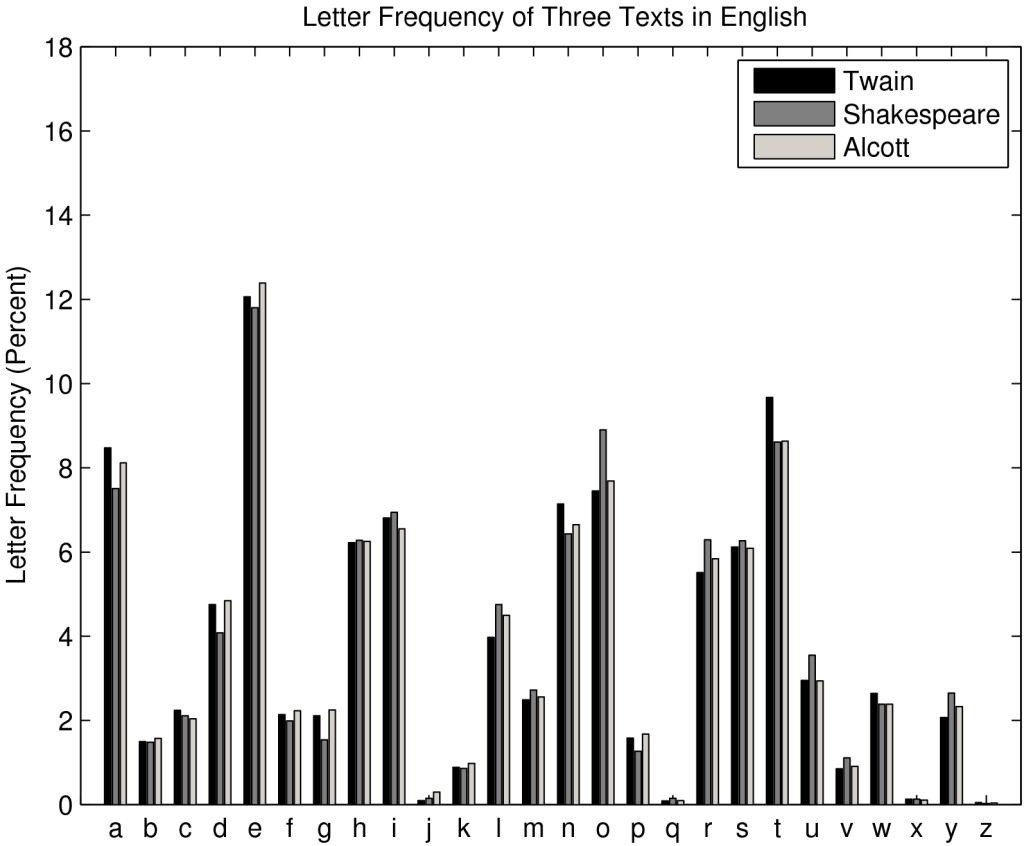
Details of the Project: Step 3

A histogram comparing texts of three languages:



Details of the Project: Step 3

A histogram comparing three English texts by different authors:



Details of the Project: Report

Each group should turn in a formal report. Your report should :

- Be typed
- Be three to five pages in length
- Be written to be understood by an engineer or computer scientist who is unfamiliar with the project.
- Contain introduction, methods, results, and conclusions sections.
- Include your source code
- Cite any references used

Example Code

```
import java.io.*;
class simpleLetters {
    public static void main (String[] args) throws IOException {
        FileReader in = new FileReader("intext.txt");
        long start, stop, elapsed;
        int c;
        start =  System.currentTimeMillis();

       for(int ii=0;ii<30;ii++)
          {
           c=in.read();
           System.out.print("Letter detected"+(char) c);
           System.out.println("Has ascii value"+c);
          };

      stop= System.currentTimeMillis();
      elapsed = stop – start;
      System.out.println("Time: "+elapsed);
    };
};
```

Example Code

Reading from a file called intext.txt:

```
import java.io.*;
FileReader in = new FileReader("intext.txt");
c=in.read();
```

When the end of the file is reached, c will be -1.

Example Code

Writing to a file called outfile.txt:

```
import java.io.*;
FileWriter outf = new FileWriter(“outfile.txt”);
outf.write(“ABC \n”);
outf.close();
```

Example Code

Timing the computation:

```
long start, stop elapsed;

start = System.currentTimeMillis();
        . . .
stop = System.currentTimeMillis();
elapsed=stop-start;
```

Example Code

Casting between int and char data types:

```
System.out.print("Letter detected"+(char)c);
System.out.prinln("has ascii value"+c);
```

Each character is represented in the computer by an ASCII value.

Example Code

ASCII Values:

64	@	74	J	84	T	94	^	104	h	114	r	124	|
65	A	75	K	85	U	95	_	105	i	115	s	125	}
66	B	76	L	86	V	96	`	106	j	116	t	126	~
67	C	77	M	87	W	97	a	107	k	117	u	127	DEL
68	D	78	N	88	X	98	b	108	l	118	v		
69	E	79	O	89	Y	99	c	109	m	119	w		
70	F	80	P	90	Z	100	d	110	n	120	x		
71	G	81	Q	91	[	101	e	111	o	121	y		
72	H	82	R	92	\	102	f	112	p	122	z		
73	I	83	S	93	]	103	g	113	q	123	{		

Teamwork

In this project, you will work in groups of two or three.

Programming is not an individual activity.

Typically, a programmer writes a program to be used by someone else.

Often in corporate environments, a team of people work on a software project.

Summary

In this project, you will write a computer program which can calculate the letter frequencies in a text and identify the language of the text.

References

[1] W. Weaver, “Translation,” *Machine Translation of Languages: Fourteen Essays*, pp. 15-23, 1949, http://www.mt-archive.info/ Weaver-1949.pdf, Date accessed: June 24 ,2009.

[2] M. P. Lewis, *Ethnologue: Languages of the World*, SIL International, 2009.

[3] Wikipedia, http://en.wikipedia.org/wiki/Wikipedia, date accessed June 2009.

[4] Worldpress, http://worldpress.org, date accessed June 2009.

[5] L. Berlin, “A web that speaks your language,” *New York Times*, May, 2009.

[6] M. Twain (S. L. Clemens), *A Connecticut Yankee in King Arthur's Court*, p. 1, 1889.

Text used for figures on slides 11, 13, 14:

M. Twain (S. L. Clemens), *A Connecticut Yankee in King Arthur's Court*, 1889.

Texts used for figure on slide 16:

English text: M. Twain (S. L. Clemens), *A Connecticut Yankee in King Arthur's Court*, 1889.

Italian text: D. Alighieri, *Divina Commedia di Dante, 1895*.

German text: E. T. A. Hoffmann, *Nachtstuecke, 1817*.

Texts used for figure on slide 17:

M. Twain, *A Connecticut Yankee in King Arthur's Court*, 1889.

W. Shakespeare, *As You Like It*, 1601.

L. M. Alcott, *Little Women*, 1867.