

ניתוח טקסט בטכנולוגיה סמנטית

אוהד אורן

טיאסג'י מערכות מתקדמות בע"מ (TSG Advanced Systems LTD.)

תוכן

3	רקע.....
3	על המחבר.....
5	הטכנולוגיה הסמנטית.....
5	הרשת הסמנטית.....
7	ייצוג תוכן.....
8	השוואת טכנולוגיות.....
8	חיפוש.....
8	קטגוריזציה.....
9	כריית טקסט.....
10	טבלת השוואה.....
10	אוראנוס – מערכת לניתוח עברית בטכנולוגיה סמנטית.....
11	מרכיבי אוראנוס.....
11	יכולות אוראנוס.....
11	חילוץ ישויות עניין מטקסט וקביעת רלבנטיות.....
11	חילוץ קשרים ואירועים מהטקסט.....
11	קטגוריזציה.....
12	חיפוש רב לשוני.....
12	סנטימנט.....
12	ניהול רשת סמנטית – אונטולוגיה.....
12	סיכום.....

באמצעים הטכנולוגיים הזמינים כיום, איסוף מידע ממקורות מגוונים בכמויות עצומות היא עובדה קיימת, הדרישה והאמצעים לאיסוף עוד ועוד מידע הולכים ומתפתחים כל העת, ומדגישים את הצורך ההולך וגבר להתמודד עם המידע ולהפיק ממנו תוצרים שישרתו את הארגון בבצוע משימותיו.

הקושי להפיק תובנות מכמויות חומר עצומות מתעצם עוד יותר כאשר רוב החומר הנאסף הוא טקסט שאינו מובנה, ועובדה זו מחדדת את הצורך להתמודד עם האתגר שלא רק באמצעים מסורתיים – זה לבד כבר לא מספיק.

טכנולוגיות אלה, בין שיהיו מבוססות מילות מפתח, או מבוססות זיהוי תבנית או מבוססות מודלים סטטיסטיים, נבחרות לעיתים קרובות בגלל הפשטות שלהן, ובהרבה מקרים זה אכן מספק. כאשר נדרש לטפל באופן מעמיק בניתוח טקסט ולייצר ממנו תשתית חזקה ליישום מערכת ניהול ידע וקבלת החלטות.

קבלת ההחלטות הסתמכו פעם על מודלים של מידע המבוסס על ERPs, או CRM's או מחקרי שוק יקרים. שווי פעילות כזאת היה לעזור להבין מה עלול לקרות בעתיד על סמך מה שכבר קרה. מודלים אלה הפכו עם הזמן לפחות יעילים, לא רק משום שהם מבוססים על נתונים היסטוריים, אלא מכיוון שהם גם לא הכלילו בתוכם את הנתונים הלא מובנים, ממקורות פנימיים, כגון תקשורת פנימית, ניהול מסמכים, הודעות דוא"ל, קבצי טקסט, מסמכי Word, מצגות, או מקורות חיצוניים, כגון זרמי מידע, דוחות ואמצעי מדיה חברתית ופתוחה שיכולים להפוך מידע תיאורי למידע תומך תחזיות.

מקורות מידע בלתי מובנים מהווים חלק קריטי במשוואה. הקושי של מינוף הידע הפנימי הקבור בפלטפורמות ייעודיות ומובנות, וההכרה בחשיבות האסטרטגית של זרמי מידע ופלטפורמות שיתוף מידע פתוח אחרות (כמו מדיה חברתית) הוא הזרז המוביל לשימוש בטכנולוגיות שיאפשרו טיפול בטקסט תוך הבנה טובה יותר של מושגים ויחסים בין מושגים, תוך שימת דגש על הבנת משמעות המילים והפרדת משמעות בהתאם להקשר. זהו בדיוק המקום שבו טכנולוגיה סמנטית בולטת וזורחת. בהמשך המאמר תגלו את הפוטנציאל של טכנולוגיה סמנטית לפתרון אחד האתגרים הגדולים של ימינו ביצירת תובנות ממידע.

על המחבר

חברת טיאסגי מערכות מתקדמות בע"מ (TSG Advanced Systems LTD) הינה כיום חברה עצמאית המוחזקת בבעלות משותפת ע"י התעשייה האווירית וקבוצת פורמולה. החברה הייתה ומוכרת בעבר כחטיבה הביטחונית של חברת נס טכנולוגיות בע"מ. החברה פועלת בעיקר בשוק הביטחוני בארץ ובחו"ל.

לחברה חטיבה עסקית המפתחת פתרונות מתקדמים ומערכות מודיעין ומיצוי מידע. חטיבה זו מנוסה בתחום של פתרונות מבוססי ניתוח טקסטואלי ומידע מכל הסוגים (structured + unstructured). לחברה 4 דורות של פתרונות ברמה הלאומית, שבליבתם שירותים מתקדמים לניתוח טקסט. בכל הפתרונות משולבים מוצרי state of the art כגון מוצרי מלינגו, בשיתוף פעולה הדוק רב שנים עם החברה.

החברה פועלת בתחום זה כבר 25 שנים מול לקוחות מתחום הביטחון בארץ כגון משרד הביטחון, אמ"ן, משרד ראש הממשלה וסוכנויות המודיעין השונות.

בניתוח מעמיק, שביצעה החברה לפני כחמש שנים לתחום הניתוח הטקסטואלי, נבדקו מנועי האחזור המובילים בעולם וכן הטכנולוגיות החדשניות בעולם בתחום זה. למול דרישת הלקוחות, זוהה הצורך לא רק למצוא תכנים על פי מלים (אחזור מבוסס מלינגו) אלא לשפר את יכולת החיפוש על פי הבנת התוכן. לקוחות ביקשו להבין "מה קשור לנושא X?", "מה שייכתי לנושא Y?", "היכן הופיעה התייחסות גם לאדם X וגם לארגון Y בהקשר לנושא Z?"

לפיכך החל משנת 2010, ועל סמך ההבנה שלא היה בנמצא פתרון שכזה, החליטה החברה להשקיע בפיתוח של פתרון איכותי חדשני, שלא רק מאחזר אלא גם מבין (הבנה = סמנטיקה) טקסט. או אז נולד **פרויקט אוראנוס** (ראה בהמשך המאמר, אוראנוס – מערכת לניתוח עברית בטכנולוגיה סמנטית).

פרויקט אוראנוס, שפותח בהשקעה משותפת עם מפא"ת הופך למבצעי בימים אלו אצל חלק מהלקוחות הביטחוניים שהוזכרו מעלה. כחלק מההשקעה המהותית של החברה הוקם מרכז מצוינות ייחודי, הכולל בלשנים, מנתחי מערכות ומהנדסי תוכנה, המתמחים

בניתוח שפה ובפיתוח תוכנה ייעודית לניתוחי שפה. מתחילת דרכו הציב פרויקט אוראנוס כמטרה את השימוש בטכנולוגיה הסמנטית לניתוח טקסט בעברית, ולשם כך נבחרה החברה האיטלקית Expert Systems (www.expertsystem.net) שפיתחה את המנוע הסמנטי Cogito כשותפה.

חברת Expert Systems שמשכנה במודנה-איטליה, היא חברה לפיתוח טכנולוגיה סמנטית, המתמחה בפיתוח מוצרים סמנטיים המסייעים לארגונים בהפקת תובנות ומודיעין מתוך המידע שברשותם. המנוע הסמנטי Cogito נמצא בליבה של כל אחד מהפתרונות אותם מפתחת החברה. לחברת ES ניסיון מוכח של כ- 15 שנה, והיא ידועה כבעלת האיכות הגבוהה ביותר להבנת תוכן במגוון שפות מכלל המנועים הקיימים בשוק.

פתרון אוראנוס שפותח ע"י TSG נהנה מעושר היכולות הסמנטיות של המנוע האיטלקי יחד עם חבילות ייחודיות למורפולוגיה וניתוח תחבירי והינה הפתרון הראשון והיחיד הידוע כיום לניתוח השפה העברית תוך שימוש בטכנולוגיה סמנטית.

הטכנולוגיה הסמנטית

מאפיין משותף של כל הטכנולוגיות המשמשות באופן מסורתי בניתוח אוטומטי של טקסט, הוא ההבנה באשר לגבול היכולות שלהם. כולם בוחרים ליצור מעקפים לבעיה כדי להיות קרובים ככל האפשר לדרך היעילה ביותר שלהם לנתח טקסט.

את הטכנולוגיה הסמנטית ניתן לסווג כתוכנה להבנה וניתוח אוטומטי של טקסט. בניגוד לטכנולוגיות אחרות, טכנולוגיה זו אינה מעבדת תוכן כרצף של תווים ואינה מנחשת את המשמעות של המילים והמושגים. תוכנה סמנטית אמיתית היא כזו המסתמכת על ניתוח סמנטי עמוק ורשת סמנטית עשירה כדי להבטיח הבנה מלאה של טקסט כפי שהאדם הקורא היה מבין אותו.

הרשת הסמנטית

היכולת לניתוח סמנטי טמונה בראש ובראשונה בקיימותה של רשת סמנטית. הרשת הסמנטית היא הפלטפורמה ליצירת אונטולוגיה לייצוג הידע של משמעות המילים (מובנים), והיחסים בין מובנים.

בתוך הרשת הסמנטית מנוהלים מילוני מובנים (Syncons) ביחד עם כל הנרדפות המייצגות את אותו Concept. כל סינקון הוא צומת ברשת הסמנטית, המקושר לסינקונים אחרים באמצעות זיקה (קשר) סמנטית במבנה היררכי. בתורה זו, כל סינקון, בנוסף לייצוג שלו עם משמעות ומאפיינים, מועשר במאפיינים של סינקונים שנמצאים "מעליו" או "מתחתיו". כל סוג של קשר ברשת הסמנטית מגדיר יחס של קשר בין מובנים בשפה, והקשרים הם מרכיבים המשמשים לארגון המובנים ברשת הסמנטית. למשל הסינקונים מאורגנים באופן היררכי מהפחות ספציפיים ליותר ספציפיים (כלי תחבורה->תחבורה ציבורית->מונית) או ע"פ הייחוס התחבירי וכדומה.

סינקונים יכולים לכלול "למות" (Lemmas), מילים מורכבות ו-Collocations. המרכיבים הראשיים של כל סינקון הם המרכיב

הדקדוקי, הקשר הסמנטי, המשמעות, הדומיין (domain), והתדירות, בסינקונים מסוימים גם הסנטימנט.

ברשת הסמנטית, המשמעות האמתית של כל סינקון, הוא השילוב של מרכיביו, והקשר שלו לסינקונים אחרים ברשת. ברשת הסמנטית ניתן לנהל זיקות מסוגים שונים, בהם הזיקות:

1. סוג של : טויטה -> מכונת
2. חלק מ : מסך -> מטלפון
3. Subject-Verb : הולנד ביקר בהולנד
4. Object-Verb : החייל בלע כדור
5. זיקה אסוציאטיבית : גנב-שוטר

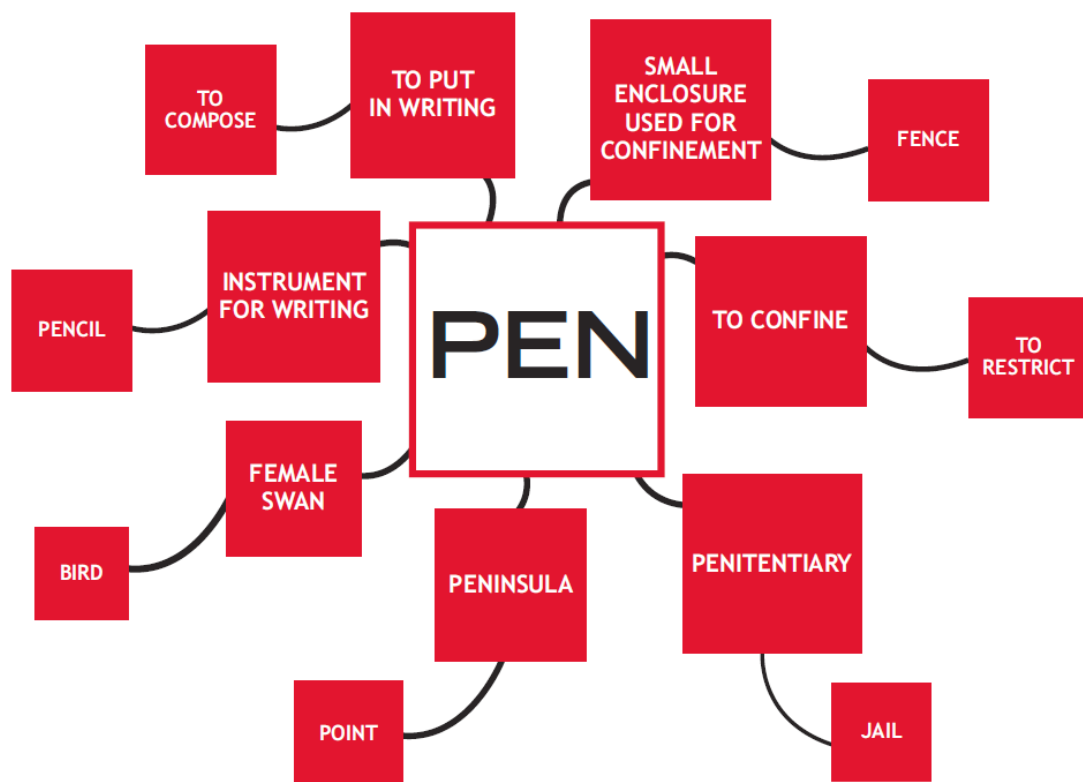
פעולת הפרדת המשמעות (Disambiguation) במנוע הסמנטי מבוצעת על משפט בודד או מסמך שלם כדי להבחין במשמעות המדויקת של כל מילה, מבין שלל המשמעויות המקושרות למושג. באופן זה, המנוע הסמנטי "מבין" את המשמעות של המילים בקונטקסט, ומשקף את מה שאדם היה עושה בקריאת הטקסט.



Figure 1 - Different meaning for "Bomb"

תוצר הניתוח הסמנטי הוא מפה קוגניטיבית-קונספטואלית של טקסט או במילים פשוטות יותר-תוצר מובנה לטקסט שאינו מובנה.

להלן דוגמא לייצוג מובנים ברשת הסמנטית באנגלית:



כאמור הלב של מנוע סמנטי הוא הרשת הסמנטית, והוא מורכב משילוב של ערכי מילונים ומידע, המייצרים יחד השמת ערך מובן מתאים לכל מילה מנותחת:

א. מילון ערכים:

ניהול מידע של כל אפשרויות המשמעות למילים ובקונטקסט המתאים הוא חיוני לעיבוד טקסט בדיוק רב. ראו להלן את כל אפשרויות המשמעות למילה driver כדי להמחיש עד כמה משמעות/מובן שונה למילה יכול להקשות על הבנת התוכן:

1. The driver of the red car was injured in the accident
2. I used the long driver
3. The driver was installed in the computer

הרשת הסמנטית חייבת לכלול את כל המובנים השונים למילה driver והזיקות שלהן למובנים אחרים ברשת על מנת להצליח בהפרדת המשמעות המתאימה למילה driver בכל משפט.

להלן דוגמאות להפרדת משמעות בשפה העברית:

1. יוסי שותף בקשר לביצוע פשע.
2. יוסי הקשר הודיע למפקד המחלקה על האירוע שדווח בקשר.

כאמור גם בדוגמא זו נדרשת הרשת הסמנטית בעברית לייצג את המובנים השונים למילה "קשר" ע"י מאפיינים וזיקות מתאימות למובנים אחרים ברשת.

ב. מאגר ידע:

מאגר ידע הוא אלמנט מרכזי להבנת הנקרא. חוסר בידע על נושא או תחום מסוימים, עלולים מהאדם הקורא שלא להבין את הטקסט בצורה שלמה.

מאגר הידע ברשת הסמנטית מאורגן ומיושם באופן הדומה לצורה שבה האדם רוכש ידע כללי חדש. כמו שהידע האישי שלנו כבני אדם מתעשר ומשתפר בלומדנו דברים חדשים, כך גם הרשת הסמנטית מתעשרת ומשתפרת ע"י שימוש במנגנונים המרחיבים את הידע הגלום בה ע"פ אונטולוגיה המתאימה לטבע הארגון. כך צפוי שמאגר הידע המיוצג ברשת הסמנטית שמנוהלת בארגון צבאי לא תהיה זהה למאגר הידע המיוצג ברשת הסמנטית של ארגון בעולם הפארמה, או הבנקאות, או הרכב וכדומה.

ג. זיכרון:

בנוסף לשימוש במאפיינים של כל מובן ורשת הזיקות של מובן ברשת, משתמש המנוע הסמנטי גם בטכניקה של זכירת הקונטקסט של הטקסט על מנת לקבוע טוב יותר את משמעות המילים.

לדוגמא אם הטקסט הוא בנושא רפואה ומופיעה המילה כדור, המנוע זוכר את ההקשר של הטקסט וייתן סבירות גבוהה יותר לבחירת הסינקון של כדור=גלולה מאשר למובנים אחרים של כדור.

ייצוג תוכן

התוצר של ניתוח סמנטי הוא מפת קונספטים המופיעים בטקסט, כאשר :

- כל קונספט המצוי בטקסט מזוהה באופן ייחודי ללא קשר למילים בהם השתמשו בטקסט כדי לבטא אותו
- כל גורם בטקסט מיוחס לפעולה
- כל אובייקט בטקסט מקושר לפעולה קשורה

באופן זה ייצוג המידע במסמך הופך למובנה ויכול להכיל את נושאי המסמך, אנשים, מקומות, ארגונים, מספרים, וכל מידע אחר שהוא בר משמעות יכול לקבל ייצוג.

מפת הקונספטים מספקת מבנה שמאפשר לשמש בכל תהליך עיבוד כגון קיטלוג, סיכום, תרגום, אינדוקס וכדומה.

השוואת טכנולוגיות

כדי להבין את ההבדל בין טכנולוגיה סמנטית לטכנולוגיות אחרות, חשוב לבצע ביניהם השוואה במושגי יעילות מזווית ראייה עסקית.

היעילות בשיטות לאחזור ואיתור מידע או במילים אחרות – הרלבנטיות, נמדדת באמצעות מדדי כיסוי ודיוק (Recall, Precision).

כיסוי – היכולת למצוא כמה שיותר אינפורמציה רלבנטית בטקסט.

דיוק – היכולת לאחזר רק את האינפורמציה הרלבנטית ללא "רעש".

בחלק הבא נסביר על היתרונות המשמעותיים שהטכנולוגיה הסמנטית מציעה לטובת המשימות החשובות של אנליזה ואחזור מידע.

חיפוש

החיפוש הוא יכולת התוכנה לאפשר אחזור מסמכים ממאגר (חיצוני באינטרנט, או פנימי בארגון). בהשוואה לשיטות אחרות,

הטכנולוגיה הסמנטית מציעה כמה יכולות ייחודיות לחיפוש:

א. חיפוש קונספטואלי:

רשת סמנטית עשירה מחזירה לא רק את המסמכים שקשורים לשאלתה של המשתמש, אלא גם מסמכים שקשורים קונספטואלית לשאלתה המשתמש. יכולת זו המתאפשרת בעזרת שימוש בסמנטיקה מאפשרת יכולת זו ללא יצירת "רעש" ושגיאות מעצם יכולתה להבחין ולהפריד את המשמעות של המילים. לפיכך התוצאות מדויקות יותר ומספקות מענה לשאלתה המתייחס להבנת המשמעות. למשל כאשר המשתמש יבחר לאחזר מסמכים בהם הופיע מכונית, המנוע הסמנטי יחזיר לא רק את המילה מכונית (על הטיותיה), אלא גם את הקונספט מכונית, כלומר ניתן יהיה לקבל בתשובה גם מסמכים המכילים בתוכם "טנדר", "משאית", "טויוטה" וכדומה, וכל זאת ללא טיפול ידני באינדקס או בשאלתה.

ד. חיפוש שפה טבעית:

המפה הסמנטית והשלשות התחביריות נושא-פעולה-נשוא הנוצרית בתהליך העיבוד התחבירי/סמנטי של הטקסט מאפשרות הבנה של כוונת השאלה של המשתמש, כאשר זו נכתבת בדרך בה אנשים מתקשרים, כולל בשימוש בסלנג, או כתיבה מקוצרת, או אפילו שגיאות כתיב.

במקום שבו יישום מסורתי היה מעבד כל מילה בנפרד, ומתעלם מהשאלה המתחבאת במשפט, המנוע הסמנטי יבין את השאלה אליה התכוון המשתמש, ויחזיר את סט המסמכים העונים לה, למשל בדוגמא: "Office hours over the wkend Brooklyn".

חיפוש תוך שימוש בשפה טבעית הוא חשוב במיוחד כאשר הוא מיושם ב-FAQ, או שירותי Wiki, או כל שירות שמאפשר למשתמש לשאול שאלות בצורה ישירה דרך ערוצי כתיבת טקסט חופשי (SMS, צ'אט, Facebook וכדומה)

ה. קריטריון מדרוג מותאם ועשיר:

העושר של התשובות תוך שימוש במנוע הסמנטי, מאפשר התאמות חכמות לפלטפורמת החיפוש, כמו גם יישום של מנגנוני מדרוג (Ranking) ע"י שימוש בקריטריונים גמישים. למשל, משתמשים של מחלקת ה-R&D יכולים לבקש את התוצאות על חיפוש מתחרים, מדורגים ע"פ האופן בו המסמך מסווג ככזה שעוסק בחדשנות טכנולוגית, בעוד שמשתמשים ממחלקת הכספים יבקשו את אותן התוצאות מדרוג ע"פ סוג התוצאות שעוסקות במיזוגים ורכישות. כל המידע זמין בשלב הניתוח.

קטגוריזציה

היצף המידע הבלתי מובנה הנאסף בארגונים, מעורר באופן טבעי את הצורך לארגנו באופן קל. השיטה הנפוצה לארגון מסמכים היא ע"י סיווגם או קיטלוגם. סיווג אוטומטי אם כן, הוא צורך בסיסי הנדרש מתהליך שמטפל בזרם מסמכים נכנס.

סיווג מסמכים לתוך טקסונומיה בנויה מראש היא שיטה מוכרת ונפוצה לניהול המסמכים בצורה יעילה.

פיתוח של מערכות סיווג אוטומטי למסמכים הפך בשנים האחרונות למדע של ממש. כמו ביישומים טיפוסיים או טכנולוגיות העוסקות בעיבוד שפה טבעית (NLP), גם לקטגוריזציה יש כמה גישות פעולה.

א. ידנית

ידוע גם כ-"שק של מילות מפתח". שיטה זו מצריכה רישום של מושגי יסוד המתארים את סוג התוכן. בהינתן רשימה מלאה, מייצרים באמצעותה מטאדטה. ע"פ שיטה זו, כאשר מסמך שויך לקטגוריה מסוימת, הוא בוודאי יכול את מילות המפתח (או חלקן) שמופו עבור קטגוריה זו.

ב. סטטיסטית

מערכות סטטיסטיות משתמשות בקורפוס אימון כדי לאפיין את הקטגוריה הנדרשת. ומפעילות אלגוריתמים שונים כדי לחלץ סט של "אירועי מפתח" במסמך על מנת לבנות חוקה משתמעת (implicit) שתשמש לקיטלוג המסמך.

ג. סמנטית

הגישה לקטלוג סמנטי מושתת בעיקרה על בניית חוקים. החוקה נכתבת כדי לתפוס אלמנטים במסמך שמשייכים אותו לקטגוריה. החוקה יכולה להיכתב באופן ידני ע"י מומחה תוכן, או להיווצר אוטומטית בתהליך הניתוח ואז לעבור בקרה ועידון ע"י מומחה תוכן. החוקה היא דטרמיניסטית – אתה יודע מדוע המסמך סווג לקטגוריה, גמישה מאוד, וחזקה בקטלוג אפקטיבי.

הגישה הסמנטית מייצרת יתרון מהותי וחשוב בקטלוג – הביצועים אינם מושפעים מגודל עץ הטקסונומיה מכיוון שהידע המיוצג ברשת הסמנטית מאפשר הבנה טובה יותר של החוקים בקונטקסט ויכולת רבה לעדכנם. זה אומר שכאשר המערכת נפרסת, המסמכים שאינם מתאימים לקטגוריה מסוימת מזוהים ומופרדים באופן אוטומטי, ובכך מאפשרים ליוצר החוקה לנתח מדוע המסמך לא סווג לאף קטגוריה. באופן זה ניתן לקבל החלטות מושכלות באשר לטיוב החוקה הקיימת, או ליצור קטגוריות נוספות עבור מסמכי תוכן שקודם לכן לא סווגו.

יתרונות הגישה הסמנטית בקיטלוג:

- עלות הנמוכה בעד חצי מהזמן הנדרש בשיטות האחרות.
- היכולת לנצל את הידע ברשת הסמנטית כדי לכתוב חוקה יעילה וללא שימוש באלמנטים מרובים כדי להגדירה.
- סמנטיקה מאפשרת לטפל ברב-משמעויות שיש בשפה, ובכך מקלה על תהליך הזיהוי והארגון של רשימת הערכים הנדרשים בחוקה.
- הגישה הסמנטית לא מצריכה הכנת סטים של אימון (שלרוב לא בהכרח קיימים בארגון, או קשה מאוד לייצרם)
- אין זמן שמושקע באימון עצמו, ובהינתן עץ טקסונומיה מורכב, זו משימה בהחלט לא מבוטלת שנחסכת בגישה הסמנטית.

כריית טקסט

כריית טקסט הוא תהליך להפקת אינפורמציה מהטקסט, נירמול שלה לפורמט עבודה נדרש, ואחסונה בארכיון לטובת שימושי יישום עתידיים.

כריית אינפורמציה מטקסט בטכנולוגיה סמנטית מאפשרת את היכולות הבאות:

א. כיסוי ודיוק גבוהים בחילוף ישויות

טכנולוגיה סמנטית מבטיחה כיסוי ודיוק גבוהים יותר הודות לעשרות אלפי הקונספטים המנוהלים ברשת הסמנטית. השימוש ברשת הסמנטית מאפשר כרייה של מידע קונטקסטואלי נוסף שמאשר את הזיהוי של המידע המחולץ מהטקסט. בנוסף, הביצועים לחילוף אינם תלויים בכך שנכיר או לא נכיר את הישויות לפני ביצוע התהליך, וגם לא נדרשת כתיבה של קוד או הוספת מושגים נוספים ללקסיקון.

ב. תאימות לסטנדרטים של W3C

ישויות העניין המחולצות במהלך ניתוח הטקסט יכולות להיות מקושרות באופן אוטומטי לישויות אחרות בקשרים מסוגים רבים ואז לשמש במאגר שלשות RDF בכל תהליך שמייצר תובנות במאגר סמנטי

ג. תיוג מדויק יותר

האלמנטים המחולצים בניתוח הטקסט, מקבלים ניקוד ע"פ הרלבנטיות הסמנטית שלהם בטקסט, וע"פ התפקיד שהם ממלאים בטקסט (נושא, פעולה וכדומה). פרקטיקה זו, בשילוב עם אלמנטים אחרים של הבנת הטקסט מבטיחים תיוג אוטומטי אפקטיבי ונכון יותר לתוכן המנותח.

ד. יישום קל לאונטולוגיות שונות

הרשת הסמנטית מייצגת את הגרעין המלא של כל הקונספטים והמובנים הקיימים בשפה. הגרעין הזה מאפשר השלמה או עדכון לכל רשת סמנטית הנפרסת אצל לקוח בעל עולם תוכן/אונטולוגיה הייחודית לו. כאשר מערכת ניתוח סמנטית תשמש ארגון בתחום האנרגיה, או לחלופין בתחום הפארמה, נוכל בשני המקרים להשתמש בגרעין של הרשת הסמנטית, וע"י עדכון של כ-20% בתוכן להתאים אונטולוגיה ע"פ אופי הארגון. התאמת אונטולוגיה באופן זה היא קצרה ואפקטיבית יותר מיצירת פתרון מ-scratch בטכנולוגיות המתחרות.

טבלת השוואה

להלן טבלה שמסכמת בנקודות מהותיות את ההבדלים בניתוח טקסט בטכנולוגיות השונות.

שיטה	מאפיינים	יכולות	חסרונות
סטטיסטית	יכולת לאמן קורפוס של מסמכים כדי ללמד את המנוע כיצד לחלץ ומה לחלץ	חילוץ ישויות	- אין יכולת לחילוץ קשרים - היכולת בעברית מוגבלת - ירידה משמעותית באיכות בשינוי הטקסט
בלשנית / מבוססת חוקים	- ניתוח ישיר של כל מלה. - נדרשת הגדרה מדויקת לכל מלה ובכל צרוף	- חילוץ ישויות - קשרים בין ישויות	נדרשת השקעה עצומה בחיבור חוקים ובתחזוקתם.
ניתוח ביטויים ומילונים	- Regular Expressions - על בסיס רשימות קיימות	חילוץ ישויות	- אין יכולת לחילוץ קשרים - היכולת בעברית מוגבלת
סמנטית	מבוסס על ניתוח עמוק המבין את השפה, את המלים, את משמעויותיהן, את הקשרים ביניהן ואת הקונטקסט שבו הן נאמרות	- חילוץ ישויות - קשרים בין ישויות - קטלוג - תמצות - חישוב סנטימנט	- מחיר התחלתי גבוה ליצירת התשתית (הרשת הסמנטית) - טקסטים שאינם נכונים תחברית "יאתגרו" את המנוע הסמנטי

אוראנוס – מערכת לניתוח עברית בטכנולוגיה סמנטית

אוראנוס הוא פתרון סמנטי שביכולתו להבין טקסטים בשפה העברית ולחלץ מתוכם מרכיבים בעלי משמעות ערכית למשתמש. אוראנוס מאפשר לחלץ את הישויות (מכל סוג) המופיעות בטקסט ואת הקשרים ביניהן כמו גם לקטלג את הטקסט ולקבוע על מה הוא מדבר. אוראנוס מאפשר לקבוע רלוונטיות של ישות למסמך, למצוא את משפטי המפתח ואף לזהות את הסנטימנט במסמך. אוראנוס הינה הטכנולוגיה הסמנטית היחידה בשפה העברית.

אוראנוס הוא מוצר ליבה לניתוח והבניית טקסט בעברית שניתן לשלב במערכות קיימות ולשמש יישומים מגוונים כגון:

- יישומים המטפלים בהבניית מידע מודיעיני המגיע כטקסט חופשי (ישויות, קשרים, מאפיינים וכדומה)
- יישומים המטפלים בחומרי ארכיון (יצירת עצי טקסונומיה ארגוניים, סיווג וקטלוג אוטומטי, תמציות)
- יישומים המטפלים בניהול ואחזור מידע ארגוני (הרחבה סמנטית למנועי חיפוש ליצירת אחזור חכם)
- יישומים המטפלים בניתוח מידע טקסטואלי ממקורות גלויים (OSINT), ובכללן רשתות חברתיות, כתבות, דו"חות ועוד
- היישומים העוסקים בניתוח ועיבוד מידע בעולמות תוכן שונים ומגוונים כמו תקשורת, כלכלה, רפואה, כלכלה משפט ועוד.

למעשה ניתן להתאים את המנוע של אוראנוס לעבודה מול כל עולם תוכן ובכל סגנון.

מרכיבי אוראנוס

אוראנוס בנוי מ-5 מרכיבים מרכזיים:

1. **מנתח מורפולוגי** - רכיב Analyzer של מלינגו המשולב בפתרון. אחראי לטוקניזציה (פירוק הטקסט למשפטים ולמילים), ולתת לכל מילה את המאפיינים המורפולוגיים שלה (חלק דיבר, צורת יסוד, זמן, מין וכו') וכן את ההסתברות של כל אפשרות.
 2. **מנתח תחבירי** – אחראי לניתוח תחבירי ברמת המשפט, מבצע חלוקה לפסוקיות וקובע לכל מילה את תפקידה במשפט – נושא, נושא, מושא וכו' וכן את הקשר ביניהם (SVO).
 3. **מנתח סמנטי** – מנוע (Cogito) האחראי לבחור לכל מילה את המשמעות שלה מתוך הרשת הסמנטית. מבוסס מודלים היוריסטיים שמסתמכים על הקשרים השונים ברשת הסמנטית ועל ההקשר הסמנטי של המילה/משפט/טקסט.
 4. **מנוע חוקים** – פלטפורמה (Cogito) לכתיבת חוקי חילוץ לישויות וקשרים בין ישויות.
 5. **רשת סמנטית בעברית*** – תשתית אונטולוגית המייצגת את מכלול המובנים והקשרים ביניהם בשפה.
- * הרשת העברית מכילה כ- 50,000 מובנים המורכבים מכ- 60,000 מילים ומעל 170,000 קשרים ביניהם.

יכולות אוראנוס

חילוץ ישויות עניין מטקסט וקביעת רלבנטיות

אוראנוס מחלץ באופן אוטומטי את הישויות: אנשים, ארגונים, מקומות, אמל"ח, תפקידים, זמנים, מספרי טלפון ועוד. מעבר לכך ניתן להגדיר חוקי חילוץ פשוטים יחסית שמבוססים על מובנים (סינקונים) ברשת הסמנטית ועל ידי כך לחלץ כל סוג ישות רצוי. בנוסף לכך אוראנוס יודע לקבוע את הרלבנטיות/בולטות של הישות בטקסט ע"פ אלגוריתמים שמתחשבים במיקום הישות בטקסט, תפקידה התחבירי במשפט, כמות מופעים ועוד.

חילוץ קשרים ואירועים מהטקסט

באופן דומה לחילוץ ישויות המערכת מאפשרת חילוץ של ישויות מורכבות יותר כמו אירועים, פגישות וקשרים בין ישויות. חילוץ קשרים מתוך הטקסט מבוסס על חוקים סמנטיים תוך התחשבות בתפקיד חלקי הדיבר בטקסט (ניתוח תחבירי). המערכת תדע לזהות פעלים בטקסט ולייצר קשר בין ישויות הנפעלות ע"י הפועל. למשל במשפט: "יוסי דיבר עם משה" יזוהו יוסי ומשה כישויות אדם, והפועל "דיבר" ישמש כעוגן לקשר בין הישויות. היות שהפתרון כולו מבוסס סמנטיקה, ניתן לבנות קשרים סמנטיים שיתארו אירועים תוך שימוש ברשת הסמנטית. למשל אפשר להחליט שרוצים לזהות בטקסט אירועי פגישה. החוק הסמנטי לזיהוי פגישה יסתמך ההגדרה הסמנטית למושג "פגישה" ולכן כתיבת החוק תהיה פשוטה ותאטר אירוע פגישה גם כאשר הטקסט יכתב באופנים שונים.

קטגוריזציה

אוראנוס מגיע עם יכולת מובנית לקטלוג אוטומטי של טקסט. יכולת זו משתמשת ברשת הסמנטית לסיווג הטקסט. אוראנוס מגיע גם עם יכולת לייצר עצי טקסונומיה מותאמי לקוח ולקטלג את מסמכי הקורפוס לתוך עצים מותאמים לאחר בניית חוקה אוטומטית מבוססת סמנטיקה.

בנוסף ליכולות האוטומטיות והיכולות לבניית עצי טקסונומיה, גם בנושא הקטלוג ניתן להשתמש בשפת החוקים כדי להגדיר קטגוריות מותאמות לקוח.

חיפוש רב לשוני

אוראנוס מציע כהרחבה גם שירות חיפוש רב לשוני. יכולת חיפוש מתקדמת המאפשרת שליפת טקסטים בכל שפה נתמכת על פי מילות חיפוש בשפה העברית או בכל שפה נתמכת אחרת. יכולת זו מתבססת על העיקרון של זהות סמנטית בין השפות שמתבטאת בקיומן של רשתות סמנטיות מקבילות הקשורות זו לזו. לדוגמא: ניתן לחפש את המילה "יהלום" ולקבל את כל הטקסטים בעברית עם המילה "יהלום" את כל הטקסטים באנגלית עם המילה "diamond" ואת כל הטקסטים בערבית עם המילה "الماس".

סנטימנט

אוראנוס מאפשר מציאת סנטימנט בטקסט. על ידי הניתוח וההבנה של המילים בטקסט המערכת מזהה את המילים עם הנימה החיובית ואלו עם הנימה השלילית וקובעת את מדד הסנטימנט הכללי הן עבור המסמך כולו והן עבור הישויות המופיעות בו. יכולת זו נמצאת בשלבי פיתוח.

ניהול רשת סמנטית – אונטולוגיה

אוראנוס כאמור, הוא פתרון סמנטי לניתוח טקסט בעברית. הפתרון הסמנטי נשען על קיימותה של רשת סמנטית בעברית שמייצגת את עולם התוכן של השפה.

כאשר נפרס פתרון אוראנוס אצל לקוח, יש לבצע התאמה אונטולוגית לרשת הסמנטית, כך שתייצג באופן מדויק את עולם התוכן של הטקסטים המנותחים אצל הלקוח.

על מנת לאפשר למידען של הארגון לנהל את הרשת הסמנטית הוא נעזר בממשק ייעודי לניהול הרשת אשר מסופק כחלק מהפתרון.

בנוסף לכך הוא יכול לצפות בקשרים השונים הקיימים ברשת הסמנטית – פעולת תחקור רשת, ובכך לראות את המובנים הקיימים ולנהל את עולם התוכן העסקי באמצעותה.

סיכום

היצף המידע ההולך וגובר עם התרבותם של אמצעי איסוף ומקורות לאיסוף, והצורך לעבד את המידע להפקת תובנות עסקיות הם חלק מהאתגרים העומדים בפני ארגונים. טכנולוגיות מסורתיות שהועדפו עכב קלות ההטמעה שלהם, לא בהכרח מסוגלות להתמודד בהצלחה עם אתגר ניתוח וניהול המידע הדרוש כיום.

הטכנולוגיה הסמנטית מציעה יתרון של הבנה קונטקסטואלית מעמיקה למידע, וזאת מבלי להתפשר על איכות במושגי כיסוי ודיוק למידע המעובד. הטכנולוגיה מבטיחה הבנה אוטומטית לתוכן באופן בולט לטובה מול טכנולוגיות מתחרות. לא רק שזו הטכנולוגיה שמספקת את הפתרון המיטבי למשימות חילוץ ישויות, קטלוג ואחזור מידע, היא גם מציעה גישה חדשה ומבטיחה ביותר לפיתוח של יישומים מתקדמים המשלבים מודלים חזויים לניתוח עסקי.