**גיליון בית 1 NLP – POS tagging– דו"ח מסכם**

זהר רימון 319010534 עדי ארבל 207919614

**אימון:**

עבור אימון שני המודלים לקחנו את סט המאפיינים שהתבקשנו לקחת (Ratnaparkhi 1996) וכן את אלו מתוך המאמר שלא חוייבנו לקחת, בנוסף לפיצ'רים מספריים ואותיות גדולות. כמו כן הרחבנו את סט הפיצ'רים על סמך מאמר נוסף ( 2003 Toutanova ) כדי לשפר את הדיוק של המודל כמפורט להלן. נשים לב כי יש לנו 44 תיוגים, אך בסט האימון לא כל קומבינציה מופיעה. לאחר כל תיאור יופיע מספר המתאר כמה פיצ'רים קיימים תחת אותה משפחה בסט האימון:

|  |  |  |
| --- | --- | --- |
| משפחת פיצ'רים | מודל 1 | מודל 2 |
| – זוגות של מלים ותיוגים שהופיעו בסט האימון | 15,315 | 1705 |
| – זוגות של סיומות מלים ותיוגן, עבור סיומות באורך 1,2,3,4 | 13,265 | 2556 |
| – זוגות של תחיליות מלים ותיוגן, עבור תחיליות באורך 1,2,3,4 | 22,393 | 3978 |
| – שלשות תיוגים סמוכים – trigram | 8150 | 1068 |
| – זוגות תיוגים סמוכים – bigram | 1060 | 316 |
| – תיוגים בודדים – unigram | 44 | 32 |
| – זוגות של תיוגים והמלים שקדמו להן בסט האימון | 32,131 | 2801 |
| – זוגות של תיוגים והמלים שעקבו להן בסט האימון | 30,793 | 2694 |
| תיוג המילה והאם מילה מכילה אות גדולה בתחילתה | 32 | 16 |
| תיוג המילה והאם כל המילה מורכבת מאותיות גדולות | 19 | 5 |
| תיוג המילה והאם קיימת אות גדולה בגוף המילה (לא כולה ולא בהתחלה בלבד) | 5 | 4 |
| תיוג המילה והאם המילה מכילה ספרה בתוכה | 6 | 4 |
| תיוג המילה והאם המילה מכילה מקף בתוכה | 15 | 7 |
| תיוג המילה והמילה שהופיעה שני מקומות לפניה | 38,050 | 3065 |
| תיוג המילה והמילה שהופיעה שני מקומות אחריה | 36,386 | 2828 |
| תיוג המילה והאם המילה מכילה נקודה בתוכה | 11 | 7 |
| תיוג המילה והאם המילה מכילה גרש בתוכה | 12 | 0 |
| **סה"כ** | **197,787** | **21,386** |

בדקנו מספר קומבינציות בתוספות למבנה הבסיסי, אך ראינו שככל שיש לנו יותר פיצ'רים הדיוק שלנו משתפר במודל. על כן בחרנו את כל הפיצ'רים שהתקבלו ובחרנו סף של 1 – כלומר פיצ'ר שמופיע פעם אחת או יותר בסט האימון יכלל בוקטור הייצוג עבור מילה בשלב ההסקה. סה"כ קיבלנו וקטור ממימד 197,787 פיצ'רים. את האתחול עשינו תחילה מוקטור אפסים אך ראינו שאתחול רנדומלי מניב תוצאות טובות יותר. כמו כן ראינו במהלך האימון שעלינו לקחת ערכי רגלוריזציה של בערך אחרי נסיון עם מספר ערכים וקבלת תוצאות לא טובות בסט המבחן למרות תוצאות אימון טובות יחסית – overfit על סט האימון.

גם עבור מודל 2 ראינו שכמות רבה של פיצ'רים מאפשרת לנו דיוק יותר גבוה ולכן לקחנו את כולם – לקחנו את סט האימון וביצענו עליו k-cross validation, עבור k=10. כיוון שהמודל הסופי יאומן על כל המידע נצפה לעלייה של הביצועים . במודל זה קיבלנו סה"כ 21,386.

**הסקה + סט בוחן:**

לא ביצענו שינויים משמעותיים באלגוריתם ויטרבי בשלב ההסקה, מלבד חישוב של כל האקספוננטים הנחוצים לחישוב ההסתברויות באופן מטריצי ברוב המקומות כדי לנצל את היעילות של המבנה המטריצי. הזמן שלוקח להסקה הוא 20 שניות למשפט בממוצע. כמו כן הרצנו 100 אפוקים וקיבלנו שב85 התוצאות היו הכי טובות על סט הבוחן.



הדיוק על סט המבחן הינו 92%. להלן מוצגת הconfusion matrix עבור 10 התיוגים בהם אנו הכי מתבלבלים

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | JJ | NN | NNP | IN | RB | VBN | DT | MD | VBD | NNPS |
| JJ | 0 | 233 | 51 | 6 | 55 | 30 | 0 | 1 | 10 | 0 |
| NN | 91 | 0 | 33 | 17 | 22 | 1 | 3 | 13 | 4 | 0 |
| NNP | 33 | 35 | 0 | 11 | 7 | 0 | 19 | 0 | 1 | 21 |
| IN | 1 | 8 | 11 | 0 | 71 | 0 | 13 | 0 | 0 | 0 |
| RB | 43 | 11 | 9 | 21 | 0 | 1 | 3 | 0 | 0 | 0 |
| VBN | 24 | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 41 | 0 |
| DT | 2 | 3 | 13 | 14 | 10 | 0 | 0 | 1 | 0 | 0 |
| MD | 5 | 19 | 0 | 0 | 2 | 1 | 0 | 0 | 12 | 0 |
| VBD | 7 | 4 | 0 | 2 | 0 | 42 | 0 | 0 | 0 | 0 |
| NNPS | 3 | 1 | 32 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

התיוגים בהם אנו הכי מתבלבלים הם JJ ו NN – במהלך העבודה חיפשנו דרכים מגוונות לנסות ולהבחין ביניהם. תוספת הפיצ'רים שכוללים תיוג + מילה מהיסטוריה יותר רחוקה (2 קדימה ו2 אחורה) שיפרה את משמעותית את היכולת שלנו לסווג נכון בין שתי קבוצות אלו. אנו מאמינים שאילו היינו לוקחים בהיסטוריה של מילה כלשהי גם תיוגים קדימה, יכולת ההבחנה בין שתי הקבוצות הללו וקבוצות נוספות תשתפר משמעותית. שינוי כזה של ההיסטוריה מצריך שינוי באלגוריתם ויטרבי, עליו נדבר בפסקה הבאה.

על מנת להכניס בחשבון תיוג אחד קדימה, השינוי בביצוע אלגוריתם ויטרבי יהיה בביצוע של ויטרבי סימטרי, כלומר ביצועו פעמיים – פעם מכיוון שמאל לימין ופעם אחת הפוך ולמשקל בין מפות החום שהתקבלו לצורך הסיווג. בגלל כיוון אחד בו התבצע ויטרבי יש חוסר סימטריה בידע שיש לתיוג על התיוג שאחריו/לפניו ועל כן במשקול שכזה נוכל לשלב את הידע הדו כיווני ולקבל החלטה יותר חכמה.

עבור מודל 2 לא היה לנו סט בוחן אך ביצענו k-fold ועל כן נוכל להעריך את הביצועים שלנו. התקבל בממוצע דיוק של 91% על כל פולד בנפרד, ולכן בסופי נצפה לערך של 92% בדיוק הסופי במידה והמודל ייבחן על טקסט מעולם תוכן דומה – בעיקר טקסט מעולם הכלכלה והעסקים.

**תחרות:**

מודל 1:

בסט האימון יש מגוון רחב של משפטים, הלקוחים מתחומים שונים ממה שראינו – ראינו משפטים משדות סמנטיים של מדע, כלכלה פוליטיקה, משפטים בכמות מכובדת מכל אחד. בקובץ התחרות ראינו בעיקר משפטים הקשורים לעולם העסקים והכלכלה. אמנם תחומים אלו הופיעו בסט האימון אך קובץ התחרות מוטה עבור תחומים אלה, ולמידה מותאמת תחום עשויה להיות יותר יעילה כיוון שמלים עשויות להיות בעלות תפקיד שונה בדיסציפלינות שונות. לכן אנו כנראה נצפה לירידה בביצועי המודל אבל לא רבה מדי כיוון שבכל זאת למדנו חלקית על תחומים אלו – ירידה של אחוזים בודדים 90% לערך.

מודל 2:

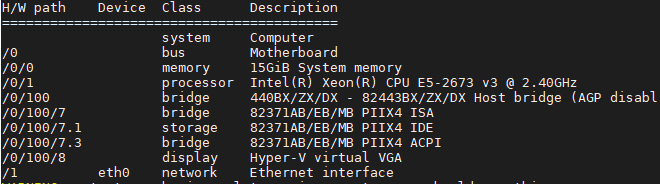
קובץ התחרות עבור מודל 2 נראה מעולם התוכן שלו ולכן נצפה לביצועים דומים ולא לירידה משמעותית.

**זמנים וחומרה**

הזמנים לצורך אימון המודלים, ההסקה ויצירת קבצי התחרות מפורטים כדלהלן:

|  |  |  |
| --- | --- | --- |
| זמן | מודל 1 | מודל 2 |
| זמן אימון כולל | 6 שעות | 5 דקות |
| זמן הסקה ממוצע למשפט | 20 שניות | 20 שניות |
| זמן הסקה לסט בוחן | 5 שעות | שעה וחצי (קרוס ולידציה) |
| זמן הסקה לקובץ תחרות | 5 שעות | 5 שעות |

השתמשנו בשרתי המחשוב שניתנו לנו בקורס לצרכי הרצה והחומרה היא כדלהלן:

****

**חלוקת עבודה**

רוב הזמן עבדנו יחד על קטעי הקוד השונים

* יצירת הפיצ'רים – כל אחד לקח כמה מהמשפחות ומימש את הפונקציות הממפיקות את הסטטיסטיקה שלהן בקובץ
* שלב הלימוד – עבדנו יחד על הפעולה הנשלחת לאלגוריתם האופטמיזציה, תוך נסיון לחשב כמה שיותר דברים פעם אחת ללא צורך בחישובים יתירים וחוזרים של ערכים שלא משתנים בין איטרציות
* שלב ההסקה – תכננו יחד את המבנה הרקורסיבי של ההסקה באלגוריתם ויטרבי ואת המקסימיזציה על כל האפשרויות.