**קורס עיבוד שפה טבעית 2018**

**פרויקט סיום**

**מטרת הפרויקט**:

ביצוע משימת של סיווג מסמכי טקסט לקטגוריות (text categorization). הנתונים המצורפים כוללים קבוצת אימון הכוללת 20000 מסמכים וקבוצת מבחן הכוללת 5000 מסמכים. כל מסמך שייך לקטגוריה אחת מתוך 22 קטגוריות. המסמכים אינם מתפלגים באופן אחיד על פני הקטגוריות השונות.

עליך להשתמש בשני אלגוריתמי קלסיפיקציה לבצע סיווג של דוגמאות המבחן: nearest centroid + אלגוריתם קלסיפיקציה אחר כראות עיניך.

**הגשה + תאריך הגשה**

העבודה תוגש בזוגות או ביחידים. תאריך אחרון להגשה 11/7/2018. ניתן ואף מומלץ להתייעץ עם המרצה בכל שלב של הכנת העבודה.

**קבצי התרגיל**

dataset\_train.txt נתוני קבוצת האימון

dataset\_test.txt נתוני קבוצת המבחן

categories.txt רשימת שמות הקטגוריות

text\_categorization\_final\_project\_2018.doc קובץ זה

**פורמט הקלט**

כל שורה בדטה מייצגת מסמך, שבנוי מכותרת ומגוף המסמך, אם כי לעיתים הגוף חסר. בקובץ קבוצת האימון המסמכים כוללים גם את שם הקטגוריה אליה שייך המסמך. שלושת השדות, כותרת, גוף, קטגוריה מופרדים זה מזה באמצעות תגיות כגון <subject> או <content>. שדה ה <content> עשוי להיות חסר (הן במסמכי train והן במסמכי test)

כל מסמך של קבוצת האימון ייוצג על ידי שורה כדלקמן

>subject> *subjectwords* </subject> <content> *contentwords* </content><maincat> c*ategory* </maincat>

מסמך של קבוצת המבחן ייוצג על ידי שורה כנ"ל אך ללא ציון הקטגוריה

> subject> *subjectwords* </subject> <content> *contentwords* </content>

בנוסף לתגיות הנ"ל מסמכי הקלט עשויים להכיל סימונים מיוחדים שאינם מילים כמו למשל &amp או &#xa שגיאות כתיב, מספרים, וסימני פיסוק כאמצעי הבעה, למשל :-), נוסחאות מתמטיות כתובות כמו x - 1 = 113 , וכמו כן מסמכים בשפות אחרות (ספרדית....)

**קלסיפיירים**

1. **Nearest centroid:** לכל אחד מהקטגוריות חשב את הוקטור הממוצע של מסמכי הקטגוריה. בשלב ה test חשב את המרחק של כל וקטור (המייצג מסמך מקבוצת המבחן) לכל אחד מהוקטורים הממוצעים. הפלט של הקלסיפייר היא הקטגוריה שעבורה המרחק הוא מינימלי. ראה גם https://en.wikipedia.org/wiki/Nearest\_centroid\_classifier
2. כל קלסיפייר אחר: SVM, KNN, Naïve Bayes וכו'. קיימים מימושים רבים ברשת לכל אחד מהללו.

**תוצרי הגשה**

* שני קובץ פלט שיכילו בדיוק 5000 שורות, בכל שורה שם של קטגוריה של מסמך של קבוצת המבחן. שמות הקבצים output1.txt (עבור קלסיפייר nearest centroid) ו output2.txt עבור הקלסיפייר האחר.
* מסמך ( word למשל) ובו תשובות לשאלות הבאות:

1. עיבוד סטטיסטי של הדטה (מבוסס על קובץ האימון):
   1. מהי השכיחות של כל קטגוריה? הצג את התשובה כטבלה ממוינת של קטגוריות, (ממוינת בסדר יורד של השכיחות) וכן כהיסטוגרמה (העמודות ממוינות באותו סדר משמאל לימין – משמאל הגבוהה ביותר ומימין הנמוכה ביותר)
   2. מהו האורך הממוצע (מס המילים) בכל קטגוריה. הצג את התשובה כטבלה.
   3. מהו אחוז המסמכים בכל קטגוריה שמכילות רק כותרת? הצג את התשובה כטבלה
   4. כמה מילים שונות יש בקבוצת האימון (גודל המילון), וכמה מילים סה"כ.
2. איזה עיבוד מקדים ( preprocessing ) בוצע על הטקסט (וראה למטה, תחת כותרת preprocessing)
3. באיזה קלסיפייר נוסף השתמשתם? באיזה מימוש? האם נעשה כיוונון פרמטרים?
4. מהי השגיאה ה**צפויה** על קבוצת המבחן (אחוז המסמכים שסווגו לקטגוריות הלא מתוך סה"כ המסמכים)
   1. של ה centroid classifier
   2. של הקלסיפייר האחר

את שלושת הקבצים יש לארוז כקובץ zip ששמו יכלול את תעודות הזהות של המגישים ולשלוח למייל המרצה.

**פורמט הפלט**:

קובצי הפלט בשם output1.txt ו output2.txt

כל קובץ יכיל 5000 שורות, שורה אחת לכל מסמך בקבוצת המבחן. בכל שורה יהיה שם הקטגוריה המנובאת. כלומר השורה הi תכיל את שם הקטגוריה של המסמך ה i בקבוצת המבחן **בצורה מדוייקת** (על פי השמות בקובץ הקטגוריות).

**מדד ביצועים**

השגיאה היחסית שתחושב על קבוצת המבחן (5000 מסמכים). השגיאה היחסית מוגדרת שכאחוז המסמכים שלא שויכו לקטגוריה הנכונה.

**הרכב ציון העבודה**

20% - ביצועים – שייקבע על ידי דירוג יחסי של הביצועים.

20% - איכות העבודה הכתובה (הצגת הפתרון, ניתוח, אנליזות, גרפים/היסטוגרמות ברורים וכו')

40% - איכות הפרויקט. היקף העבודה, עבודה שיטתית ומקיפה של preprocessing , כיוונון פרמטרים, מהו החלק היחסי של עבודה מקורית לעומת שימוש במודולים מוכנים.

20% - ציון מפגש. בקיאות בחומר העבודה.

**Preprocessing**

ניתן לבצע עיבוד מקדים מתוך הרשימה לעיל (או אחרים כראות עיניכם) על מנת לשפר את ביצועי הקלסיפייר. בכל מקרה עליכם לדווח מה מימשתם בהתאם לדרישה בתוצרי ההגשה.

* ניקוי (של תגי html, סימנים מיוחדים כמו &#xa וכו)
* ביצוע של lowercase
* ניקוי של סימני פיסוק
* טוקניזציה (חלוקה של הטקסט לtokens רצפים של תוים האמורים לייצג מילים)
* ניקוי של stopwords
* Stemming (אפשר להשתמש בכל אחד מהמימושים הרבים ברשת למשל בדף הבא <http://tartarus.org/martin/PorterStemmer/>)
* טיפול במחרוזות הכוללות מספרים (שלמים, עשרוניים, תאריכים וכו...)
* ניקוי של מילים נדירות (מילים המופיעות פחות מ k=1,2,3… פעמים)
* הוספת מאפייני מסמך שאינם קשורים לשפה עצמה: כמו למשל יש או אין גוף מסמך, מה מספר התווים הכולל, האם יש נוסחאות, האם המסמך בשפה אחרת? האם יש שימוש ניכר בסימני פיסוק (!!!!! או ....) וכו.
* תיקון שגיאות כתיב
* טיפול מיוחד בביטויים מתמטיים / תאריכים