מי אתה ? יצירת חתימת טקסט יחודית לכותב ברשתות חברתיות

חגית בן שושן

מנחה: ד"ר אסנת (אוסי) מוקרין

משתמשיהן של רשתות חברתיות מתקשרים ומביעים את דעותיהם באמצעות טקסט. על מנת לזקק סגנון ייחודי של כותב אנו מציעות להתייחס אל רשימת המילים הממושקלת של אותו כותב כאל רשימה ממוינת ולהשוות אותה אל רשימת המילים הממושקלת בה השתמשו כל יתר הכותבים באותה רשת חברתית אשר עוסקת באותו תחום. בבואנו לחפש שיטת מדידה המתאימה לדרישתנו סרקנו לא מעט אלגוריתמים בתחום של מרחק בין רשימות ממוינות, אולם לא מצאנו שיטה הולמת אשר תאפשר לנו למדוד מרחק של משתמש יחיד מן הכלל.
השיטה חייבת לקחת בחשבון משקל שונה לכל אלמנט ברשימה ויכולת להשוות בין שתי רשימות אשר אינן זהות באורכן, מכאן שקיימת חפיפה חלקית בין האלמנטים המושווים בין שתי הרשימות. כמו כן חשוב לנו לשקלל את משקלו של פריט חסר ברשימה, לתת משקל גבוה יותר לאי התאמות הנמצאות בראש הרשימה לעומת תחתיתה.
השיטות המקובלות בתחום מסוגלות לבצע חישוב מרחקים בין תוצאות של מנועי חיפוש, או לבצע השוואה בין רשימות זהות באורכן או רק עבור N פריטים הנמצאים בראשן .
לצורך עבודה זו פיתחנו שיטה חדשה המאפשרת לחשב מרחק של כותב (קטגוריה) מן הרשימה המשותפת של כל הכותבים (DVR) על סמך המילים מהן השתמש הכותב (אלמנטים) ומשקלן היחסי בכל טקסט probabilitie).
לצורך מימוש השיטה אנו משתמשים בכלים העומדים לרשותנו בעולם ה Big Data, המאפשרים להריץ אלגוריתם שאינו חסכוני על סט נתונים גדול. באמצעות השימוש בטכנולוגיות אלו אנו מבטלים את האילוץ לבצע חישובי הסתברות שאינם מדויקים במלואם, ואינם לוקחים בחשבון את משקלם השלילי של אלמנטים חסרים.
עקב כלליותה של השיטה המוצעת, מצאנו כי היא מתאימה לעולמות רבים שעיסוקם לאו דווקא בעיבוד טקסט. אנו נסביר ונדגים את השימוש בשיטה בעולמות תוכן שונים.