תיאור הבעיה והדרך לפתרון

מטרת העל של הפרויקט היא השגת דרך לייצוג מילים על ידי וקטורים בצורה הטובה ביותר. כלומר על ידי שימוש בכלים הקיימים כיום בעולם המחשבים, w2v, glove וכל כיוצא בהם, לתת ייצוג וקטורי לכמה שיותר מילים ובצורה המדויקת ביותר.

החזון שפרויקט זה בא להשיג הוא בכך שכאשר נצליח לתת ייצוג נכון ומדויק לכל מילה קיימת יהיה בידנו יכולת לנתח טקסט מכל סוג שהוא ואף לסווג את מילותיו במכנה משותף קרי אשכולות או סיווג של חלקי משפט נושא נשוא וכו' מה שיאפשר הבנת טקסט באופן תכנותי.

האתגר הגדול של הפרויקט הוא בכך ששני הכלים אינם מכסים את כל המילים ולכן דרך הפעולה של הפרויקט תהיה על ידי בדיקת תוצאת כל כלי בפני עצמו, שילוב הכלים על ידי אופרטורים שונים כגון איחוד, חיבור וכו' (בדגש על יתרונותיו וחסרונותיו של כל אופרטור) ובסדר מסוים. כמו כן, נוכל לבחון מילים מסוימות להן לא מצאנו התאמה לווקטור על ידי הסרת רישות וסיפות שמשנות רק את יחס המילה אך לא את משמעותה כגון אכל ואכלתי שהשוני בין המילים הוא רק בגוף המדבר. על ידי כך נוכל להציע כיסוי גדול יותר של מילים ועדיין להישאר במשמעות נכונה של מילה.

לוח זמנים משוער

"כל ההתחלות קשות" (רש"י על מעמד הר סיני) ועל כן סביר להניח שבתחילה נצטרך להתגבר על קשיי ההסתגלות לכלים איתם נעבוד ולשיטת העבודה אותה נצטרך לאמץ בשביל להגיע לפתרון המיוחל. ברם, אנו מאמינים שקצב העבודה יגדל באופן אקספוננציאלי שכן לנסיון שנצבור בשימוש הכלים יתווסף העניין שיש לנו בפרויקט ובכך יערב לנו מאז ולהבא כך שבעזרת השם נוכל להציג תוצאות חלקיות לפחות כבר בסוף הסמסטר הנוכחי.

תוכן עניינים משוער

* מבוא.
* מילים לוקטורים (הסבר על תהליך הכלים, שימושם ומימושם).
* כלים עם חסרונות ויתרונות (הסבר על הכלים מאיפה הגיעו וכו').
* כיסוי המילים על ידי הכלים כל אחד בפני עצמו.
* דרכים לשפר את הכיסוי (תתי נושאים רבים).
* זיהוי אשכולות על פי הדרך האופטימלית.
* זיהוי נושא על פי הדרך האופטימלית.
* סוף דבר.
* ביבילוגרפיה.