המכללה האקדמית אחוה
אחזור מידע ומנועי חיפוש
מטלה מספר 2

חלק ראשון: תיאורטי

עיינו בטבלת השכיחות ($tf_{t,d}$) של עבור שלושת המסמכים Doc1, Doc2, Doc3 באיור 2.

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

א. השלימו את הטבלה למטה. חשבו את משקלות tf-idf עבור המלים car, auto, insurance, best עבור כל מסמך. השתמשו בטבלת הבאה, כאשר N= 806,791. (תזכורת – tf $idf_{t,d} = tf_{t,d} * idf_t$)

term	df_t
car	18,165
auto	6723
insurance	19,241
best	25,235

	Doc1	Doc2	Doc3
car			
auto			
insurance			
best			

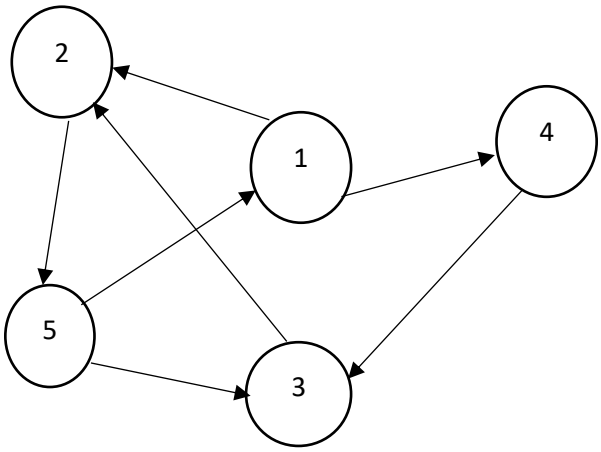
צריך לחשב את הערכים של idf_t קודם. נא להציג אותם בטבלה נפרדת

ב. השלימו את הטבלה עם נירמול של הערכים לפי האורך

ג. עם הטבלאות מסעיפים 3, דרגו את המסמכים עבור השאילתא car insurance.

שאלה 2:

להלן מבנה הדפים הבא



אם נוסיף קישור **מ**צומת **2 ל**צומת **4**, האם וכיצד יושפעו ערכי ה Authority ו- HUB, על פי אלגוריתם HITS **על צומת 5,** כלומר האם יהיו גבוהים יותר, נמוכים, ללא שינוי. ענו **מבלי לחשב את הערכים**

להלן מבנה של דפים:

1. דף A מצביע על D, C,B
2. דף B מצביע על E
3. דף C מצביע על B ו- D
4. דף D מצביע על E
5. דף E מצביע על A ו- C

א. הריצו את אלגוריתם ה authorities and hubs על הגרף הנתון. הראו את ערכי hub and authorities אחרי 2 איטרציות. לכל עמוד בכל שלב בחישוב. הציגו את התוצאות עבור A, B, C, D, E.

ב. עבור אותו הגרף, חשבו את ה PageRank של כל דף אחרי שתי איטרציות

שאלה 3

להלן אוסף של ביקורות על סרטים יחיד עם הסיווג: ביקורת חיובית **+** או ביקורת שלילית **−**

Review	**Class**
Boring and Predictable	**−**
Excellent movie	**+**
Predictable Extremely mediocre	**−**
A pathetic attempt at a romcom	**−**
Good movie with great actors	**+**
Fantastic job	**+**

על ידי שימוש Naïve Bayes Classifier, חשבו את הסיווג של הביקורת הבאה **:**

It is a good job, but it's extremely predictable

אל תתחשבו ב stop words מומלץ לעשות קירוב עד ארבעה מספרים אחרי הספרה העשרונית. הראו את כל שלבי החישוב. השתמשו בנוסחה המתוקנת עם **smoothing technique**

שאלה 4: הערכה

נתון מנוע חיפוש שמחזיר תוצאות לא מודגות. למנוע החיפוש ניתנה השאילתה הבאה

Mountain Biking Adventures

התקבלו:

Mountain Biking: The Ultimate Guide
Adventure Sports: Mountain Biking Edition
The Best Hiking Trails - Outdoor Adventures
Worldwide Mountain Biking Tours

לא התקבלו

Mountain Biking Basics and Techniques
Extreme Mountain Biking: A Thrill Seeker's Guide
Biking Gear for Road Biking
Essential Mountain Biking Skills for Beginners

חשבו את Recall ואת precision של התוצאות

חלק שני: מימוש אלגוריתם PageRank ב-Python

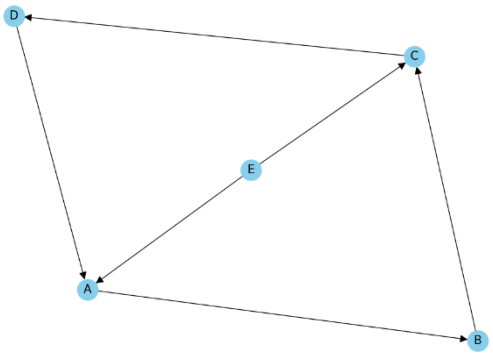
בחלק זה אתם תממשו אלגוריתם PageRank ב Python. המימוש יכולול שתי גרסאות: גרסה שלמדנו בכיתה וגרסה משופרת שתוסבר בהמשך.

Input:

Graph Representation: Your implementation should accept a representation of the website graph as an adjacency list.

Adjacency List: An adjacency list representation for a graph associates each vertex in the graph with the collection of its neighboring edges.

```
graph = {
    'A': ['B'],
    'B': ['C'],
    'C': ['D'],
    'D': ['A'],
    'E': ['C', 'A']
}
```



Output:
PageRank Scores: a list of ranking values.

למטלה מצורף שני קבצים.
קובץ pagerank.py שצריך להשלים
קובץ util.py שמסופק במלואו.

הסבר על calc_pagerank_with_damping: עליכם להשלים לבד את המושג damping factor. אבל להלן הסבר יחד עם הנסוחה

The damping factor in the PageRank algorithm is a probability value (usually set around **0.85**) that represents the likelihood of a user continuing to click on links versus jumping to a random page.

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

where $p_1, p_2, \ldots, p_N$ are the pages under consideration, $M(p_i)$ is the set of pages that link to p_i, $L(p_j)$ is the number of outbound links on page p_j, and N is the total number of pages.

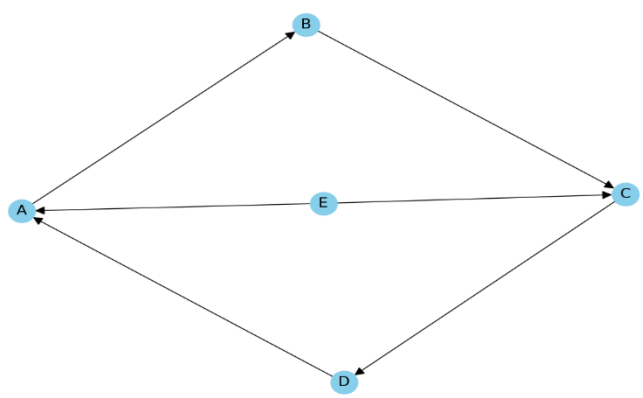
Running Exampels

```
graph = {
    'A': ['B'],
    'B': ['C'],
    'C': ['D'],
    'D': ['A'],
    'E': ['C', 'A']
}

pr = PageRank(iterations=10)
util.draw_graph(graph)

page_rank_scores = pr.calc_pagerank(graph)
site_scores = util.toString(page_rank_scores)
print(site_scores)

page_rank_scores = pr.calc_pagerank_with_damping(graph)
site_scores = util.toString(page_rank_scores)
print(site_scores)
```



{'Website A': 0.4, 'Website B': 0.2, 'Website C': 0.2, 'Website D': 0.2, 'Website E': 0}

{'Website A': 0.3228250000000001, 'Website B': 0.2, 'Website C': 0.2255, 'Website D': 0.221675, 'Website E': 0.03}