Tomer Wolfson <tomer.wolfson@gmail.com>

עדכון אחרון ML

1 message

Saph N <saph102@gmail.com> Wed, Feb 3, 2016 at 12:36 PM
To: Tomer Wolfson <tomer.wolfson@gmail.com>

יש כאן את העדכונים האחרונים שלי.

מה עשיתי בגדול:

1. הגדלתי את מאגר המילים שלנו.

2. חילקתי את הפיצ'רים ל-bigrams המכילות מילים מהמאגר ואלו שלו וכנ"ל עבור unigrams (לעבוד רק עם bigrams לא נתן לי תוצאות טובות מספיק).

3. במקום סף אני לוקחת פיצ'רים על סמך הסתברות כאשר אני קובעת לכל קבוצה מהו הסף שהיא צריכה לעבור כדי להתקבל כפיצ'ר.

במקרה זה, bigrams ו-unigrams המכילות מילים מהמאגר צריכות להוות אחוז קטן יותר מסה"כ ההופעות על מנת להתקבל,

בעוד אלו שלא מכילות מילים מהמאגר צריכות להופיע מספר רב של פעמים על מנת שיתקבלו (סף הסתברות גבוה יותר).

הכוונון והמשחק עם האחוזים השונים משפר את התוצאות על הסקלה שבין overfitting להתאמה גרועה.

כרגע אני חושבת שמצאתי דרגה טובה לכל קבוצה כדי להישאר עם מעט פיצ'רים אבל עם התאמה גבוהה.

4. כמו כן bigrams שמורכבות **רק מ-stop words** (כלומר שתי המילים הן stop words) לא נכללות - שיפר את התוצאות בהרבה!

5. לכל ביקורת יש דירוג של כוכבים כאשר ביקורת עם 1 היא הכי גרועה ו-10 הכי טובה.
הוספתי את הדירוגים כמשקלות לנקודות שלנו כך שביקורת גרועה מאוד או טובה מאוד מקבלת משקל גבוה יותר.

6. עשיתי std ל-svm.

בינתיים זהו. אני לא חושבת שיש בתיקיה את כל מה ששמת אלו רק הדברים שעבדתי איתם (הקוד, הרשימות והביקורות) אני ממשיכה לחשוב על רעיונות.

ML_new.zip

1839K