**עיבוד שפה טבעית – תרגיל רטוב 1**

מגישים:

איתי כספי –

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, אימון המודל המורכב לוקח כ-3 שעות (תלוי בלמבדה)

למודל המורכב הוספנו את הפיצ'רים הבאים:

* משפחות 100-107 כפי שנלמדו במצגת (כל הפיצ'רים שמערבים מילים – העברנו ל-lowercase).
* האם המילה הנוכחית היא עם אות גדולה, למרות שאינה בתחילת המשפט או אחרי נקודה.
* האם המילה הנוכחית היא מספר/range של מספרים (עם regexp מתאים)
* האם במילה יש את התו '
* האם במילה יש את התו "-" שמפריד בין לפחות 2 מילים או מילה ומספר
* האם המילה היא מילה נדירה (כלומר – האם המילה הופיעה פחות מ-t פעמים ב-training). ה-t שבחרנו הוא 1.
* מהי המילה שמופיעה לאחר המילה הבאה

אלגוריתם ההסקה:

ניתוח תוצאות קובץ המבחן:

סיכום המשימה:

תחרות:

חלוקת תפקידים:

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

דור – מימוש תהליך הלמידה והפיצ'רים.