### Voynich Manuscript

Malika, Grace, George

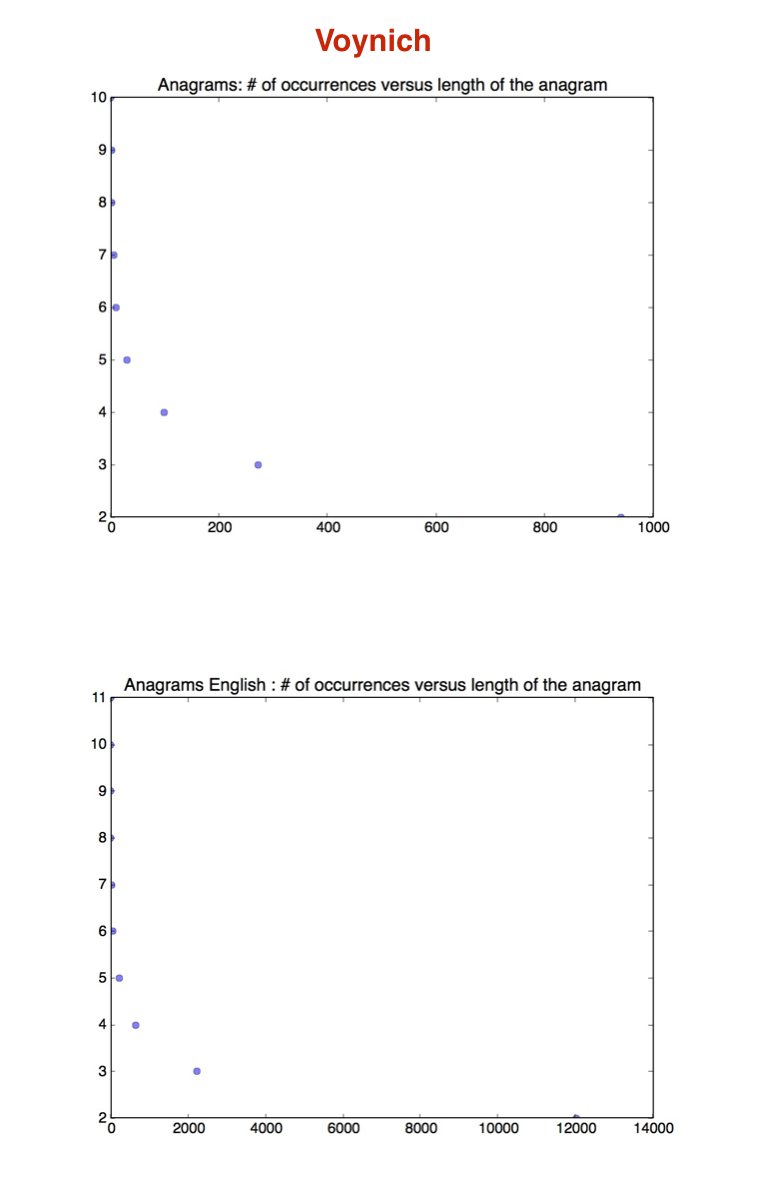
### Motivation

The Voynich Manuscript provides an interesting opportunity to apply the study of computational linguistics to find insights into the unknown language used. Through word length comparisons, anagram analysis, hidden Markov model, and arithmetic compression, the language used in the Voynich Manuscript mimics a real one.

### Word Length Analysis

Based on comparison of three sets of words from different languages, Voynich and two other languages (French and Latin), there is a very similar structure among them when looking at word lengths. The peak occurs around 5 < length < 10, and the histogram is downward sloping with a right skew after length > 10. This again supports the theory of the Voynich manuscript being a real language.

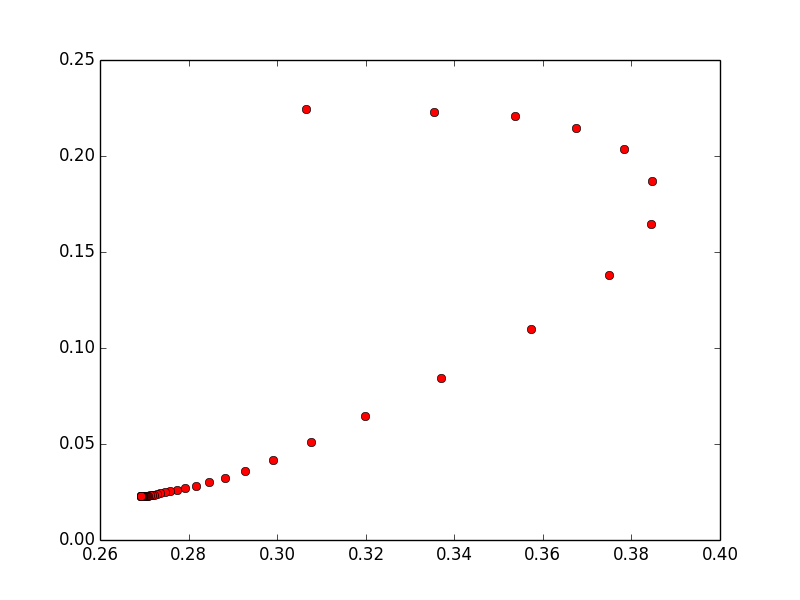
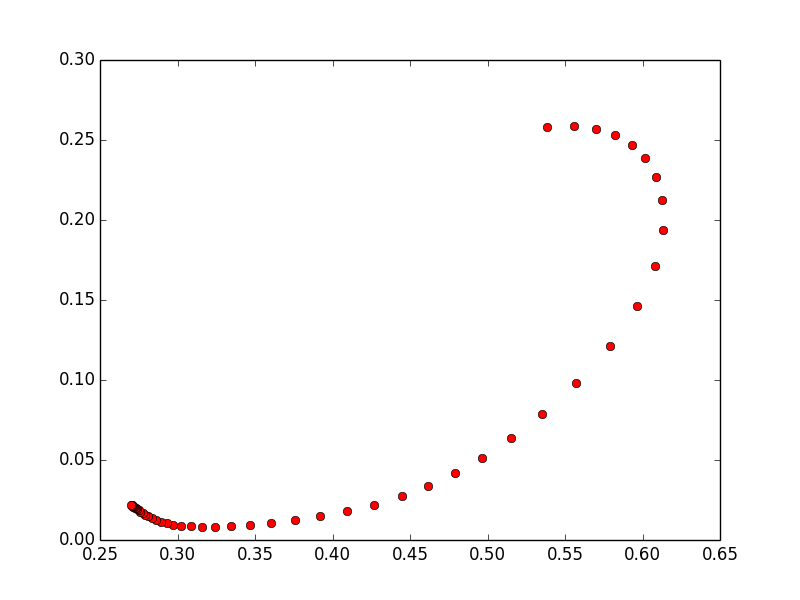
### Anagrams Analysis



Here, the Voynich manuscript is compared to English words based on anagrams of length > 2. The top graph is the former, and the bottom one represents the latter. Both graphs are downward sloping and again have a similar shape.

Note: The difference of the x-axis of course amounts to the fact of the set size. The set of English from Homework #2 was much larger.

### HMM



### Log ratios of emissions from the 2 states

### Positive:

### q 82.3436572959

### 4 63.3855617313

### 5 13.1085603862

### c 9.06943759915

### z 8.90469510507

### p 7.94272359696

### f 5.80681761037

### h 5.46311573334

### s 4.99603600306

### t 4.63075251894

### d 3.64349546882

### o 2.79787590094

### 3 1.28370990834

### 0.775906189335

### - 0.168477602574

### Negative:

### w -243.894698224

### m -112.886720195

### n -107.920317862

### ? -88.6771630732

### # -68.2031683708

### i -63.6249715347

### l -61.3645495985

### j -60.7669051777

### 9 -47.7265553148

### k -44.6522382539

### 7 -4.71462327515

### r -3.22548245287

### a -3.02556248326

### 6 -2.91718501653

### x -2.72971983557

### g -1.22569951688

### y -1.10215798215

### 8 -0.595375433845

### = -0.526399311092

### e -0.210765957404

### 2 -0.0867359585459

---------------------

Sample of Results Table (without plog) – probability of each word

---------------------

-4ODCOE# 2.4693970113e-10

-4ODOE# 1.86560202665e-09

-8TOR# 2.19520087882e-08

-GDCIIIR# 2.07185794879e-12

-ODAN# 8.11957725704e-09

-ODCC# 8.22659301245e-08

-ODCCG# 1.1889502434e-08

-SCDG# 2.09420332248e-08

2# 0.00258577136387

2-2# 5.50002370659e-07

2-2SO# 6.14389786614e-10

2-8AN# 2.85142075445e-09

2-8G-8G-GDCCTG# 1.83229592137e-21

2-AM-SAPTC8G# 1.57055670205e-20

2-GHOM# 6.45723789934e-11

2-OD# 7.70156850407e-08

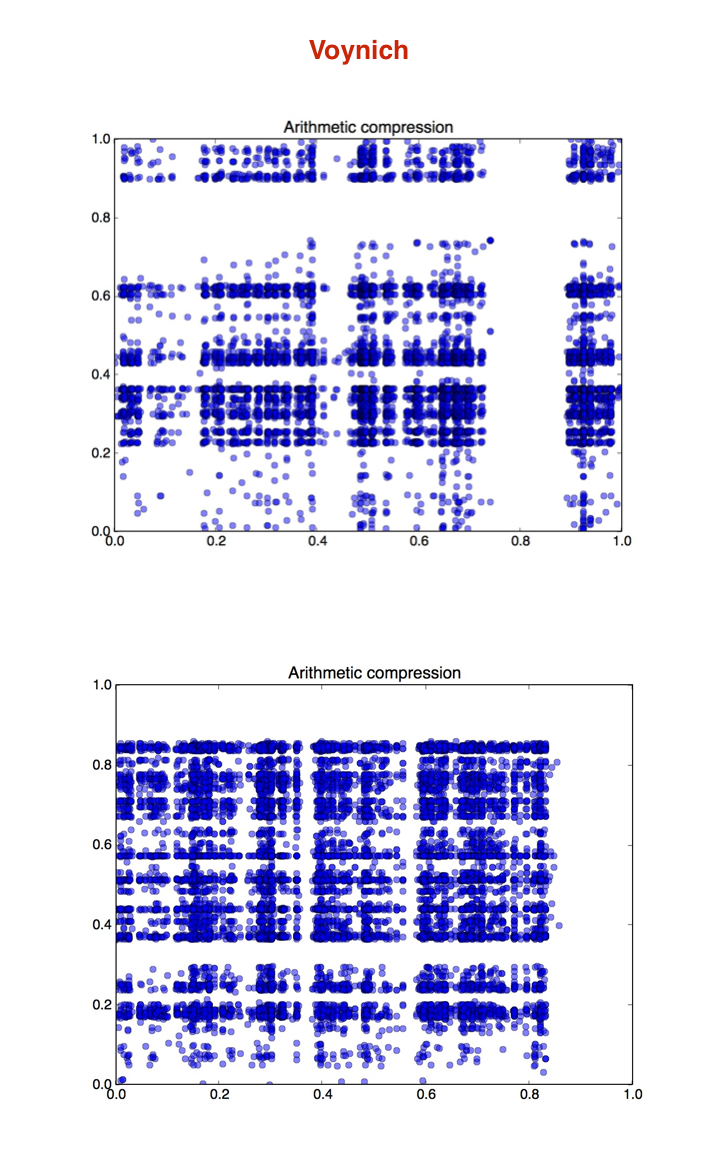
2-ODOE# 1.1121047553e-09

2-OHAM# 2.20418568349e-10

2-SCDG# 3.00146353279e-10

Running the HMM Model over the Voynich document produced these results. The two graphs compare the state transitions a(0,1) and a(1,0) after respective runs of 50 iterations each. Similar to the results run over a dictionary of English words, the plot moves toward the upper right corner as the maximization algorithm finds optimal probabilities for each of the two states. The log emission ratios further show a distinction between two sets of characters in the languages represented by the Voynich document. In English, the distinction between positive and negative log ratio emissions corresponded to one set being vowels and the other being consonants. The same could apply to the Voynich with one group indicating vowels of the language and the other the consonants. In this case, the language would have a fairly equal ration between vowels and consonants with a similar number of characters in the positive and negative log ratio emission groups. The sample of the final results table of the logs of each word further shows the similarly of the Voynich language to a real one given the words all have low probabilities yet a wide range of values, showing some to be more frequent than others without one that truly dominates the set.

### Arithmetic Compression

Most interestingly, by running the arithmetic compression algorithm on Voynich manuscript in comparison to English it seems like the similarities are quite obvious. As per Homework #7, the x-axis indicates the probability of starting with that phoneme, while y axis indicates probability of ending with it. In fact, we can see on the graph clear concentration of points that almost form a line in both graphs. For example, points at (x, 0.2) form almost straight perpendicular line. That means that phoneme is a common end phoneme. That makes sense for English since a lot of words have similar endings such as all adjectives with -ed at the end. However, it is quite surprising for Voynich manuscript.

**Morphological structure**

The graphs of arithmetic compression potentially suggest a morphological structure of Voynich. Thick lines of endings and beginnings on certain letter might indicate suffixes and prefixes.

### Punctuation

From my observation on the Voynich document provided, it seems like there are two letters that occur only at the end of the sentences and that is “K” and “L”. But those also occur in the vocabulary, so there seems to be no punctuation, at least not in a common sense.

Examples:

!%%%\*!OM.OHCCG.OHCAR.ROEOHG.HZAAR.8AM.ODAM.OR.ODAL-

DTRG.TO8AM.OE.OEHTCG.TAR.FZAR.AK-

### Conclusion

Based on comparison of Voynich manuscript versus natural languages, analysis of this assignment seems to indeed indicate number of similarities from word length to arithmetic compression to the structure of anagrams.