# מסמך מערכת

# Stack**overflow** Toolkit

## כותבים :

ענבל עמרני, יזהר גלבוע, ערן הירש

# מוטיבציה

1. ניתן לראות שקיים צורך למערכת שמאפשרת למשתמש לקבל שאלות שיידע לענות עליהן.  
   <http://meta.stackoverflow.com/questions/214650/is-there-a-question-recommendation-mechanism>
2. ניתן לראות שקיים צורך למערכת שמאפשרת בהינתן שאלה למצוא את המשתמש המתאים לענות עליה.
3. באתר Q&A גדול אחר שנקראת Quora,יש מנגנון שמאפשר בהינתן שאלה למצוא את המשתמש המתאים לענות עליה, ב-stackoverflow אין.
4. קיים שוק חזק של אפליקציות מעל stackoverflow ,ניתן לראות את כל האפליקציות בשוק בקישור הבא :   
   http://stackapps.com/questions/tagged/app  
   המערכת שלנו תשתלב היטב בשוק הקיים.

# דרישות

הקדמה  
הפרויקט שלנו מתמקד באתר stackoverflow כאשר נרצה לתת סט כלים על גבי האתר.  
הכלי הראשון שאנחנו רוצים ליצור הוא כלי Answerer Recommendation.  
  
Answerer Recommendation - backend  
- בהינתן "שאלה חדשה" תחזיר את המשתמשים שהכי מתאימים לענות עליה.  
- בהינתן משתמש תחזיר את השאלות שהוא עשוי לענות עליהן.   
  
Answer recommendation - UI  
דף חיפוש  
- תיבת טקסט (חיפוש) שמכניסים אליה קישור לשאלה ב-stackoverflow.  
- הפלט של החיפוש יהיה משתמשים שמדורגים לפי המשתמש שהכי מתאים לענות עליה. בנוסף למשתמש הכי מתאים ניתן להראות על איזה שאלה דומה הוא ענה.  
דף רישום  
- תיבת טקסט שמכניסים אליה username של משתמש ו-email.  
- המשתמש ייקבל למייל כל ערב רשימה של השאלות הכי מתאימות לו לענות עליהן.

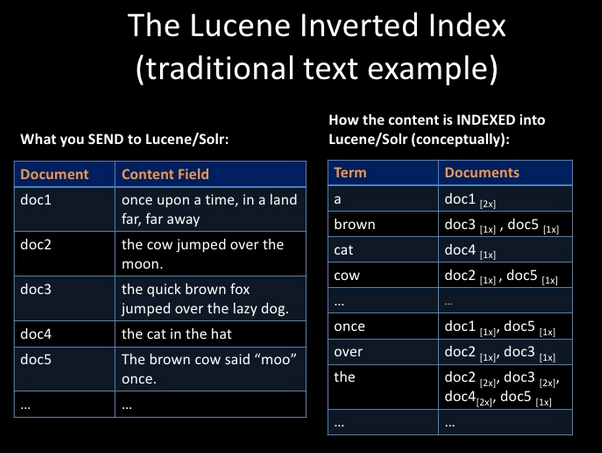
# סקירה טכנולוגית

## Solr

Solr הוא שרת חיפוש/אינדוקס המבוסס על ספריות Lucene לחיפוש/אינדוקס Full-Text.

## Lucene

Lucene היא ספריית חיפוש טקסטואלי בעלת ביצועים גבוהים, אך ניתן להתייחס אליה כספריית Token Matching ודירוג תוצאות בעלת ביצועים גבוהים, המסוגלת לבצע חיפוש טקסטואלי. אם נתרגם זאת למונחי Machine Learning, ניתן להסיק מכך ש Lucene Index הוא בעצם מטריצה רב מימדית שעבור שדה מסוים שנאנדקס נקבל מטריצה שממפה כל Token שאונדקס לכל המסמכים המכילים אותו.  
כמו כן עבור כל מסמך ניתן לקבל את ה"וקטור" של ה Tokensהמרכיבים אותו.



# תחילת עבודה

* הורדת המאגר של SOF , שהוא כ-25GB דחוסים של טקסט.
* הקמת Solr על שרת וירטואלי מ-Google Compute Cloud.
* אינדוקס 15 מיליון שאלות ותשובות של משתמשי StackOverflow מ 2008 ועד 2014. (קובץ Posts.xml)
  + בניית רכיב אינדוקס שמפרק את ה XML של הטקסט ובונה SolrDocuments, ושולח ל Solr ע"מ שמנוע ה Lucene יפרק אותם ל Tokens ויאנדקס.
* בניית מנוע בסיסי שבהינתן מסמך (שאלה ב SOF) מוצא פריטים דומים. כרגע הדימיון מחושב לפי הפרמטרים של תוכן השאלה וכותרת השאלה, בלי התחשבות בפרמטרים נוספים כגון תאריך, Tags וכו'.
* מציאת משתמשים שענו לשאלות הדומות.
  + בעצם זוהי תשתית לביצוע בדיקות אופטימיזציה שבאמצעותה נוכל להשוות את המשתמשים שקיבלנו למשתמשים שענו בפועל לשאלות.