עיבוד שפות טבעיות - תרגיל בית רביעי

פרטי הגשה

מגישים

אורי אריאל ניפומניאשצי	318261468
ישי גרוניך	208989186

מערכת הפעלה

Windows 7

שאלות בעקבות כתיבת הקוד

סעיף א

מהו מספר הדוגמאות שיש מכל קטגוריה של המילה "line" ב train set?

קטגוריה	מספר דוגמאות
Formation	299
Division	324
Cord	323
Phone	379
Product	2167

סעיף ב

מהו random baseline accuracy לסיווג זה? כלומר, אם נסווג כל דוגמא בצורה אקראית, מה יהיה ה-accuracy שנקבל?

אם נסווג כל דוגמא באופן אקראי, תוחלת הערך accuracy יהיה 1 חלקי מספר המשמעויות, ובפרט 0.2. הסיבה לכך היא שכל דוגמא תסווג נכון בהסתברות 0.2. באופן מתמטי, אם יש n דוגמאות וכל אחת מהן מסווגת נכון בהסתברות 0.2, נסמן משתני אינדיקטור $a_1..a_n$, כאשר a_i שווה 1 אם הדוגמה i סווגה נכון, אחרת 0.

$$E[a_1 + ... + a_n] = E[a_1] + ... + E[a_n] = P[instance\ i\ was\ true] + ... + P[instance\ n\ was\ true] = 0.2 \cdot n$$

סעיף ג

מהו majority baseline accuracy לסיווג זה? כלומר, אם נחליט שכל הדוגמאות שייכות לקטגוריה בעלת מירב הדוגמאות, מה יהיה ה-accuracy שנקבל?

אם נסווג כל דוגמא בתור הקטגוריה עם מירב הדוגמאות בקורפוס האימון, אחוז ההצלחה במסווג יהיה אחוז הדוגמאות מקטגוריה זו שנמצאות בקורפוס הבדיקה. מאחר שלכל הקטגוריות אותו מספר הדוגמאות בקורפוס הבדיקה, וישנן 5 קטגוריות, נקבל שערך זה הוא 20%.

סעיף ד

מהי הקטגוריה עם ה-precision הגבוה ביותר? עם ה-recall הגבוה ביותר? האם יש קשר בין precision ו recall בקטגוריות אלה? נסו לאמוד את טיב הקשר ולהסביר את סיבות ה"הצלחה".

להלן החלק הרלוונטי מתוך פלט התכנית:

cord: precision: 0.7058823529411765, recall 0.96
division: precision: 0.7454545454545455, recall 0.82
formation: precision: 0.7435897435897436, recall 0.58
phone: precision: 0.7169811320754716, recall 0.76
product: precision: 0.9428571428571428, recall 0.66

הקטגוריה עם הערך precision הגבוה ביותר היא product.
הקטגוריה עם הערך recall הגבוה ביותר היא cord.

ניתן לראות בערכים הנ"ל קשר הפוך בין precision לבין recall - כלומר, לקטגוריות עם precision גבוה יחסית יש recall נמוך יחסית ולהפך. ניתן היה לצפות לכך עקב ההגדרות של precision ושל recall. נניח, שמאיזושהי סיבה, יש קטגוריה c שיש לה נטייה להבחר הרבה על ידי המסווג - הוא "מעדיף" אותה על פני קטגוריות אחרות. אזי סביר שבאחוז גבוה מהדוגמאות של אותה הקטגוריה c, המסווג אכן יזהה אותה בצורה נכונה (אחוז גדול TP מתוך TP+FP). מצד שני, סביר שהמסווג גם יחשוב שדוגמאות אחרות הן מהקטגוריה c (כלומר אחוז יחסית גדול FP מתוך FP+TP). לסיכום, קטגוריה c שכזו תקבל ערך recall גבוה אך ערך precision נמוך. דוגמה לכך היא הקטגוריה cord. ניתן כמובן להתייחס לקטגוריה מסוג הפוך, שיש לה נטייה להבחר מעט על ידי המסווג, ולכן תקבל precision גבוה (רק כשהמסווג "בטוח" לגביה) אבל recall נמוך (לא כל הדוגמאות תזוהנה).

לפני שהרצנו את התכנית, ציפינו לקבל יחס הפוך בין precision לבין recall, כפי שאכן קיבלנו, אבל ציפינו שדווקא המחלקה product תקבל ערך גבוה של recall וערך נמוך של precision, מאחר שהמסווג ייטה לבחור אותה (כתוצאה מהערך prior הגבוה שלה). התוצאות היו הפוכות - 35 דוגמאות בלבד סווגו למחלקה product בעוד שלמחלקה cord סווגו 68.

המשך -->

מעניין לחקור מדוע ישנן קטגוריות עם נטייה להבחר הרבה על ידי המסווג, וקטגוריות עם נטייה להבחר מעט. כאמור, ציפינו שהגורם שיכריע בעניין זה יהיה הערך prior, שנקבע לפי מספר הדוגמאות מכל קטגוריה, אך הופתענו לגלות שכמעט לא היתה לו השפעה. השערתנו היא שתופעה זו נובעת מהעובדה שלקטגוריה עם מספר רב של משפטים, כמו הקטגוריה product, יש גם מספר רב של types. לכן, בין היתר בגלל smoothing, ההסתברות של המילים הפופולריות ביותר בשפה יורדת (מרחב ההסתברות מתחלק על יותר טוקנים שונים). בנוסף, מכיוון שיש הרבה types וגם הרבה tokens, ההסתברות של טוקן לא מוכר (על פי שיטת laplace smoothing) היא קטנה לעומת קטגוריות אחרות. לכן בהנתן דוגמה מקורפוס הבדיקה, ההסתברויות של הטוקנים בה תהיהנה קטנות לעומת ההסתברויות בקטגוריות אחרות.

סעיף ה

הציעו שינויים ותוספות ל feature set אשר עשויים לשפר את דיוק הסיווג (אין צורך לממש).

תוספות ל feature set שעשויים לשפר את דיוק המסווג:

- המיקום היחסי של המילה line במשפט - למשל, נגדיר שלוש קטגוריות מיקום: התחלה (עד 33%), אמצע (33%-67%) וסוף (67%-100%), בהנחה שיש משמעויות למילה line שסבירות יותר באיזור מסוים במשפט. לתכונת המיקום היחסי ישנה בעיה קטנה אם המילה line מופיעה יותר מפעם אחת בתוך אותו instance. על מנת לפתור בעיה זו, אפשר להגדיר 3 features, שכל אחד מהם הוא ערך בוליאני - האם המילה line הופיעה בין היתר בתחילת אחד המשפטים, באמצע אחד המשפטים, או בסוף אחד המשפטים, ודבר זה מאפשר לסמן יותר מאפשרות אחת.
- מספר המופעים של המילה line - ייתכן שבחלק מהמשמעויות למילה line סביר היא תופיעה מספר פעמים באותו context. על מנת שהתכונה הזו תשפר את הדיוק, ייתכן שצריך קורפוס יותר גדול, מכיוון שבקורפוס הנוכחי רוב הדוגמאות בעלות מופע אחד של המילה line. ניתן גם לספור בנפרד את מספר המופעים של המילה lines, שהיא סבירה יותר במשמעויות מסוימות.
- חלקי הדיבר הצמודים למילה line - מילים שקרובות למילה line הן בעלות הרבה יותר אינפורמציה מאשר מילים רחוקות. על כן ניתן להתייחס במיוחד לטוקנים הצמודים למילה line. התייחסות לטוקנים עצמם עלולה לגרום למצב של overfitting ולכן אפשר להתייחס לחלקי הדיבר שלהם.

שינויים ל feature set שעשויים לשפר את דיוק המסווג:

- בהמשך הדו"ח מפורטים מספר דרכים בהם ניסינו (מימשנו) לשנות מעט את ה feature set כדי לשפר את רמת הדיוק.

ניסיונות שיפור

ניסינו לשנות מעט את ה feature set של המסווג על מנת לשפר את רמת הדיוק: הורדנו מילים שהן stop words וכמו כן גם סימני פיסוק ומספרים, וזאת מתוך הנחה שהם מהווים רעש. בנוסף, העברנו כל טוקן לצורת הבסיס שלו (למשל, Doors הופך להיות door) באמצעות lemmatizer (שלפניו נדרש שלב מוקדם של POS Tagging) ובאמצעות העברת הטוקנים לכתיב lowercase. ההעברה לצורת הבסיס באה מנקודת הנחה שהחשיבות של המילים היא המשמעות שלהן ולא הנטייה שבה הן מופיעות. הנחה זו אינה נכונה באופן כללי (למשל, בסעיף קודם הצענו לספור בנפרד את מספר המופעים של line ושל lines), אבל היא מועילה כאשר הקורפוס אינו מאוד גדול.

ניסיונות אלו משפרים את התוצאות, אך לא באופן משמעותי. זמן הריצה מתארך כתוצאה מהסיבוכיות של POS Tagging. ניתן לראות את ההבדל על ידי שינוי הפונקציה `TextBayes.break_down` - בתוכה ישנן שתי תת פונקציות `break_down_strong` `break_down_weak` שניתן לבחור מביניהן (כאשר הפונקציה `break_down_weak` היא הדיפולטית, והפונקציה `break_down_strong` מכילה את ניסיונות השיפור אך זמן הריצה שלה ארוך יותר).