Ex-02 answars.

2.2 (10)

as we will choose a smaller epsilon, our results will be more accure, and on the other way the oposite.

.5

a.

if we will use the train path for testing too, we will get 100% correct results if we will get a english train set and franch test set, the accure will be close to zero. its becouse the machine is very dependent on the train, and gets all the information from there, thats wat it will not get an information to mark franch mail.

b.

התוכנית אשר כתבנו אכן מתייחסת לכל מילה באופן עצמאי, זאת ניתן לראות ע"י איסוף המילים לתוך Hash הצורה נפרדת, והזנת פרמטרים כמותיים והסתברותיים ספציפית לכל מילה בהתאם להתנהגותה במהלך הלימוד ואגירת המידע. בעקבות כך אנו יכולים להבחין כי האלגוריתם מאוד נאיבי ואוסף מידע בצורה ישירה אשר אינה משקפת את תרבות הכתיבה שלנו בני האדם. עיקרון מאוד חשוב שנתבסס עליו הוא העובדה כי לא ניתן לסווג ע"י ספירה האם הודעה מסויימת היא "דואר זבל" או דואר רגיל. הרי דוגמא מצומצמת על סיווג של משפט (אפילו לא הודעה אשר מורכבת ממספר משפטים), המשפט "פה תגלה איך תוכל להרוויח הרבה כסף תוך כמה ימים" והמשפט "איך תוכל להעביר את הכסף? ותוך כמה ימים?" מכילים יותר מילים משותפות מ-לא משותפות, אך ברור שהמשפט הראשון הוא משפט אשר כלול בהרבה "הודעות זבל" והמשפט השני בהודעה רגילה ואפילו חשובה. כמו בהוכחות מטמתיות, דוגמא נגדית שוללת טענה ולכן אין להסתמך על צורת החישוב של האלגוריתם לתוצאות נכונות בסבירות גבוהה, במיוחד לאחר שאני הצלחתי בקלות להרכיב שתי משפטים אשר תומכים בשלילה. בנוסף לכך קיימות מילים מסויימות בעלות יותר ממשמעות אחת, או שימושים מגוונים בצורות כתיבה "סלאנגים", פרמטר נוסף אשר משבש את שיטת החישוב הנ"ל. לבסוף עלינו לזכור כי המורכבות של הודעת דואר אלקטרוני ממספר רב של מילים גורמת להתייחסות למילים עצמם באלגוריתם הסיווג לשגיאות (זאת ניתן לראות גם בתוצאות הסיווג של התוכנה שכתבנו) והשאלה האם דואר מסויים הוא "דואר זבל" או דואר רגיל צריכה לכלול את הבנת המלל והכוונה בדואר בכדי להענות בצורה נכונה, דבר שאינו יכול להתממש ע"י ספירת מילים.