מבוא למערכות לומדות
האקתון

רן שחם - 203781000 - ransha ; ברק הלה - 305007361 - barak.halle

23 ביוני 2017

1 של מי הכותרת הזאת בכלל?

1.1 רקע ועיבוד מקדים

במשימות סיווג של מידע מילולי חסר מבנה (שהוא סוג המידע שקיבלנו) ישנה חשיבות רבה לאופן בחירת הפיצ'רים המתארים באופן המוצלח ביותר את הנתונים, לאור משימת הסיווג. על אף הדמיון לכאורה שבמשימה שלנו למשימות סיווג מסמכים לנושאים, המטרה שהוצבה בפנינו היא למעשה שונה בתכלית - שכן עלינו לדעת להבדיל בין סגנונות שנים על אף הדמיון שבתוכן.

כפתרון בסיסי לבעיה, פנינו לבחינת אוצר המילים של כל אחד מהעיתונים. בבחינה ראשונית זיהינו כי בין קבוצת המילים שבהן נעשה שימוש בכותרות הארץ לבין זו של ישראל היום הבדלים רחבים - אם נסמן ב־H את קבוצת המילים שבכותרות "הארץ" וב־I את זו של "ישראל היום", מצאנו כי:

$$|H \backslash I| \approx |H \cap I| \approx |I \backslash H|$$

בעקבות תובנה זו, הסקנו שאפילו מדד פשוט כמו ההרכב הלקסיקלי של כותרת יכול להעיד על מקורה. לכן, הייצוג הראשון אותו בחרנו לממש הוא ייצוג Bag of Words הממפה כל כותרת לוקטור ב־$\{0,1\}^d$ כאשר d הוא מספר המילים בלקסיקון (הלקסיקון הוא סט המילים בכל נתוני האימון) ובקורדינטה ה־i מופיע 1 אם"ם המילה ה־i מופיעה בכותרת.

בשלב הבא, ניסינו לשפר את המודל ע"י נורמליזציה של הוקטורים באמצעות תהליך ה־TF_IDF אותו למדנו בכיתה. בנוסף לכך, הרחבנו את מרחב הפיצ'רים בהוספת צמדי מילים (bi-grams) כפיצ'רים, וזאת על מנת לשמר אלמנט של סדר המילים המופיעות בטקסט, ולא רק את עצם הופעתן.

בהמשך ישיר לשלב האחרון, רצינו להמשיך ולבטא בצורה נאמנה יותר את המאפיינים הסגנוניים בכתיבת הכותרות. קבענו את תפקידה התחבירי של כל מילה בכותרת וניסינו להשתמש במידע זה בשתי דרכים - הראשונה היא הצמדה של מילה ותפקידה התחבירי (ע"מ לחדד את הפיצ'רים הנוכחיים ובכך לשפר ביצועים), השנייה היא הוספת תפקידיהן התחבירי של המילים כפיצ'ר בפני עצמו (ובכך לאפשר זיהוי תבניות תחביריות המאפיינות את העיתונים השונים).

1.2 פרטי המימוש הנבחר

להלן תיאור גס של התהליך אותו מימשנו:

1. קריאת הנתונים וערבוב סדר השורות
2. חלוקת המדגם לשני חלקים: מדגם אימון ומדגם מבחן
3. המרת המידע הגולמי למידע עם מבנה:
   (א) יצירת מרחב של מילים וצמדי־מילים
   (ב) תיוג חלקי הדיבר במשפט וצימודם למילים (למשל, "...Netanyahu said..." → "...netanyahu-noun, said-verb...").

(ג) מתן ערך מספרי לכל מילה/צמד בכל כותרת (על ידי TF-IDF).

4. אימון המודל: ביצענו Grid-search[1] עם cross-validation כדי להתאים מודל מסוג Linear SVM לנתוני האימון

5. בחינת המודל: בדקנו את ביצועי המודל המאומן על מדגם המבחן (כאשר הפלט בשלב זה הוא $L_V(h) = \frac{1}{m}\sum_{i=1}^{m} \ell^{0-1}(h,(\mathbf{x}_i,y_i))$ ו־$V = \{(\mathbf{x}_i,y_i)\}_{i=1}^{m}$ הוא מדגם המבחן).

1.2.1 תוצאות הערכת המודל

בבדיקת המודל המאומן על מדגם המבחן קיבלנו דיוק של $\approx 83\%$. כמוכן, לשם שעשוע, השתמשנו במודל המאומן לסווג כותרות עדכניות מהעיתונים הרלוונטים ומצאנו שאחוזי הדיוק המתוארים לעיל משקפים נאמנה את הצלחת המודל (נכון להיום בבוקר) – ☺.

1.3 כיוונים אחרים שניסינו

1.3.1 Bag of Words

בשימוש במודל הפשוט של bag of words הגענו לתוצאות די טובות, אם כי מעט פחות מהמודל הסופי אותו מימשנו. עם זאת, ניכר שעיקר ההצלחה של האחרון נובע מהשימוש במאפיינים הללו (הוספת הפיצ'רים האחרים אחראית לתוספת קטנה לדיוק).

1.3.2 חלקי הדיבר כפיצ'רים נפרדים

בשלב זה בחנו את האפשרות לקבוע את זהות הכותרת על סמך חלקי הדיבר המרכיבים אותה [1]. זיהינו את חלקי הדיבר באמצעות שימוש במודל מאומן מספריית עיבוד השפה הטבעית nltk [2]. ניסינו לחזות את זהות הכותרת הן באמצעות חלקי הדיבר בלבד והן ע"י הוספת חלקי הדיבר כפיצ'רים נוספים **ונפרדים מהמילים** אולם אף אחת מן האפשרויות לא הביאה תוצאות טובות יותר מהמודל הסופי ולכן העדפנו את המודל הפשוט יותר.

1.3.3 מפריד SVM לא לינארי (Radial base filter)

השימוש במודל זה הביא לתוצאות פחות טובות באופן משמעותי, כנראה בגלל overfitting שכן המצב השתפר כשתפעלנו את פרמטר הרגולריזציה. שוב, בהתאם להעדפתנו את המודל הפשוט, נמנענו מהשימוש בשיטה זו.

1.3.4 רשתות קונבולוציה (CNN)

בהשראת מאמרו של קים יון [3] (וממה שלמדנו בכיתה, כמובן), ניסינו לגשת לבעיה באמצעות רשתות קונבולציה. בדומה לבעיה שלנו, יון ניסה לפתור בעית סיווג בינארי של משפטים באמצעות CNN. השתמשנו בארכיטקטורה בה הוא השתמש תוך כיוונון הפרמטרים והתאמתם לנתוני הבעיה שלנו. בדומה לתוצאות המאמר, הגענו לאחוזי דיוק של $\approx 74\%$ - אחוז נמוך בהרבה משהתקבל במודל הפשוט.

References

[1] Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822, 2007.

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition, 2009.

[3] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[1] על מנת להתאים פרמטר רגולריזציה λ