

תרגיל 3

סיוג

מבוא

משימות סיוג נפוצות מאוד בתחום עיבוד השפות, בעיקר כי בעיות רבות ניתן לתרגם לעיבוט סיוג. משימות סיוג רבות שנראות קשות לעין האנושית יכולות להתבצע באופן מעולה ע"י אמצעי מיידת מכונה פשוטים. בתרגיל זה ננסה בסיווג טקסט מתוך פרוטוקולי הכנסת על-פי הדובר. כאמור, בניית תכנית שתאמן מסובגים שונים לסיווג ייחודי טקסט לשני דוברים. לצורך תרגיל זה השתמש בקורסוס ה-כנסת שבנווים בתרגילים הקודמים. כמו כן, [יעזר באובייקטיבים מספריית scikit-learn](#).

שלב 1 : הגדרת המחלקות

בתרגיל זה השתמש בקורסוס הכנסת שיצרתם בתרגילים הקודמים. להזיכרכם, בקובץ `labeled_sj` של הקורפוס שלנו יש שדה המציין את שם הדובר. מצאו את שני הדוברים עם המספר הכי גדול של משפטים בקורסוס שלכם.

דוברים אלו יהוו את המחלקות שלכם לסיוג באופן הבא :

1. **משימת סיוג בינארית binary classification** : עליכם לחלק את הדטה לשתי מחלקות - כל אחד משני הדוברים מהויה מחלוקת בפני עצמו. עברו כל מחלוקת השתמשו בכל הרשומות של הדובר המתאים ורק בהן. יחידת הסיוג במשימה זו תהיה משפט אחד. כתבו בדו"ח מי הם שני הדוברים שיהוו את המחלוקות שלכם.

2. **משימת סיוג מרובה מחלוקת multi-class classification** : במשימה זו יהיו לכם 3 מחלקות : אחת עברו כל אחד משני הדוברים שמצאתם, כמו במשימה 1, והשלישית תהיה מחלוקת של "אחר" בה יהיו כל הרשומות הנותרות של שאר הדוברים. יחידת הסיוג במשימה זו תהיה משפט אחד.

הערה : אותו דובר בקורסוס שלכם יכול להופיע במספר דרכים שונות. למשל "בניין גן" ו"בני גן", או הבדלים שנובעים מטעויות בחילוץ וקיון השמות. עברו בחרית המחלוקות, אינכם נדרשים לאחד את השמות של אותו דובר. בחרית המחלוקות תתבצע רק על פי מחרוזת השם, וכל מחרוזת שונה תחשב בשם דובר אחר. יחד עם זאת, על מנת למקסם את כמות המשפטים לאימון ולהמנוע *overfitting*, יש לנסות לאחר בחרית שתי המחלוקות הראשיות (שני הדוברים בעלי מספר המשפטים הרב ביותר), לכלול כמה שיותר משפטיים של שני דוברים אלו תחת המחלוקת שלהם, גם אם מחרוזת השם שלהם כתובה בדרךים שונות בקורסוס. הסבירו בדו"ח איך התמודדותם עם דרישת זו.

שלב 2 : איזון המחלקות

על-מנת לסווג באופן מיטבי, נרצה שהמחלקות תהינה מאוזנות. לשם כך, עברו כל אחת מהמשימות עשו -*mosop sampling* (רנדומלי) למחלקות הגדולות. כאמור, בחרו באופן רנדומלי פריטים מהמחלקה/ות הגדולהות

כמספר הפריטים במחלקה הקטנה וזרקו את יתר הפריטים במחלקה, כך שיתקבלו שתי מחלקות באותו הגודל עבור המשימה הבינארית ושלוש מחלקות באותו הגודל עבור המשימה רבת המחלקות.
כתבו בדוחך מה היה מספר הפריטים בכל מחלקה לפני ואחרי sampling-down שבייצעתם.

שלב 3: ייצרת וקטור מאפיינים (feature vector)

עליכם ליצור שני וקטורים שונים באופן הבא :

1. עבור כל משפט יצרו וקטור W_{Bo} כוקטור מאפיינים. ניתן להשתמש ב- [Tfidf](#). ניתן גם לבחור להשתמש ב- [CountVectorizer](#).
2. צרו וקטור משלכם, עם מאפייני סגנון ותוכן. לשם כך, אתם יכולים להסתמך על הדאטה שיש לכם ולהשוו מה יכול לעזור בסיווג. פיצרים יכולים להיות למשל אורך המשפט, סימני פיסוק, צירופי מילים מסוימים וכיו"ב. הנכם מוזמנים להשתמש כתוכנות גם בעמודות אחרות בדאטה, מלבד עמודת הטקסט. **בוקטור זה אסור לכם להשתמש בוקטור W_{Bo} .**

שלב 4: אימון

1. על מנת לסווג את שני סוגי וקטורי המאפיינים שלכם, אמןו שני סוגי מסווגים :
 - i. [KNearestNeighbors](#)
 - ii. [LogisticRegression](#)
 2. הערכו את דיקום המסווגים ע"י [5-fold Cross Validation](#)
 3. הוסיפו לדוחך המפרט את תוצאות ההערכה עבור כל משימה, עבור כל מסווג ועבור כל וקטור מאפיינים.
- הערה :** למסווגים השונים יש פרמטרים שונים שאתם יכולים לנפג או להשאיר את ברירות המחדל, לפי בחירתכם. [פרטו והסבירו את החלטותיכם בדוחך.](#)

שלב 5: סיווג

לתרגיל מצורף קובץ בשם `knnesset_sentences.txt`, המכיל בכל שורה משפט מתוך טקסטים של הכנסת. עליכם לسوוג כל משפט לאחת המחלקות : הדובר הראשון שלכם (זה יהיה בעל מספר המשפטים הכי גדול בקורס) כפי שמצאתם בשלב 1), הדובר השני (בעל מספר המשפטים השני בגודלו בקורס), או "אחר", בעזרת אחד מהמודלים שאמנתם, לבחירתכם. עליכם לכתוב את הסיווגים לקובץ בשם `classification_results.txt` כל שורה בקובץ תתייחס לשפט שבאותה שורה בקובץ המקורי, ותכליל רק את תוצאה הסיווג "first", "second", "other". **לא שורות רזות.**

למשל :

first

first
other
second
...
...

הערות:

1. שימושו לב, שבקובץ הקלט בשלב 5 מופיעים רק הטקסטים עצם ולא ערכיהם התואמים לעמידות אחרות, לכן, אם השתמשתם באלו בוקטור המאפיינים שיצרתם, לא תוכלו לסוג את הדוגמאות האלה בעזרתו. בחרו מודל שכן מתאים למשימה.
2. לאורך הקוד יש מספר מקומות בהם יש מידת אקראיות. עליהם להשתמש ב- `(random.seed() ו-`
`numpy.random.seed()`. על מנת לקבע את התוצאות שלכם, אחרת הן ישנה בכל ריצה. לשם כך, הוסיפו בתחילת הקוד :

```
import random
import numpy as np
random.seed(42)
np.random.seed(42)
```

שאלות

ענו בדוחיך על השאלות הבאות :

1. מה הם האתגרים שיכולים להיווצר בשימוש בחלוקת "אחר" במשימת הסיווג?
2. נניח שאתם משתמשים בתחרות מודלים לחיזוי בינהiri שבה אם המודל שלכם יזכה נכון את כל הדוגמאות של הדבר הראשון, תקבלו פרס כספי גדול, ואם המודל שלכם יטעה על לפחות דוגמה אחת של הדבר הראשון תקבלו קנס כספי גבוה.
מבחן המודדים המופיעים בclassification report, איזה ממד תרצו למקסם? איזה מהמודלים שאימנטם תבחרו למטרה זו? הסבירו.
3. ענו שוב על 1 כאשר שינו את החוקים בתחרות וicut אם המודל שלכם יסוווג נכון את **כל** הדוגמאות של שני הדברים תקבלו פרס כספי גבוה, אבל אם המודל שלכם יסוווג לפחות דוגמה אחת בצורה לא נכוונה, תקבלו קנס כספי גבוה.
4. הסבירו מה היתרונות והחסרונות של שיטת המסתורינון cross validation על פני חלוקה פשוטה לחלוקת אימון ובדיקה. איזו משיטה ההערכה אמינה יותר לדעתכם?
5. יחידת הסיווג בתרגיל היא משפט אחד. אם במקומות זאת, היינו מחייבים על יחידת סיווג שמאחדת יחד מספר משפטים מאותה מחלוקת, כך שכל דוגמה לסיווג הייתה מקבץ של משפטים. מה היו היתרונות והחסרונות בכך? התיחסו בתשובהיכם ליחידות סיווג של 2, 5, 10, 100 משפטים.
6. איזה גודל של יחידת סיווג עדיף לדעתכם (1, 2, 5, 10, 100) בנסיבות שלנו? הסבירו.

הערות כלליות

1. על הקוד שלכם להיות מסוגל להתמודד עם שגיאות בכל שלב בתהיליך ולא לקרו. השתמשו ב-`try` Except blocks
2. שימו לב, בבדיקה תרגילי הבית בקורס ניתן משקל גדול מהኒקוד הן על **הז"ח**, ההסברים והידע שהפוגנים בחומר הנלמד והן על **הקוד**, אופן המימוש, יעילותנו, קרייאותו ועמידותנו. בפרט, הרבה מהבדיקות הן אוטומטיות ולכן עליכם להקפיד על קוד תקין שרצ' לא שגיאות ועל עמידה **מדויקת** בפלט הנדרש וביתר הנקודות.
3. ניתן לשאול שאלות על התרגיל בפורים המועד במודול. למעט מקרים איסיים מיוחדים, אין לשלו שאלות הקשורות לתרגיל הבית במיל.
4. על אחריותכם לעקוב אחר הודעות הקורס במודול (בלוח הودעות ובפורום) ולהיות מעודכנים במידה ויהיו שינויים בהנחיות.

ספריות מותרכות לשימוש

אתם יכולים להשתמש ב Pandas, Numpy, scikit-learn ובכל ספרייה סטנדרטית של python.python. אתם יכולים לחפש שם של ספרייה ב <https://docs.python.org/3/library/index.html> על מנת לבדוק אם זו ספרייה סטנדרטית. לא יהיה מענה על שאלות לגבי שימוש בספריות ספציפיות.

- למען הסר ספק, `mosz` היא ספרייה סטנדרטית של `python`.
- מומלץ להשתמש כל פרוייקט בסביבה וירטואלית virtual environment חדשה משלו על מנת להיות בטוחים שאתם משתמשים רק בספריות מותרכות ולמנוע קונפליקטים עם ספריות קודמות שהתקנתם בעבר. [ראו ממצת על כך במודול](#)

אופן ההגשה

1. ההגשה היא בזוגות בלבד.
 2. עליכם להגשים קובץ `zip` בשם `hw3_id1_id2.zip` (כאשר `hw3` הם מספרי תעודות הזוחות של הסטודנט הראשון והשני בהתאם), המכיל את הקבצים הבאים :
- a. קובץ `python` בשם `knesset_speaker_classification.py` המכיל את כל הקוד הנדרש כדי למש את שלבים 5-1.
 - i. - הקלט לקובץ יהיה נתיב לקובץ הקורפוס, נתיב לקובץ משפטים לסייע, נתיב לתיקייה פלט - הפלט יהיה קובץ הסיוגים כפי שתואר בשלב 5 שמור בתיקייה שהתקבלה בקלט. [בשלב הגשת התרגיל על הקובץ לא להדפיס שום דברelman. נטרלו את הדפסת classification reports](#).

ii. על הקובץ לזרץ תחת הפקודה (לא הסימוניים <>) :

python knesset_speaker_classification.py <path/to/corpus_file.jsonl> <path/to/sentences_texts_file.txt> <path/to/output_dir>

.b. קובץ text בשם **classification_results.txt** כפי שתואר בשלב 5.

c. קובץ PDF בשם **hw3_report.pdf** ובו דוח המפרט על הקוד, על ההצלחות שקיבלתם במהלך העבודה על התרגילים, מהן על השאלות וכל מה שתתבקשتم לפרט לאורך התרגילים. אל תשחחו לצין בתחילת הדוח את שמותיכם ותעודות זהות שלכם בעברית.

יש להקפיד על עבודה עצמית, צוות הקורס יתייחס בחומרה להעתיקות או שיתופי קוד, כמו גם שימוש בכללי AI chatGPT.

ניתן לשאול שאלות על התרגילים בפורום הייעודי לכך במודול.

יש להגיש את התרגילים עד לתאריך 25.12.24 בשעה 59:23.

בצלחה!