Sentiment analysis - methodology

# איסוף המידע

המשפטים שמשמשים אותנו לבניית הדאטהסט, נקצרו מאתרי חדשות בארץ בחלוקה לפי קטגוריות: חדשות(פוליטי-מדיני), כלכלה וספורט.

מכיוון שחלק גדול מהמידע לא מתאים לניתוח סנטימנטים - מתייגים עברו על כל המשפטים, מחקו משפטים לא תקינים(ע"פ מסמך ההגדרות למתייגים), וניסחו מחדש משפטים תקינים.

# תיוג

בשלב הראשון כל משפט מתוייג ע"י 2 מתייגים. אם אין הסכמה בין המתייגים, אנחנו מתייגים את המשפט ע"י 3 מתייגים נוספים.

מבין המשפטים שיש עליהם הסכמה של 2 המתייגים הראשונים - אנו לוקחים 10% ומתייגים אותם ע"י מתייג שלישי (זאת על מנת לקבל כמות משפטים מסוימת שיש עליהם קונצנזוס רחב יותר)

# מתייגים

כלל המתייגים שמשתתפים בפרויקט הינם מתמללים בחברת סאמיט, עברו מבחני קבלה, והם דוברי עברית ברמה גבוהה. מבחני הקבלה כוללים ידיעת השפה, בדיקת איות ודקדוק, כתיבה משמיעה וקצב הקלדה(חלק מהמבחנים לא קשורים בהכרח למשימת התיוג אבל הם עוברים אותם כחלק מתהליך הon-board לחברה). המתייגים שנבחרו לעבוד בפרויקט הם מתמללים שעובדים אצלנו בתמלול פרוטוקלים ויש להם "קרדיט" גבוה אצלנו

כל מתייג שנכנס לעבודה מקבל כמות משפטים קטנה יחסית בפעמים הראשונות, והתיוגים שלו עוברים בקרה:

1. על ידי מנהלי הפרויקט
2. על ידי השוואה לתיוגים של מתייגים אחרים

כל מתייג יקבל פידבק ע"י מנהל הפרויקט בסיום המנות הראשונות

מתייגים שהתיוגים שלהם שונים באופן גבוה מהצפוי - לא יוכלו לקבל משפטים נוספים, והתיוגים הקודמים שלהם יוסרו מהמאגר

# תיוג וניתוח

* שלב א' - כל המשפטים עוברים סינון או תיקון ע"י המתייגים
* שלב ב' - המשפטים מחולקים למתייגים ומתוייגים פעמיים ע"י מתייגים שונים
* שלב ג'
  + מסננים את כל המשפטים שסומנו כ"לא תקין"
  + לוקחים את כל המשפטים שאין עליהם הסכמה ומעבירים אותם לתיוג ע"י 3 מתייגים נוספים
  + מבין המשפטים שיש עליהם הסכמה - לוקחים 10% מהמשפטים ומתייגים אותם ע"י מתייג שלישי