תרגיל כיתה

סווג אי-מייל/ספאם (e-mail/spam).

- בתרגיל זה נשתמש באלגוריתם בייס הנאיבי Naïve Bayes algorithm כדי לתכנן מסווג אי-מייל/ספאם (e-mail/spam) שילמד לבצע הבחנה בין e-mail "אמיתי" לספאם. קבוצת האימון כוללת קבוצת e-mail וקבוצת ספאם. על-ידי שימוש בתוכן המכתב ובשורת הנושא נלמד להבחין בין spam ו- non-spam .
- הדוא"ל כבר עבר קדם-עיבוד (pre-processing) כדי להתאים אותו לשימוש על-ידי האלגוריתם. אפשר להתרשם מדוגמא מקורית של spam בקובץ spam_sample_original ומקובץ ספאם לאחר קדם עיבוד : spam_sample_preprocessed. אפשר למצוא את הקבצים כמו גם קבצי e-mail (לא ספאם למעשה קבצים מתוך) לפני ואחרי עיבוד במחיצת התרגיל במודל.

```
Subject: biz offer
Mime-Version: 1.0
Content-Type: text/html; charset="us-ascii"
Content-Length: 1594
Status: RO

<html>
Dear Sir, <br>
We have come to know that your honorable organization dealing surgical
instruments. We would like to introduce Daska Surgical Corp as
manufacturer and exporter of best quality surgical instruments. <br>
We have complete setup in this field and producing all kind of
instruments.We have our own furnus material 10 Hammers for forging,
milling plants and polishing units.That why the prices of our products
are much low than others. <br>

We are coating here prices of some instruments for your kind information.
<br>
Bakhuas Towel Clamp 5,1/2&quot; <i>US$ 0.70</i> <br>
Bakhuas Towel Clamp 3,1/2 <i>US$ 0.</i>65 <br>
MOSQUITO STR SS US$:0.60<br>
MOSQUITO CUR SS US$:0.90 with quallity<br>
KELLY 14CM SS US$:0.60<br>
CRILE FORCEP SS 14CM US$:0.60<br>
PEAN 14CM SS US$:0.65<br>
DRESSING SCISSOR14CM US$:0.50<br>
TOWEL SS 9CM US US$:0.50<br>
MAYO HAGER NEEDLE/H 14CM US$:0.65<br>
It would be pleasure for us to start business with your prestigious
company. We hope you will consider our price compatibility and anticipate
for future business dealings. <br>
```

איור 1: מקטע מדוגמא מקורית של spam מתוך הקובץ spam_sample_original

```
biz offer
dear sir,
we have come to know that your honorable organization dealing surgical
instruments. we would like to introduce daska surgical corp as
manufacturer and exporter of best quality surgical instruments.
we have complete setup in this field and producing all kind of
instruments.we have our own furnus material NUMBER hammers for forging,
milling plants and polishing units.that why the prices of our products
are much low than others.
we are coating here prices of some instruments for your kind information.
bakhuas towel clamp NUMBER/NUMBER  usDOLLAR
bakhuas towel clamp NUMBER/NUMBER  usDOLLAR
mosquito str ss us$:NUMBER
mosquito cur ss us$:NUMBER with quallity
kelly NUMBERcm ss us$:NUMBER
crile forcep ss NUMBERcm us$:NUMBER
pean NUMBERcm ss us$:NUMBER
dressing scissorNUMBERcm us$:NUMBER
towel ss NUMBERcm us us$:NUMBER
mayo hager needle/h NUMBERcm us$:NUMBER
```

```
it would be pleasure for us to start business with your prestigious
company. we hope you will consider our price compatibility and anticipate
for future business dealings.
```

איור 2: מקטע לאחר קדם עיבוד עבור הדוגמא מאיור 1.

- השורה הראשונה במסמך המעובד היא התגית שלו ואינה חלק מהמסר.
- כתובות מייל (EMAILADDR), כתובות רשת (HTTPADDR), מטבע (DOLLAR) ומספרים (NUMBER) הוחלפו על-ידי קידוד מתאים למלים (בסוגריים) כך שיתאימו לתהליך הסווג.
- העבודה של מיצוי וקטורי התכונות מהמסמכים כבר נעשתה, כך שמטריצות המלים (design matrices or document-word matrices) כבר מכילות את כל הנתונים.
- במטריצת מלים כנ"ל, השורה ה- i-ית מייצגת את המסמך או ה- e-mail ה- i, והעמודה ה- j מייצגת את המילה (לעתים משתמשים במונח סימן או token) ה- j במסמך. לפיכך הרכיב ה- (i,j) במטריצה מייצג את מספר ההופעות של המילה ה- j במסמך ה- i.
- עבור בעייה זו קבוצת המלים שנבחרה (כלומר המילון) היא קבוצת המלים בעלות תדירות לא נמוכה מדי (נדירות מדי) או גבוהה מדי (מלים כמו the, of, and המכונות content free, הנמצאות בכל מסמך או מייל ולהופעתן אין ערך עבור המודל. בנוסף נעשה גיזום למלים באמצעות אלגוריתם גיזום סטנדרטי כך שמלים כמו price, prices, priced מיוצגות על-ידי "price", כך שאפשר להחשיב אותן כאותה מילה. רשימת המלים מופיעה במסמך tokens_list.

- רשמו כמה מלים מכילה הרשימה? _____.

- מאחר ומטריצת המלים היא מטריצה דלילה, כלומר מכילה הרבה אפסים (sparse matrix), היא שמורה בפורמט מיוחד דחוס כדי לחסוך במקום. הפונקציה readMatrix מאפשרת לקרוא את המטריצה הפרושה המקורית מתוך המטריצה הדלילה.
- לדוגמא רשמו ב- Matlab:

```
[spmatrix, tokenlist, trainCategory] = readMatrix( 'MATRIX.TRAIN' );
trainMatrix = full(spmatrix);
numTrainDocs = size(trainMatrix, 1); % number of training ducuments
numTokens = size(trainMatrix, 2); % number of training tokens
```

- לדוגמא אם נקרא 50 ערכים מתוך מטריצת האימון trainMatrix נקבל את המטריצה הבאה:

```
trainMatrix(1:5,1:10)
```

0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

-
- מהו מספר הערכים שאינם 0 במטריצה trainMatrix?

_____.

- מהו אחוז הערכים שאינם 0 במטריצה trainMatrix?

_____.

- המטריצה MATRIX.TRAIN היא מטריצת מלים/מסמכים עבור האימון של מסווג ה-NB. מטריצת המבחן היא MATRIX.TEST.
- כמה מסמכים יש בקבוצת האימון? _____.
- כמה מסמכים יש בקבוצת המבחן? _____.
- מהי המילה העשירית ב - tokenlist? _____.
- מהו הערך של המילה החמישית במסמך העשירי של המטריצה trainMatrix (מטריצת המלים)? _____.
- מהו הערך של המילה ה- 101 במסמך ה- 502?
- באמצעות הפקודה imagesc ציירו את מטריצת האימון של מאה המלים הראשונות במאה המסמכים הראשונים.
- המטריצה מכילה את המלים הנמצאות בשורה הראשונה של כל מסמך (התגיות), כלומר את המלים spam ו- news. מצאו באיזה עמודה נמצאות המלים והוציאו אותן מהמטריצה.

- הפכו את המטריצה trainMatrix למטריצה בינארית trainMatrixbin.
- השתמשו שוב בפקודה imagesc וציירו את מטריצת האימון הבינארית.
- איזה מילה היא הנפוצה ביותר (מופיעה בהכי הרבה מסמכים)? בתשובתכם רשמו מהו האינדקס של המילה _____, בכמה מסמכים היא מופיעה?

_____.

- בכמה מסמכים מופיעה המילה ה- 700?
- מהי המילה המופיעה במסמך אחד מספר רב ביותר של פעמים? _____.
באיזה מסמך? _____ מה מספר הפעמים? _____.
- מהי המילה? _____

- חישובי הסתברויות (אימון)

- מהו מספר מסמכי הספאם? _____

מאחר והמטריצה trainCategory היא מטריצה דלילה, השתמשו בפקודה full (ראו ב- help) כדי להפוך את המטריצה למלאה.

- חשבו את p1, ההסתברות האפריורית לספאם p1=_____

- חשבו את p0 ההסתברות האפריורית למייל "אמיתי" p0=_____.

- חשבו את ההסתברות המותנה (ההסתברויות המותנות) של המילה j בהינתן שהמסמך הוא ספאם. מהי הסתברות המותנית של המילה השניה ? P(j=2|i=1)= _____.

- חשבו את ההסתברות המותנה (ההסתברויות המותנות) של המילה j בהינתן שהמסמך הוא לא ספאם. מהי הסתברות המותנית של המילה השניה ? P(j=2|i=0)= _____.

- מה ההסתברות המותנה שהמילה j לא מופיעה במסמך בהינתן שהמסמך הוא ספאם? _____.

- חישובי הסתברויות (חיזוי)

- קראו את המטריצה MATRIX.TEST והפכו אותה למטריצה מלאה.
- הפכו את המטריצה המלאה למטריצה בינארית.
- חשבו את ההסתברות $p(x \mid y=1)=\prod_{j=1}^{N} p\left(x_j \mid y=1\right)$ עבור המסמך הראשון.
- תארו את הבעייה בה אתם נתקלים (בעייה זו נקראת underflow).
- כיצד אפשר לפתור את הבעייה? (רמז – שימוש בלוגריתם).
- חשבו את ההסתברות שהמסמך הוא ספאם, כלומר חשבו את:

$$p(y=1 \mid x)=\frac{p(x \mid y=1)\cdot p(y=1)}{p(x)}=\frac{\prod_{i=1}^{n} p(x_i \mid y=1)\cdot p(y=1)}{\prod_{i=1}^{n} p(x_i \mid y=1)\cdot p(y=1)+\prod_{i=1}^{n} p(x_i \mid y=0)\cdot p(y=0)}$$

- אפשר להשתמש בטרנספורמציה הבאה:

$$p = \log(p), \quad \tilde{q} = \log(p)$$
$$\tilde{r} = \log(p+q)$$
$$\tilde{r} = \tilde{p} + \log\left(1 + e^{\tilde{q}-p}\right)$$

- חזרו על החישובים עבור כל המסמכים וחשבו את אחוז השגיאה.
- באופן אינטואיטיבי, קיימות מלים אינדיקטיביות יותר מאחרות על האם מכתב שייך למחלקה כלשהי (מייל "אמיתי" או ספאם). אפשר למצוא מלים כאלה (העשויות להצביע על ספאם) באופן לא פורמלי על-ידי חישוב:

$$\log \frac{p(x_j = i \mid y = 1)}{p(x_j = i \mid y = 0)}$$

ככל שערך זה חיובי וגבוה יותר, המילה כפי הנראה נמצאת בספאם.

מצאו את חמש המלים האינדיקטיביות ביותר לכך שהמכתב הוא ספאם.