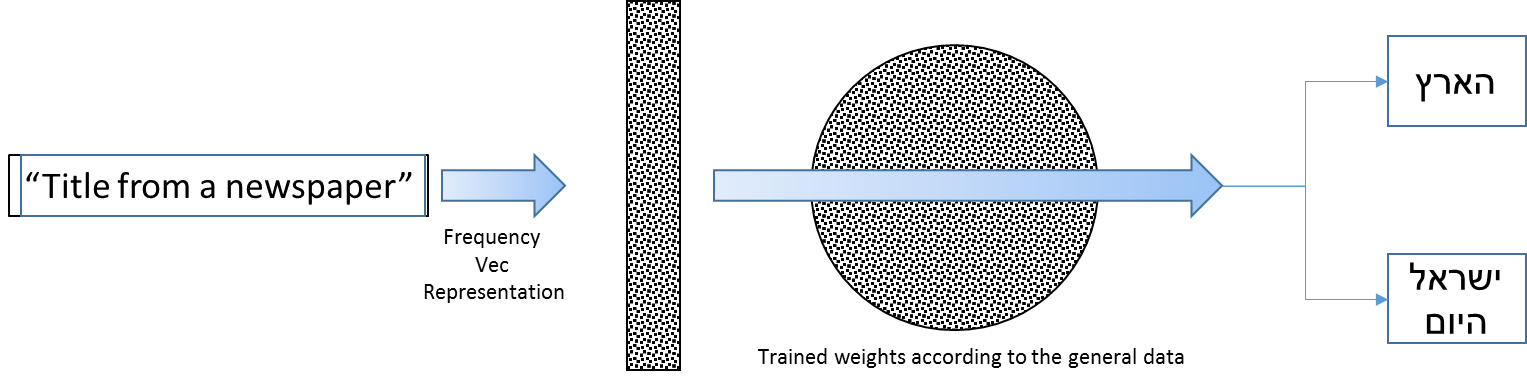
האקתון IML - משימה 1

בתחילת הדרך שקלנו לייצג את משפטי הכותרות באמצעות וקטור של פרמטרים (כמו בצורת bag of words קרי וקטור לפי המילים, או באמצעות שימוש ב word2Vec שאומן קודם לכן), ולהזין זאת לרשת נוירונים שתלמד משקולות רלוונטיות למילים החשובות. חקרנו באינטרנט וגילינו שלמען עיבוד שפה טבעית רשתות במבנה RNN מתאימות מאוד. מימשנו רשת ע"פ מבנה שמצאנו באינטרנט, והיא התכנסה אך עם overfitting, ולבסוף הבנו שככל הנראה אין לנו מספיק דוגמאות למען אימון הרשת עד התכנסות גם בולידציה. לפיכך, בנינו crawler ושלפנו עוד כ-1000 כותרות מ"הארץ", אך זה לא הועיל (ככל הנראה היינו צריכים עוד דוגמאות רבות, או לשנות את מבנה הרשת – אך לא היה לנו מספיק ידע בנושא זה).

מכאן, החלטנו לשנות כיוון מרשת הנוירונים, והחלטנו לפתח k-means כך שנזהה clusters של מילים קרובות (לאחר שהומרו לוקטורים ע"י words2Vec מאומן), ואז לנסות ללמוד משפטים חדשים ע"ב שיוך המילים לוקטור, למשל ע"י מפרידים לינארים. החלטנו לשפר את האלגוריתם ע"י הוספת features נוספים לוקטור המאפיין מילים מלבד הcluster אליו הן שייכות, ולהשתמש ב Model selection ולהריץ SVM ואלגוריתמי למידה דומים כך שנוכל לבחור את הטוב מביניהם. הfeatures השונים שבחרנו ללמוד היו: וקטור של features המבוסס על המילים שמגיעות במשפט (מתוך וקטור של 1000 המילים הנפוצות ביותר), אורך המילה הממוצע, אורך המשפט הממוצע, ציון משפט ע"ב נרמול המילים המופיעות בו לעומת הופעת המילים בכל אחד מהdatasets , וכד'. נתקלנו בשתי בעיות: הראשונה, מתברר שwords2Vec מאומן דורש כ-3 ג'יגה בייט זיכרון והנחנו שלא נוכל להגיש אותו, ולצערנו כבר מימשנו את האלגוריתם k-means אך ללא הועיל. שנית, הfeatures שבנינו לא הצליחו לייצג את המדגם טוב יחד עם כל אלגוריתמי הלמידה שניסינו. לבסוף, מיקדנו את עבודתנו לתצורה הנוכחית של הפרויקט: מעין bag of words מתוחכם, שיפעל כמתואר בפסקה הבאה.

ה-bag of words מתבסס על וקטורים שמייצגים את המשפטים ע"ס המילים שמוכלות בהם. החלטנו לייצג את כל המילים שלנו באופן sparse אשר מיוצר ע"י ספרית scikit-learner לאחר שנותנים לה את הdata שלנו. בהתחלה פרסרנו את המילים כך שנתעלם מstop-words (מילים נפוצות שככל הנראה לא מלמדות על המדגם, כגון 'the' ), וגם ניקינו אותם מסימנים לא נחוצים. אז, הפעלנו אלגוריתם SGD ולאחר שויתרנו על סינון ה stop-words הוא עבד באחוזים די גבוהים (כמובן, שלא אימנו על כל המדגם שקיבלנו, החלטנו לאמן על 80% מהdata, ולהשאיר 10% לולידציה ואת 10% הנותרים לtest).

בשלב זה, הצלחנו להגיע לכ-78% הצלחה והתחלנו לטייב את הפרמטרים לפיהם האלגוריתם הלמידה רץ (בין אם הפרמטרים של יצירת וקטור הfeatures או הפרמטרים או האופטימיזציה וכו'). בעיקר על סמך ניסיון וטעיה, הצלחנו למצוא את פונקציית הloss הטובה ביותר עבורנו, ועוד פרמטרים דומים אשר הביאו לתוצאות הטובות ביותר, ולשיפור של כ8% נוספים.

להלן תרשים של ההפעלה על כותרת חדשה, לאחר סיום האימון: