עיבוד שפה טבעית – פרויקט סיום

מגישים:

אמיר קרן, 039666144

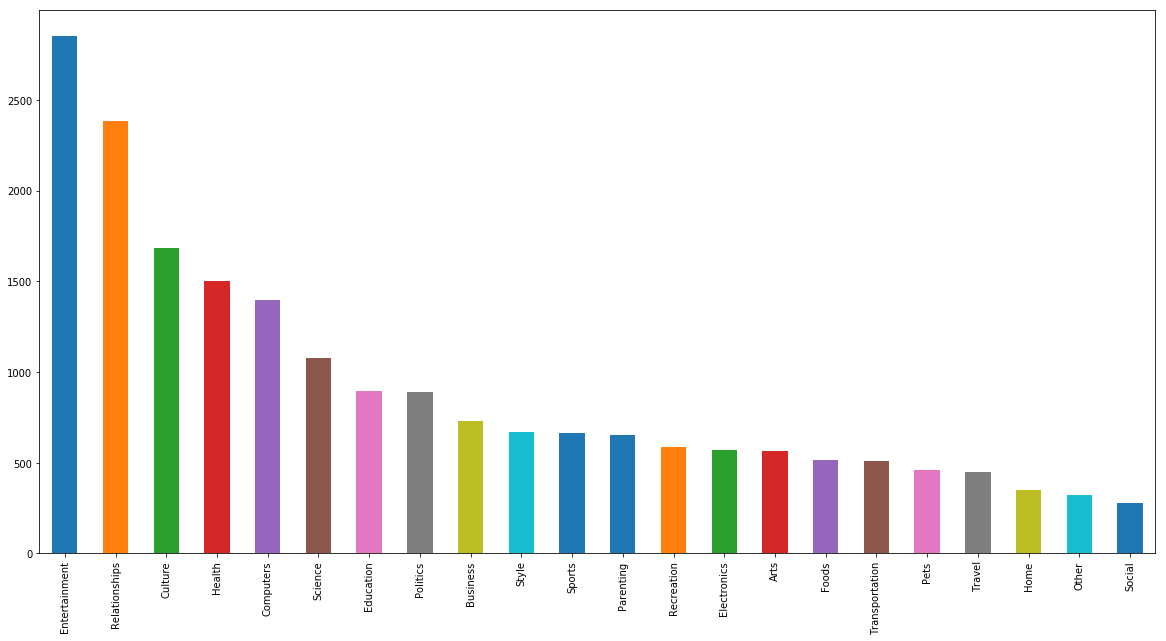
נאור דוד, 204510077

1. עיבוד סטטיסטי של הדטה (מבוסס על קובץ האימון):
   1. מהי השכיחות של כל קטגוריה? הצג את התשובה כטבלה ממוינת של קטגוריות, (ממוינת בסדר יורד של השכיחות) וכן כהיסטוגרמה (העמודות ממוינות באותו סדר משמאל לימין – משמאל הגבוהה ביותר ומימין הנמוכה ביותר)
   2. מהו האורך הממוצע (מס המילים) בכל קטגוריה. הצג את התשובה כטבלה.
   3. מהו אחוז המסמכים בכל קטגוריה שמכילות רק כותרת? הצג את התשובה כטבלה
   4. כמה מילים שונות יש בקבוצת האימון (גודל המילון), וכמה מילים סה"כ.

כל התשובות מתבססות על קוד הפייתון המצורף

a. שכיחות הקטגוריות:

|  | Counts |
| --- | --- |
| Entertainment | 2853 |
| Relationships | 2381 |
| Culture | 1685 |
| Health | 1501 |
| Computers | 1396 |
| Science | 1078 |
| Education | 896 |
| Politics | 888 |
| Business | 731 |
| Style | 667 |
| Sports | 666 |
| Parenting | 654 |
| Recreation | 584 |
| Electronics | 569 |
| Arts | 563 |
| Foods | 516 |
| Transportation | 511 |
| Pets | 462 |
| Travel | 450 |
| Home | 347 |
| Other | 324 |
| Social | 278 |



b. אורך ממוצע של משפט

| sentence\_length |  |
| --- | --- |
| category |  |
| Arts | 44 |
| Business | 53 |
| Computers | 49 |
| Culture | 57 |
| Education | 41 |
| Electronics | 42 |
| Entertainment | 46 |
| Foods | 35 |
| Health | 54 |
| Home | 46 |
| Other | 43 |
| Parenting | 73 |
| Pets | 63 |
| Politics | 55 |
| Recreation | 41 |
| Relationships | 87 |
| Science | 41 |
| Social | 55 |
| Sports | 42 |
| Style | 46 |
| Transportation | 47 |
| Travel | 35 |

c. אחוז המשפטים ללא content.

|  | Percent |
| --- | --- |
| category |  |
| Arts | 36.234458 |
| Business | 32.694938 |
| Computers | 29.226361 |
| Culture | 30.801187 |
| Education | 36.049107 |
| Electronics | 31.458699 |
| Entertainment | 31.861199 |
| Foods | 38.565891 |
| Health | 28.314457 |
| Home | 36.023055 |
| Other | 33.950617 |
| Parenting | 18.042813 |
| Pets | 21.645022 |
| Politics | 31.756757 |
| Recreation | 27.397260 |
| Relationships | 16.043679 |
| Science | 37.384045 |
| Social | 30.575540 |
| Sports | 34.534535 |
| Style | 26.536732 |
| Transportation | 35.616438 |
| Travel | 38.888889 |

d.

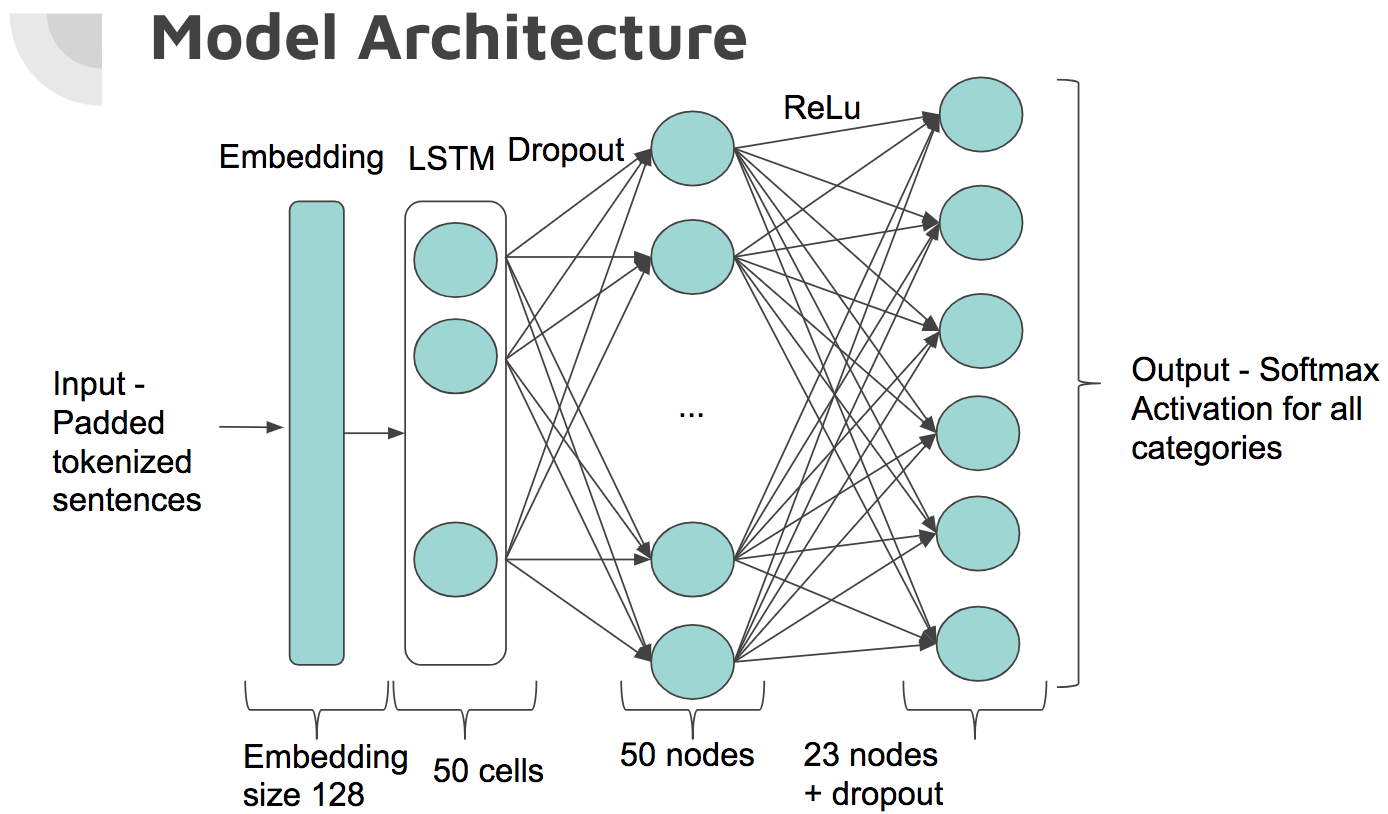
מספר המילים הכולל הוא 1073401

מספר המילים הייחודיות הוא 45715

1. על הטקסט בוצע העיבוד המקדים הבא –
   1. ניקוי (של תגי html, סימנים מיוחדים כמו &#xa וכו׳)
   2. ביצוע של lowercase
   3. ניקוי של סימני פיסוק
   4. טוקניזציה (חלוקה של הטקסט ל- tokens רצפים של תווים האמורים לייצג מילים)
   5. ניקוי של stopwords
   6. stemming
   7. קיצור אורכי מילים לא תקינות (מקרים כמו: looooong לדוגמא)
   8. תיקון שגיאות כתיב
2. הקלסיפייר הנוסף שבחרנו להשתמש בו מבוסס על רשתות ניורונים, ספציפית על LSTM.

אנחנו מזינים לרשת את המשפטים מרופדים ב- 0 לפי גודל מקסימלי לתוך שכבת Embedding ומשם לתוך שכבת LSTM. לבסוף הדטה עובר בשתי שכבות Fully Connected ומופעל Softmax כדי לנרמל את ערכי התוצאה ולקבל את הערך הבודד.

להלן ארכיטקטורת הרשת -



ביצענו תהליך של Hyperparameter Tuning על הפרמטרים והערכים הבאים –

batch\_size = [20, **32**, 64, 128]

embedding\_size = [32, 64, **128**]

lstm\_size = [**50**, 100]

dense\_size = [**50**, 100]

סה״כ **48** הרצות שונות (grid search) כדי להגיע לדיוק המירבי.

1. השגיאה הצפויה על קבוצת המבחן היא -
   1. של ה- centroid classifier – על סמך K-fold cross validation (הFOLD עם ה-validation accuracy הכי גבוה) – 52.63%
   2. של ה- LSTM – 6.16%