

תרגיל 1 : מידול דיאגרמות ישויות קשרים

תאריך הגשה : 23: 55, 22/05/2024

הוראות הגשה:

בתרגיל זה אתם נדרשים להגיש קובץ zip בודד שיכלול את הקבצים הבאים :

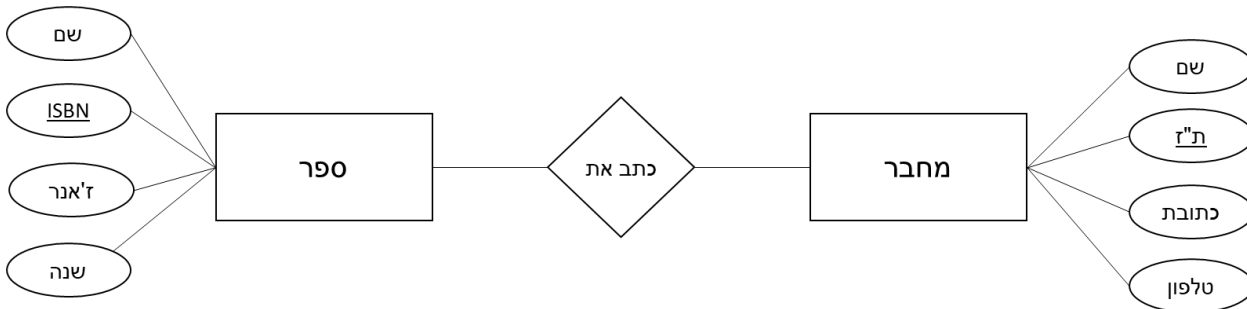
- ex1.pdf עם התשובות לשאלות להלן.
- create.sql
- drop.sql
- ex1.py
- README שמכיל שורה בודדת ובו ה-login של הסטודנט שמגיש את התרגיל. אם התרגיל מוגש בזוגות, על שורה זאת להכיל את שני ה-login מופרדים בפסיק.

שימו לב:

- התרגיל צריך להיות מוגש כ-PDF מוקלד.
- את החלקים בתרגיל שבהם אתם נדרשים לצייר דיאגרמות, תוכלו לצייר באופן ידני, לסרוק באיכות טובה, ולהדביק במקומות המתאימים בתוך ה-PDF של הפתרון.

שאלה 1:

נתונה דיאגרמה בסיסית של מסד נתונים שמכיל מידע אודות ספרים ומחברים.
לכל ספר יש שם, מספר מזהה ISBN, ז'אנר, ושנת הוצאה.
לכל מחבר יש שם, מספר ת"ז, כתובת וטלפון.
כמו כן, שומרים מידע על מי המחבר של איזה ספר.



בכל סעיף יש לצייר בדיאגרמה רק את המידע הנדרש באותו סעיף ולא להסתמך על סעיפים קודמים.

- (א) איך היית משנה את הדיאגרמה הבסיסית אם ידוע שלכל ספר יש בדיוק מחבר אחד? (בכל סעיף יש לצייר מחדש את הדיאגרמה)
- (ב) איך היית משנה את הדיאגרמה הבסיסית אם ידוע שיש ספרים המשתתפים בסדרה, ורוצים לשמור לכל ספר בסדרה את הספר הקודם לו? (ספרים שונים בסדרה יכולים להיכתב ע"י מחברים שונים)
- (ג) איך היית משנה את הדיאגרמה הבסיסית אם ידוע שקיימות חברות הוצאה לאור שונות, לכל הוצאה לאור יש שם, כתובת וטלפון. כמו כן, ידוע כל ספר יכול להיות מוצא לאור ע"י יותר מחברה אחת, בשנים שונות, ובכל הוצאה לאור, יש לספר מאייר אחד. כמו כן, למאייר יש שם, ת"ז כתובת וטלפון. ידוע גם שאותו מוציא לאור יכול להוציא את אותו ספר רק פעם אחת.

ד) איך היית משנה את הדיאגרמה הבסיסית אם ידוע שיש ספרים המיועדים לילדים, ויש פרסים הניתנים לספרי ילדים מוצלחים במיוחד. לכל פרס יש שם, וספר ילדים יכול לזכות בפרס לכל היותר בשנה אחת. יש לשמור גם את שנה בה הספר זכה בפרס.

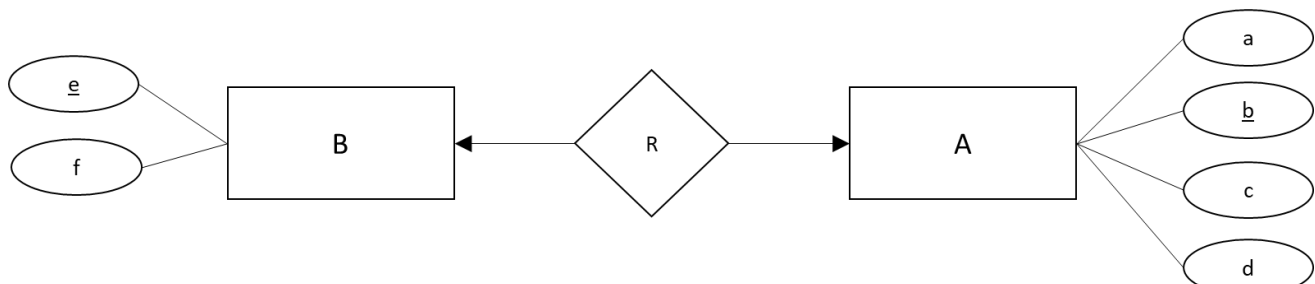
שאלה 2:

בכל סעיף:

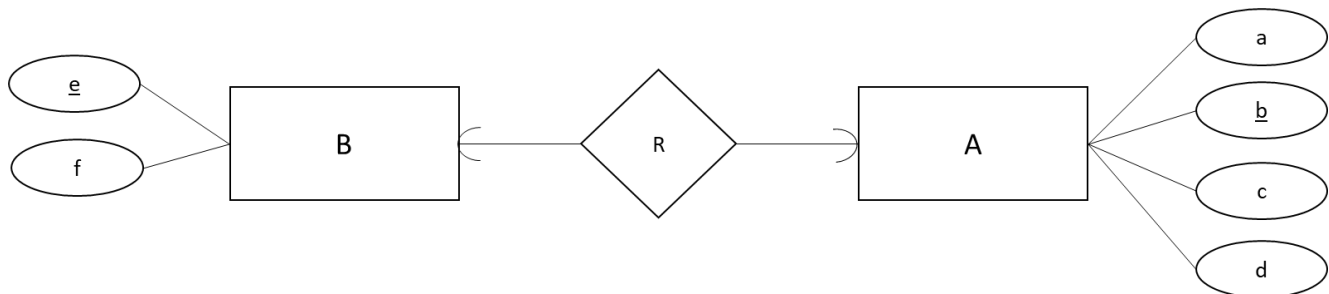
(i) יש לתרגם את הדיאגרמה ליחסים ולציין את השדות של כל יחס, ואת המפתחות. אם יש כמה אפשרויות למפתח, ציינו את כולן. אם יש ירושה (isA), תרגמו בשיטת E/R style.

(ii) נסמן ב-|A| את מספר הישויות בקבוצת הישויות A. מה ניתן לומר על מספר הישויות בקבוצת A לעומת מספר הישויות בקבוצת B? יש להתייחס לשתי קבוצות אלו בלבד ולהשתמש בסימנים: <, >, <=, >=, =. במקרה שלא ניתן לקבוע יש לציין "לא ניתן לקבוע".

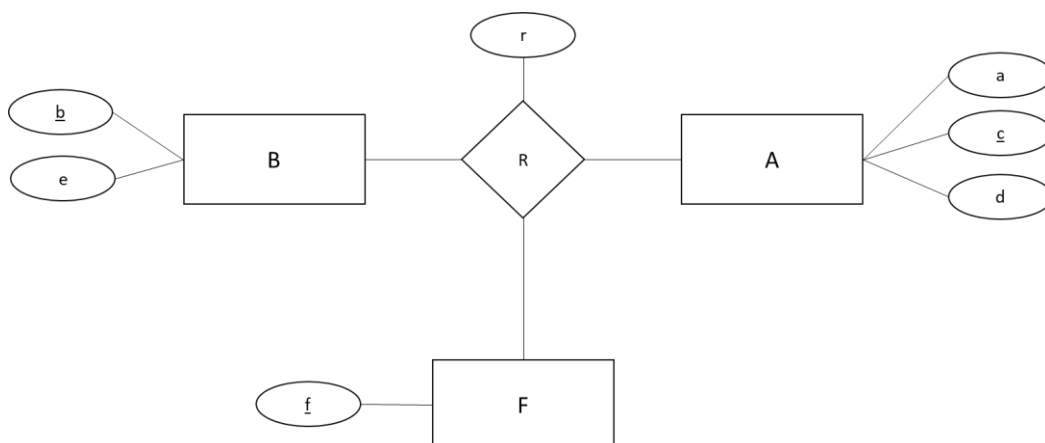
א.



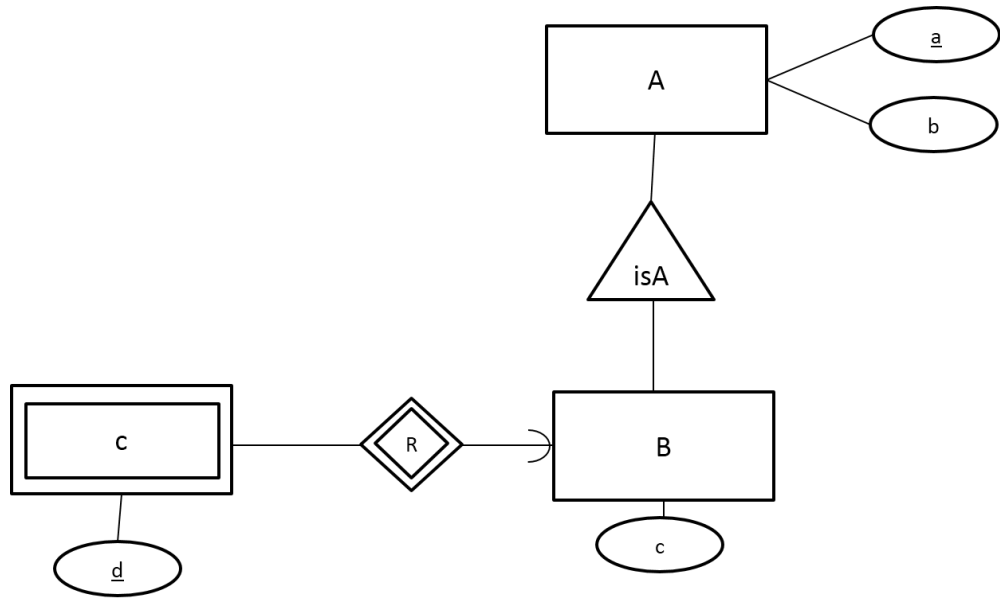
ב.



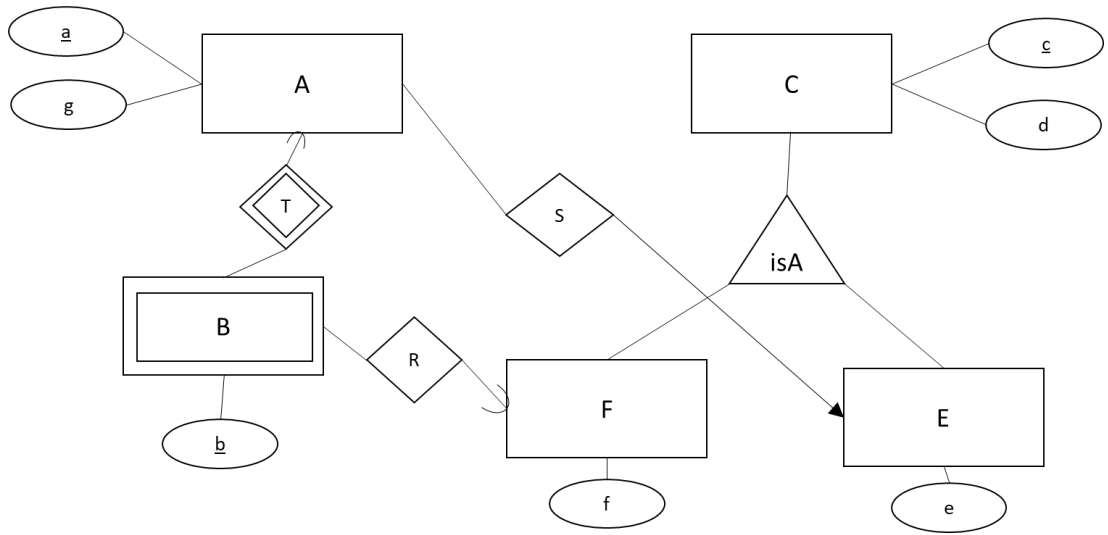
ג.



.7



.7



שאלה 3:

אתם נדרשים לתכנן ולבנות מסד נתונים עבור מאגר מידע אמיתי של נתוני מועמדות וזכייה בטקסי האוסקר. כדי להשיג את המידע תשתמשו באתר <http://www.kaggle.com>. האתר Kaggle מהווה קהילה למדעני דאטה ולמידה חישובית. היא מאפשרת למשתמשים למצוא ולפרסם מאגרי מידע, לבנות מודלים חישוביים ולעבוד עם מומחים אחרים כדי לגלות תובנות. מכיוון שכך, זה אתר שחשוב להכיר.

תרפרפו קצת באתר של Kaggle כדי לראות אלו סוגים של מידע אפשר למצוא. אנחנו נשתמש במידע של oscar-movies שנמצא בכתובת <https://www.kaggle.com/martinmraz07/oscar-movies>.

כאשר תסתכלו בדף תראו רק חלק מהשורות ומהעמודות. על מנת לראות את כל 30 העמודות, יש לבחור `select all` בלחיצה על מספר העמודות:

Detail

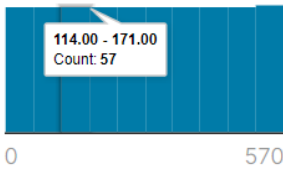
Compact

Column

10 of 30 columns

About this file

This data contains the Oscar Best Picture winners and nominees. Additionally, the data contains IMDB and Rotten Tomato ratings. The inspiration for this dataset is to eventually develop a classifier to identify winners of future Best Picture awards.

#	Film	Oscar Year	Film Studio/Produ..
Index	Title of Film	Year of award show. Earlier years were grouped together. Source: Wiki	Film Studio/Producer Film Source: Wiki
	564 unique values	1934 2% 1935 2% Other (547) 96%	Metro-Goldwyn-Ma... Warner Bros. Other (511)
0	Wings	1927/28	Famous Players-La

לצורך התרגיל שלנו, אנחנו לא נשתמש במידע הגולמי כמו שהוא מופיע באתר של Kaggle, מכיוון שכמו בהרבה מקרים, המידע באתר מכיל כל מיני טעויות שיכולות להקשות על ביצוע התרגיל. אנחנו נספק לכם העתק שעבר preprocess לצורך ניקוי של מידע שגוי. בנוסף, בקובץ שאנחנו נספק לכם, מופיעות רק חלק מהעמודות שמופיעות במידע במקורי כדי לפשט את הפתרון.

תוכלו למצוא את הקובץ מכוון בפורמט zip באתר של הקורס, וגם במערכת המחשבים במעבדה בתיקה:

~db2/data/ex1/oscars.zip

ניתן להעתיק אותו לתיקיה שלכם. הקובץ הזה מכיל טבלה אחת ענקית עם כל המידע על אירועי האוסקר, מועמדויות וזכיות.

בטבלה הנתונה מופיעות העמודות הבאות:

Index	מספר השורה, ניתן להתעלם מנתון זה בהמשך התרגיל
Film	שם הסרט
Oscar year	שנה שבה הסרט היה מועמד לזכייה באוסקר
Film studio/ producer	הסטודיו שהפיק את הסרט
Award	האם הסרט זכה באוסקר (Winner) או רק היה מועמד ולא זכה (Nominee)
Year of release	שנת ההוצאה של הסרט
Movie time	אורך הסרט בדקות
Movie genre	רשימת הז'אנרים אליהם שייך הסרט
IMDB rating	דירוג הסרט באתר IMDB
IMDB votes	מספר ההצבעות לדירוג הסרט באתר IMDB
Content rating	דירוג תוכן הסרט (G, PG, PG-13, R, NR)
Directors	הבמאי או במאים של הסרט
Authors	הכותב או הכותבים של התסריט
Actors	רשימת השחקנים שמשחקים בסרט
Film ID	מזהה ייחודי של הסרט

שימו לב : לצורך בניית הדיאגרמה ניתן להניח שכל המידע שאמור להופיע בטבלה קיים למרות שבנתונים עצמם חסר מידע.

במסד נתונים אמיתי לא כדאי לשמור את המידע בצורה כזאת, כי יש בו שדות שהם רשימת ערכים ולא ערך בודד, וזה מאוד מקשה על שליפת נתונים באמצעות שאילתות .
בתרגיל זה אנחנו ננסה למדל את הנתונים בצורה נכונה בעזרת דיאגרמת ER, ובהמשך לטעון את הנתונים לתוך הטבלאות הנגזרות מהדיאגרמה.

א) ציירו דיאגרמת ישויות קשרים מתאימה הממדלת את המידע בעמודות של הקובץ oscar.csv. מומלץ להוסיף תיאור מילולי של הדיאגרמה המכיל את כל הידע. אם הסתמכתם על הנחות שלא נאמרו במפורש, חובה לציין אותן. ייתכן שבדיאגרמה לא תצליחו למדל את כל ההנחות שמתקיימות בנתונים. במקרה כזה, ציינו אילו הנחות הדיאגרמה שלכם איננה ממדלת.

ב) תרגמו את כל הדיאגרמה ליחסים רלציוניים. לכל יחס ציינו את האטריביוטים שהם המפתח. אם יש מספר אפשרויות למפתח מספיק לבחור מפתח אחד.

את סעיפים א' וב' יש להגיש בקובץ ex1.pdf ביחד עם התשובות לשאלות 1 ו-2.

בחלק הבא תשתמשו במסד הנתונים Postgres ובקוד python כדי לבנות טבלאות ולטעון את הנתונים לתוך הטבלאות. הסבר על הגישה לחשבון משתמש שלכם במערכת Postgres מצורפת בסוף התרגיל.

שימו לב! יש לוודא שהקבצים שלכם רצים על מחשבי המעבדה. לא יינתנו נקודות לתשובות שנכשלות בטעינה לתוך מסד הנתונים.

(ג) בסעיף זה, תתנסו ביצירת טבלאות, טעינת נתונים ומחיקת טבלאות בעזרת קבצי עזר. **שימו לב: הסעיף הזה להתנסות בלבד. אין תוצר להגשה מסעיף זה.**

הורידו מאתר הקורס את הקבצים : ex1.py, create.sql, drop.sql :

- create.sql מכיל פקודה אשר יוצרת במערכת ה Postgres טבלה אחת בשם Oscars הזהה בצורתה לטבלה המקורים של המידע.
- drop.sql מכיל פקודה המוחקת את הטבלה הנ"ל.
- ex1.py מכיל קוד השולף מתוך קובץ המידע המכוון (תחת השם oscar.zip) את שורות המידע, וכותב אותן לתוך קובץ חדש, oscar.csv ע"י שימוש בפונקציה process_file. שימו לב – קובץ זה רץ באמצעות python3 ומעלה בלבד (במחשבי המעבדה השתמשו בפקודה python3 כדי להריצו).

כעת :

- הריצו את הקוד בקובץ ex1.py וודאו שנוצר לכם הקובץ oscar.csv.
- התחברו למערכת postgres מתוך התיקיה שבו שמרתם את כל הקבצים על ידי הפקודה : (ההוראות המצורפות בסוף התרגיל, אבל גם רשומות כאן באופן חלקי לנוחיותכם.)

```
psql -h dbcourse public
```

- הריצו את הקובץ create.sql ליצירת הטבלה Oscars בעזרת הפקודה

```
\i create.sql
```

- התנתקו מהמערכת בעזרת הפקודה

```
\q
```

- טענו את הנתונים לתוך הטבלה שיצרתם בעזרת הפקודה

```
cat oscar.csv | psql -h dbcourse public -c "copy oscar from STDIN DELIMITER ',' CSV HEADER"
```

אחרי הרצת הפקודה הפלט אמור להיות :

```
COPY 571
```

כלומר 571 שורות הועתקו לתוך הטבלה.

- התחברו שוב למערכת postgres והריצו את השאילתה הבא המחזירה את כל השורות בטבלה שיצרתם :

```
SELECT COUNT(*) FROM Oscars;
```

השאילתה מחזירה את מספר השורות בטבלה enrollment , כך תוודאו שאכן הנתונים נטענו לטבלה כראוי.

- הריצו את הקובץ drop.sql כדי למחוק את הטבלה

```
\i drop.sql
```

ד) כעת אתם נדרשים לעדכן את הקבצים `create.sql`, `drop.sql` כך שייצרו את הטבלאות המתאימות ליחסים שהגדרתם בסעיף ב. ניתן לשנות מעט את הגדרות הטבלאות על מנת לנצל את תכונות מסד הנתונים (למשל, המסד מאפשר ערכי null). אם בסעיף זה בחרתם ליצור טבלאות שונות מאלו שהגדרתם בסעיף ב, הוסיפו בקובץ `ex1.pdf` הסבר עבור השינויים שבחרתם לעשות. שימו לב שבפועל חלק מהמידע שהנחנו שקיים בשלב בניית הדיאגרמה חסר בנתונים, ויש עמודות שמופיעים בהן ערכי null.

- כתבו פקודות `create table` בתוך הקובץ "`create.sql`" היוצרות את הטבלאות שלכם. בפתרון וודאו שכללתם את כל התנאים והמגבלות (`key`, `foreign key`, `check`, etc). שיכולות להיות מוגדרות על הטבלאות. אתם יכולים להניח שכל נתון טקסטואלי הוא באורך מקסימלי 100.
- כתבו פקודות `drop table` בקובץ "`drop.sql`" שמוחקות את כל הטבלאות שייצרתם.

התחברו למערכת `postgres` וודאו שהפקודות שלכם רצות ללא הודעות שגיאה.

ה) לבסוף, בסעיף זה אתם נדרשים לשנות את הקוד בקובץ `ex1.py` כך שיפצל את המידע לקבצים שונים, בהתאם להגדרות הטבלאות שלכם.

סיפקנו לכם בקוד פונקציה בשם `split_list_value()` המקבלת ערך בעמודה המכילה רשימות, למשל `directors`, ומחזירה `list` של כל השמות ברשימה (על בסיס הסימן `&&` שהתווסף בשלב העיבוד המקדים שסופק).

בקובץ `ex1.py`, אתם נדרשים לבצע את השלבים הבאים:

- עבור כל טבלה צרו קובץ עם סיומת `csv` הנקרא באותו שם כמו הטבלה. יש להקפיד על שם זהה כולל אותיות גדולות וקטנות באנגלית.
 - עדכנו את הפונקציה `process_file` כך שתרושום את המידע הרלוונטי מכל שורה לתוך קבצי ה-`csv` של הטבלאות השונות. הקפידו לסגור את כל הקבצים שפתחתם בקוד אחרי שאתם מסיימים לכתוב אליהם!
 - עדכנו את הפונקציה `get_names` כך שתחזיר רשימה עם שמות כל הטבלאות שהגדרתם. השמות צריכים להיות תואמים גם לשמות טבלאות שהגדרתם בסעיף ד, וגם לשמות קבצי ה-`csv` שהגדרתם בקוד.
- שימו לב:** יש להחזיר את שמות הטבלאות לפי הסדר הנכון לטעינת נתונים. כלומר, אם יש טבלה A עם אילוף מפתח זר לטבלה אחרת B, יש להחזיר קודם את B ורק אח"כ את A ברשימה.

שימו לב! המידע בקבצי ה-`csv` שאתם מייצרים צריך להופיע בלי שורות שחוזרות על עצמם כדי שניתן יהיה לטעון את הנתונים באופן תקין לטבלאות מבלי להפר אילוצי מפתח.

כעת, תבדקו שניתן לטעון את הנתונים לכל אחד מהטבלאות בהצלחה. כלומר, תייצרו שוב את הטבלאות. הריצו פקודה של טעינת שורות עבור כל אחד מהטבלאות, לפי אותו סדר שהחזרתם בפונקציה `get_names`. תוודאו, על ידי שאילתות, שהנתונים נכנסו כראוי. לבסוף תמחקו את הטבלאות.

יש להגיש את הקבצים `create.sql`, `drop.sql`, `ex1.py` בתוך zip ההגשה שלכם.

Appendix: Using Postgres

You can access your database account with the command:

```
psql -h dbcourse public
```

in the computer labs. After running this command, you can enter queries and DDL commands directly into the command line prompt.

In this exercise it will be more useful for you to write your create and drop table commands in a file, and then this file can be loaded into the database for execution. To do so, use the command

```
\i a.sql
```

within the prompt of the database, assuming your commands are in the file “a.sql”. Some other useful commands are:

- \q exit psql
- \h [command] help about ‘command’
- \d [name] describe table/index/... called ‘name’
- \dt list tables