Markov Networks as Prior Distribution for

Multi-Label Classification (in Wikipedia)

By: Alon Wolf (318336930) and Aviv Hadar (pi)

תקציר *// בסוף*

תאוריה

* מודל מרקובי pairwise *// אביב*
* אלגוריתם צ'או ליו *// אביב*
* שימוש אלגוריתם צ'או ליו כ-prior *// אביב*
* "הכללה" לקורלציות שליליות *// אביב*

דאטא *// אלון*

* ויקיפדיה ו-infobox

את המידע חילצנו מאתר ויקיפדיה. השתמשנו בספריות wikipediaapi, wptools על מנת לחלץ את הטקסט וה-infobox של ערכי הויקיפדיה (דוגמה לערך כזה ניתן לראות ב*איור 1*).

נשאלת השאלה אילו ערכים לקחת – יכלנו לעשות רשימה של ערכים ולבחור מתוכם אך על מנת שההתפלגות של הערכים תהיה זהה (שהערכים ייבחרו על ידי התפלגות אחת) לקחנו את כלל הערכים מתוך קטגוריה אחת: מוזיקאים (<https://en.wikipedia.org/wiki/Category:Musicians>).



*איור 1 – ערך ויקיפדיה של אלטון ג'ון. באדום – הערך (טקסט), בכחול – עובדות מהירות. מטרת הפרויקט היא לחזות את החלק הכחול בהינתן האדום.*

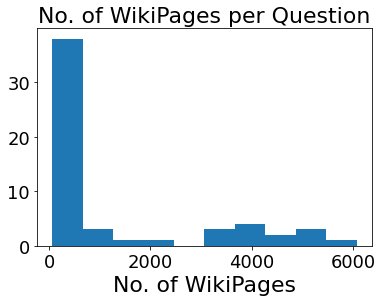
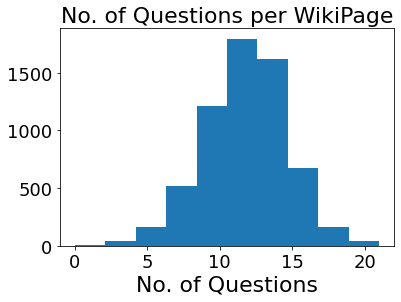
לקחת את הטקסט כמשתנה הבלתי-תלוי ואת השאלות של ה-infobox כמשתנה התלוי.

את הטקסט העברנו טוקניזציה פר מילה כך שלכל מילה ייצרנו מימד חדש על האם הטקסט הכיל את המילה. לכל מילה הכנסנו ציון לפי אלגוריתם TF-IDF כך שככל שמילה מופיע יותר בטקסט כך היא מקבלת ציון גבוה יותר, מנורמל לפי מס' הפעמים שהמילה הופיעה בכלל הטקסטים. הורדנו מילים שחזרו במס' בודד של טקסטים כיוון שלא רצינו שהמודל ילמד קורלציה שרק קורלטיביות למילים בודדות.

כמו כן, מתוך השאלות גם כן סיננו שאלות לא נפוצות (מופיעות עבור מס' בודד של ערכים) כיוון שעבור דוגמאות כאלו הציונים הסופיים מאוד משתנים לפי החלוקה של קבוצת האימון וקבוצת המבחן ולכן מאוד לא יציבים.

כפי שניתן לראות מהדוג', הבעיה הינה MultiLabel כיוון שלכל ערך ישנו מס' משתנה של שאלות.

את התפלגות מס' השאלות לכל ערך ניתן לראות באיור מס' 2.



תוצאות *// אלון*

* לכל מטריקה [roc, f1]
  + לכל מודל
    - לכל prior
      * מה התוצאות
* הסברים (+ גרפים) *//* נכתוב ביחד

סיכום ומסקנות *// בסוף*