**ניקוי הדאטה**

**סיומות ותיקונים פשוטים:**

סיומות של מדינות- איך נכון לעשות? להפנות לרשימת מדינות שאת הכל יוריד? איך ידע להתייחס לקיצורים? לשים לב לא להוריד שמות ערים אלא רק מדינות. מצרפת את מה שמצאתי:

* UK
* AU
* EMEA
* EUROPE
* USA
* US
* AUSTRALIA
* CANADA
* NEW ZEALAND
* SINGAPORE

מדינות שנכתבות בכמה צורות (לאו דווקא בסיומות אלא לפעמים בשם ההוצאה ממש). אולי שווה שכל נקודה תהפוך לרווח או תימחק? כבר יפתור חלק מהבעיה, אבל רק חלק.

* US
* U S
* U.S.
* U.S

סיומות בתוך סוגריים (לאו דווקא שמות מדינות) – יש איך להוריד? אולי אם השורה נגמרת בסגירת סוגריים שיחתוך עד שהוא פוגש את הסוגריים הפותחים?



סיומות כלליות שכדאי להוריד – עלול אולי לפגוע בשם החברה, אבל יותר מידי פעמים זה גרם להבדלה בין אותה החברה וכן כדאי למחוק

* INC
* LTD
* INTERNATIONAL
* LIMITED
* CO

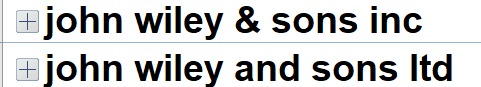
סיומות שלפעמים מפריעות אבל כדאי לא להוריד כי ייצור מלא בלאגן והתאמה בין דברים לא זהים – BOOKS, PUBLISHER, PUBLISHING

סיומות כפולות שצריך להוריד את שתיהן – נגיד UK LTD. אולי כל מילה להריץ שוב על הבדיקות עד שעוברת את כל הבדיקות? ככה יוריד את 2 הסיומות. כנל כשיש סיומת ואחריה סימן שצריך להוריד.

סיומות שלא מופרדות ברווח – למשל INC,US. איך יידע להעיף את כל המילה? אולי שווה להגדיר שכל פסיק יהפוך לרווח ואז נעיף את כל הפסיקים ויהיה יותר פשוט?

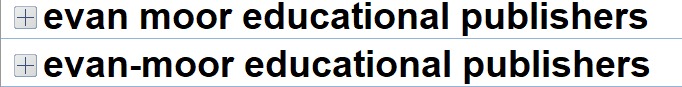
סיומת של סימן – נגיד מקף, פסיק, נקודה וכאלה. יכול להיווצר גם אחרי הורדת סיומת אחת ונשאר סימן אחריה. אז להעיף סימנים בסיומות.

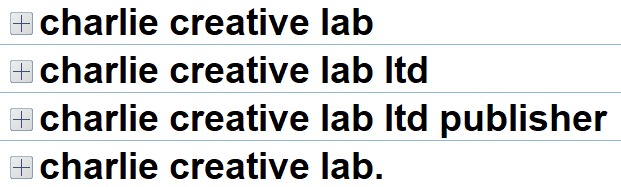
AND ו-& שצריך להחליף שיהיו אחידים (מצרפת דוג')

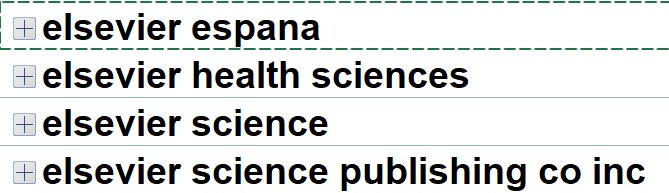


**דוגמאות ספציפיות לבעיות שצריך לחשוב איך לפתור אותן:**

1. מקף בין 2 מילים בשם ההוצאה

1. סיומת PUBLISHER\BOOKS שתוקעת (כל השאר היה מסתדר להיות אחיד)  
   
2. אותה הוצאה אבל עם סקשנים שונים לחלוטין. יש לנו מה לעשות?



1. אותה הוצאה בינלאומית (שיש תחתיה המון ספרים) שכתובה ב700 דרכים. בדקתי שזו אכן אותה חברה (אולי שווה נקודתית להגדיר עליה שכל הוצאה שמתחילה במילה SPRINGER תיקרא פשוט כך בלי התוספות).



1. סיומות שלא מופרדות ברווח מהשם של ההוצאה (בשונה ממה שכתבתי קודם שהסיומות מוצמדות אחת לשנייה). אם כל פסיק באמת יהפוך לרווח יכול לפתור גם את זה לא?)  
   