Real-World Engineering Projects

Language Identification Programming Project

Summary Lecture

Outline

- Summary of the Project
- Trade-offs
- Frequency Analysis
- Conclusion
- Future Directions
- References

Summary of Project

Write a computer program which can identify the language of a text.

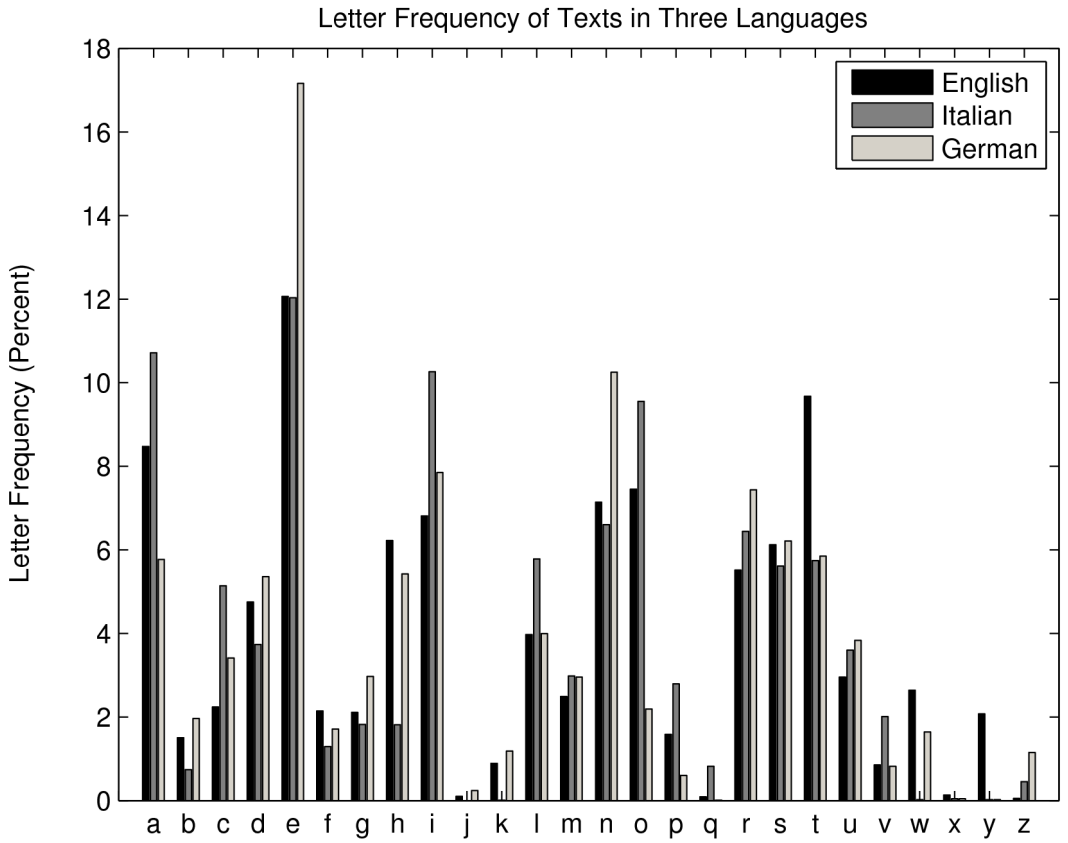
Step 1: Calculate a histogram of letter frequencies

Step 2: Analyze example texts

Step 3: Identify the language of the text

Summary of the Project

Example histogram of texts in three languages:



Summary of the Project

An example algorithm to distinguish between English, Italian, and German texts:

```
If ((norma < 9.2) && (normo > 5.0))
   {System.out.println(“English”);};
If ((norma < 9.2) && (normo < 5.0))
   {System.out.println(“German”);};
If (norma > 9.2)
   {System.out.println(“Italian”);};
```

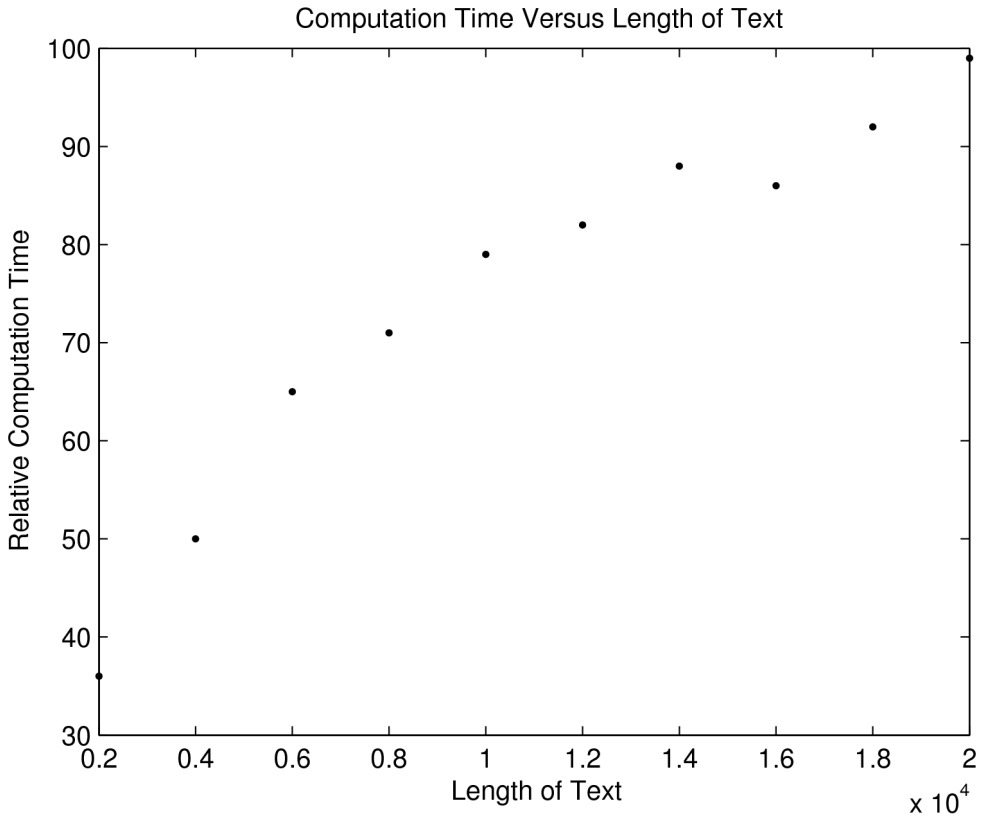
norma is the percent of 'a's.
normo is the percent of 'o's.

Trade-offs

Engineers always face trade-offs when designing products whether the products involve hardware, software, or both.

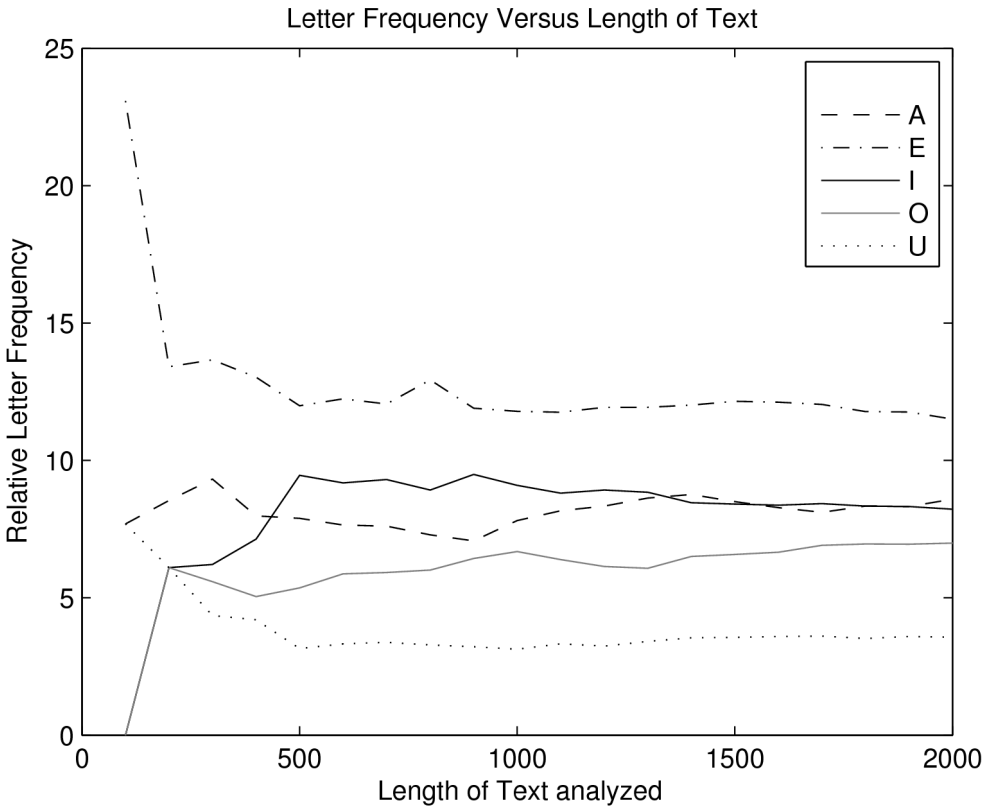
Trade-offs

A trade-off exists involving the length of text analyzed.



Trade-offs

A trade-off exists involving the length of text analyzed.



Trade-offs

The frequency of the letter 'a', for example, is not a precise value. It varies from text to text and within a text.

A random variable is a quantity which cannot be predicted. The first letter appearing on a page of text is a random variable.

A random process is a function of a random variable. The frequency of the letter 'a' as a function of the length of a text is a random process.

Trade-offs

It is not possible to exactly predict the next letter of a text or the the frequency of the letter 'a' in a text.

Most functions, or signals, encountered in the real world involve some amount of randomness or noise.

Electrical and Computer engineers often study signals which are random processes.

Trade-offs

The Central Limit Theorem says that a random process, which represents the normalized sum of random variables with the same probability, will converge as the size length of the random process increases [1].

Each letter has roughly the same probability of being an 'a'. The frequency of the letter 'a' converges as the length of text analyzed increases.

The Central Limit Theorem is discussed further in a class on Probability and Statistics or Random Processes.

Frequency Analysis

This project involved calculating the **frequency** of letters in a text and using the result to find out information about the language of the text.

Frequency analysis is a fundamental mathematical tool.

Electrical and Computer engineers study the mathematics of frequency analysis in Digital Signal Processing and other courses.

Frequency Analysis

- In the 1600s, Newton developed Calculus and introduced the concept of the frequency spectrum of light.
- In 1807 Fourier studied heat flow in solids. He developed the mathematical theory for studying signals as functions of frequency. His work has been called [2] "one of the greatest advances in the history of mathematics."
- In 1949, Weaver proposed using frequency analysis for for translating texts [3].
- In 1965, Cooley and Tukey developed the FFT, a fast
- algorithm for numerically computing the Fourier transform [4].

Frequency Analysis

Frequency analysis has been used to identify:

- The author of a piece of literature [5], [6]
- The author of historical documents [7],[8]
- The artist of a painting [9]
- The programmer of a piece of computer code [10]

Frequency analysis has also been used to determine whether audio files [11] or images [12] have been illicitly altered.

Conclusion

- In this project, you wrote a computer program to identify the language of a text.
- Your program is a tool which helps aid communication.
- The project involved computer programming and frequency analysis.

Future Directions

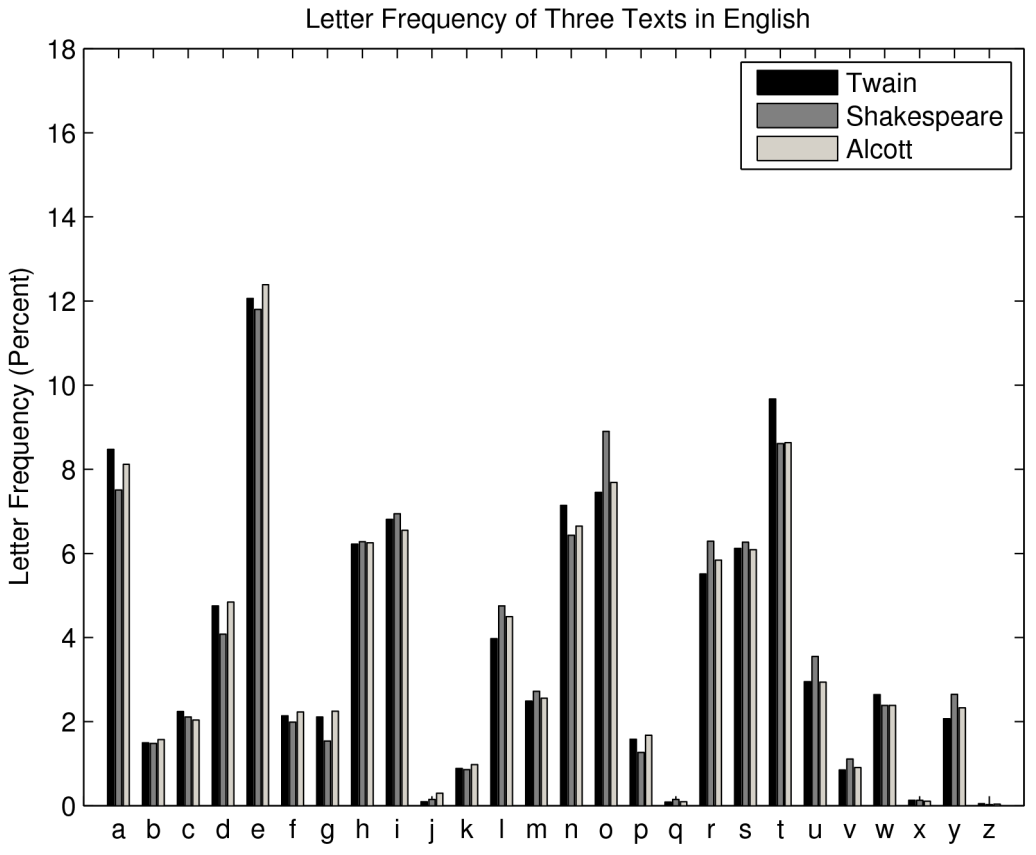
This project could be continued or expanded in many ways. The code could:

- Differentiate between more than three languages
- Be optimized to run faster or more efficiently
- Identify the language of a website instead of a file
- Verify that a text was written by a given author
- Not only identify the language of text, but also translate

Future Directions

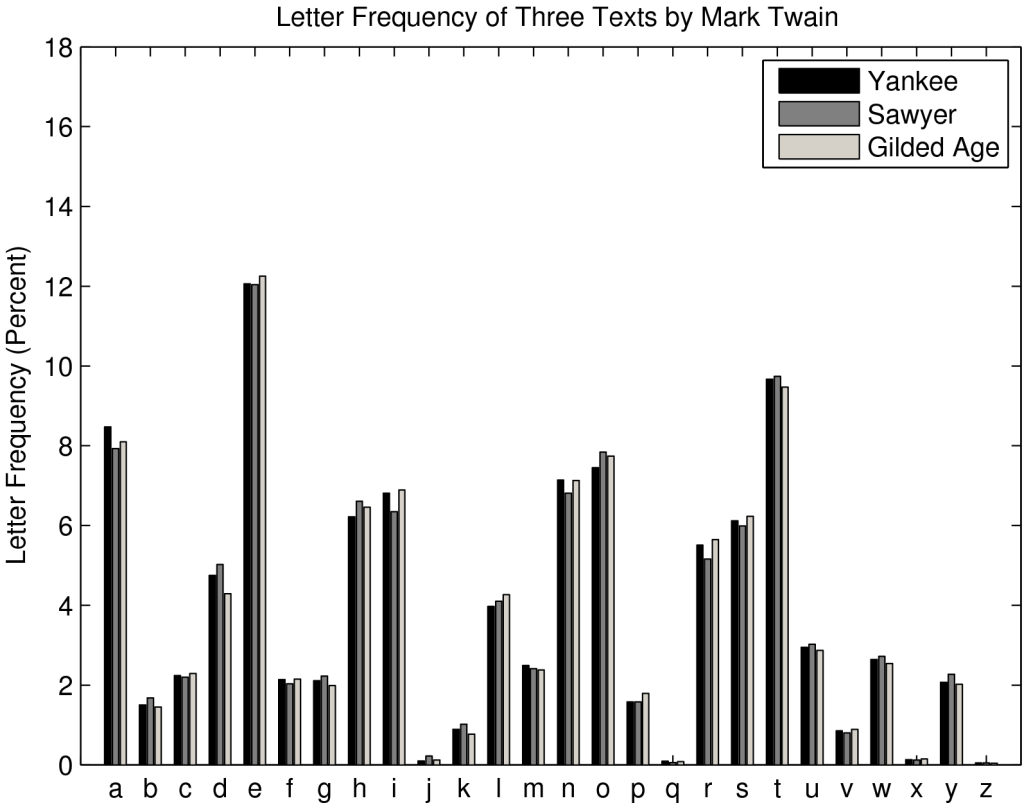
With more precision, the code could be made to verify that a text was written by a given author.

Example histogram of three English texts by different authors:



Future Directions

Example histogram of three English texts by the same author:



Future Directions

Identifying a text is the first step of translating a text.

Early machine translation programs took a dictionary approach. More recent, and more successful programs use frequency analysis and concepts from random processes.

References

[1] G. R. Cooper and C. D. McGillem, Probabilistic Methods of Signal and System Analysis, 3rd edition, Oxford 1999, p. 73.

[2] E. A. Robinson “A Historical Perspective of Spectrum Estimation”, *Proceedings of the IEEE*, vol. 70, no. 9, Sep. 1982, pp.885.

[3] W. Weaver, “Translation,” *Machine Translation of Languages*, pp. 15-23, 1949.

[4] J. W. Cooley and J. W. Tukey, “An algorithm for the Machine Calculation of Complex Fourier Series”, *Mathematics of Computation*, vol. 19, No. 90, Apr. 1965, pp. 297-301.

[5] D. A. Graves, “Vocabulary profiles of letters and novels of Jane Austen and her Contemporaries,” *Persuasions On-line*, vol. 26, no. 1, 2005.

[6] R. Thisted and B. Efron, “Did Shakespeare write a newly-discovered poem,” *Biometrika Trust*, vol. 74, no. 3, pp. 445-455, 1987.

[7] F. Mosteller and D. L. Wallace, “Interference in an authorship problem.” *Journal of the American Statistical Association*, vol. 58, no. 32, pp. 275-309, June 1963.

[8] C. S. Brinegar, “Mark Twain and the Quintus Curtius Snodgrass letters, a statistical test of authorship,” *Journal of the American Statistical Association,* vol. 58, no. 301, pp. 85-96, Mar. 1963.

[9] S. Lyu, D. Rockmore, and H. Farid, “A digital technique for art authentication,” *Proceedings of the National Academy of Science*, vol. 101, no. 49, pp. 17006-17010, Dec. 2004.

[10] P. Sallis, A. Aakjaer, and S. MacDonell, “Software forensics: old methods for a new science,” *Software Engineering: Education and Practice*, pp. 48-485, 1996.

[11] L. Hunyadi, K. Abari, and E. Toth, “Forensic linguistics, its contribution to humanities Computing,” *Literary and Linguistic Computing*, vol. 8, no. 1, pp. 49-62, 2003.

[12] H. Farid, “Detecting digital forgeries using bispectral analysis,” *Technical Report AIM-1657*, 1999.

References

Texts used for figure on slide 4:

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*.

Text used for figure on slides 7 and 8:

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

Texts used for figures on slide 17:

M. Twain, (S. L. Clemens), *A Connecticut Yankee in King Arthur's Court*, 1889.
W. Shakespeare, *As You Like It*, 1601.
L. M. Alcott, *Little Women*, 1867.

Texts used for figures on slide 18:

M. Twain, (S. L. Clemens), *A Connecticut Yankee in King Arthur's Court*, 1889.
M. Twain, (S. L. Clemens), *The Adventures of Tom Sawyer*, 1876.
M. Twain, (S. L. Clemens), *The Gilded Age*, 1873.