בס"ד

ספר משתמש

CLASSIFIER RAMBAM

מטרה

מטרתו של המסמך הינו להסביר את דרך התוכנית ואת הפעלתה על כל האפשרויות הניתנות על ידה.

המסמך יציג את מרכבי התוכנית וכן את מאגרי המידע הנלווים אליה ואשר מהווים חלק אינטגרלי מהותי לצורך הפעלת התוכנית.

התקנה

התוכנה מופעלת על ידי הcommand line ולפיכך אינה דורשת שום דבר נוסף.

התוכנה נוסתה על מחשב ווינדוס.

ספריות נדרשות

- python v.3.7
- argparse v.1.1
- sklearn v.0.20.1
- numpy v.1.15.4

הקדמה

מטרת הפרויקט הייתה ליצור מסווג טקסט אשר ייקח שאלה מספר שו"ת וידע לסווג אותה למצווה עליה היא מדברת. לאחר מכן ניתן יהיה ליצור מאגר של שאלות אשר משותפות לכל מצווה ובכך ליצור אפשרות חיפוש וקישור מהירות בין שאלות. בצורה זו יוכל הלומד לראות ספרים ושאלות נוספות אשר קשורות לחומר אותו הוא לומד כעת בלי צורך לחפש שאלות דומות, מה גם שלעיתים בתוצאות החיפוש נמצא שאלות אשר מזכירות את הנושא רק בתור תזכורת או דוגמא אך אינן קשורות לנושא עצמו.

ליצירת המסווג נבחר הרמב"ם אשר משנתו מקיפה את התורה ובפרט עוסקת בכל אחד מהנושאים היכולים לצוץ בספרי השו"ת. ולכן בחירה בו מהווה את האפשרות הטובה ביותר עבור המסווג.

הרחבה על הרמב"ם והבחירה בו תמצא במסמך המסביר על הפרויקט.

מאגרי המידע

לצורך יצירת מסווג אנו נדרשים תחילה לאתר ולהשתמש במאגרי מידע אשר ישמשו עבור האימון והניסוי.

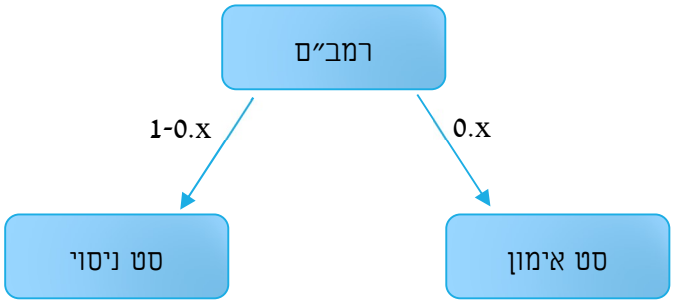
כל מאגר מחולק מלכתחילה לפי מספר מילים תוך התעלמות מסעיפים ופרקים דהיינו התוכנית תיקח את כל המקומות בהם דובר על נושא מסויים ותחלק אותו לפי סך מילים ולא לפי סעיפים.

עם זאת במאגרי המידע שאינם הרמב"ם ישנה אפשרות לחלוקה לפי סעיפים.

אימון

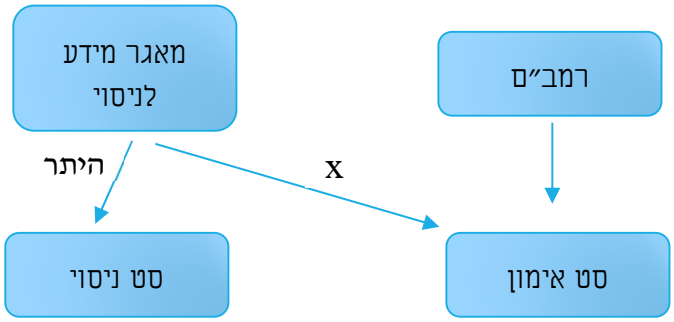
עבור האימון התוכנית משתמשת בספר משנה תורה של הרמב"ם אשר מהווה את הבסיס עבור כל אחת מצורות האימון האפשריות של המסווג.

צורת האימון הראשונה היא חלוקת הרמב"ם לשתי חלקים חלק לאימון וחלק לניסוי לפי יחס מסוים הנע בין 0.50 ל0.99.



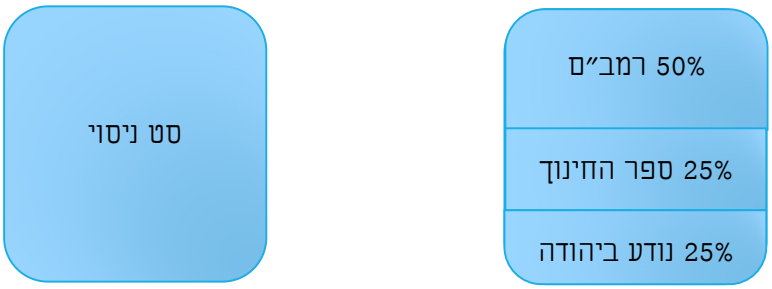
צורת האימון השנייה היא השמת הרמב"ם כולו לאימון ושימוש באחד ממאגרי המידע של הניסוי לניסוי.

צורת אימון שלישית היא השמת הרמב"ם כולו לאימון ובנוסף חלוקת סט הניסוי לשני חלקים כאשר חלק ממנו יצורף לסט האימון. בניגוד לצורה הראשונה בצורה זו נשתמש במספר דוגמאות ולא ביחסים כדי שהמסווג לא ילמד את סט הניסוי עצמו.



צורת האימון הרביעי דומה לשלישית אלא שכעת הדוגמאות שילקחו לאימון יהיו דוגמאות אשר מוזכר בהן הרמב"ם.

צורת האימון החמישית היא שימוש בשני מאגרי מידע נוספים המשמשים לאימון. האחד הינו ספר החינוך והשני שו"ת נודע ביהודה. השימוש בהם יעשה בצורה הבאה: סט האימון יורכב מ50% רמב"ם, 25% ספר החינוך ו25% שו"ת נודע ביהודה.



ניסוי

ישנם שלוש מאגרי מידע עבור הניסוי.

הראשון הינו טור.

השני הינו הבן איש חי.

השלישי הינו קיצור שולחן ערוך.

מיקום

בברירת המחדל ממוקמים מאגרי המידע בתוך תיקייה בשם data כאשר מאגרי המידע של האימון יהיו בתת תיקייה train_data והניסוי בתת תיקייה test_data.

אפשרויות המסווג

בתוכנית נעשה שימוש בשתי אופציות של מסווג

הראשונה והיא גם ברירת המחדל היא LinearSVC **והשנייה** SVC.

בעוד ששתי האפשרויות טובות האפשרות הראשונה הראתה תוצאות טובות יותר עבור האימונים השונים.

ניתן לקבל את המסווג בשלוש צורות שונות.

הראשונה שימוש בדגלון fit ייתן את המסווג עליו נוסו אפשרויות האימון השונות והוא נותן את התוצאות הטובות ביותר.

השנייה שימוש בדגלון load אם קיים מסווג הנשמר על ידי pickle ניתן לשנות את שמו לpipe.pickle ולשים אותו בתיקייה הראשית והתוכנית תיקח אותו כמסווג ולא תאמן אחד משלה.

השלישית שימוש בדגלון search יריץ חיפוש וימצא את התוצאות הטובות ביותר עבור סט האימון.
הערה יש לקחת בחשבון כי אופציה זו לוקחת זמן אשר ישתנה בהתאם למערכת בו רצה התוכנית.

פעולת התוכנית

התוכנית פועלת בצורה הבאה:

1) התוכנית מקבלת את מאגרי המידע הלא ערוכים.
2) התוכנית עורכת אותם לסטי אימון וניסוי ושומרת אותם בתיקייה.
3) התוכנית מאמנת על סט האימון מסווג.
4) התוכנית בודקת את המסווג על סט האימון ומדפיסה את אחוזי ההצלחה.

הפעלת התוכנה

הפעלת התוכנה נעשית על ידי הרצת הפקודה

```
Python rambam_classifier.py
```

דגלונים

התוכנה מאפשרת הוספת דגלונים המאפשרים לקבל את כל האפשרויות השונות לאימון המסווג.

דגלון	ברירת מחדל	אפשרויות	הסבר
-h	-	-	מראה הסבר על הדגלים.
—train_path	./data/train_data/	-	המיקום של מאגר האימון הלא ערוך.
--test_path	./data/test_data/	-	המיקום של מאגר הניסוי הלא ערוך.
--set_path	./data_set/	-	המיקום של סטי האימון והניסוי.
--no_shuffle	False	-	אם דגלון זה מופיע התוכנית לא תערבב את הסט לפני שתחלק אותו לאימון וניסוי במידה ואכן נבחר לחלק את המאגר לאימון וניסוי.
--train_ratio	1.	1. – 0.5	אחוז הסט שיעבור לאימון כאשר היתר לניסוי. מיועד עבור צורת האימון הראשונה. אם 1. כל המאגר ילך לאימון.
--test_amount	20	-	כמות הדוגמאות שהולכת לאימון. מיועד עבור צורת האימון השלישית והרביעית

--only_rambam	False	-	הוספת הדגלון לוקחת רק את הרמב"ם לסט האימון ללא ספר החינוך והנודע ביהודה. עבור כל צורות האימון חוץ מהחמישית.
--test_sorter	fit	full, amount, rambam	fit – ישים את כל סט הניסוי אצל הניסוי. amount – מפעיל את אופציית האימון השלישית. rambam – מפעיל את אופציית האימון הרביעית.
--test_source	rambam	rambam, ben, kizur, tur	rambam – נותן ניסוי עבור הרמב"ם עצמו. נועד לאופציית האימון הראשונה. יתר האפשרויות נותנים את יתר סטי הניסוי.
--classification	LinearSVC	LinearSVC, SVC	בוחר את האופציה עבור המסווג.
--pipe	fit	fit, load, search	fit – נותן את המסווג המומלץ על ידי התוכנה. load – טוען מסווג. search – מחפש את מאפייני המסווג האופטימלי ביותר עבור סט האימון.
--show_results	False	-	מראה את התוצאות של חיפוש המאפיינים האופטימליים עבור המסווג.
--save_pipe	False	-	שומר את הpipe לשימוש נוסף. נועד עבור --pipe עם האופציה "search"
--print_report	False	-	מדפיס את הטבלה הנסוי המלאה עבור כל תג ותג.