פרויקט גמר בקורס מבוא לעיבוד שפה טבעית  
זיהוי ותיקון של שיכול אותיות  
אביעד פינס: 307958165

הרעיון הזה הגיע בעקבות אקטון שהשתתפתי בו לפני כשנה.  
בין 5-20% מהאנשים בעולם סובלים מליקות למידה, האחוזים בטווח כזה גדול מכיוון שיש המון דעות חלוקות לגבי מה נחשב ומה לא נחשב לקות למידה.

לקויות הלמידה שקשורות לטקסט מהוות חלק גדול מאחוז האנשים שיש להם לקות למידה, ואחת הלקויות היותר נפוצות היא שיכול אותיות.  
שיכול אותיות נמצא בדיסגרפיה ובדיסלקסיה, הלקות גורמת לכותב(דיסגרפיה) או לקורא(דיסלקסיה)

המטרה שלנו היא לזהות

זיהוי ותיקון של לקות למידה מסוג שיכול אותיות בכתיבה(דיסגרפיה).

אתגרים בדרך:

* הכנת קורפוס של משפטים עם שיכול אותיות.
* זיהוי שגיאה אפשרית מסוג שיכול אותיות.
* תיקון השגיאה למילה שאליה התכוון הכותב במידה וקיימת שגיאה.

הכנת המידע:  
הקורפוס נלקח מאתר מילה, ויקפדיה 2013.

* יצירת מילון מכל הtoken שקיימים בקורפוס, בנוסף ליצירת המילון אני גם סופר כמה פעמים הופיעה כל מילה בקורפוס.
* חילוץ bigram מכל המשפטים בקורפוס.
* הכנת משפטים עם שיכול אותיות (בגלל שלא הצלחתי לארגן מידע מתאים):
  + לקחתי כ150000 משפטים בשביל הtest/
  + בשביל להכין את המשפטים הייתי צריך דרך למצוא מילים אפשרויות לשיכול אותיות.  
    ולכן במקום להשתמש במודל של הערוץ הרועש, לוקחים את כל הערבוב של הסדר של האותיות עבור כל מילה, ועבור כל האפשרויות שמתקבלות בודקים איזה אפשרויות תקינות לפי המילון שיצרתי, ככה אני מקבל את כל האפשרויות למילה שהיה לי שיכול אותיות.
  + אני מגריל במקום משפט(סתם בשביל שהשיכול לא תמיד יהיה באותו מקום למרות שזה בכלל לא משנה), ורץ מהמקום המוגרל עד סוף המשפט, עבור כל מילה אני בודק אם קיים לה שיכול אפשרי במידה וקיים אני בוחר גם בהגרלה אחד מבין השיכולים ומחליף את המילה האמתית במילה המשוכלת כמובן שאני שומר על המילה האמתית בנוסף כדי שיהיה אפשר להשוות.

ככה יצרתי 150000 משפטים עם שיכול אותיות כדי לבדוק את המודל.

הכנת המודל:

* עבור כל משפט שאנחנו רוצים לבדוק האם קיים בו שיכול אותיות, אנחנו בודקים את כל האפשרויות לשיכול בדומה לבדיקה שביצענו עבור ההכנה של הtest.
* במידה וקיים שיכולים אפשריים אנחנו בודקים לאיזה מבין השיכולים(והמילה האמתית כמובן) יש הסתברות יותר טובה, ובוחרים את המילה בעלת ההסתברות הטובה ביותר.

תוצאות:

* מתוך מאגר של 102757 משפטים עם שיכול אותיות (שאני הכנתי)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| סוג | החלקה | True | False | Probabilities | הערות |
| BaseLine | ללא | 0 | 102757 | 0 | פשוט רנדומלי עשיתי כמה הרצות וזה פשוט כל פעם שינה מילים שלא הוחלפו במשפט המקורי ולכן הכל נהרס לו |
| unigram | UNK | 65914 | 36843 | 64.145% |  |
| Bigram | ללא | 48218 | 54538 | 46.9242% |  |
| Bigram | Unk + Additive Smoothing | 51562 | 51195 | 50.1785% |  |
| Bigram | Unk + Additive Smoothing | 98347 | 4410 | 95.7083% | המודל משתמש בbigram עבור המילה הקודמת והמילה הבאה בו זמנית. |

חקירת שגיאות:

בשביל מחקר השגיאות פתחתי פקד שעוזר להסתכל על השגיאות, ואומר לנו אלו מילים השתנו נכון אלו מילים השתנו למרות שלא היו צריכות, ואלו מילים לא השתנו בכלל.

הסבר על התוכנה מצורף בreadMe של הגיט.  
הדבר הזה תרם לי רבות(לצערי עשיתי אותו רק בסוף), קודם כל אפשר לראות כל מיני שיכולים מעניינים שנוצרו בצורה רנדומלית (הוסבר למעלה את הדרך המלאה) וקצת חששתי מהתוצאות אבל יש דברים באמת נחמדים מאוד, הנה כמה דוגמאות:

* לתאר ⬄ לאתר
* מרצים ⬄ מצרים
* הריאה ⬄ האריה
* תירשם ⬄ תרשים

ויש עוד הרבה, מה שאפשר להגיד שאולי לא התעסקנו עם מידע אמיתי לגמרי אבל הוא בהחלט מתקרב לשם.

שגיאות:

* יש משפטים שהמילה המשוכלת תוקנה כמו שצריך אבל מילה אחרת נשכלה במקום, לדוגמא משפט 45, המילה "מרצים" התחלפה ל"מצרים" מה שלא היה נכון, אבל המילה "התקיעה" הוחלפה ל"העתיקה" כמו שצריך, הדוגמא הזו מעניינת כי אפשר להבין את המילה בתור העתיקה(משהו ישן) ובתור העתיקה( העתיק במבחן) שזה יכול להסתדר יפה גם עם מצרים וגם עם מרצים(כי לשניהם יש דברים קשורים).  
  הדוגמא הזו מראה שיש כמה
* מילים שפחות מוזכרים צריכים הקשר יותר רחב של המשפט כאשר הם מופיעות בתחילת המשפט, מכיוון שבמידה ויש להם שיכול אפשרי שהוא יותר נפוץ השיכול הזה יהיה יותר חזק בתחילת המשפט כי סביר להניח שמידה ושניהם לא מופיעים אותה כמות של פעמים אז הסיכוי שהמילה הנפוצה יותר תהיה יותר חזקה. (לדוגמא משפט 150, "האיברים" כנראה קבוצה של עמים, הוחלפה לה "האבירים" שזה בהחלט נשמע הגיוני ליד המילה "היו".
* כאשר יש מילה עם שיכול אפשרי אחרי מילות יחס זה סיכוי לטעות, יש כמה משפטים שנפלו על זה.
* שיכול שהמשמעות שלו זהה למילה המקורית רק בזמן אחר, לדוגמא: וקיבל ⬄ ויקבל(משפט 204)
* המון שגיאות של המילים: "העברית" ⬄ "הערבית"