מבוא למערכות לומדות
הקאתון 2020
משימה 2 - של מי הקוד הזה?

שמות: ליאת ויסמן, גיא קורנבליט ומתן אלישר

12 ביוני 2020

הקדמה

מסמך זה מפרט את תהליך העבודה של צוותינו סביב בעיית קלסיפיקציה של טקסט. המשימה שבחרנו הייתה לבנות מסווג אשר מתייג קטעי קוד קצרים (בין שורה לחמש שורות) לפרויקט המקור שמהם נלקחו, מבין שבעה פרויקטים שונים ופעילים שנלקחו מ- GitHub.

שאלות ראשונות שהציפו אותנו בשלב ההכנה המקצועית, עוד בטרם תחילת פתרון הבעיה, וקיווינו לקבל עליהן תשובה לאורך העבודה:

- כיצד מפצלים את הדאטה שברשותינו לקבוצות למידה, וידוי ומבחן בצורה נבונה?
- כיצד מבצעים את תהליך חקירת הדאטה בצורה טובה? איך ממירים את המידע שאנו מפיקים מהויזואליזציה להחלטות פרקטיות בשטח?
- דגימות חסרות - איך לטפל בהן? האם ע"י ניקוי, החלפה בממוצע או חציון, או תהליך אחר בכלל?
- אם הרגולריזציה באופן טבעי "מנקה" פיצ'רים שפחות משפיעים על התיוג האם עליי להתייחס לכך בעבודה על הפיצ'רים?
- ועוד...

תהליך ראשוני: הגדרת הבעיה והבנת סוג המידע

הבעיה שקיבלנו הינה Batched Supervied MultiClassification problem . את הדאטה קיבלנו מחולק לשבעה קבצי טקסט המכילים את כל קבצי הפרויקט שניתנו לנו, משורשרים אחד אחרי השני. כל פרויקט מכיל קבצים ממגוון שפות תכנות (פייתון, ג'אווה, גו ועוד), אך הרוב המוחלט כתוב בפייתון.

בחרנו ב- MissClassification כקריטריון הביצוע שלנו, שכן הוא יעיל ואינטואיטיבי לעבודה בבעיה מסוג זה, ולא מצאנו כי קיים הבדל בחומרת טעות סיווג בין הפרויקטים השונים.

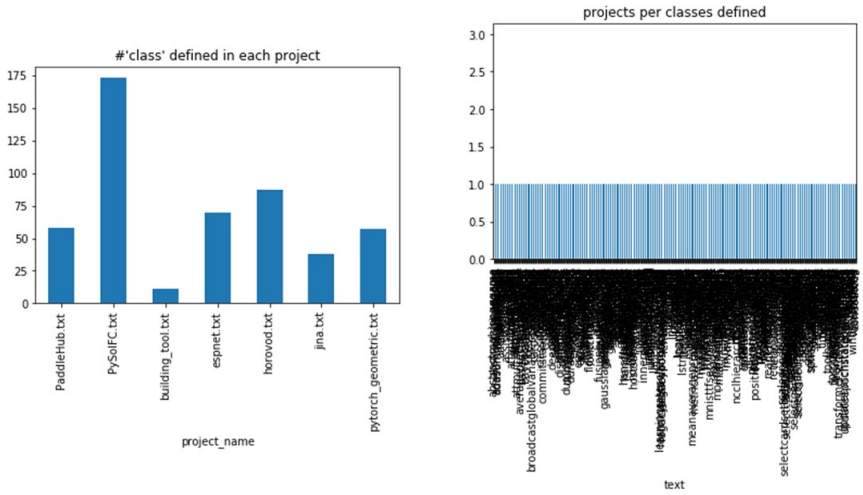
סוג המידע איתו התעסקנו היה המקור הראשוני לבחירתינו במשימה זו על פני השניה. העבודה עם מידע טקסטואלי הינה תחום ענף בעולם הבינה המלאכותית והיא חדשה עבורנו. מהר מאוד הבנו כי עלינו לבצע למידה ראשונית של השיטות הבסיסיות לעיבוד טקסט ויצירת פיצ'רים. חקרנו את טכניקת ה- Bag of words והשימוש ב- TfiDf.

עיבוד מידע: חלוקת הדאטה לקבוצות למידה וידוי ומבחן, חקירה מעמיקה של הדאטה ועיבוד מקדים (Preprocessing).

את המידע חילקנו באופן ידני לקבוצות אימון, וידוי ומבחן, ביחס של $70, 15, 15$ אחוזים מכל פרויקט. מאחר ולא היו לנו פיצ'רים קיימים, שלב החקירה המעמיקה של הדאטה התבטא בחיפוש מאפיינים אינדיקטיבים בדאטה, שעל בסיסם נוכל לבנות פיצ'רים בהמשך. חלק מהרעיונות שעלו היו:

- סוגי הקבצים בכל פרויקט, והיחס ביניהם.
- סגנון כתיבה - שמות ארוכים למשתנים או פונקציות, $CamelCase$ לעומת שימוש בקו תחתון וכו'.
- תיעוד - האם הקטע שאנו בוחנים כעת הינו קטע תיעוד, האם זה מסייע לנו לסווג אותו

רעיון נוסף שהעלנו היה שימוש בשמות המחלקות בקבצים. ביצענו ניתוח של הדאטה בכיוון זה: השתמשנו ברג'קס על מנת להפיק את שמות כל המחלקות בקבוצת האימון על מנת ליצור מאגר שישמש אותנו. חיפשנו להבין את תדירות השימוש והאינדיקטיביות של מידע זה.



איור 1: ניתוח שמות מחלקות בקבוצת האימון

בגרף השמאלי ניתן לראות את התפלגות שמות המחלקות בקבוצת האימון בין הפרויקטים השונים. בגרף הימני ניתן לראות, כפי שציפינו, שמידע זה הינו מאוד אינדיקטיבי. נדון בבחירת הפיצ'רים הסופית בחלק הבא.

בחלק של העיבוד המקדים של נתקלנו בשאלה הבאה: האם עלינו לבצע את הלמידה על קבוצה גדולה של דגימות, המורכבות כל אחת מקטע קוד של שורה עד חמש שורות,

או האם ביכולתינו ללמוד מאפיינים כללים מכלל הטקסט שאותם נוכל ליישם על קבוצות המבחן. במבט ראשוני לא ראינו יתרון בחלוקת קבוצת האימון לדגימות, שכן ראינו לנגד עינינו מסווג שמגיע עם ידע מקדים על הטקסט, למשל ע"י מילון שכיחויות לפי קובץ, מילון שמות מחלקות וכו', ומבצע התאמה של קטע קצר על בסיס ידע זה. כמו כן ניסינו להבין כמה השפעה תהיה לאופן החלוקה שלנו לדגימות קטנות על תוצאת הסיווג. אחד הכיוונים שהיינו בודקים לשיפור תוצאות, אם העבודה הייתה ממשיכה, הוא נושא זה חיתוך הדגימות. היינו מנסים למשל למצע את המסווג על פני חיתוכים שונים רנדומליים.

כך או כך הבנו שעלינו ליצור מנגנון חיתוך כזה על מנת להביא את החלקים שהוקצבו לקבוצת הוידוי והמבחן לצורה הדומה לאופן בו ניבחן על הצלחת המסווג. לכן יצרנו פונקציה גנרית והחלטנו לבדוק את ההבדל בביצועים בהמשך. בדיעבד הסתבר שהמודל שבחרנו עבד בצורה משמעותית טוב יותר על קבוצת אימון שמחולקת להרבה דגימות מייצגות, מאשר שבעה וקטורים גדולים של כל פרויקט.

פיצ'רים: יצירה / בחירה

אז הבנו שאנחנו חייבים פיצ'רים. נושא בחירת הפיצ'רים הינו הנושא המרכזי ביותר איתו התמודדנו בתהליך. בניגוד לבעיות קודמות בהן נתקלנו בקורס, לראשונה נפגשנו במידע שאין עבורו טבלת פיצ'רים ברורה שצריך לעבוד מתוכה. החלטנו לעבוד בשני מישורים: יצירת פיצ'רים הקשורים בהיבטים שונים בטקסט (על בסיס חקירת הדאטה) ושימוש ב־Tfidf כדי להמיר את המילים בטקסט לפיצ'רים. המחשבה הייתה שנוכל לשלב את שתי הגישות, או לבחור את הטובה מביניהן. בפועל גילינו שיצירת פיצ'רים "עצמאיים" כרוכה בעבודת מחקר מעמיקה בהרבה, אשר דורשת זמן עבודה רב שלא היה לנו, שעיקרה שימוש ברג'קס וכלים נוספים לעיבוד טקסט. לכן בחרנו להתמקד בשימוש בכלי של Tfidf. לאחר שביצענו וקטוריזציה ראשונית של קבוצת הלמידה, הבחנו שעלינו לשקול לבצע ניקוי נוסף של הדאטה. כיוון נוסף להמשך שיפור ביצועים הוא בחינת מדוקדקת של המילון הנוצר מתהליך זה, ולנסות להבין אם ביכולתינו להוריד את מימד וקטור המילים, על מנת להקטין את d.

בניית המודל: יצירת אלוגירתם בייסליין, העלאת מועמדים למודל מתקדם ובחירת הטוב ביותר

בתהליך בחירת המודל השתמשנו באובייקט Pipeline מתוך החבילה sklearn. אובייקט זה מכיל שלושה אובייקטים :

1. אובייקט CountVectorizer אשר אחראי על ספירת המילים בארכיון ויצירת Document-Term-Metrix (DTM).
2. אובייקט TfidfTransformer אשר אחראי על המרה ויצירת של DTM על פי Tfidf.
3. אובייקט מסווג של sklearn.

ליצירת אלגוריתם בייסליין בחרנו להשתמש במסווג בייס נאיבי. האלגוריתם השיג דיוק של 0.82 , שהיווה סף גבוה כאלגוריתם התחלתי. לאחר מכן הרצנו סדרה של בדיקות על המועמדים הבאים:

- מסווג SVM עם פרמטר רגולציה של $0.01, 0.001$ (ב־ sklearn הפרמטר נלקח ע"י ההופכי שלו).
- מסווג RandomForst עם עומק מקסימלי של בין 500 ל־ 5000.

כמו כן שילבנו את הפרמטרים הנ"ל עם שימוש/אי־שימוש במדד ה־ idf בייצירת ה־ DTM , וכן פרמטר ngram של מילה אחת, או קבוצות של מילים בודדות וצמדי מילים, ביצירת וקטור השכיחויות של המילים בארכיון, באובייקט CountVectorizer.

להפתעתנו מסווג ה־ SVM לא הצליח להתעלות על הבייסליין שלנו, לעומת זאת כאשר פעלנו עם מסווג RandomForst קיבלנו השתפרות משמעותית ככל שהגדלנו את העומק המקסימלי של העץ. עם עומק מקסימלי של 3000 וללא שימוש ב־ idf הגענו לדיוק של 0.895 .

כל האימונים בוצעו על קבוצת האימון והבדיקות של קבוצת הוידוי.

יישום המודל הנבחר: כתיבת הקוד, הערכת ביצועים ואופן השימוש.

לבסוף בחרנו להשתמש במסווג מסוג RandomForst ולאמנו על כלל הארכיון. את הקוד מימשנו בקובץ פייתון עם מחלקות שונות האחראיות על התהליכים השונים. המסווג מגיע מאומן ומשתמשים בו ע"י יצירת אובייקט של המסווג עם הקבוצה אותה רוצים לתייג.

אנו מעריכים כי המסווג עומד על דיוק של 0.9.

דיווח תוצאות:

בבדיקה הסופית על קבוצת המבחן ששמרנו הגענו לדיוק 0.85.

סיכום:

נסכם את העבודה בנקודות הבאות:

- בחוויה הכללית שלנו תהליך הלמידה היה לא מספיק ממצה. מלאכת יצירת הפיצ'רים הייתה מסובכת בהרבה משציפינו, ולכן החלק המהותי ביותר של הלמידה, בתחושתנו, שהינו האקספלורציה, והבנת קורלציה בין פיצ'רים לא קיבל את הנפח שלו. ברור לנו שבעיות אמיתיות כוללות בתוכן הרבה הליכה בערפל, אבל אנו מרגישים שבתחום של עיבוד שפה לנו היה קשה להבין את הדרך ליישם את הכלים שלמדנו בכיתה.

- עם זאת בנימה חיובית ואופטימית, החוויה הזאת נתנה תחושה מאוד חזקה של עבודת לימוד עצמי, עבודה בצוות, ובעיקר העלאת שאלות חקירה להמשך.