

ניתוח חוות דעת שליליות וסיווג לקטגוריות באמצעות כריית נתונים

סהר יעקב 314741851

מבוא

- נושא: ניתוח וסיווג חוות דעת שליליות לזיהוי מגמות.
- מטרה: שיפור חוויית לקוח ותהליכים עסקיים באמצעות עיבוד טקסט.
- כלי עיקרי: אלגוריתם K-Means.
- חשיבות הניתוח בעידן הדיגיטלי.
- הבנת הקשר בין חוות דעת ללקוחות.

INTRODUCTION

חשיבות עסקית

→ ערך: זיהוי בעיות נפוצות.

→ שיפור מוצרים ותמיכה.

→ יישומים: ניתוח ביקורות, חוויית משתמש, ושיפור תהליכים.

→ הבנת צרכי הלקוחות.

→ שיפור נאמנות לקוחות.



K-Means - חיבור לעולם כריית נתונים

→ יישום קל יחסית על נתונים גדולים.

→ מהות: סיווג חוות דעת לקבוצות דומות לפי דמיון מילולי.

→ שימוש ב-TF-IDF לייצוג טקסט בעזרת חלוקה לוקטורים.

→ יתרונות: פשטות ומהירות.

→ בחירת קלאסטרים: Elbow Method.



LDA - חיבור לעולם כריית נתונים

→ מהות: זיהוי נושאים עיקריים בטקסטים וחלוקתם לפי תוכן משותף.

→ שיטה: המודל מזהה נושאים סמויים על בסיס המילים בטקסט.

→ יתרונות: מבנה פשוט, מתאים לטקסטים גדולים, לא דורש תוויות מראש.

→ בחירת נושאים: שימוש במדדים כדי לקבוע כמה נושאים כדאי לבחור.

שימושים: ניתוח מסמכים, סיווג טקסטים וזיהוי תמות עיקריות.

עיבוד נתונים

- מקורות: חוות דעת מאתרי נתונים כמו קגל.
- עיבוד: ניקוי טקסט והסרת רעשים.
- המרה לוקטורים מספריים (TF-IDF).
- כלים: Python וספריות pandas, nltk, scikit-learn.
- עיבוד המוקדם לדיוק התוצאות ומציאת חילוק לקלאסטרים
- מינמלים לפי שגיאה מינימלית וחישוב דיוקים למודל.



DATA

תוצאות

→ מדדים: score-F1 , דיוק.

→ תוצאות: הצלחה בסיווג קטגוריות עיקריות כמו "איכות מוצר" ו"שירות לקוחות".

→ שיפורים: שיפור בתתי-קטגוריות מורכבות.

→ ניתוח תוצאות והסקת מסקנות.

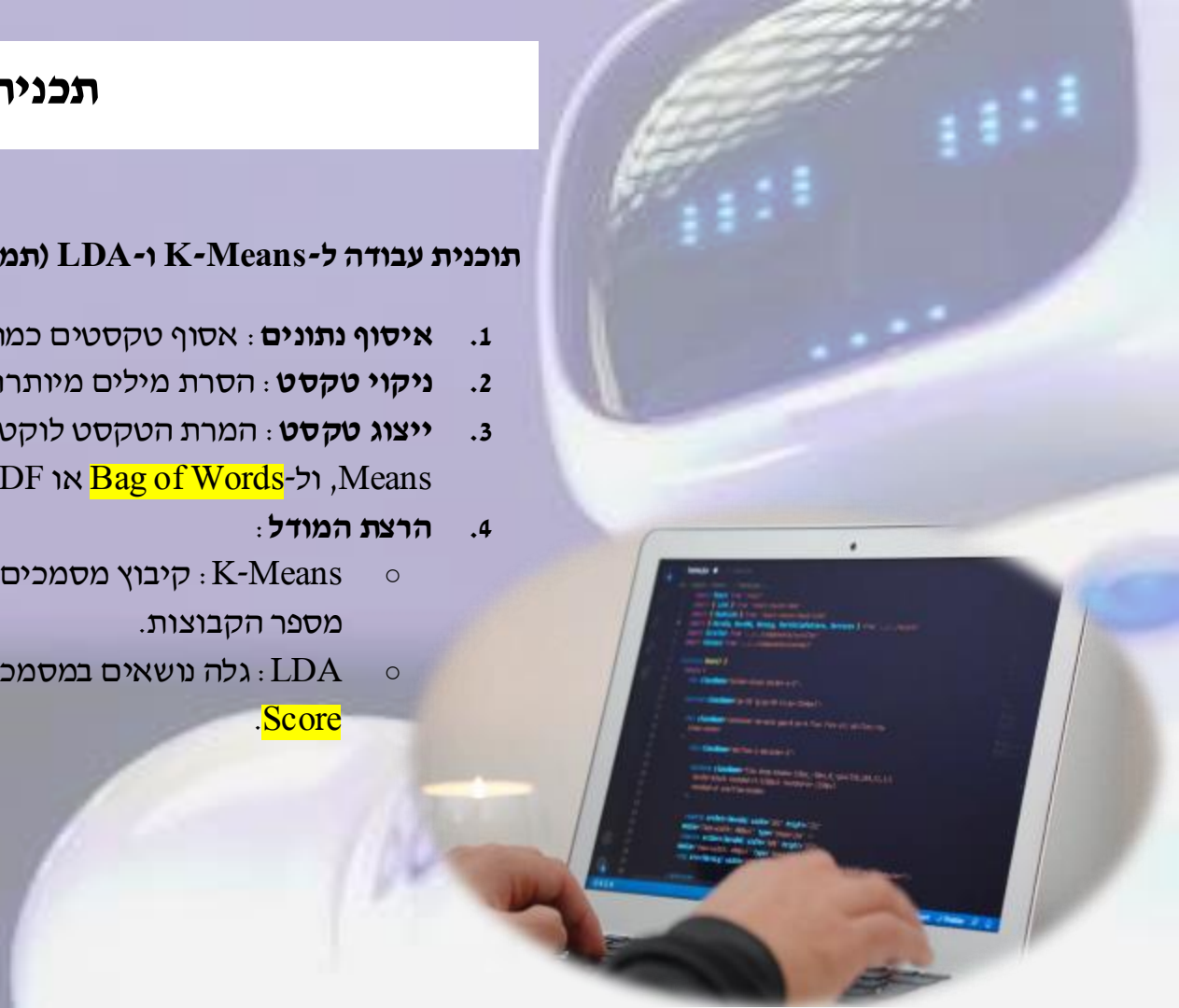
→ הצגת נתונים בצורה גרפית.



תכנית עבודה

תוכנית עבודה ל-K-Means ו-LDA (תמציתית):

1. **איסוף נתונים** : אסוף טקסטים כמו מסמכים או חוות דעת.
2. **ניקוי טקסט** : הסרת מילים מיותרות וערכים חסרים, פיסוק ונירמול טקסטים.
3. **ייצוג טקסט** : המרת הטקסט לוקטורים (TF-IDF או Embeddings) עבור K-Means, ול-**Bag of Words** או TF-IDF עבור LDA.
4. **הרצת המודל** :
 - K-Means : קיבוץ מסמכים לפי דמיון בעזרת Elbow Method לקביעת מספר הקבוצות.
 - LDA : גלה נושאים במסמכים ובחר מספר נושאים בעזרת **Coherence Score**.



טרמינולוגיה

☐ Bag of Words - המרת טסקטים למספרים לעבודה מתמטית.

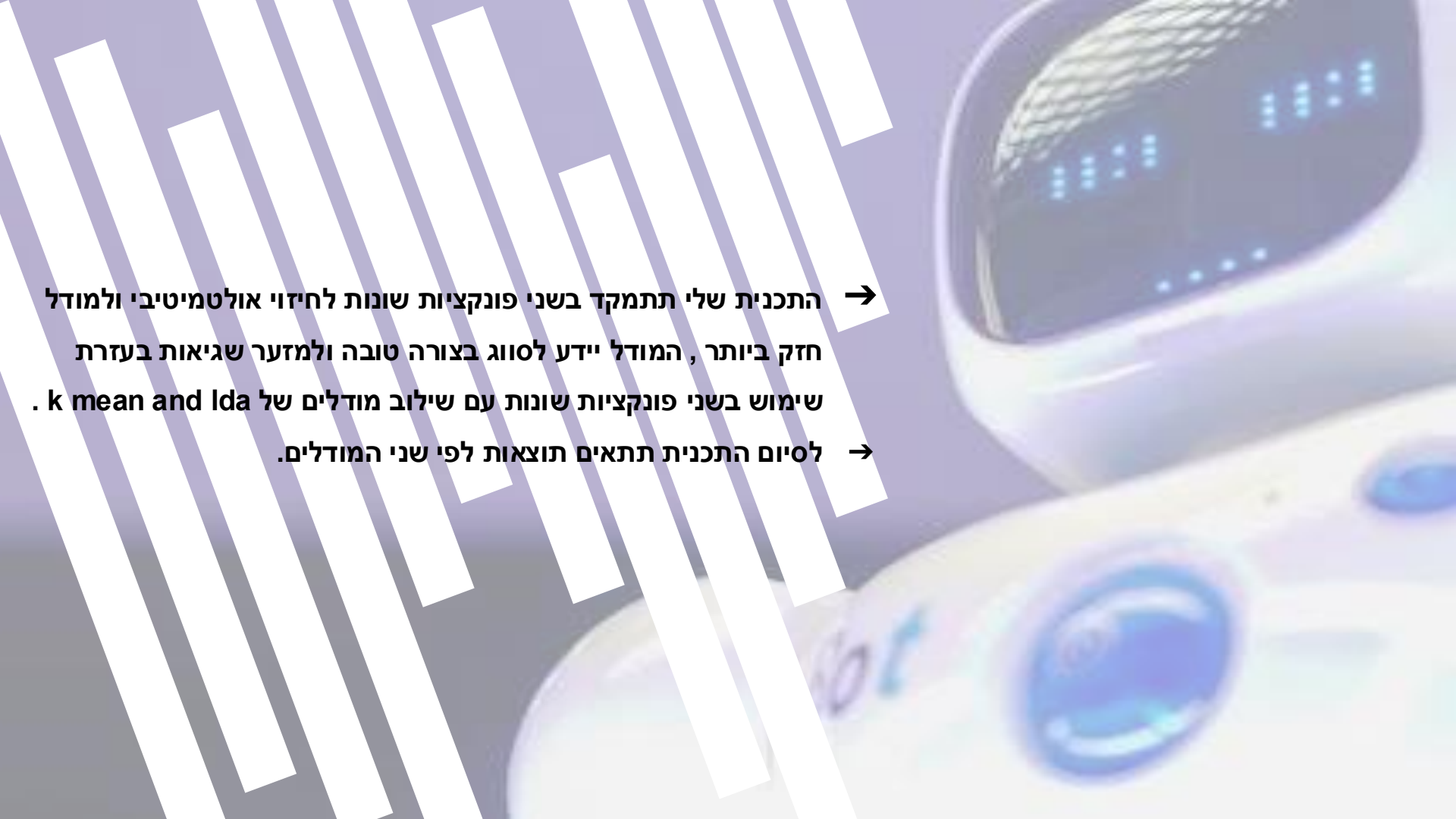
יתרונות – מתאים למודלים לינאריים ונאיב בייס

חסרונות – מתעלם מהקשרי המילים.

☐ Coherence Score – מודד קשר בין מילים בטקסט.

יתרונות – מעריך את איכות הנתונים ובוחר נושא בצורה אופטימלית.

חסרונות – תלות בנתונים , רגיש למספר נושאים , ממוקד רק במילים מרכזיות



→ התכנית שלי תתמקד בשני פונקציות שונות לחידוי אולטמיטיבי ולמודל חזק ביותר , המודל יידע לסווג בצורה טובה ולמזער שגיאות בעזרת שימוש בשני פונקציות שונות עם שילוב מודלים של k mean and lda .

→ לסיום התכנית תתאים תוצאות לפי שני המודלים.

אתגרים

- איכות נתונים
- מודל LDA ולימוד שלו
- התמודדות עם ערכים חסרים
- התמודדות עם שגיאה בסיווג ומדדים נמוך

impossible



MAYBE

מסקנות

- K-Means מתאים לסיווג ראשוני אך דורש שיפורים.
- הרחבת שימוש במודלים NLP מתקדמים תספק תוצאות מדויקות יותר.
- חשיבות ההתאמה בין מודלים לנתונים.
- המלצות להמשך מחקר.
- הבנת מגמות עתידיות בתחום.

