**עיבוד שפה טבעית – תרגיל רטוב 1**

מגישים:

איתי כספי – 201239480.

דור זהר – 200951135.

אימון המודלים:

מכיוון שלכל משפחת פיצ'רים יידלקו פיצ'רים בודדים (לרוב פיצ'ר אחד בלבד), במקום לבצע מכפלה סקלרית בין v ל-f המלאים שמרנו את הפיצ'רים הדולקים, וחיברנו את האיברים ב-v לפי האינדקסים של פיצ'רים אלה.

ביצענו הרבה חישובים מקדימים על מנת לעבור כמה שפחות על ה-data כפי שהגיע:

* שמירת האינדקסים שכל מילה ב-training מדליקה עבור ההיסטוריה בה היא נראתה + כל טאג אפשרי.
* ה-Empirical Counts ב-gradient קבוע לכל ה-training, ולכן חישבנו אותו מראש והוספנו אותו בכל איטרציה של האופטימיזר.

ניסינו להימנע מפעולות כבדות ב-Python כמו קריאות מרובות לפונקציות. אימון המודל הבסיסי לוקח כ-3/4 שעה על לפטופ, אימון המודל המורכב לוקח כ-3 שעות (תלוי בלמבדה) על לפטופ וכחצי שעה על שרת עם 28 ליבות.

הפיצ'רים איתם עבדנו במודל המורכב הם –

* משפחות 100-105 כפי שראינו בכיתה. כל הפיצ'רים שמערבים מילים הומרו ל-lowercase.
  1. משפחה 100 – 13831 פיצ'רים
  2. משפחה 101 – 12074 פיצ'רים
  3. משפחה 102 – 19334 פיצ'רים
  4. משפחה 103 – 7871 פיצ'רים
  5. משפחה 104 – 1049 פיצ'רים
  6. משפחה 105 – 45 פיצ'רים
* פיצ'ר 106 – המילה הקודמת באותיות קטנות – 28876 פיצ'רים
* פיצ'ר 107 – המילה הבאה באותיות קטנות – 29245 פיצ'רים
* פיצ'ר 108 – האם המילה הנוכחית מתחילה באות גדולה – 33 פיצ'רים
* פיצ'ר 109 – האם המילה הנוכחית מכילה ספרות - 6 פיצ'רים
* פיצ'ר 110 – האם המילה הנוכחית מכילה גרש – 12 פיצ'רים
* פיצ'ר 113 – האם המילה הנוכחית מכילה מקף שמפריד בין שני מילים או מילה ומספר – 11 פיצ'רים

פיצ'רים נוספים שניסינו ולא נמצאו מועילים הינם –

* פיצ'ר 111 – אינדקס המילה – 2160 פיצ'רים
* פיצ'ר 112 – אורך המילה הנוכחית – 304 פיצ'רים
* פיצ'ר 114 – המילה שלפני המילה הקודמת – 33914 פיצ'רים
* פיצ'ר 115 – האם המילה הנוכחית היא מספר – 4 פיצ'רים
* פיצ'ר 116 – האם המילה הנוכחית מתחילה באות גדולה אף על פי שאינה המילה הראשונה במשפט או אחרי נקודה – 29 פיצ'רים
* פיצ'ר 117 – האם המילה הנוכחית נדירה (הופיעה פחות מ-t פעמים ב-training set עבור t=2) – 25 פיצ'רים
* פיצ'ר 118 – האם המילה הנוכחית היא חלק מרצף מילים שמתחילות באות גדולה ושמסתיים במילה w – 1672 פיצ'רים
* פיצ'ר 119 – האם המילה מכילה מאפיינים של תואר כמו סיומת er ו-est – 3 פיצ'רים
* פיצ'ר 120 – האם המילה מכילה מאפיינים של פועל כמו סיומת ing, ies, ied, ed, es – 12 פיצ'רים
* פיצ'ר 121 – האם נסיון הטייה של המילה להווה מניב מילה תקינה לפי הקורפוס – 17 פיצ'רים
* פיצ'ר 122 – המילה שלאחר המילה הבאה – 35342 פיצ'רים

את הלמבדה קבענו על ידי ביצוע cross validation. הפרדנו את סט האימון לסט בגודל 4500 משפטים איתו התאמנו וסט קטן בגודל 500 משפטים איתו בדקנו את התוצאות. בדקנו למבדות בסדרי גודל שונים (0.0001 עד 10 בכפולות של 10) כאשר הלמבדה המיטבית הינה 1.

אלגוריתם ההסקה:

השינוי היחיד אשר ביצענו באלגוריתם ויטרבי הבסיסי הוא שימוש ב-log ההסתברויות של q על פני ההסתברויות באופן ישיר. המוטיבציה לכך היא התמודדות עם בעיות נומריות שנובעות מעבודה עם מספרים קטנים מאוד אשר מתקבלים בוקטור ההסתברויות. בנוסף, עבודה עם log מאפשרת לבצע סכימות ולא מכפלות, מה שאמור להיות מעט יותר יעיל.

מלבד זאת, האלגוריתם ממומש בצורה מקבילית כאשר כל הגדרנו thread pool שבו כל thread פועל על משפט ב-test set.

ניתוח תוצאות קובץ המבחן:

**עבור המודל הבסיסי** (משפחות פיצ'רים 100,103,104) – 90.1% דיוק. קובץ המשקולות שמור בתיקיית הקוד בשם vector\_simple.py

**עבור המודל המורכב** (משפחות פיצ'רים 100-110+113) – 95.96% דיוק. קובץ המשקולות שמור בתיקיית הקוד בשם vector\_complex.py

ההבדלים בין התוצאות במודל הבסיסי ובמודל המורכב נובעים מהוספה של מספר פיצ'רים שמאפשרים לזהות הטיות שונות של פעלים וסיומות / תחיליות המאפיינות עצמים ותארים כמו משפחות 101 ו-102. בנוסף, פיצ'רים כמו פיצ'ר 105,106 משפרים את הדיוק על ידי התייחסות לתגים לפי מודל Unigram שמוסיף אינפורמציה שלא קיימת בלעדיו. פיצ'רים אחרים שהוספנו מאפשרים לזהות תבניות מוכרות של שמות עצם כדוגמת מספרים, או פיצ'רים כדוגמת משפחה 113 שמזהה תארים על ידי שילוב מילים עם מקף (33-year-old man כאשר 33-year-old הוא תיאור).

מטריצת בלבול בין 10 התגים עבורם המודל טועה הרבה (מטריצות מלאות ניתן למצוא בתיקיית הקוד) –



התמודדות עם בלבול בין זוגות תגים –

1. **Noun vs Adjective** – ראינו כי הרבה Adjectives מכילים מקף כחלק מהמילה בעוד ש-Nouns אינם מכילים מקף לרוב. לכן הוספנו פיצ'ר שבודק המצאות מקף על ידי חיפוש תת מחרוזת בתבנית הבאה – ספרה-אות, אות-ספרה ואות-אות.
2. **Noun vs Proper Noun** – Proper Noun הם שמות ולכן יתחילו תמיד באות גדולה. לכן הוספנו פיצ'ר שבודק אות גדולה בתחילת מילה ובודק אם המילה בתחילת משפט.
3. **Verb Past Tense vs Adjective** – בדקנו האם הוספה או הורדה של סיומות המאפינות פעלים בזמן עבר כמו ied, ed, d מניבה מילים שראינו אותם בעבר ב-train set. שיטה זו לא עזרה בפועל ולכן לא נכנסה לפיצ'רים הסופיים.
4. **Verb vs Adjective** – בדקנו האם הוספה או הורדה של סיומות המאפיינות פעלים בזמן הווה כמו ing, ies, es מניבה מילים שראינו אותם בעבר ב-train set.

סיכום המשימה:

בעזרת הוקטור הבסיסי הבנו את ה-data איתו אנו עובדים (בעזרת אחוז הדיוק, ה-Confusion Matrix ומדגם על התגים הלא נכונים) והתקדמנו לוקטור משופר. בתהליך איטרטיבי ניסינו מספר פיצ'רים על מנת לשפר את הדיוק, תוך כדי שכלול אלגוריתם הלמידה (זמן ריצה) וההסקה (קבלת נתונים רבים יותר מה-data).

תחרות:

אחוז הדיוק שאנו צופים לקבל הוא נמוך מעט מאחוז הדיוק על סט האימון (לדוגמא 90%). הסיבה לכך היא שקובץ התחרות לקוח מדומיין מעט שונה מזה של קובץ האימון והמבחן. קובץ התחרות מכיל בעיקר תכנים ביולוגיים וכלכליים בעוד שקבצי האימון והמבחן מכילים בעיקר תכנים כלכליים ומעט מאוד תכנים ביולוגיים.יש אומנם חפיפה בין הדומיינים אך בקובץ התחרות יש מילים רבות שאינן מוכרות ואופי כתיבה מעט שונה. לכן הוספנו פיצ'רים שאמורים להתמודד עם מילים שהופיעו מעט פעמים בסט האימון, פיצ'ר שאמור לעזור עבור התחרות אך בפועל גרם לירידה בדיוק על פני ה-test set.

חלוקת תפקידים:

איתי – מימוש אלגוריתם ויטרבי, הכולל הסקה, אחוז דיוק ושמירת ה-Confusion Matrix.

דור – מימוש תהליך הלמידה והפיצ'רים.