זיהוי כותרות - פרויקט האקטון

אנו גאים להגיש לכם את מזהה הכותרות שלנו.
הגענו לדיוק גבוה מאוד על הנתונים שברשותנו (98% על test-data נפרד) ואנחנו מצפים מהמתייג שלנו שיוכיח עצמו גם בשבועות הקרובים.

במהלך העבודה הבנו שהאתגר העיקרי אינו הפעלת אלגוריתם כזה או אחר אלא הכנה, מיצוי וטיוב של המידע שברשותנו.

Feature extraction

בשלב ה - pre-proccessing, אנחנו "מנקים" מהמידע רווחים כפולים, סימני פיסוק וכו'.
בשלב הבא אנחנו קוראים את המידע ומייצגים אותו כאובייקט DATAFRAME, בו כל עמודה מייצגת מילה, וכל שורה מייצגת כותרת של כתבה. המספר בתא(כותרת, מילה) הוא מספר הפעמים שהמילה הופיעה בכותרת. למעשה, קיומה של מילה במשפט הוא הפיצ'ר הראשון והבסיסי ביותר של כותרת.

```
In [18]:  1 print(tmp.head())

   000  007  10  100  100b  100m  100th  101  102  11  ...    zionist  \
0    0    0   0    0     0     0      0    0    0   0  ...          0
1    0    0   0    0     0     0      0    0    0   0  ...          0
2    0    0   0    0     0     0      0    0    0   0  ...          0
3    0    0   0    0     0     0      0    0    0   0  ...          0
4    0    0   0    0     0     0      0    0    0   0  ...          0

   zionists  zoabi  zombie  zone  zones  zoo  zuckerburg  zurich  zweig
0         0      0       0     0      0    0           0       0      0
1         0      0       0     0      0    0           0       0      0
2         0      0       0     0      0    0           0       0      0
3         0      0       0     0      0    0           0       0      0
4         0      0       0     0      0    0           0       0      0

[5 rows x 8853 columns]
```

בשלב הבא אנחנו מוסיפים למידע פיצ'רים נוספים.
כבר בשלב קריאת המידע נוספות עמודות המייצגות מספר ההופעות של זוגות מילים במשפט.
הפיצ'ר הבא שהוספנו היה אורך המשפט, ואחריו באו אורך מילה ממוצע, מספר ההופעות של נקודה במשפט, וגולת הכותרת - מספר הפעמים שכל חלק דיבר מופיע במשפט.

התאמת מודל

לאורך הלילה ניסינו מספר רב של classifiers מתוך ספריית scikit-learn, וככל שהלילה התקדם ונוספו פיצ'רים נוספים למידע כך הclassifiers הביאו תוצאות טובות יותר.
בשלב מסוים ניסינו להשתמש בVoteClassifier שעושה היתוך מידע לתוצאות של כמה Classifiers אחרים שגילינו כמוצלחים במיוחד.
לבסוף בחרנו להשתמש בMLPclassifier, שהוא למעשה רשת נוירונים. הוא אכל את כל השאר בלי מלח מבחינת ביצועים, והתברר כפשוט לשימוש.