1.

Pw = 0.75

Pb = 0.25

W W W B B W W W

Arithmetic coding

Interval?

A = 0.75\*0.75\*0.75\*0.25\*0.25\*0.75\*0.75\*0.75 = 0.01112

H = -log 0.0112 = 6.49 bits => 7 bits

Bit sequence?

WWWBBWWW : high = 0.406629 low = 0.39551 (did the counting in my textbook by hand)

0.5 –> 0

0.25 –> 1

0.375 –> 1

0.4375 –> 0

0.40625 –> 0 -> we are inside here, buffer bits

0.390625 –> 1

0.3984 –> 1

0110011, the code length is 7 (as expected)

3.

(Did the extraction with my c# program, and saved the results to a .txt file)

My conclusion:

Depending on the language, we always have different probabilities and entropies, since languages have their own properties, what kind of words they use. For example if a language adds the letter ‘s’ if there’s more of something, then it’s more likely to have a high occurrence. For the Hungarian language, we have a lot of ‘e’ letters in our words, the second most common letter is ‘a’ or ‘t’. So yes, depending on the language we have different results, and that is the reason this could be used to detect languages.

5.

Started doing the cross entropy calculation, but I didn’t get the summation part exactly, so it’s not done 100% percent, when I have the right idea it won’t take long to finish this task.

This is what I found on Wikipedia: , I don’t see how many elements we need to Sum, I thought we do this element by element. (From the text A probabilities we try to use the text B’s probabilities).