ליאור רזניק 307712224 אלכס בוייב 314393158

מבוא

החלטנו לבנות מערכת היברידית אשר בנויה משתי מערכות נפרדות:

- מערכת רגרסיה מבוססת matrix factorization אשר נבנתה בעזרת TensorFlow2 , מערכת זו מנסה למפות למרחב נסתר (Latent Space) משותף את המשתמשים ואת העסקים. וכאשר נכניס לה משתמש עם כל העסקים, היא תוציא לנו את ההתאמה של המשתמש לכל אחד מהעסקים (ערכים בין 0 ל-2).
- מערכת opinion mining אשר מקבלת טקסט מהמשתמש ופולטת מספר בין 0 ל-2, המערכת למעשה ממפה את ביקורת המשתמש על עסק מסוים לרייטינג.

ליאור רזניק 307712224 אלכס בוייב 314393158

פרק 1: הכרת המידע

השתמשנו במאגר נתונים אשר נלקח מ-Yelp, כלומר מסוג של רשת חברתית אשר נועדה להציע עסקים מקומיים למשתמשים, ברשת זו המשתמש יכול לתת פידבק על העסק, לדרג אותו ולהמליץ על דברים. בנוסף, ברשת זו מופיעים דברים חשובים כמו: חנייה, שעות פתיחה, האם העסק מתאים לילדים, משפחות וכו'.

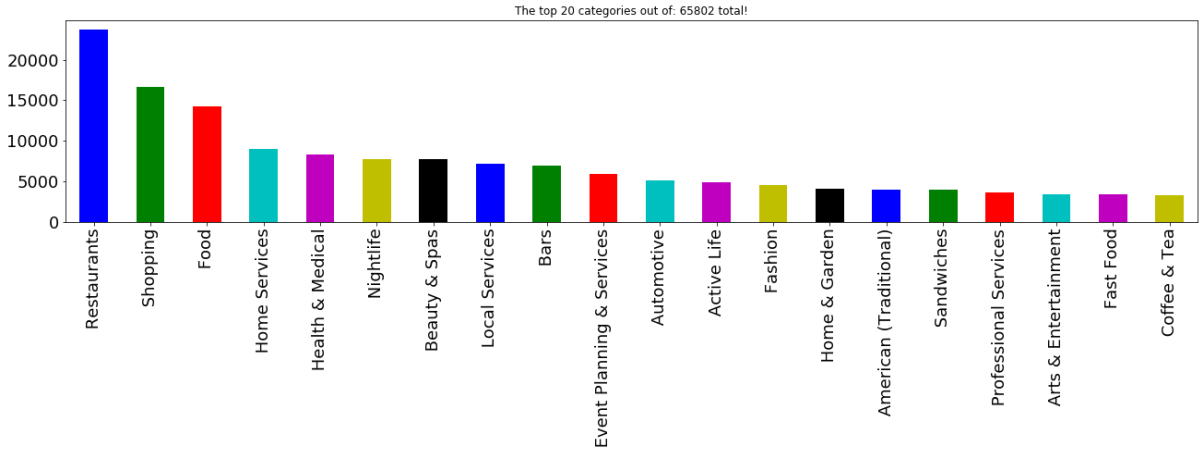
המאגר נתונים מורכב מ:מידע על הלקוח, מידע על עסקים, ודירוג עסקים על ידי הלקוחות.

את מאגר המידע ניתן להוריד מ-https://www.yelp.com/dataset/download.

מאפייני המאגר

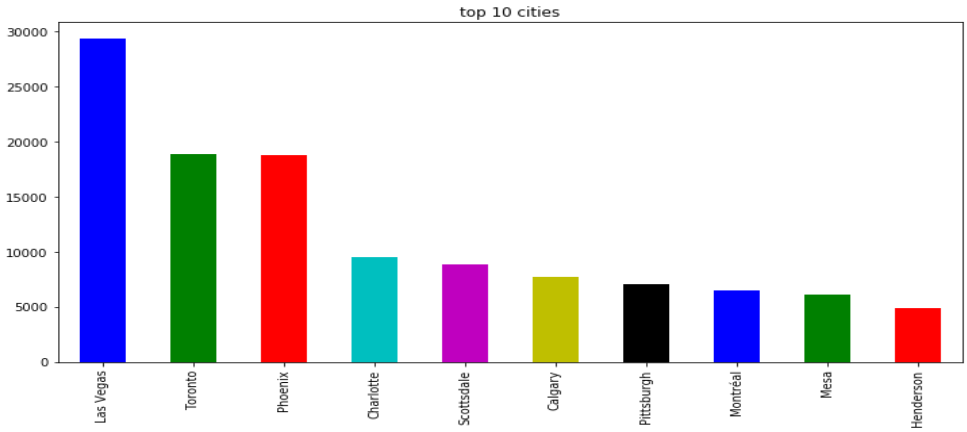
ביצענו חקר מקדים של המידע אשר ברשותנו:

1. 20 הקטגוריות הפופולריות ביותר הם:



- נשים לב כי הקטגרויות אינן זרות לחלוטין.

2. 10 הערים בעלות מספר העסקים הרב ביותר:



3. הערים עם הריטיניג החציוני הטוב ביותר:

city	review_count	stars
Las Vegas	1965651	3.696170
Phoenix	706761	3.646808
Toronto	510856	3.444409
Scottsdale	372805	3.932500
Charlotte	299233	3.539594
Pittsburgh	218776	3.602893
Henderson	210596	3.760221

מחקר המידע, החלטנו להתמקד בעיר לאס וגס (שכן, בא יש את הכי הרבה עסקים).

בנוסף, בחקר שביצענו ראינו התפלגות של ריטינג מ-1 ל-5 , אנחנו המרנו אותו ל-0 עד 2 , כאשר 0 זה דירוג גרוע, 1 זה דירוג ניטרלי ו-2 זה דירוג מעולה.

פרק 2: הסבר על המערכת

כאמור, המערכת שלנו מחולקת לשתי חלקים: מערכת רגרסיה להמלצה על תכנים לפי העבר של המשתמש ו- מערכת סיווג לסיווג הביקורת של המשתמש.

מערכת רגרסיה

עיבוד מקדים

- קריאת הנתונים והשארת הנתונים של עסקים פעילים בלאס וגס (20360 עסקים).
- שילוב בין frame של המשתמשים והביקורות לבין ה-frame של העסקים (לפי id של העסק).
- מחיקת כל העמודות הלא רלוונטיות (כלומר, כל מה שלא id של משתמש, עסק או דירוג).
- מתן מספר ייחודי לכל ID של משתמש ולכל ID של עסק בעזרת Label encoding . אנחנו עושים זאת משתי סיבות: אנו צריכים ייצוג מספרי בכדי להכניס למודל ואנחנו צריכים אותו רציף, כלומר שלא יהיו חורים בין הייצוגים(אנחנו לא רוצים ייצוג דליל מניסבות זיכרון ודיוק).

התוצאה: frame של כ-1,726,332 שורות.

- חלוקת המידע ל-סט אימון וסט בדיקה ביחס של 0.33. X שלנו הוא [user_id,bis_id] ומשתנה המטרה הוא הרייטינג.

המודל עצמו

כאמור המודל שלנו מנסה לדמות את תהליך ה-Matrix Factorization וזאת בעזרת ביצוע של embedding למשתמש ולעסק(תרגום של אובייקט ל-latent vector space כלומר, ניסיון לתרגם את האובייקט למרחב שבו אובייקטים דומים , נמצאים קרוב יותר זה לזה . לצורך הדוגמא: אם היינו לוקחים שתי שולחנות שונים ושתי כיסאות שונים, במרחב ה"גלוי", זה שאנחנו רואים בו, השולחנות יראו שונים זה מזה וכך גם הכסאות, אך במרחב ה"נסתר" השולחנות יהיו קרובים יותר זה לזה והכיסאות יהיו קרובים יותר זה לזה , שכן יש להם פיצרים דומים (במקרה של שולחנות: פלאטה, 4 רגליים).

1. המודל הבסיסי אשר מהווה עבורנו baseline הוא המודל אשר מבצע את השלבים הבאים:
   - ביצוע embedding למשתמש.
   - ביצוע embedding לעסק.
   - Dot production בין הייצוג של יוזר לבין הייצוג של הסדרה.
   - חיזוי משתנה המטרה (רייטינג).

למעשה, אנו מעוניינים למפות את המשתמשים והעסקים למרחב נסתר משותף.

2. ניסיון ראשון לשיפור:
   - אתחול משקל ה-embedding לפי התפלגות HE.
   - שימוש ב-regularizer מסוג L2 על מטריצת המשקלים של שכבת ה-embedding , הסיבה לכך היא הרצון "להעניש" אנומליות גבוהות מדי, כלומר יכול להיות מצב שיש מספר פיצריים שיקבלו משקל גבוה מאוד אך המספר פיצריים הנ"ל חשוב למקרה מאוד ספציפי ובמקרים אחרים עלול להכניס רעש רב למערכת ולהפריע לחיזוי מדויק, בשל כך אנו "מענישים" את אותם הפיצרים.
3. ניסיון שני לשיפור:
   - הוספת bias , משתמשים נוטים להגזים כלומר, הרייטינג אשר ניתן לא תמיד משקף את הביקורת האמיתית (לרעה או לטובה), לכן, החלטנו להוסיף הטייה מסויימת למערכת. ניסיון זה הראה שיפור עליו נדבר בהמשך.

ליאור רזניק 307712224 אלכס בוייב 314393158

מערכת סיווג רגשות

עיבוד מקדים

- קריאת הביקורות והדירוג המתאים להן, מחיקת כל הביקורות העולות על 100 מילים.
- שינוי הדירוג לפי הסקלה המצויינת מעלה.
- contraction expanding בכדי למנוע כפילויות והגדלה מלאכותית של המילון.
- מחיקת מילות עצירה למעט not, nor ו-no.
- למטיזיציה (הגעה לבסיס משותף).
- בניית מילון (id ייחודי לכל מילה).
- Padding של ביקורות ל-100 מילים.
- בניית מטריצת משקלים המאותחלת בצורה הבאה: לכל מילה במילון(לפי id) ->וקטור של FastText מאומן מראש אם קיים. ואם לא, וקטור רנדומלי. (300 מימדים).
- חלוקת המידע ל-train ו-test ביחס של 33 אחוז.

התוצאה של העיבוד המקדים:767,139 ביקורות, מתוכם:513,983 לאימון ו-253,156 לבדיקה.

המודל עצמו

אנו מעוניינים למפות תגובה/ביקורת אשר המשתמש מכניס למספר.

1. המודל הבסיסי אשר מהווה עבורנו baseline הוא המודל אשר מבצע את השלבים הבאים:
   - שכבת embedding המאותחלת בעזרת מטריצת המשקלים.
   - LSTM דו-כיווני (דו- כיווני בשביל שיהיה לו את היכולת לעבור על המשפט משתי הצדדים, דבר אשר יכול לתרום להבנה וחיזוי של מילים).
   - שכבת fc בגודל 32 עם פונקציית אקטיבציה RelU ובחלק מהמקרים ניסיון לעבור ל-Leaky_ReLU בגלל בעיית הניורונים המתים.
   - שכבת fc כפלט (גודל 3).
2. ניסיון ראשון לשיפור:
   כפי שנראה בפרק הבאה, המודל הבסיסי הראה overfitting ולכן, ניסיון ראשון להפחית אותו היה להוסיף שכבת dropout , שכבה אשר בזמן האימון מוותרת באופן רנדומלי על אחוז מסוים מהחיבורים בין הניורונים.
3. ניסיון שני לשיפור: כפי שנראה, עדיין היה Overfitting ולכן, היה ניסיון להוסיף Regulation על ה-embeddings ועל היציאה של ה-LSTM.

פרק 3: הערכת המודל

טכניקות אשר השתמשנו בהם לשתי המערכות

1. Validation set: מכונות למידה עמוקה סובלות לעיתים מבעיית Overfitting , בעיה הנגרמת מכך שהמכונה לומדת טוב מדי את המידע שהיא מתאמנת עליו אך נכשלת כאשר היא מקבלת מידע חדש שלא ראתה מקודם. בכדי לראות אם יש לנו Overfitting ולנסות ולטפל בבעיה חילקנו את סט האימון לסט אימון וולידציה (לפי חלוקה של 0.2) כך שבסוף כל אפוק, הרשת מנסה לחזות על סט הוולידציה , כלומר על סט שהיא לא ראתה מקודם. אם קיים הבדל משמעותי בין תוצאות הסט הזה לתוצאות של סט האימון אז קיים Overfitting .
2. ModelCheckpoint: יכול לקרות מצב, שהמודל הכי טוב הוא לאו דווקא המודל אשר נוצר באפוק האחרון, בשל כך , אנו מבצעים מוניטורינג על (val_loss -בבעיית הרגרסיה ועל val_acc בבעית הסיווג) כאשר הוא משתפר המודל נשמר, דבר אשר מבטיח לנו שמירה של המודל הטוב ביותר.

מערכת רגרסיה

בגלל שמדובר בבעיית רגרסיה כלומר, בבעיה שהתוצאה שלה הוא מספר רציף או למעשה פונקציה איננו יכולים להשתמש במדדים כגון:F1,ACC וכו'.

בשל כך, השתמשנו אך ורק ב-MAE,MSE,,RMSE ו-HUBER.

כאשר:

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \text{ , } \text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \text{ .}$$

לשתי הפונקציות הללו ישנה בעיות:

- MSE (RMSE) :
  -אין התייחסות לכיוון השגיאה.
  - בגלל ההעלאה בריבוע, ניתן משקל רב לתצפיות החריגות.
- MAE: קשה יותר לחישוב גרדיאנט ולכן, עלול להשפיע על העבודה ברשתות ניורונים.

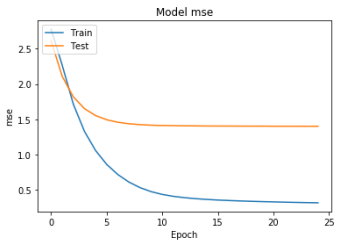
בשל בעיות אלו הומצאה פונקציית HUBER שכן, מדובר בשילוב בין שניהן:

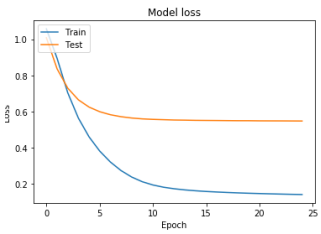
- כאשר השגיאה גדולה וישנן חריגות אז נשתמש ב-MAE .
- כאשר אנו מתקרבים למינימום אז נשתמש ב- MSE.

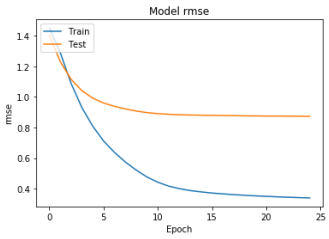
ולכן: חישוב ה-LOSS והלמידה של המכונה נעשתה בעזרת HUBER אך המשכנו לבצע מוניטורינג גם על כל השאר.

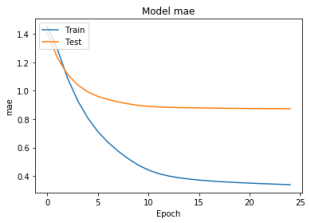
תוצאות:

<u>עבור המודל הראשון:</u>





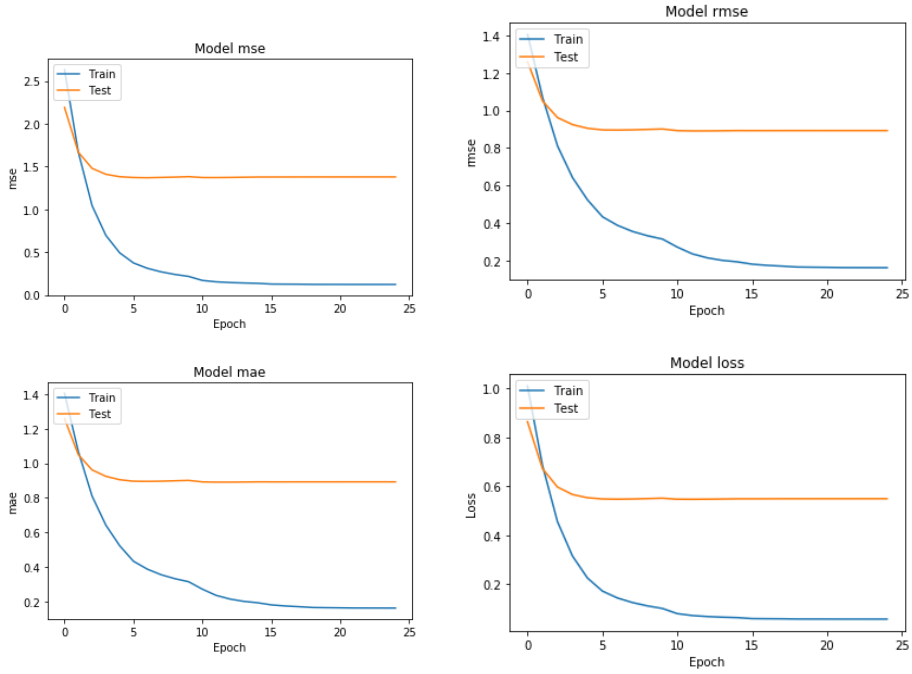




- ניתן לראות Overfitting.
- הערך הכי טוב: 0.5483263560826029- ערך סביר שכן מרחב הטעות שלנו הוא 0-2 אך יש מקום לשפר.

ליאור רזניק 307712224 אלכס בוייב 314393158

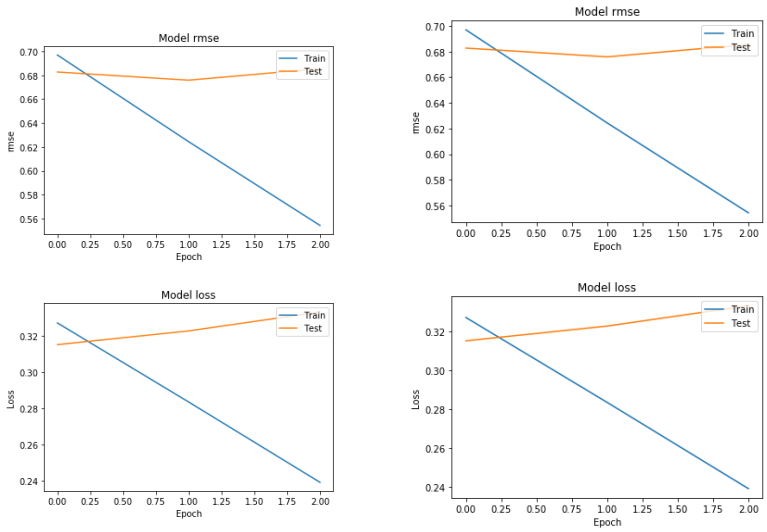
עבור המודל השני:



לא ניתן לראות שיפור.

עבור המודל השלישי:

- הגבלנו ל-3 אפוקים שכן, הריצה לוקחת זמן.
- הערך הטוב ביותר: 0.3150707400550556 שיפור משמעותי מההרצות הקודמות.
- ערכי Test: [0.3126610851339925, 0.67161924, 0.69032454, 0.69032454]
- עדיין ישנו Overfitting.

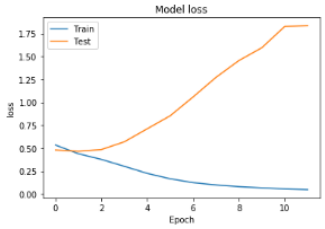


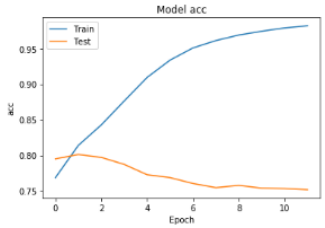
ליאור רזניק 307712224 אלכס בוייב 314393158

מערכת סיווג

במקרה זה מדובר בבעיית סיווג ולכן, המטריקות שבחרנו לעבוד איתן הן: ACC, F1.

מודל ראשון:





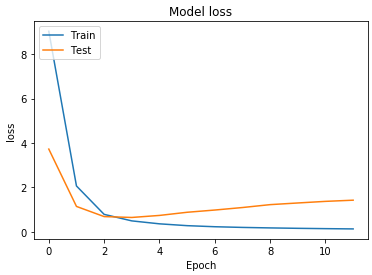
ניתן לראות Overfitting.

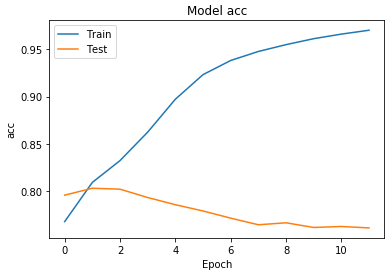
התוצאה הטובה ביותר: val_acc של 0.8012393.

תוצאות על סט הבדיקה:

	precision	recall	f1-score	support
0	0.79	0.78	0.79	85301
1	0.64	0.66	0.65	81698
2	0.83	0.81	0.82	86157
accuracy			0.75	253156
macro avg	0.75	0.75	0.75	253156
weighted avg	0.75	0.75	0.75	253156

מודל שני:



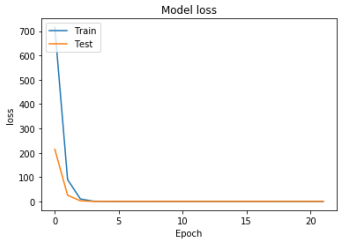


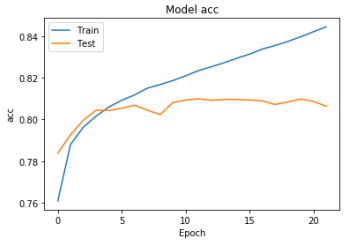
התוצאה על סט הבדיקה:

	precision	recall	f1-score	support
0	0.79	0.79	0.79	83749
1	0.68	0.65	0.67	87366
2	0.81	0.84	0.83	82041
accuracy			0.76	253156
macro avg	0.76	0.76	0.76	253156
weighted avg	0.76	0.76	0.76	253156

ליאור רזניק 307712224 אלכס בוייב 314393158

מודל שלישי:





ניתן לראות שיפור מבחינת Overfitting בנוסף, ניתן לראות שיפור בתוצאות על ה-test set.

	precision	recall	f1-score	support
0	0.84	0.82	0.83	86605
1	0.73	0.72	0.72	86023
2	0.84	0.88	0.86	80528
accuracy			0.80	253156
macro avg	0.80	0.81	0.81	253156
weighted avg	0.80	0.80	0.80	253156

q .

פרק 4 : מסקנות

הצלחנו לבנות מערכת היברדית המורכבת משתי תתי-מערכות נפרדות. למרות התוצאות הלא רעות בכלל, עדיין יש מקום לשיפור.

אחד האתגרים העקריים שבהם נתקענו בו הוא חוסר כח חישובי, דבר אשר גרם לכך שנצטרך להתפשר על איכות המודלים וכמובן מנע מאיתנו לבדוק מגוון רחב של מודלי למידה עמוקה.

שיפורים אפשריים

מודל רגרסיה

1. הוספת עומק למודל (שכבות fc נוספות).
2. המודל עדיין סובל מ-Overfitting ולכן כדי לנסות להוסיף Dropout.
3. לנסות גדלים שונים למרחב הנסתר.
4. להוסיף מאפייני metadata של העסק (לשם כך, בשל באלגן בסט המקורי לנסות ולבנות מודל נוסף אשר יעשה למידה לא מונחת ויחלק לקלאסטרים את כל העסקים).
5. לנרמל (min-max או כל נרמול אחר) את משתנה המטרה.

מודל סיווג

1. להוסיף שכבת CNN לפני שכבת ה-LSTM למטרת feature extraction.
2. משחק עם אורך התגובה.
3. משחק עם Hyperparameters.
4. שינוי המודל למודל Self-Attention , מודל אשר מהיר יותר ממודל RNN ומדוייק יותר שכן פותר את בעיית התלויות הארוכות.
5. הוספת התייחסות ל-Cold Start
   - עבור עסק: על ידי בניית קבוצות לעסקים ובדיקת המרחק של עסק חדש מכל הקבוצות ושיוכו לקבוצה הכי קרובה.
   - עבור משתמש: המלצה על חמשת העסקים האהובים ביותר.

דוגמאות הרצה:

```
In [127]: rec['recommend']([1])
{'Delmonico Steakhouse', 'Chewy Boba Company', 'Puls', 'Starbucks'}

In [128]: rec['recommend']([8])
{'Acres Cannabis', 'Vesta Coffee Roasters', 'Chicas Bonitas'}

In [129]: rec['sentiment']("Went here last weekend and was pretty disappointed. They did not have one thing that was pictured and recommended on yelp as being good.
We started off with the steak grilled skewers which were just Ok, nothing special. My freind got the lasagna and I got some special chicken dish. They were both
pretty bland and lacking that kick. Our waitress was really nice and got the manger to switch out our dishes. My freind got the hamburger and I got the shrimp scampi.
Her hamburger was better then the lasagna but was still lacking flavor. My scampi was better then the chicken but was also still under seasoned and the noodles were a
bit under cooked. With such a big name attached to this restaurant and going on such an empty stomach we had such high hopes. The service was great which is why i
gave it three stars.")
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\liorr\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\liorr\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
0.22199947
```

[קישור לדרייב](#)

הדרייב מכיל: קוד לכל אחד מהמודלים, קוד להרצה של מודלים מוכנים, היסטוריה.