תרגיל 1 – NLP יעקב חיימוף 318528510

הסקריפט מקבל simlex ו corpus ואת 20k המילים הכי נפוצות ובונה 4

טבלאות.

2 מטריצות של frequency count ו2 מטריצות של ppmi.

השורות בטבלה מורכבות מה- simlex והעמודות מה-20k המילים הכי נפוצות.

אחרי שנחשב את הטבלאות נחשב וקטור של similarity בעזרת הcosine- לכל זוג מילים מה- simlex.

את התוצאות האלה ניתן לראות ב-4 קבצים הנוספים שהגשתי.

לאחר מכן חישבנו את הדמיון בין הקבצים לבין קובץ ה- simlex עצמו בעזרת קורלציה.

Correlation of window 2: 0.07028212885553342

Correlation of window 5: 0.07185720368905786

Correlation of ppmi 2: 0.10090702257764907

Correlation of ppmi 5: 0.11126133672239152

התוצאה הטובה ביתור קיבלנו בחלון 2.

Text, application, chat or text message

Description automatically generated