תרגיל בית 1

מיאס גנטוס -213461692

פאיסל עומרי –

שלב 1 :

1. קיבלתי את כל המסמכים הנמצאים באותה תיקיה דרך ואז עשיתי ניתוח על השם של הקובץ ושמרתי אותו ב list of dictionary , כך ש הוא המספר XX,

הוא אומר אם הוא מליאה או ועדה, הוא התוכן ו טקסט שעוזר בתכנות ולמצאו איפה השגיות.

פונקציה get\_all\_docx\_in\_current\_foleder .

1. אנחנו משנים את המדבר כאשר מופיע הסימן ":" וגם כל השורה היא underline(יש לנו בעיה בו כי זה אומר שאם אין ":" אז אנחנו לא נשנה את המדבר אפילו אם יש מדבר חדש ואנחנו לא יכולנו להתמודד עם זה אבל לרוב המקרים יש ":" ו underline שזה לא ישפיע כל כך על התוצר הסופי).

אם מופיע משפט מסוג הזה אבל יש לו אז בהסתפרות גבוהה זה לא דובר חדש ולכן אנחנו משייכים אותו לדובר בקודם.

אם המסמך ועדה:

ראשון אנחנו נמצא את מה הדבר הראשון שידברו עליו כי אנחנו נקלוט טקסט משמה עלה ,עושים את זה על ידי ההופעה הראשונה של 'סדר' ו'יום' באותו שורה ,כי אז זו הסתפרות גבוהה שזה אומר סדר הדיבור של המסמך.

אם המסמך הוא מליאה:

לא עשינו את השלב הזה (של הסעיף הקודם) כי במסמכים האלה אין להם בפתיחה מה שהוא שמזיק על האלגוריתם שלנו (נקודתיים עם underline).

מה הבעיה אם לא עשינו ניקוי?

זה אומר אם השם כתוב בצורות שונות (למשל בן ימין נתניהו – ראש הממשלה או בן ימין נתניהו (ראש הממשלה) ) אז הוא לא יזה שזה אותו בן אדם.

מה הבעיה אחרי הניקוי?

יכול להיות שעשינו לאיזה שהוא שם ניקוי לא נכון ואז גם לא נזהה שזה אותו בן אדם.

איך עשיתי ניקוי לשם?

1. אם יש לך סוגריים אז לא תקרא את הסוגריים ומה יש בניהם (בעיה אם השם האמתי בין הסוגריים).
2. אם יש לנו מה שהוא שמפריד אבל עם רווחים (כמו " – ") אז תקרא רק את החלק הראשון (בעיה אם השם הוא בחלק השני, אם אין רווחים ואם השם עצמו יש בו סמל מפריד עם רווחים).
3. אם יש קיצורים (הקיצורים יהיו תמיד עם הסמל ) אז לא תקרא.
4. אם יש לך מספר מספור בעברית (כמו " היו"ר ר' שם) אז תחסיר את המספור ומה שיש לפניו. (בעיה אם אין מספור או שהשם לפני התפקיד שלו או אין את הסימן " " " בקוד עשיתי שהמספר המקסימאלי של המספור הוא 2 תווים אבל יוכל להיות יותר וגם בעיה אם השם יש לו פחות משני תווים ויש " " בסוף).
5. תעשה split לפי התווים .؟!:; אבל אם הseparator בתוך גרשיים (""\") אז לא תעשה split לפי הseparator הזה, אלא תכלול אותו כולו במשפט הנוכחי.

עשיתי את זה בגלל שלפעמים הטקט בין הגרשיים הוא מה שהוא פירוט על הטקט שלפנו ולכן אני רוצה את הטקט שבין הגרשיים שלם.

אבל אם יש שני טקסטי גרשיים אחד בתוך השני אז אנחנו לא יכולים לזהות את זה ולכן זה בעיה שלו יכולים להתגבר עליה.

1. ניקוי משפטים:

פשוט עשיתי אם האות לא בעברית או שהיא סמני פיסוק אז תמחק ותשווה עם הטקסט המקורי אם הוא שווה זה אומר שהוא יכול להיות תקין אם הוא לא שווה אז תמחק את כל המשפט. ואם המשפט אין בה עברית אז תמחק , אם יש ' - - -' אז תמחק המשפט כולו.

(לא יודיע לנקות את התאריך אם צריך)

1. לפעמים יש סמני פיסוק שמופעיות בטקט אבל הם לא באמת סמני פיסוק למשל אנחנו יכולים להשתמש בגרש כדי לעשות קצור (כמו יו"ר) או להשתמש ב "-" בשם (כמו: בית-ספר)

ולכן אילה לא יהיו טוקנים.

איך אני מפריד בין אילה לסימן פיסוק רגיל?

ניקח את התא אם יש רווח או סימן פיסוק מימין או משמול אז אני סימן פיסוק אמתי (תפקיד של פיסוק) זה אומר שאנחנו נבדוק מהתחלת המילה קדימה עד שנמצא תא בעברית, וגם מסוף המילה אחורה עד שנמצא תא בעברית. ונסים לכל סימן פיסוק טוקן.

1. בכל פרוטוקול לכל דובר עשיתי את ה data frame שלו ואז איחדתי כל ה data frames.