מנועי אחזור טקסט בעברית - רשימת ספקים / מוצרים
גרסת מאי 2015

עפר דרורי
offerd@gmail.com

מבוא
מערכות מידע בעבר טיפלו בעיקר בניהול רשומות בתוך בסיסי נתונים כאשר רוב המידע ברשומות היה מידע מפורמט בשדות נתונים בעלי אופי מוגדר מראש (הן בגודל שדות המידע והן בפורמט שלהם). מזה שנים רבות מערכות מידע נדרשות לטפל גם במידע שאיננו מפורמט כמו טקסטים, תמונות, קבצי קול ועוד. גם בהווה בו קיימים סוגי מדיה שונים מרכיב הטקסטים במערכות המידע הוא גדול ביותר. מכיוון שטיפול בטקסטים הוא משימה מחשבית מורכבת מתקיים בתחום הנוהג שמפתחי מערכות אינם מפתחים מנועי אחזור טקסט למערכות המידע שהם כותבים בדומה לכך שלא נהוג לפתח תכונות של עיבוד טקסט המתקיים במעבדי תמלילים. בנוסף מוצרי תשתית רבים בתחום מערכות המידע כוללים בתוכם מנוע חיפוש מובנה דבר שלעתים מייתר את הצורך במנוע חיפוש חיצוני.

"עברית שפה קשה" אמר המשורר ובכל הקשור לטיפול ממוחשב בשפה העברית על אחת כמה. נהוג לדרג את השפות בעולם על פי הקושי הנדרש בטיפול ממוחשב בהן. בתחתית הסולם נמצאת השפה הסינית שבה אין הטיות ואין רב משמעות למילים. אחרי הסינית מבחינת הסיבוכיות נמצאת האנגלית, אחריה צרפתית כאשר העברית והערבית נחשבות כשפות הקשות ביותר לטיפול ממוחשב מכיוון שהן מכילות הטיות רבות, מורפולוגיה מורכבת וריבוי משמעויות.
את התכונות הנדרשות בשפה העברית ניתן לחלק לטיפול בטקסט ולטיפול בממשק (במידה והוא מסופק עם המוצר).
טיפול בטקסט מתייחס לתכונות כמו: כיווניות שפה, מורפולוגיה (שהיא ייחודית לשפה), אחזור על פי שורש מילה (השונה מהותית בשפה העברית משפות לועזיות אחרות), צליל (סאונדקס), גדומים (אשר יש להם משמעות רבה יותר בשפה האנגלית מאשר בשפה העברית), טיפול בתזאורוס המותאם לשפה ועוד.
טיפול בממשק מתייחס לשפת התפריטים, כיווניות השפה המוקלדת בעת ביצוע שאילתת החיפוש, להצגת המידע, לעזרה המקוונת ועוד. כאמור יש להתייחס למרכיב זה כאשר המוצר כולל ממשק.

מטרת מסמך זה להציג את רשימת הספקים והמוצרים הקיימים בתחום, התומכים בשפה העברית וניתנים להשגה בארץ. מסמך זה נילווה למסמך "**קריטריונים לבחירת מנוע אחזור טקסט - גרסה 5**" ואשר יכול לסייע בתהליך בחירת מנוע מסחרי מסוים מתוך רשימה של מספר מנועים. המסמך עצמו נמצא כאן [1]

בשנים האחרונות חלו תמורות בשוק מנועי החיפוש. מנועים חיפוש רבים נבלעו ע"י מנועים אחרים. חברות גדולות רכשו חברות קטנות יותר אם לשם קבלת המוצר בהיותו חסר בסל המוצרים שלהן ואם לשם "הריגת" המוצר המתחרה. סופו של התהליך מראה כי היקף המוצרים בתחום הצטמצם משמעותית, אולי למספר הקטן ביותר של מוצרים בשוק הישראלי מזה שנים.

לצד צמצום החברות והמוצרים בתחום חל גידול בשוק ה Open source בו מוצרים טובים ואפילו מוצרים טובים מאוד עומדים זמינים לכל אחד ומספקים את כל התכונות שהמוצרים המקצועיים מספקים. אם בתחילה היה חשש משימוש במוצרים חופשיים שאין להם "אבא" עם כתובת בארץ והתמיכה העברית בעייתית, המצב היום שונה. לשני המוצרים המובילים בתחום יש תמיכה עברית, הן מסורתית של חברות

[1] http://www.sigtrs.org/a342803-%D7%92%D7%9C%D7%95%D7%A4%D7%94-%D7%9C%D7%9E%D7%A1%D7%9E%D7%9A-%D7%A7%D7%A8%D7%99%D7%98%D7%A8%D7%99%D7%95%D7%A0%D7%99%D7%9D-%D7%9C%D7%94%D7%A9%D7%95%D7%95%D7%90%D7%AA-%D7%9E%D7%A0%D7%95%D7%A2%D7%99-%D7%97%D7%99%D7%A4%D7%95%D7%A9-%D7%92%D7%A8%D7%A1%D7%94-5-%D7%9E%D7%90%D7%99-2009

והן של מוצרים חופשיים שלא נופלים משמעותית מהמוצרים הנרכשים. בנוסף קמו חברות ובודדים המספקים שרותי תמיכה למוצרים החופשיים כך שהן אלטרנטיבה שחייבים לקחת אותה בחשבון בעת החלטה על רכישת מנוע חיפוש לארגון.

היבט נוסף שגם אותו חובה להזכיר הוא השינוי התפיסתי של ארגנים בהקשר של מנועי חיפוש. בעבר הארגון רכש מנוע כזה כדי לאנדקס את המידע שלו ולאפשר למשתמשי הארגון להגיע למידע הנדרש במהירות. בשנים האחרונות, היקף המידע בארגונים גדל מאוד והוא מאוחסן בסביבות עבודה רבות ומגוונות. גם הצרכים של הארגונים גדלו והבקשות ממנועי חיפוש גדלו בהתאמה. כל זה הוביל למצב בו חברות אינטגרציה נכנסו לתחום והן מציעות "שרות" מלא בתחם האחזור הכולל מוצר חיפוש שנילווה בשירותים רבים היקפיים. לארגונים גדולים זהו שרות שיש לשקול מבחינת עלות מול תועלת ולזכור שהחיפוש היום במאגרי הארגון הוא משימה מורכבת שרק מומחים יכולים לתת לה טפול יעיל.

להלן רשימת הספקים והמוצרים הנתמכים בארץ וכוללים טיפול מסוים בשפה העברית. כפי שנאמר הטיפול בעברית יכול להיות בכמה רמות ועל הארגון הבוחר את המוצר לתת את הדעת לנושא זה כמו לתכונות האחרות של המוצר. הרשימה כוללת מוצרים שניתן להפעילם על פלטפורמות שונות ושאינם מוגבלים לעבודה מול בסיס נתונים מסחרי אחד.

רשימת המנועים מעודכנת למאי 2015 באדיבות היצרנים והנציגים, אם הנך נציג מוצר התומך באחזור טקסט בעברית או אם אתה משתמש ומכיר מוצר כזה אנא העבר לי פרטיו כדי שאוכל לשבץ אותו בטבלה לתועלת הציבור המתעניין בתחום. ניתן להעביר את הפרטים באמצעות דוא"ל ל- offerd@gmail.com. עדכונים למוצרים עצמם, גרסאות או פרטים מזהים אחרים יתקבלו בברכה.

מסמך זה הוא גרסה מעודכנת **ותשיעית** למסמך המקורי שיצא לאור לראשונה בשנת 2002. ממסמך זה הושמטו מספר מנועים שאינם פעילים יותר בשוק הישראלי, ראה הערות בסוף המסמך.

פרטים מזהים של הספקים

שם המוצר	Active Intelligence Engine™ (AIE)	GSA
שם קודם		
גרסה נוכחית	4.3.1	7.4
שם החברה המפתחת	Attivio	Google
כתובת החברה	ארה"ב	ארה"ב
שם הנציגות בארץ	AIS	DoIt ויעל תוכנה
כתובת הנציגות	http://www.active-is.com	http://doit-intl.com
אתר אינטרנט של המוצר	www.attivio.com	https://www.google.com/work/search/products/gsa.html
הערות	מבוסס לוסין	פתרון אחזור הכולל שרתים
מעודכן לתאריך	מאי 2015	מאי 2015

שם המוצר	ElasticSerch	Solr
שם קודם		
גרסה נוכחית		5.1
שם החברה המפתחת	elastic	אפאצ'י
כתובת החברה		
שם הנציגות בארץ		
כתובת הנציגות		
אתר אינטרנט של המוצר	https://www.elastic.co/products/elasticsearch	http://lucene.apache.org/solr/
הערות	Open source מבוסס לוסין	Open source מבוסס לוסין
מעודכן לתאריך	מאי 2015	מאי 2015

מנוע מורפולוגי (לעברית וערבית)

שם המוצר	**מורפיקס**
שם קודם	
גרסה נוכחית	בהתאם למנועים השונים
שם החברה המפתחת	מלינגו
כתובת החברה	תוצרת הארץ 16 תל אביב
שם הנציגות בארץ	כנ״ל
כתובת הנציגות	כנ״ל
אתר אינטרנט של המוצר	www.morfix.co.il
הערות	מוצר ישראלי, גרסאות למנועים השונים
מעודכן לתאריך	מאי 2015

הערות לגבי מוצרים שיצאו מהרשימה בשנים האחרונות

1. Verity - נרכשה ע״י אוטונומי והמוצר הוטמע בתוך Idol
2. Idol ירד מהרשימה בגלל פעילות נמוכה
3. Retrieval Ware - נרכשה ע״י Fast
4. XRS – הופסקה מכירה של המוצר לפני כשנתיים+
5. Fast – המוצר שולב תחת SP וסוף חיי המוצר כבר הוכרז לשנים הקרובות
6. DTSearch – המוצר לא נתמך יותר בארץ
7. WizDoc - המוצר לא נמכר כמוצר חיפוש עצמאי