חיפוש באינטרנט – איך זה עובד?

פתח דבר – מידע למדריך

מטרת השיעור:

- הצגה מופשטת של מנועי חיפוש ברשת – יכולות, אתגרים בפיתוח, ארכיטקטורת המערכות.
- הצגת אופן פעולת מנוע החיפוש של גוגל כמקרה ייחודי (המנוע המפורסם והידוע מכולם).
- גירוי קהל היעד המשתתף בנושאים "חמים" בעולם החיפוש – Geolocalized Search, "למידת" המשתמש וסינון תוכן עבורו.
- באם מתאפשר לקיים בחדר מחשבים – למידת עקרונות חיפוש מידע, חיפוש מתקדם בגוגל.

מבנה ההדרכה - ראשי פרקים:

- הצגת הצורך במנועי חיפוש ברשת ה-WEB.
- פירוט הדרישות שלנו כמשתמשים ממנוע חיפוש.
- **הצגת מנוע החיפוש כמפעל המורכב ממבנים וצוותי עובדים מיומנים ומסונכרנים.**
- **Case Study - מנוע החיפוש של גוגל –**
  - Drill-down לקרביו של מנוע החיפוש הדומיננטי והחזק בעולם כיום.
  - מערכת דירוג האתרים – המתכון הסודי של גוגל, או "מי יותר נחשב?". ההיגיון מאחורי שיטת דירוג האתרים של גוגל.
  - סינון תוכן למשתמש – כל משתמש ותחביביו, תחומי העניין שלו – נדרש לתוכן המותאם לצרכיו. כיצד לומדים את הצרכים הללו? כיצד מגישים תוכן רלוונטי?
  - יכולות חדשות וייחודיות למנוע החיפוש – שילוב מידע גאוגרפי, ניתוח מילות חיפוש להבנת "מגמות" עולמיות, חיפוש בשפה טבעית...

"דגשי השף" למדריך:

- הפעילות מתמקדת במתווה טכנולוגי כללי. המלצתי היא למצותו ביעף על-מנת לגרות דמיון המשתתפים, ובהמשך לצלול לפרטים הטכניים – עפ"י תחושה ובהתאם לחוזק הקבוצה.
- מומלץ לחבר דוגמאות מהמציאות (אקטואליה) לאורך הפעילות, ולשלב "נושאים חמים" שיחזירו תלמידים שהלכו לאיבוד בחזרה אליכם.
- למערך השיעור נוסף נספח "ביצוע הפעילות בכיתת מחשבים", לשימושך.

בהצלחה! (ובהנאה...)

יאיר סלע

פרק א' – התנעה

טריוויה קצרה –

ש: כמה בני-אדם חיים בעולם?
ת: 7,093,514,370 (מעודכן ל-23/1/2013)

ש: כמה מתוכם משתמשים ברשת האינטרנט?
ת: 2,484,547,870 – 35% מאוכלוסיית כל העולם! (מעודכן ל-23/1/2013)

ש: כמה אתרים (לא סתם עמודים) יש ברשת האינטרנט?
ת: 625.3 מיליון (מקור – WolframAlpha.com, מעודכן ל-11/2012)

ש: כמה חיפושים מבוצעים בגוגל **ביום** ממוצע?
ת: 4,000,000,000 (מעודכן ל-23/1/2013) – כמעט שני חיפושים בממוצע לכל אדם המחובר!

מה אפשר ללמוד מהנתונים האלו?

- האינטרנט תופס חלק מרכזי בחיים שלנו – עסקים, בילויים, חברויות, קניות, ניווט ומה לא?!

מישהו יכול למצוא מתכון לאפיית פשטידת קישואים בלי שימוש במנוע חיפוש?
איך היינו יכולים לחפש מתכון שכזה ללא מנועי חיפוש?

- לזכור שמות של אתרי מתכונים או תכניות בישול ולנווט בהם.
http://www.mevashlim.com/Tag/9222001.asp - ממש פשוט לזכור לא? רעיון רע!
- לרשום את האתרים שאנחנו מכירים כבר בקובץ וורד לחפש בו כל פעם – אבל אז הקובץ לא מתעדכן, ויכולים להיות שינויים. מה גם שכל הזמן עולים אתרים חדשים לרשת – ולא נשמע עליהם בכלל, הקובץ יהיה גדול ומסורבל.... רעיון רע!
- אנחנו צריכים מערכת פשוטה, נוחה, שתמצא במהירות אתרים בהתאם לבקשתנו באותו הרגע, שתהיה עדכנית, שתציג את תוצאות החיפוש המתאימות ביותר מול עינינו ואת הפחות-טובות מאחוריהן, שתתמוך בעברית ובאנגלית, שתאפשר חיפוש משפטים שלמים ולא רק "פשטידה", שתציג תמונות ווידאו! – **אנחנו צריכים מנוע חיפוש!!!**

פרק ב' – התיאוריה

אנחנו מכירים מסך עם תיבת טקסט, כותבים בה מספר מילים ולוחצים "Search", מקבלים תוך 20 מאיות השנייה רשימת תוצאות שכוללת לעיתים 3 מיליארד עמודים ומתחילים לשוטט. מאחורי הקלעים – היכולת לחפש מידע במהירות גבוהה, להציג תוצאות מתאימות ועדכניות – מצריכה השקעה רבה וגם המון יצירתיות בפיתוח.

נתבונן במנוע החיפוש כאילו היה **מפעל גדול** המורכב מבניינים וצוותים. נתמקד כל פעם בצוות אחר במפעל כדי להבין את תפקידו ולא-פחות חשוב; כיצד הצוותים השונים מתחברים ביחד למנוע חיפוש שלם. ככל שנתקדם נחשוף עוד חוליה/מבנה עד הצגת ***ארכיטקטורת*** המערכת השלמה.

<u>צוות #1 "המשוטטים" (Crawler,Spider) –</u> הצוות שדואג להשיג עוד מידע! שהרי מנוע חיפוש שלא מתעדכן ומנסה כל הזמן להגדיל את מספר האתרים שהוא מכיר הוא לא רציני! צוות המשוטטים מונה מאות תוכנות מחשב חרוצות המשוטטות ברשת האינטרנט, ביום ובליל (בלי קשר לחיפושים שמישהו עושה), ומחפשות אתרים חדשים. המשוטטים מקבלים מצוות **הסוחרים** (**URL Server,** לא מפורט בהרחבה) רשימת אתרים גדולה וידועה (בד"כ אתרי חדשות מרכזיים, אתרי פרסום, רשתות חברתיות), וממנה ממשיכים לכל היפר-קישור (Link) בו הם נתקלים בעמודי האתרים. כל עמוד חדש שנמצא – יידחס ויישלח **לצוות המאחסנים** שיכניסו אותו **למחסן**.

<u>צוות #2 "המאחסנים" (Store Server) –</u> הצוות מקבל עמודי אינטרנט חדשים מצוות **המשוטטים** (אחרי שאלו בדקו שהאתרים באמת חדשים), ומכניס אותם כמו שהם, בלי שום מחשבה, סדר או ארגון למדפי **המחסן**. המאחסנים צריכים לעבוד מאוד מהר, שכן צוות המשוטטים מייצר כמויות אדירות של עמודי אינטרנט חדשים בכל שנייה (מעל 100 <u>אתרים</u>, שהם בערך 1000 עמודי אינטרנט, <u>בשנייה</u>).

<u>בניין #1 "המחסן" (Repository) –</u> במחסן נאגרות כמויות אדירות של מידע גולמי (לא מנותח), שהוא בעצם עמודי האינטרנט דחוסים עם כל המידע בהם (תמונות, טקסט וכו'). המחסן נקרא בשפה של אנשי-מחשבים "*בסיס נתונים*", והוא המבנה המרכזי במפעל. הדחיסה מאפשרת להכניס כמויות אדירות של אתרים שלמים, מבלי להיכנס לבעיות של נפח אחסון.

<u>צוות #3 "הגנזכים" (עובדי המחסן) (Indexer) –</u> צוות מסור, יסודי ואחראי. השליטים הבלתי-מעורערים במחסן. תפקידם הוא לקחת את המידע הגולמי בו, לנתח ולעבד אותו. הגנזכים מגיעים **למחסן**, שולפים עמודי אינטרנט דחוסים מהמדפים, מנתחים אותם לעומק – מייצרים אוסף מילות מפתח המאפיינות את העמוד (Hits), כולל מיקומם בעמוד (גבוה יותר=רלוונטי יותר), גודל הגופן בעמוד (גדול יותר=רלוונטי יותר), ופרטים נוספים שיכולים להעיד על מידת ההתאמה של העמוד למילת המפתח. לדוגמא – המשוטטים סיפקו את האתר <u>www.amazon.com</u>; בדיקה שטחית תסיק כי מדובר בחנות המוכרת ספרים ואלקטרוניקה, אולם גנזך יסודי יבחין כי בעמוד החברה התווספה אפשרות חדשה של רכישת מוזיקה – שכן Amazon (חברת ענק אמריקאית, ולה רק חנויות וירטואליות) בחודשים האחרונים נכנסה גם לשוק המוזיקה ומוכרת קבצי MP3, ולכן ידידינו הגנזך יקטלג את עמוד הבית של החברה לצד המילים "ספרים, אלקטרוניקה, וגם <u>מוזיקה</u>". עבודת הגנזך מורכבת, ולכן מתכנתי גוגל כתבו שיטת פעולה מפורטת הכוללת את השיקולים המלאים – שיטה זו נקראת "*אלגוריתם*" בשפת אנשי מחשבים. בסוף שלב ניתוח כל עמוד-אינטרנט, הגנזכים מעבירים את העמוד ואת מילות המפתח שאספו עבורו **לאגף החביות**, ממנו כבר אפשר לשלוף במהירות את המידע לפי מילות מפתח בשלב החיפוש.

<u>בניין #2 "אגף החביות" (Barrels) –</u> אגף החביות כבר מכיל את עמודי האינטרנט אחרי שעברו עיבוד, ולצדם מילות המפתח המאפיינות אותם (דוגמא למידע במחסן הוא תוכן עמוד-הבית של האתר <u>http://www.one.co.il</u> המוצמד למילות המפתח "כדורגל, כדורסל, ליגת-העל, ניר דוידוביץ" ונוספות). ברגע שמשתמש מחפש במנוע החיפוש, סריקה יעילה של אגף החביות תוך שימוש במילות המפתח שסיפק המשתמש בחיפוש, תספק רשימת תוצאות של אתרים מועמדים. גם אגף החביות הוא "*בסיס נתונים*", אולם הוא בסיס נתונים שנבנה לגישה מהירה מאוד לשליפת מידע. רשימת העמודים שתתקבל מהחביות איננה מסודרת לפי רלוונטיות, אלא רק לפי התאמת מילות-המפתח לעמוד. לדוגמא חיפוש המילה facebook בגוגל, תשלוף מהחבית גם את עמוד הבלוג של

דניאלה שבמקרה כתוב בו בפרק 150 שחבר שלה שלח לה הודעה חמודה ב-facebook. האם זה מעניין אותנו? האם זה העמוד שנרצה לראות בצמרת רשימת תוצאות החיפוש? לא ממש... אנחנו זקוקים למישהו שיסדר לנו את תוצאות החיפוש **לפי מידת העניין שלנו** (למרות שהוא לא מכיר אותנו). וכאן בדיוק נכנס **צוות המדרגים**.

צוות #4 "המדרגים" (PageRank) – אצל גוגל, זהו **הצוות המיוחד והחשוב ביותר** בכל המפעל, ובלעדיו ככל-הנראה גוגל היה סתם עוד מנוע חיפוש. מסתבר שהאתגר הכי גדול של מנוע החיפוש הוא לאו-דווקא מציאת העמודים, אלא הצגתם למשתמש באופן בו האתרים המתאימים ביותר לחיפוש מופיעים למעלה, וכל שיורדים למטה בעמוד התוצאות, ההתאמה הולכת ויורדת. מנוע חיפוש טוב נדרש להתאים לכל עמוד בבסיס הנתונים שלו (**אגף החביות** במפעל) **ציון** בהתאם למילת המפתח – הציון מושפע ממיקום המילה בעמוד, גודל הגופן שלה וכמה פעמים היא מופיעה בעמוד, אבל לא רק! הציון מושפע גם **ממידת פופולריות העמוד**.

כדי להסביר את תפיסת הפופולריות, אתן דוגמא[1] פשוטה (אפשר לבצע כמשחק בקבוצה – יש מעסיק ויש מחפשי-עבודה): אתם בעלים של חברת הזנק (Startup) מחפשים מתכנתים לתפקיד אצלכם בחברה. מגיע אליכם בחור צעיר, נראה רציני, עם קורות-חיים והמלצה מהמעסיק הקודם שלו ובה נאמר כי הבחור רציני ואחראי. נראה טוב לא?! למחרת מגיעה בחורה להתראיין לתפקיד, עם מסמך קורות-חיים דומה והמלצה מ-10 מעסיקים קודמים שלה שהם מנהלים בחברות הייטק ישראליות. מי נראה לכם יותר מתאים עכשיו? ביום השלישי מגיע בחור אחרון, עם מסמך קורות-חיים מרשים, ועם המלצה **אחת בלבד** (זהו?!), אבל מנשיא חברת Microsoft העולמית, **ביל גייטס**. ביל כותב עליו שהוא עובד מסור, אחראי ומהרציניים שפגש. במי תבחרו עכשיו?! הבחור השלישי מנצח בגדול! למה? בגלל ביל! כך פועל ההיגיון מאחורי מנגנון דירוג הפופולריות של גוגל!

פופולריות היא תפיסת עולם בגוגל – לפיה יש אתרים פופולריים, המדורגים בצמרת הרשימה, ויש אתרים פחות פופולריים המדשדשים בתחתית. כאשר **צוות המדרגים** בוחן עמוד חדש – הוא מסתכל על מספר האתרים שקיימים כבר בבסיס-הנתונים והמכילים קישורים לעמוד החדש; ככל שמספר האתרים המקשרים אליו גבוה יותר, כך דירוגו עולה. אבל זה לא נגמר כאן – צוות המדרגים משקלל בציון הסופי גם את **מידת הפופולריות של האתרים** שקישרו לעמוד החדש. כך לדוגמא עמוד חדש שהעליתי לרשת אשר זכה לכתבת-שער עם קישור באתר רשת CNN יקבל דירוג גבוה יותר מעמוד שני שבניתי, שקיבל הפנייה מהאתר של אחי הגדול.

צוות #5 - "המחפשים" (Searcher) ומעגל החיפוש השלם – זהו צוות השיווק שפוגש הלקוח אשר נכנס למפעל (שהוא בעצם אנחנו, כאשר נכנסנו לאתר www.google.com). הלקוח בוחר מה הוא מעוניין לחפש (שאילתת חיפוש), צוות השיווק מפרק את השאילתה למילים ורץ מהר **לאגף החביות** כדי לדוג את כל העמודים הקשורים למילים הללו, ובעדיפות עמודים שמכילים את כל מילות השאילתה יחדיו. משם ממשיך צוות המחפשים עם רשימת האתרים **לצוות המדרגים**, אשר הכין מבעוד מועד את רשימת הדירוגים לכל עמוד ברשימה, ועל-פי הקריטריונים לדירוג שהגדרנו קודם, מחזיר לצוות המחפשים את רשימת העמודים באופן ממוין עפ"י התאמה ופופולריות – הכי טובים למעלה! צוות המחפשים רץ בחזרה ללקוח ומחזיר לו את הרשימה הממוינת, עם חיוך! (כל התהליך שתיארנו עכשיו אורך בממוצע 0.2 שניות!!!! ומחזיר לעיתים רשימה של 2 מיליארד תוצאות!)

[1] באדיבות אוהד ברזילי, אוניברסיטת ת"א

בשלב זה מומלץ להציג את ארכיטקטורת המערכת המלאה (נספח ב')

פרק ג' – דיון בקבוצה, היבטים מתקדמים בחיפוש

האם התהליך שתיארנו מתאר את מנוע החיפוש האולטימטיבי?

מה אפשר לשפר?

- אתרים חדשים שלא מקבלים מספיק תשומת לב נזרקים לתחתית הרשימה, ומנוע החיפוש יציג אותם רק אחרי משך זמן ארוך של "חדירה לתודעה" וגם אם הם בדיוק מה שהמשתמש חיפש. אפשר לחשוב על מנגנון שימזער את הפגיעה באתרים כאלו, אם הם באמת טובים.
- **"היכרות אישית"** עם המשתמש – כל משתמש ותחומי-העניין שלו, מקליד בדרך-כלל מילות חיפוש מהן ניתן ללמוד עליו המון! כל מילת חיפוש שאנחנו מריצים בגוגל מתועדת, ובאמצעותה נבנה "פרופיל" לכל משתמש, עליו מתבסס מנוע-החיפוש בהתאמת תוכן רלוונטי המותאם לצרכיו (כך לדוגמא- אוהד כדורגל ישראלי המריץ מספר-רב של חיפושים לתוצאות משחקים בארץ יקבל בראש רשימת החיפוש את האתרים המתאימים ביותר בנושא, לדעת גוגל). בנוסף החברה מוסיפה פרסומות לצד רשימת החיפוש התואמות לפרופיל המשתמש – פרסום הינו מרכיב מרכזי ברווחיות חברת גוגל.
- **חיפוש מבוסס מקום (geo-localized search)** – לפעמים אנחנו מחפשים תשובות שקשורות למיקומנו הגיאוגרפי (לדוגמא: חנות משחקי-המחשב הקרובה אליי). זו שיטת חיפוש שהפכה נפוצה מאוד מרגע כניסת הטלפונים החכמים המצוידים ב-GPS, ומנועי החיפוש הגדולים התאימו עצמם אליה.
- **זיהוי "מגמות" חיפוש** – לדוגמא בבחירות 2013; משתמש שחיפש "בנימין נתניהו" ככל-הנראה התעניין בביבי בהקשר הבחירות ופחות בהקשרים אחרים (ביבי הלוחם בסיירת מטכ"ל, או ביבי כסטודנט בארה"ב בשנות העשרים לחייו). זיהוי חיפושים המבוצעים ע"י משתמשים רבים בפרק-זמן קצר מאפשר התאמת תוצאות **אקטואליות** למשתמש.
- **חיפוש בצילום –** מה יותר נוח מצילום תמונת לוגו של חברה וקבלת תוצאות חיפוש מיידית עבורה ברשת? אפליקציות שכאלו קיימות כבר כיום בטלפונים חכמים וכוללות מנגנון מורכב לניתוח התמונה ופירושה, שכן היא למעשה מהווה את מילת החיפוש שלנו.

נסכם בשני סיפורים מצחיקים מימיו הראשונים של מנוע החיפוש "גוגל" –

- המשוטטים פועלים כמו נחיל-דבורים אדיר; מבקרים תוך דקות במעל לחצי-מיליון שרתים, ובתוך כל אתר משיגים כל עמוד שהצליחו לשלוף. ביקור כל-כך מאסיבי באתרים מסויימים – הבהיל את אחראי האתרים וגם להם להתקשר למפתחי מנוע-החיפוש כדי להבין מדוע הם כל-כך התעניינו באתר שלהם...
- תוכנת השיטוט הגיעה במקרה לאתר ענק של משחק-רשת, ביקרה באלפי עמודים בו תוך מספר רגעים, פעולה שהובילה ליצירת שחקנים חדשים במשחק והעלאת מספר-רב של הודעות זבל שקפצו לכל השחקנים מול העיניים במהלכו! לא נעים....

נספח א' – ביצוע הפעילות בכיתת מחשבים

ביצוע הפעילות בכיתת מחשבים מאפשר לקבוצה לבחון בעצמה את החומר שנלמד ולהכניס **היבטים יישומיים בחיפוש מידע באינטרנט**. הקבוצה מתפזרת בכיתת המחשבים ומבצעת את הפעילות בהנחיית המדריך.

פעילות 1 – הדרך הנכונה לחיפוש מידע ("מידענות"):

נכנסים לאתר: http://info.org.il/search_keywords.shtml של המידען חנן כהן, ובו נלמד על ארבעת החוקים המרכזיים לחיפוש נכון במנועי-החיפוש ברשת האינטרנט:

1. לא כותבים את השאלה כמו שהיא, אלא מילות מפתח שעשויות להופיע בתשובה.
2. מילות מפתח לא יכילו מילות שאלה, מילות חיבור, אותיות חיבור וסימני פיסוק.
3. יותר מילות מפתח – פחות תוצאות חיפוש.
4. הבנת הקשר השאלה.

מומלץ לאפשר זמן קצוב לקריאת העמוד הקצר לפני שממשיכים הלאה.

פעילות 2 – חיפוש מתקדם בגוגל:

הצגת העמוד לחיפוש מתקדם בגוגל - http://www.google.com/advanced_search

סקירת אפשרויות החיפוש המתקדמות, המאפשרות שליטה על תוצאות החיפוש בהגדרת מילים מסוימות (וקשר בוליאני - AND/OR/NOT), סילוק תוצאות המכילות מילים מוגדרות, מספרים בטווח מוגדר ועוד, צמצום תוצאות לשפה מוגדרת, לאתרים במדינות מסוימות, לאתרים מסוימים, לאזורים מוגדרים בעמודי התוצאה (כותרת, שורת הכתובת - URL), פורמט קבצים מסוים ועוד.

מומלץ לעבור ביחד על מבנה העמוד ולספק דוגמאות למצבים בהם נרצה להשתמש בכל אפשרות.

פעילות 3 – תחרות "חפש את המטמון":

בחלק זה מקיימים תחרות בין המשתתפים, ובה בכל פעם המדריך מבקש תשובה לשאלה אותה הוא רושם על הלוח או מקריא בקול. הראשונים למצוא תשובה לשאלה זוכרים בפרס צנוע, וזוכים לעלות ולתאר את תהליך הגדרת השאילתה לחיפוש שהביאה אותם לתוצאה המהירה.

- המדריך מוזמן להתערב ולהזכיר עקרונות חיפוש שנלמדו בפעילות 1,2 ליישום בהמשך.
- מומלץ להציג שאלות אשר תהליך החיפוש שיביא לפתרונן יהיה ברמת קושי עולה – ויצריך מהמשתתפים (בשלב המתקדם) שימוש ברוב הכלים שנלמדו בפעילות.

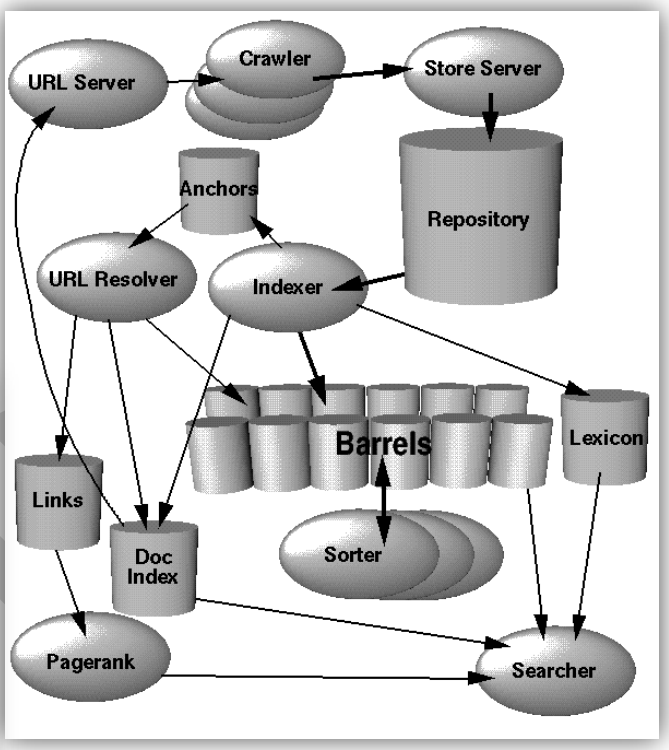
פעילות 4 – "מבט לעתיד":

הצגת מנועי חיפוש מתקדמים –

- המשקפיים של גוגל (אם יש למדריך אייפון/אנדרויד).
- חיפוש מבוסס מיקום באפליקציות שונות (b144, גוגל-מפות ב-iOS).
- חיפוש תוך שימוש בשפה-טבעית (www.wolframalpha.com).

נספח ב' - איורים

ארכיטקטורת מנוע החיפוש של "גוגל":



חוות-שרתים של "גוגל" בארה"ב (ענקית!):