Joel Youngblom
Nat. Lang. Proc.
May 11, 2004

VOYNICH MANUSCRIPT ANALYSIS PROGRAM

BACKGROUND

The Voynich Manuscript became popular in recent times after Wilfrid M. Voynich, an American collector, brought it back here from Frascati, Italy (Zadonelia, 2001). When Voynich discovered the manuscript at this time, there was a letter inside dated 1666 from the rector of the Charles University of Prague submitting the book to be deciphered in Rome (Zandbergen, 2004). In the letter, he mentioned that the Roman Emperor Rudolph II of Bohemia had bought the manuscript for six hundred ducats (Whitfield, 2003). Six hundred ducats is approximately fifty thousand US dollars (Whitfield, 2003).

The Voynich Manuscript is written on 6x9 inch paper with some larger pieces folded to fit inside script (Zandbergen, 2004). The 234 pages of the manuscript are all written in an unknown. The language has been dubbed "Voynichese," for obvious reasons ("Another twist in the tale," 2004). Some of the characters almost resemble letters of the Roman alphabet, while other characters are unrecognizable. Since no other samples of this language are known to exist, punctuation and sentence structure of "Voynichese" is hard to determine. People have been trying to decipher this text since the mid-1600s. The manuscript didn't become popular until Voynich brought it to the United States and began distributing copies. In the last 90 years, a few have claimed discover the meaning behind the scribbles. All have later been found to be false. The resistance to decryption has lead Voynich hobbyists to a few possible explanations. One possibility is that the manuscript is encrypted with some unknown method. Some have also speculated that the language represents the Chinese language written down phonetically. Another option is that the manuscript is just gibberish. Some believe that the manuscript was created using an ancient encryption tool (Whitfield, 2003). The Roman Emperor Rudolph II of Bohemia is known for buying many ancient texts, some of them phonies. It's not too hard to believe that someone fabricated this book in hopes of procuring $50,000.

There are many different pictures accompanying the text. Hobbyists analyzing the manuscript have divided it into a number of apparent sections. The largest of these sections appears to be dedicated to plants and herbs. The pictures drawn in this section are mostly unrecognizable. Only a few of the plants depicted resemble any real plant life. The next section is filled with signs of the Zodiac and other apparently astronomical illustrations. One of the illustrations resembles the Andromeda Galaxy. Next section has illustrations resembling body parts and tubes such as blood vessels. More sections follow, each with different types of illustrations. Throughout most of the text there are drawings of pot-bellied naked women. If they were clothed, that could provide some clue to the origins of the manuscript. The art accompanying the text is crude. Zadonelia (2001) mentions a 12-year old could have drawn the art if it weren't for the subject matter.

METHODS

I used the Takeshi Takahashi transcription. This file is available online at http://www.voynich.com/pages/PagesH.txt. I wrote programs to translate both this file and English text files to a readable and consistent format. In this format, each line of text in the file represents a sentence. Words are separated by spaces only—all punctuation has been removed. This helps to make more generalized programs instead of writing specific programs for each type of text that will essentially perform the same tasks.

I had to make certain assumptions in editing the transcription file. These assumptions are the basis for the editing programs. There are wildcard characters present in this transcription that denote unreadable characters. The editor program removes these wildcards and whatever "word" it occurs in. All line and page comments were also removed from the transcription.

These are the methods used to analyze the Voynich Manuscript:

1. Zipf's Law
2. Entropy
3. Word Family Recognition

Zipf's Law states that in a large sample of text, a few words will occur frequently and many words will occur infrequently. This relationship has proven true for other applications as well, such as comparing the colors in a photograph. (Black, 2002) In other words, if the words in a sample of text are sorted from most frequent to less frequent this property would be observed:

$$\text{Rank(word)} \times \text{Frequency(word)} = K; \quad K \text{ is a constant}$$

This property is not exactly true. It is more an approximation. In actual tests, a somewhat stable K value is reached after an initial period of stabilization. A list of K values isn't altogether very helpful in determining a relation to actual human languages, so zipf.pl finds the average K value and the standard deviation of the 'constant.' Here are the calculations:

$$K_{avg} = \text{Sum}(k_i) / \#\text{ of types}$$

$$K_{std.\ dev.} = \text{Sqrt}(\text{Sum}((k_i - K_{avg})2)/\#\text{ of types})$$

To make further sense out of this information the standard deviation will be taken as a percentage of the average K. This will show how much the K value varies relative to the average value.

The entropy of the manuscript will be analyzed in a few different ways. *VMC_EDITOR.PL* creates two edited versions of the Voynich Manuscript. One file, suffixed "_w," is the product of parsing the text assuming that each segment of text separated by a "." is an individual word. The other file generated, suffixed "_l," assumes that each period separated segment is a sentence. Therefore, each letter represents a word. This would correspond to languages such as Chinese where individual characters are words in themselves. The English editor also creates two text files. Obviously Roman characters are letters that make up words. This file will prove useful, however, in analyzing the cross and relative letter entropy.

Probability of x:

$$P(x_i) = \text{freq}(x_i) / \#\text{ of words}$$

Entropy of x:

$$H(x) = -\text{Sum}(P(x_i) \times \log_2(P(x_i)))$$

Relative Entropy of x relative to q:

$$D(x||q) = \text{Sum}(P(x_i) \times \log(P(x_i)/Q(x_i)))$$

Cross Entropy of x and q:

$$H(x,q) = H(x) + D(x||q)$$

Entropy is a measure of how much information is carried with data. Predictable data has a lower entropy value. Relative entropy shows how similar two samples of data are together. It is used to compare a proposed model, q, to a "true" model, x. The relative entropy value is between 0 and 1. A relative entropy value near 0 means the two samples are nearly identical. Conversely, a value near 1 means the samples have little in common. Cross entropy is the combination of the two. Cross entropy that is close to the entropy of x means that the two samples of data are similar and can be substituted for each other (Jurafsky and Martin, 2000).

Cross entropy may be the most useful calculation. The value obtained can directly relate the two samples of text. Unfortunately, in this case the cross entropy must be calculated on a letter by letter basis. Cross entropy is set up in such a way that there has to be a probability for each x_i to occur in both models. The occurrence of Voynichese words in English text is not likely and vice versa. Even after using the letter model for both samples, some Roman letters apparently were not present in the transcription. If the cross entropy program finds letters with a zero frequency it switches from maximum likelihood estimation to Witten-Bell smoothing. Witten-Bell smoothing is used to assign probabilities to zero frequency events (Jurafsky and Martin, 2000):

$$P_{WB}(x_i) = freq(x_i) / (T + N), \text{ when } freq(x_i) != 0$$
$$P_{WB}(x_i) = T / (Z \times (T + N), \text{ when } freq(x_i) == 0$$

This assigns zero probability events very low probabilities.

A Voynich hobbyist has created a sort of stemming guide to the manuscript. This page is available at http://www.dcc.unicamp.br/~stolfi/voynich/Notes/015/pages-html/index.html. The author has identified words that vary only by a few letters. Here are the basic functions used to relate similar words:

```
map 'p' to 't' and 'f' to 'k'
map 'k' to 't' and 'f' to 'p'
replace 'ei' by 'a'
map 'a' and 'y' to 'o'
delete the word-initial 'q'
delete any embedded spaces
```

From this web site I constructed a dictionary file: "vmcdict.txt." "wfr_editor.pl" takes the original Voynich file and constructs another version of the manuscript, but this time if it finds a match to any variation of a word in the dictionary file, that word from the dictionary will be substituted into the output file.

EXPERIMENTAL RESULTS

Test 1 output:

```
EVALUATING   vmc.txt   AGAINST   vmc.txt

Based on NON-word family text:
    Average score: 100%
    Grade: A++
Based on word family text:
    Average score: 100%
    Grade: A++

EVALUATING   vmc.txt   AGAINST   vmc.txt
```

```
vmc_l.txt vs. vmc_l.txt
    Relative Entropy: 0
    Grade: A++

vmc_lfr.txt vs. vmc_lfr.txt
    Relative Entropy: 0
    Grade: A++
```

Test 2 output:

```
EVALUATING   holmes.txt   AGAINST   vmc.txt

Based on NON-word family text:
    Average score: 93.1813825484838%
    Grade: A
Based on word family text:
    Average score: 93.1813825484838%
    Grade: A

EVALUATING   holmes.txt   AGAINST   vmc.txt

holmes_l.txt vs. vmc_l.txt
    Relative Entropy: 0.753521588744275
    Grade: Fail!

holmes_l.txt vs. vmc_lfr.txt
    Relative Entropy: 0.722943372360331
    Grade: Fail!
```

Test 3 output:

```
EVALUATING   sonnets.txt   AGAINST   vmc.txt

Based on NON-word family text:
    Average score: 85.5754865444513%
    Grade: B
Based on word family text:
    Average score: 85.5754865444513%
    Grade: B

EVALUATING   sonnets.txt   AGAINST   vmc.txt

sonnets_l.txt vs. vmc_l.txt
    Relative Entropy: 0.813499951403095
    Grade: Fail!

sonnets_l.txt vs. vmc_lfr.txt
    Relative Entropy: 0.771108120181863
    Grade: Fail!
```

Test 4 output:

```
EVALUATING   sonnets.txt   AGAINST   iliad.txt

Based on NON-word family text:
    Average score: 86.5456327186745%
    Grade: B
Based on word family text:
    Average score: 86.5456327186745%
    Grade: B
```

```
EVALUATING   sonnets.txt   AGAINST   iliad.txt

sonnets_l.txt vs. iliad_l.txt
    Relative Entropy: 0.00809706155945599
    Grade: A++

sonnets_l.txt vs. iliad_l.txt
    Relative Entropy: 0.00809706155945599
    Grade: A++
```

Test 5 output:

```
EVALUATING   tale.txt   AGAINST   treasure.txt

Based on NON-word family text:
    Average score: 97.4061182209654%
    Grade: A+
Based on word family text:
    Average score: 97.4061182209654%
    Grade: A+

EVALUATING   tale.txt   AGAINST   treasure.txt

tale_l.txt vs. treasure_l.txt
    Relative Entropy: 0.00186336839314783
    Grade: A++

tale_l.txt vs. treasure_l.txt
    Relative Entropy: 0.00186336839314783
    Grade: A++
```

EVALUATION

In order to assign a grade to the two samples of text in the "test1.pl" program, two factors were taken into consideration. Values from the standard deviations relative to the average Zipf value and the entropies were averaged. All values used to calculate the relation between the two samples are fractions. Therefore the further they differ from 1 the less similar the two are. By finding this difference, subtracting it from 1, and multiplying by 100 you can find a percentage. I assigned grades based on the following scale:

```
100-97 A++
 93-96 A
 90-92 A-
 87-89 B+
 83-86 B
 80-82 B-
 77-79 C+
 73-76 C
 70-72 C-
 60-69 D
  0-59 Fail!
```

The "test2.pl" program just assigns grades based on the letter cross entropy. This time the closer to 0 the relation is the more similar the files are. To find percentage, subtract the relative entropy from 1 and multiply by 100. The grades follow the same scale.

REFERENCES

Another twist in the tale; The Voynich manuscript. (2004). *The Economist (US), 370,* 71.

I found this article to be very concise and informative. It provides a good history of the manuscript and also provides a list of likely origins of the text. The article goes into detail about the possibility that the manuscript is a fraud. This article was helpful for background information on the Voynich Manuscript.

Black, Paul E. (2 Aug 2002) Zipf's Law: definition. National Institute of Standards and Technology. Retrieved April 25, 2004, from http://www.nist.gov/dads/HTML/zipfslaw.html.

Jurafsky, Daniel, & Martin, James H. (2000) Speech and Language Processing. In M. Horton (Ed.), *Entropy* (pp. 223-227). Upper Saddle River, New Jersey: Prentice-Hall, Inc.

This section of the textbook was helpful for development of my entropy, relative entropy and cross entropy algorithms.

Jurafsky, Daniel, & Martin, James H. (2000) Speech and Language Processing. In M. Horton (Ed.), *Witten-Bell Discounting* (pp. 210-213). Upper Saddle River, New Jersey: Prentice-Hall, Inc.

This section of the textbook was helpful for development of my Witten-Bell Smoothing algorithm. I used that algorithm for analyzing cross entropy of the letters in the manuscript. When all letters of the Roman alphabet are not present in either of the models submitted to that program, it switches from maximum likelihood estimates to probabilities after Witten-Bell Smoothing.

Whitfield, John. (17 Dec. 2003) World's most mysterious book may be a hoax. *Nature Science Update.* Retrieved April 5, 2004 from http://www.nature.com/nsu/031215/031215-5.html.

Zandbergen, René. (8 Mar. 2004). Voynich MS – Short Tour. Retrieved April 5, 2004, from http://www.voynich.nu/s_intro.html.

Zadonelia, Catherine. (2001) Book of riddles: Are we on the brink of decoding the most mysterious document in the world. *New Scientist, 172,* 36-39.

Along with a descriptive history of the manuscript, Zadonelia goes into detail about the pictures and characters used. She next delves into various attempts to decipher the text. This section was very helpful for deciding what methods to use to test the manuscript and how to evaluate its properties. It doesn't go into great detail about the methods used to evaluate the text, but this is where I got the idea to analyze word length and entropy.