Natural Language Processing

Project Proposal:

Voynich Manuscript

By: Scott Daniels

4/14/04

Introduction

The problem that I am attempting to solve is trying to distinguish whether the Voynich Manuscript is a human language or not. The Voynich Manuscript contains hundreds of ancient pages with many strange writings and pictures of flowers, mythical lands, and naked women. Found in the mid 1600's, the Voynich Manuscript has been transcribed, or rewritten, into English letters so that we can try and find a pattern or a solution to the mysteries that the manuscript holds. There are many theories as to what the manuscript could possibly entail or even theories as to why the manuscript was written. Some believe that the manuscript is a giant hoax written by a man in order to fool Emperor Rudolph II of Bohemia out of lots of money. Rudolph II was a great collector of manuscripts in his time and he was known to spend large sums of money for manuscripts that are now known to be counterfeit.

The Voynich Manuscript has been studied extensively by cryptologists, linguists, and many other language experts so much evaluation has already been done by far smarter people than I. Many experts believe that the Voynich Manuscript is of European decent because the pictures of humans in the manuscript all depict the styles and fashions of European culture at the time it was theorized to have been written. Other experts believe that the Voynich "language" has close ties with the Chinese language in how the suffixes and prefixes of the words are composed. The translation of this manuscript is one of the most sought after tasks in all of language processing and cryptology fields simply because it has never been deciphered. No one knows what secrets that it might hold or if the hundreds of pages retain nothing but mindless blather. I hope in my research to at least answer if the Voynich Manuscript is a human language or not so that many people don't waist their time deciphering one of the biggest hoaxes every written.

Pre-Experimental Research

A source that can assist my search for the answer of whether the Voynich Manuscript is a language or not is a dissertation titled “Maximum Entropy Models For Natural Language Ambiguity Resolution” written by Adwait Ratnaparkhi. There are many topics brought up in this dissertation, but the key ideas are to come from his maximum entropy framework discussion. I used the overall formula for my entropy calculations from this text and I also used the author. I also took an extended look at the authors ideas about Non-Overlapping Features because I knew that my cryptology attempts were going to be using two completely different texts so there would be many non-overlapping features in my calculations. In fact, the author states that the maximum entropy framework reduces to a very simple type of probability model when the features do not overlap so my calculations will not have to be that difficult after all (Ratnaparkhi 33). This article will mainly help me deal with my tree structure in my single substitution cryptology attempt (discussed in detail later).

Another source that can help me with my research is entitled “Can Zipf Analyses And Entropy Distinguish Between Artificial And Natural Language Text?” written by Cohen, Mantegna, and Havlin. This article deals with how you can use Zipf’s Law and Entropy calculations to see if a text is real or not. The part of the article that I will focus on will be about Zipf’s Law. The article describes the “necessity” for a text to follow Zipf’s Law and it also describes how a variation of Zipf’s Law, called the inverse Zipf analysis (not used in my research), could be a better estimator of linguistic tendencies between two texts (Cohen 13).

Overview of Approach

My approach in solving this problem will be from multiple angles. While focusing on the main objective, determining whether or not the manuscript is a human language, I also will attempt some very basic code breaking tactics. Although I know that experts have been doing this for years, I figure I'd give a try at cracking the code if it is in fact a code.

To answer the question of whether the manuscript is a human language, I will use the Profiler program I designed to see if Zipf's law indeed holds for the text. I will split on the clearly appointed word boundaries (periods) and use that word vocabulary to see if there is a rank-to-frequency propensity. If there is a correlation, I will push towards the fact that this is, in fact, a human language. Also, I will use the Ngram program programmed in Lab4 to construct unigrams, tri-grams, and five-grams and then test those sentences on the Profiler program to see if there is a correspondence with Zipf's Law. In addition to these tests, I plan to do an entropy calculation (including cross-entropy) based on the characters in the manuscript to see if there are any similarities with a well known text from Italian literature, Dante's Divine Comedy. This entropy value range will describe the text's strength of being a human language or not.

As far as trying to crack the Voynich code, I intend to use basic cryptology techniques to format the text and then run the same Profiler and entropy tests on that formatted text. I will be using a single letter substitution algorithm that takes the text and substitutes all 26 letters in for a character and then take that new text and find the cross entropy or straight scoring algorithm with the Divine Comedy. The highest "score" from the scoring algorithm will be the "most likely" substitutions and the algorithm will pass to the next letter until all the letters are decoded. If the score for a new text segment is equal for all of the 26 substitutions, I plan to use a random character for the substitution and move on with the algorithm. I realize that this is pretty far fetched idea, but I think it will be interesting to try to use some of these random cryptology tactics and see what I can get back from the results

Evaluation Plan

To evaluate my findings, I plan to use the "gold standard" Italian text known as Dante's Divine Comedy.. This paper was written by Dante in the early 1300's. The time and the location that the Voynich Manuscript was found in matches rather well the text so I deduced that this would be a good standard. The results that I gather from the Profiler data will be used to relay the differences between the Voynich Manuscript and this data. For example, if the K value in the regular corpus seems to be leveling off at a constant value (which we have seen that most corpora do according to Zipf's Law) I will calculate the difference between the individual K values of each word and also the overall average of the leveled off K value. This should give me some correlation as to how closely Zipf's Law holds for the Voynich Manuscript.

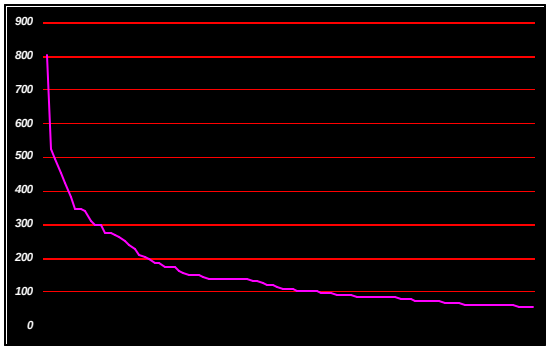
As far as the Entropy of the Voynich Manuscript goes, I will use the entropy and cross entropy formulas in tandem with the "gold standard" corpora as described above to see the real differences between the two texts. In that algorithm, I will be using the smallest cross entropy to continue down the list of single letter substitution. Along with the cross entropy, I will also do tests with normal entropy. If the difference in the entropy between the Voynich Manuscript and the "gold standard" is remarkably high, then I can conclude that this Voynich Manuscript is indeed a candidate for fallacy. I will also calculate the cross entropy of two known texts to make sure that I am comparing the data in a correct manner. The entropy and cross entropy factors will probably be the strong point in my conclusions as to whether the Voynich Manuscript is a human language or not.

Experimental Results

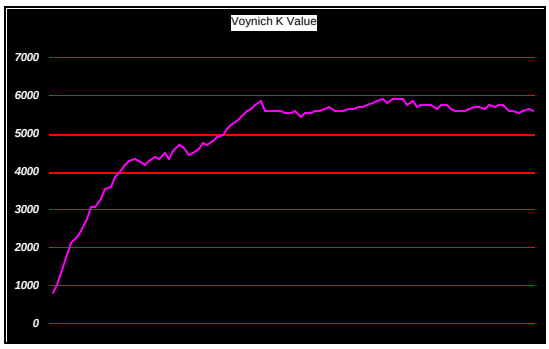
The three approaches that I took brought some very interesting results. The first approach, the Zipf's Law analysis, came back with results that strengthen the theory that the Voynich Manuscript is indeed a human language. Zipf's law, the distribution of word rank times the frequency which that word occurred, can show whether a text has human qualities or not. I ran the Profiler program on the Sherlock Holmes text as well as the Voynich Manuscript and I made these graphs from the data I received:

Voynich Manuscript

(Rank * Frequency) vs. Rank

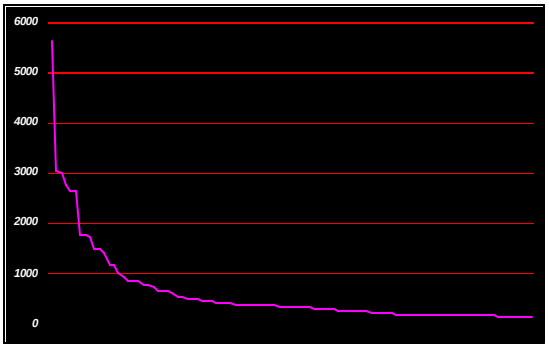


K-value vs. Rank

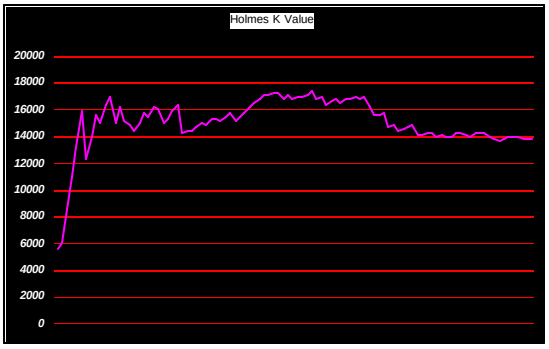


Sherlock Holmes

(Rank * Frequency) vs. Rank



K-value vs. Rank



From this data, it seems that the Voynich text shows some similarities to the data received from the Holmes text. The rank * frequency graphs are remarkably similar between the two texts and the K-value shows similar characteristics such as a data stabilization point. The biggest difference between the two texts is that the K-value for the Voynich Manuscript seems to grow a lot slower than the Holmes text. From the data that I gathered from the Zipf's Law analysis, for the most part; I can say that the Voynich Manuscript shows a strong relationship to human text.

The second approach taken, the entropy and cross entropy calculations, made the Voynich Manuscript seem like a human language as well. The data gathered from entropy calculations are as follows:

	Entropy Calculation
Voynich Manuscript (with stars)	**10.5579814914084**
Voynich Manuscript (no stars)	**10.5375691704889**
Sherlock Holmes	**10.0666574711316**
Dante's Divine Comedy	**10.9058119575507**

	Cross Entropy (Divergence) Calculations
Holmes vs. Voynich (no stars)	**.808615736555144**
Holmes vs. Divine Comedy	**.926722686971558**
Divine Comedy vs. Voynich (no stars)	**.942933364344184**
Divine Comedy vs. Divine Comedy	**0.00000000000000**

Let's start with the Entropy calculations. The entropy values for all four of the texts that I tested were surprisingly similar, all between 10 and 11. The Italian text (Divine Comedy) scored the highest of the four texts that were tested and the Sherlock Holmes text scored the lowest. The Voynich Manuscripts (with or without stars) scored about the same at around 10.5. This is well within the bounds of a normal language text which was theorized to be between 9 and 11. This data strengthens the claim that the Voynich Manuscript is a human language. The Cross Entropy (Divergence) data does not show as much as the Entropy value. Basically this value represents how different two texts are. The numbers that I got for the English vs. English texts did not surprise me as the numbers were very low. On the other hand, the values that I got back from the Italian vs. Voynich and the English vs. Voynich text calculations were startling. It almost seems that Sherlock Holmes is more closely related to the Voynich Manuscript than it is to the Divine Comedy text. To test to see if my values were calculated correctly, I ran the Divine Comedy against the Divine Comedy and the result was a sharp zero difference which is the correct value. The Divergence calculations strengthened the theory that the Voynich Manuscript is a human language.

The data that I received from the third approach that I took, the cryptology experiment, was not as clear cut as the other approaches. Basically the algorithm works like this:

1. Read the text to be "decrypted" in, lowercase the letters, and put each letter into a giant array (lowercase letters signifies letters yet to be "decrypted")
2. While not at the end of the array and the letter to be "decrypted" is not a space or a capital letter:
    A. Substitute the letter globally with one of the 26 capital letters in the alphabet (A-Z)
    B. Take that newly formed text and throw it threw one of the two scoring algorithms I devised (descriptions below) and assign this value it to the capital letter in a hash
        1. Divergence (explained in section above)
        2. Straight Score
            a. Return the number of words in the text that are words in the dictionary specified

C. Once all 26 capital letters have been assigned values in the hash, sort the hash from largest value to smallest and pick the capital letter at the top of the list (if all 26 values are the same, perl will pick a letter at random). Set the array as the text with the chosen letter replaced globally

D. Keep track of capital letters that have been decrypted and make sure the program does not pick the same "decoded" letter twice.

E. Reset all data and go to the next letter

3. When all the letters are "decrypted" (or capital) print out the final string that was decrypted and end the program

If this sounds confusing too confusing to follow I will attempt to provide an easY, English example of the algorithm.

Example

Sentence = "a street is where the crime has happened"
Dictionary = English (Holmes text)
Scoring Algorithm = Straight Score

- Take letter at index 0 (a) and replace it with every capital letter and get score:
  "A street is where the crime hAs hAppened" → 8
  "B street is where the crime hBs hBppened" → 7
  "C street is where the crime hCa hCppened" → 7
  …..”Z street is where the crime hZs hZppened" → 7
  (Sentence A has 7 English words in it while sentence B only has 6 because the substitution in the word "happened" made that word invalid)
- Program picks letter A as best choice, replaces Sentence, and moves on
- Take letter at index 1 ( ) and advance because it is a space
- Take letter at index 2 (s) and replace it with every capital letter and get score:
  "A Btreet is where the crime hBs hBppened" → 7
  (notice it doesn't use A again)
  … "A Street is where the crime hAs hAppened" → 8
  ….”A Ztreet is where the crime hAs hAppened" → 7
- Program picks letter S as best choice, replaces Sentence, and moves on
- ……
- Take letter at index (40) and replace it with every capital letter and get score
  "A STREET IS WHERE THE CRIME HAS HAPPENB" → 7
  …”A STREET IS WHERE THE CRIME HAS HAPPENED" → 8
  …”A STREET IS WHERE THE CRIME HAS HAPPENEZ" →7
- Program picks letter D as best choice, replaces Sentence, and ends

"Decrypted" Sentence = A STREET IS WHERE THE CRIME HAS HAPPENED

Obviously, this approach to "decrypting" is very naive. When given a correct text (such as in the example), the algorithm is almost always going to work, but the problem is that the "encrypted" sentences don't always come so cut and dry. For example, if the

Voynich text is specified to be the sentence to be "decrypted", the program does not always make the best decisions at the beginning of the algorithm. Because of the extensive computational time it takes to run one of these linear programs to completion, a tree structure (which would the best structure) would take an eternity to complete, but would yield better results because the program could correct some of it's mistakes made earlier on in the selection process. This early error propensity can be seen in the following example:

Sentence = "a street is where the crime happened"
Dictionary = English (Holmes)
Scoring Algorithm = Divergence

First letter picked : T
→ "T street is where the crime hTppened"
(Notice that the last word "happened" is unfixable now that a mistake was made)
Second letter picked : X
→ "T Xtreet iX where the crime hTppened"
(Notice that the word "street" is unfixable because of errors, but the program sees "iX" as the roman numeral "4" so it picks X as the next letter)
…
Last letter picked : F
→ "T XSREES IX WHERE SHE PRICE HTODEQEF"
(Notice that the words that were deemed unfixable before are just a garbled mess, but all of the other words in the text are, in fact, English words and close to what they are supposed to be).

Here are the results of some tests that I ran:

	Matched Words Before	**Matched Words After**
One Page of the Voynich Divergence Scoring Italian Dictionary	**2**	**5**
One Page of the Voynich Straight Scoring Italian Dictionary	**2**	**9**
Full Voynich Divergence Scoring Italian Dictionary	**48**	**50**
Full Voynich Straight Scoring Italian Dictionary	**48**	**116**

I am measuring my success and failure on how many words are Italian words in the ending "decryption". I couldn't find a more suitable measure because this cryptology endeavor is such a shot in the dark. Basically, if the program doesn't work perfectly, I can classify it as failure because one little mistake in choosing the next letter will destroy

the whole thing. Keep in mind that this will only work if indeed the Voynich text is written in Italian which is a very high improbability. This approach was just a shot in the dark and it turned out that it failed because of the complexity restrains that my program had and the fact that the Voynich Manuscript had to be Italian for it to work.

Through all of my research of the Voynich Manuscript, I have witnessed a lot of data that supports the claim that the manuscript is a human language. I certainly haven't found any data that says it is not. The Zipf's Law analysis, Entropy, and Cross Entropy calculations all strengthen this claim, but my cryptology research didn't really push the strength of the claim one way or another. The cryptology portion of my research was basically done to take a shot in the dark at cracking the Voynich Manuscript. Obviously, I am not really any closer to solving the riddles of the Voynich Manuscript, but I did get a lot of information about the ancient pages that I did not have before.

References

Ratnaparkhi, Adwait. Maximum Entropy Models For Natural Language Ambiguity Resolution. Diss. U of Pennsylovania, 1998. 20 Apr. 2004 http://citeseer.ist.psu.edu/31701.html.

Cohen, A., Mantegna, R.N., Havlin, S. Can Zipf Analyses And Entropy Distinguish Between Artificial And Natural Language Text? Retrieved April 20, 2004, from Department of Physics, Bar-Ilan University. <http://citeseer.ist.psu.edu/cache/papers/cs/9017/http:zSzzSzory.ph.biu.ac.ilzSz~havlinzSzPSzSzcmh289.pdf/can-zipf-analyses-and.pdf/>