מערכת לאגירת חדשות

במטלה זו ישנו דגש על OOP הפרדה למחלקות וירושות.
בחרתי להשתמש בתבנית עיצוב MVC (Model View Controller).
הסבר על MVC:

תבנית זאת מתחלקת לשלושה חלקים: Model, View, Controller ובכך מפחיתה את התלות בין ממשק המשתמש לשאר חלקי התוכנה.

1. Model-משתמש כbackend שמכיל את כל הלוגיקה של הנתונים.
2. View- משמש כחזית או ממשק המשתמש הגרפי.
3. Controller- משמש כ"מוח" של המערכת הוא שולט על אופן הצגת הנתונים.

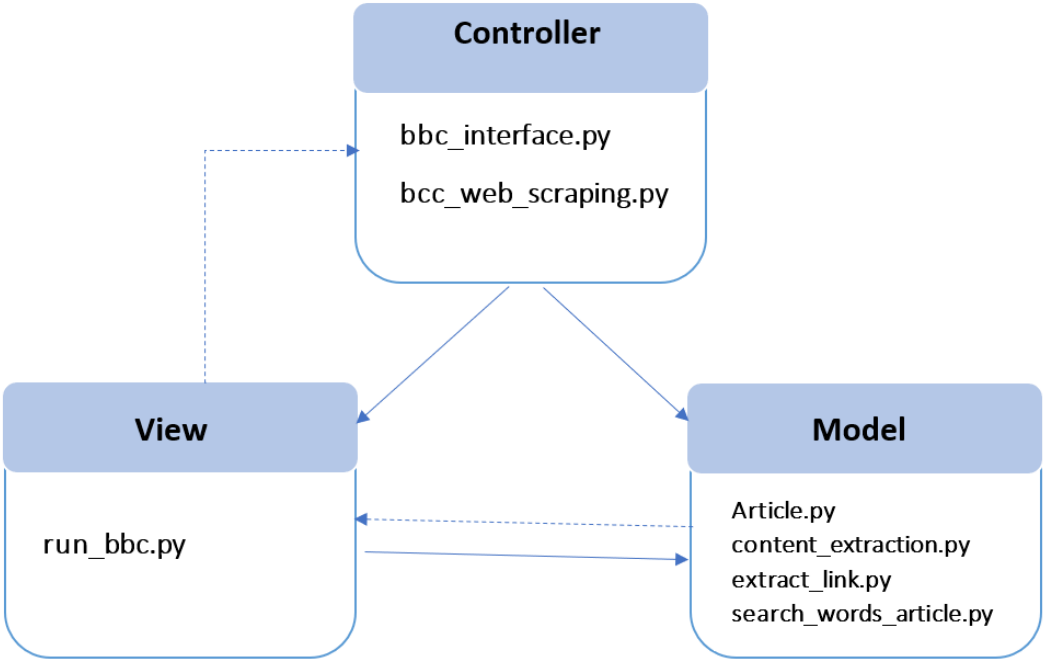
מערכת כתבות BBC

תחילת התמקדתי בפתרון בבניית מערכת המורידה חומרי גלם מאתר:

https://www.bbc.com/

אתר זה מכיל כתבות בנושאים מרובים: חדשות, ספורט, תזונה ועוד..

התבנית MVC מתבטאת במערכת של האתר BBC באופן הבא:



הסבר על המחלקות:

המחלקה Scraping_BBC הנמצאת בקובץ: bbc_web_scraping.py:

היא המחלקה הראשית במימוש המערכת על אתר Bbc.

מחלקה זו מממשת את הממשק bbc_interface.py ויורשת webdriver.Chrome.

הפונקציות במחלקה עוסקות במימוש תתי משימות המפורטות בדרישות המטלה:

הורדת המאמרים מהאתר, אגירת חומרי גלם וחיפוש מילה או קומבינציה של מילים במאמרים.

להלן הפונקציות המחלקה:

א. land_page – טוענת בdriver את האתר BBC.

ב. Install_articles –
פלט- קבצי JSON המכילים את כל חומרי הגלם מהמאמרים.
ראשית, הפונקציה שומרת ברשימה את כל הקישורים לכל הכתבות הנמצאים באתר.
לאחר מכן היא עוברת על כל קישור נכנסת לכתבה ולוקחת את חומרי הגלם.
(חומרי הגלם שבחרתי הם: כותרת, קישור , ותוכן הכתבה)

ג. Search_from_articles-
פלט- מדפיסה טבלה המכילה את שם הכתבה והלינק שבו נמצאת המילה או קומבינציה של מילים.
נעבור על כל קבצי הJSON (כל קובץ = כתבה) ונחפש אם קיים בו המילה/צירוף אם כן נוסיף לרשימה.
לבסוף נדפיס את הרשימה.

הקובץ constants.py:

<u>קושי</u>- יתכן אי סינכרון בין הניתוב על המחשב שלכם לשלי.
ניסיתי לפתור אך ללא הצלחה.
הוא קובץ המכיל את כל הניתובים לתיקיות, דרייברים ואתרים.

המחלקה ExtractLink הנמצאת בקובץ: bbc_web_scraping.py:

מחלקה זו אחראית על חילוץ כל קישורי הכתבות הנמצאות בדף הראשי של האתר והשימוש בה נעשה רק במחלקה הראשית **Scraping_BBC**.

<u>קושי:</u> היה עלי לחפש תג html אשר זהה בין כל קישורי הכתבות.
תחילה חשבתי שמצאתי ואז כשספרתי כמה כתבות יש בפועל באתר וכמה קישורים מצאתי ראיתי שכתבות REEL עם תג שונה ולכן בקוד ניתן לראות שהתייחסתי לשניהם.

הבנאי שלה מקבל את הDriver של אתר הבית אשר בעזרתו נחלץ את התג לקישורי הכתבות.

פונקציה אחת שמומשה: pull_all_links_articles – אשר עוברת על "הילדים" של התג ששמרנו כאובייקט המחלקה ומחפשת את התכונה 'href' אשר ידוע בhtml שמסמל קישור.

המחלקה ContentExtraction הנמצאת בקובץ:
content_extraction.py:

מחלקה זו אחראית על הורדת חומרי הגלם והשימוש בה נעשה רק במחלקה הראשית **Scraping_BBC**.

<u>קושי:</u> נתקלתי בכך שלא כל קבצי הJSON נשמרו טוב.
לאחר הבנה מדוע דווקא כתבות ספציפיות לא נשמרו טוב מצאתי שיש תווים כמו: ? גורמים לבעיה.

<u>להלן פונקציות המחלקה:</u>

א. pull_all_information_articles-
פונקציה זו מושכת את המידע מדף הHTML בעזרת תגיות

המשותפות לקטגוריות הכתבה.
לדוגמא, מצאתי שכל הכתבות מסוג News יש להם תג המשותף להן וכך שלפתי את תוכן המאמר.

ב. extract_article_content-
פונקציה זו נקראת ע"י הפונקציה מסעיף א.
אנחנו עוברים שורה אחר שורה ומחפשים את התג עם התכונה של טקסט משרשרים את כל תוכן הכתבה לאובייקט string.

ג. is_exit-
לאחר הפעלה חוזרת של המערכת איננו רוצים שישמרו כפילויות של קבצי JSON לכן לא נוודא שלא נוריד כתבה שכבר קיימת לנו במאגר.

ד. write_to_json-
פונקציה זו פותחת את קובץ הJSON ומגדירה בו את שם הכתבה, הלינק ותוכן הכתבה.

ה. remove_unnecessary_letters_from_name-
פונקציה זו פותרת את הקושי שציינתי למעלה.

המחלקה SearchWords הנמצאת בקובץ: search_words_article.py:

מחלקה זו אחראית על חיפוש מילה/צירוף בכל הכתבות שהורדנו והשימוש בה נעשה רק במחלקה הראשית **Scraping_BBC**.

<u>פונקציות המחלקה:</u>

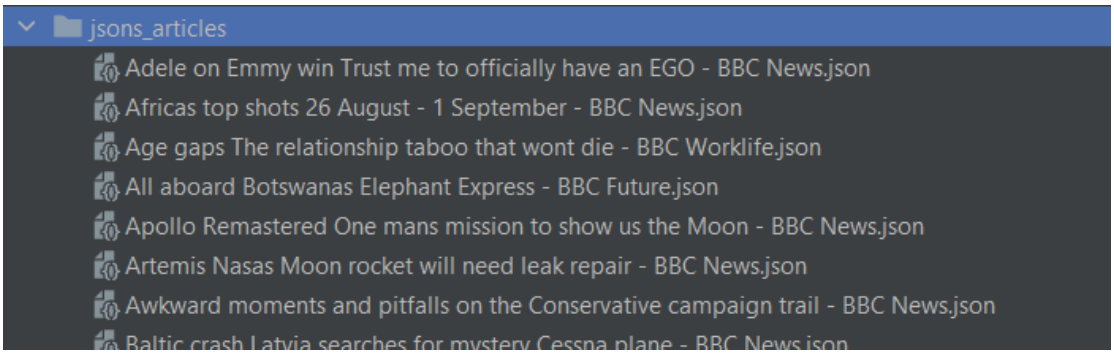
א. print_data-
פונקציה זו מדפיסה טבלה מסודרת עם שמות העמודות שהגדרתי מראש.

ב. is_exit_in_art-
פונקציה שבודקת האם מילה/צירוף של מילים נמצא בטקסט מסויים.

ג. search_in_all_articles-
פונקציה זו עוברת על כל הקבצים המסתיים בסיומת של json.
בתיקייה ספציפית ובודקת האם קיים טקסט המכיל מילה מסויימת/ צירוף.
אם כן, נשמור את כל הקבצים המכילים זאת באובייקט מסוג רשימה.

צילומי מסך ודוגמאות הרצה:

לאחר הרצה נשמרו כל הכתבות בתיקיה: jsons_articles שם הקובץ הוא שם הכתבה.



התוכן שנשמר הוא: שם הכתבה, קישור לכתבה ותוכן הכתבה.

לאחר הרצה של פונקציית החיפוש על המילים הבאות:

1. USA

Article Name	Article Link
Age gaps The relationship taboo that wont die - BBC Worklife	https://www.bbc.com/worklife/article/20220317-age-gaps-the-relationship-taboo-that-wont-die
All aboard Botswanas Elephant Express - BBC Future	https://www.bbc.com/future/article/20220816-the-botswana-buses-tackling-human-elephant-conflict
Australian artist removes Ukraine and Russia mural after backlash - BBC News	https://www.bbc.com/news/world-australia-62791089
Chinese motorcyclists join Chongqing wildfire battle - BBC News	https://www.bbc.com/news/world-asia-china-62770119
In pictures The life of ex-Soviet leader Mikhail Gorbachev - BBC News	https://www.bbc.com/news/world-europe-62732829
New prime minister Seven big questions for the new leader - BBC News	https://www.bbc.com/news/uk-politics-62769937
Tamil Nadu India womans food paintings you will want to eat - BBC News	https://www.bbc.com/news/world-asia-india-62763005
The spongy cities of the future - BBC Future	https://www.bbc.com/future/article/20220823-how-auckland-worlds-most-spongy-city-tackles-floods
UK looks to Sweden for a solution to nuclear waste - BBC News	https://www.bbc.com/news/business-62677534
Why its time to talk about poo - BBC Future	https://www.bbc.com/future/article/20220830-the-new-science-of-recycling-human-poo

2. UK businesses

Article Name	Article Link
Gas prices soar 26% after Russia keeps key pipeline closed - BBC News	https://www.bbc.com/news/business-62789675

3. a

Article Name	Article Link
Adele on Emmy win Trust me to officially have an EGO - BBC News	https://www.bbc.com/news/entertainment-arts-62793207
Africas top shots 26 August - 1 September - BBC News	https://www.bbc.com/news/world-africa-62751874
Age gaps The relationship taboo that wont die - BBC Worklife	https://www.bbc.com/worklife/article/20220317-age-gaps-the-relationship-taboo-that-wont-die
All aboard Botswanas Elephant Express - BBC Future	https://www.bbc.com/future/article/20220816-the-botswana-buses-tackling-human-elephant-conflict
Apollo Remastered One mans mission to show us the Moon - BBC News	https://www.bbc.com/news/science-environment-62662685
Artemis Nasas Moon rocket will need leak repair - BBC News	https://www.bbc.com/news/science-environment-62758482
Australian artist removes Ukraine and Russia mural after backlash - BBC News	https://www.bbc.com/news/world-australia-62791089
Awkward moments and pitfalls on the Conservative campaign trail - BBC News	https://www.bbc.com/news/uk-politics-62753199
Baltic crash Latvia searches for mystery Cessna plane - BBC News	https://www.bbc.com/news/world-europe-62789877
Boris Johnsons next move Making millions or a comeback - BBC News	https://www.bbc.com/news/uk-politics-62547853
Canada stabbings Police hunt suspects after killing spree in Saskatchewan - BBC News	https://www.bbc.com/news/world-us-canada-62790980
Chile constitution Voters overwhelmingly reject radical change - BBC News	https://www.bbc.com/news/world-latin-america-62792025
China quake Deadly tremor rocks Sichuan city in lockdown - BBC News	https://www.bbc.com/news/world-asia-china-62764223
Chinese motorcyclists join Chongqing wildfire battle - BBC News	https://www.bbc.com/news/world-asia-china-62770119
Crypto.com pulls out of Uefa Champions League deal - BBC News	https://www.bbc.com/news/technology-62764654
Flexible firms take break from fixed bank holidays - BBC News	https://www.bbc.com/news/business-62491321
Foo Fighters Girl, 12, drums at Taylor Hawkins memorial gig - BBC News	https://www.bbc.com/news/uk-england-suffolk-62793175
Gas prices soar 26% after Russia keeps key pipeline closed - BBC News	https://www.bbc.com/news/business-62789675
How to stay cool the Japanese way - BBC News	https://www.bbc.com/news/business-62659199
In pictures The life of ex-Soviet leader Mikhail Gorbachev - BBC News	https://www.bbc.com/news/world-europe-62732829
Japans female bosses mapping a course for other women - BBC News	https://www.bbc.co.uk/news/business-62719779
Legionnaires suspected cause of Argentina pneumonia deaths - BBC News	https://www.bbc.com/news/world-latin-america-62785348
[illegible] 2022 Meet the esteemed film festival jury - BBC Culture	https://www.bbc.com/culture/article/20220826-[illegible]-2022-meet-the-esteemed-film-festival-jury

מערכת לוח טיסות נתב"ג

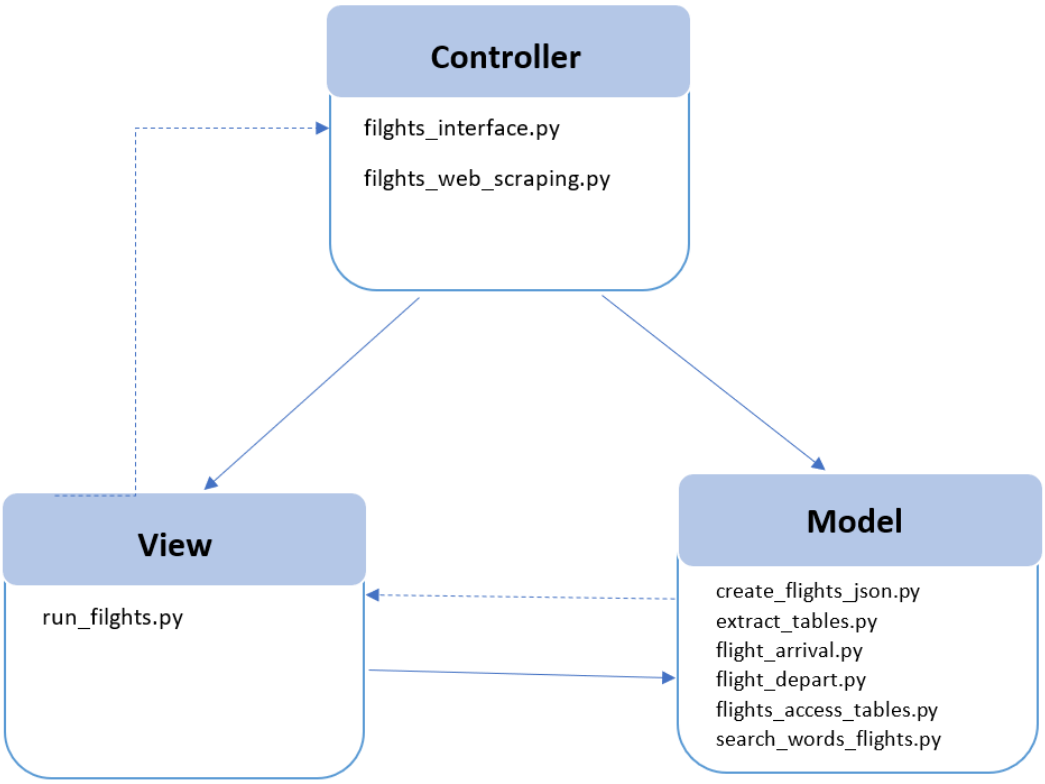
שנית עסקתי בבניית מערכת המורידה חומרי גלם מאתר:

http://www.iaa.gov.il/he-IL/airports/BenGurion/Pages/OnlineFlights.aspx

יש לציין שאתר זה נותן מידע על טיסות אשר יוצאות/נכנסות בשדה תעופה בן גוריון ישראל.

בחלק זה התבקשנו לאסוף את נתוני הטיסות בReal-Time, נעזרתי באתר הבא: לינק

התבנית MVC מתבטאת במערכת של האתר הנ"ל באופן הבא:



הסבר על המחלקות:

המחלקה Scraping_Flights הנמצאת בקובץ:
flights_web_scraping.py:

היא המחלקה הראשית במימוש המערכת על אתר הנ"ל.

מחלקה זו מממשת את הממשק flights_interface.py ויורשת webdriver.Chrome.

הפונקציות במחלקה עוסקות במימוש תתי משימות המפורטות בדרישות המטלה:

הורדה של נתוני לוח טיסות בזמן אמת, אגירת חומרי גלם וחיפוש מילה או קומבינציה של מילים במידע של הטיסות שאגרנו.

<u>להלן הפונקציות המחלקה:</u>

א. land_page – טוענת בdriver את האתר טיסות בבן גוריון.
ב. install_flights-
פלט- קבצי JSON המכילים את כל חומרי הגלם מלוח הטיסות.
ניתן לראות כי בGUI יש שני קטגוריות לטיסות (נחיתות והמראות) ובנוסף הטבלה לא מוצגת במלואה וישנו כפתור אז פותח עוד ועוד פרטים על טיסות.
אני רוצה לאגור את **כל** נתוני הטיסות בלוח הטיסות לכן עלי לפתוח את לוח הטיסות בשני הקטגוריות ולהציגו במלואו.
לכן, ראשית התחלתי עם לוח הנחיתות פתחתי עד סוף את הרשימה.
לאחר מכן, עצרתי את ה"עדכון אוטומטי" מאחר שבזמן אגירת פרטי הטיסה אינני רוצה לפספס שום פרטים שהתעדכנו.
וכך עשיתי גם בלוח ההמראות.
לבסוף, אספתי את כל המידע על כל טיסה הנמצאת בלוח הטיסות.
ג. search_from_flights-
פלט- מדפיסה טבלה המכילה את שם חברת התעופה, מס טיסה, יעד, מס טרמינל, זמן הגעה, זמן עדכני וסטטוס שבו נמצאת המילה או קומבינציה של מילים.
נעבור על כל קבצי הJSON (כל קובץ = פרטי טיסה) ונחפש אם קיים בו המילה/צירוף אם כן נוסיף לרשימה.

לבסוף נדפיס את הרשימה.

המחלקה AccessExtractTables הנמצאת בקובץ:
flights_access_tables.py:

מחלקה זו אחראית על חילוץ כל פרטי הטיסה בלוח הטיסות של נתב"ג והשימוש בה נעשה רק במחלקה הראשית **Scraping_Flights**.

הבנאי שלה מקבל את הDriver של אתר הבית אשר בעזרתו נחלץ את התג לכפתורים שנרצה ללחוץ עליהם ("הצג תוצאות נוספות" ,"עצור עדכון אוטומטי")

<u>להלן פונקציות המחלקה:</u>

א. <u>get_tables-</u>
פונקציה זו היא הפונקציה הראשית אשר אחראית על חילוץ חומרי הגלם.

ב. <u>get_arrival_table-</u>
ראינו כי לוח הטיסות בעמוד הראשי קצר.
על מנת לראות את כל הטיסות עלינו ללחוץ על כפתור שפותח עוד ועוד פרטים על טיסות.
לכן פונקציה זו נועדה לפתוח טבלת הטיסות הנכנסות.
בנוסף, עצרתי את הלוח מלהתעדכן כדי שבזמן אגירת המידע לוח הטיסות ישתנה.

ג. <u>get_depart_table-</u>
באופן זהה לפונקציה מסעיף ב רק שפה אנחנו צריכים ללחוץ על כפתור אשר מציג את לוח הטיסות היוצאות ולהמתאים שיפתח.

המחלקה ExtractTables הנמצאת בקובץ: extract_tables.py:

מחלקה זו אחראית על הורדת חומרי הגלם והשימוש בה נעשה רק

במחלקה **AccessExtractTables**.
הבנאי שלה מקבל את הDriver של אתר הבית אשר בעזרתו נחלץ את כל המידע על הטיסה בעזרת התגים המתאימים.

<u>להלן פונקציות המחלקה:</u>

א. extract_flights-
הפונקציה הראשית של המחלקה אשר משתמשת בפונקציות עזר שאפרט מטה.
אחראית על הוצאת המידע מהטיסות ושולחת למחלקה שתפקידה לשמור את המידע שהוצאנו בקבצי JSON.

ב. pull_flight_arrival_information/ depart -
פונקציה זאת עוברת שורה אחר שורה בטבלת הטיסות הנכנסות והיוצאות ושולחת לפונקציית עזר שתשלוף את הטקסט מהשורה.

ג. get_arrival_row_data/ depart -
מכניסים לאובייקט טיסה נכנסת/יוצאת את המידע שנשלוף בעזרת פונקציות הGET.

ד. פונקציות Get למיניהם-
פונקציות אלו עוברת על לוח הטיסות בדף הHTML ולפי תגים שבהם נמצאים הטקסט מחזירות את המידע הרצוי.

המחלקה CreateJson הנמצאת בקובץ: create_flights_json.py:

מחלקה זו אחראית על שמירת נתוני הטיסות בקבצי ג'ייסון.
והשימוש בה נעשה רק במחלקה ExtractTables.

הבנאי שלה מקבל רשימה של טיסות וסוג טיסה (המראה או נחיתה).

<u>להלן פונקציות המחלקה:</u>

א. <u>write_arrival_to_json / write_depart_to_json –</u>
מגדירה את המבנה שבו ארצה שהקובץ JSON יבנה (key:value).

ב. <u>create_file-</u>
בפונקציה זו נפתח קובץ JSON שאת שמו בחרתי להיות מספר הטיסה.

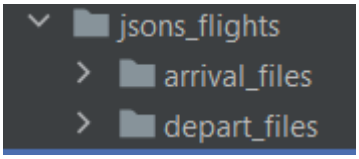
המחלקה SearchWordsFlights הנמצאת בקובץ:
search_words_article.py:

מחלקה זו אחראית על חיפוש מילה/צירוף בכל פרטי הטיסות שהורדנו והשימוש בה נעשה רק במחלקה הראשית **Scraping_Flights**.

מחלקה זאת דומה למחלקת חיפוש הכתבות BBC השוני היחידי הוא שפה נעבור על כל העמודות בקובץ הJSON של טיסה.

<u>צילומי מסך ודוגמאות הרצה:</u>

לאחר הרצה נשמר תוכן לוח הטיסות בתיקיה: jsons_flights שם הקובץ הוא מספר טיסה.
התוכן שנשמר הוא: שם חברת תעופה, מספר טיסה, מאיזה/לאיזה מדינה, מספר טרמינל, זמן מתוכנן, זמן עדכני וסטטוס טיסה .



קבצי הJson מסודרים באופן הבא:

לאחר הרצה של פונקציית החיפוש על המילה: "Blue" קיבלנו-

1. נחיתות

Airline Name	Flight Num	From -> Israel	Terminal	schedule_time	updated_time	Status
blue air aviation	0B 170	בוקרשט	3	05/09 22:55	23:25	בזמן
blue bird airways	BZ 112	לרנקה	3	06/09 09:20	09:20	בזמן
blue bird airways	BZ 702	אתונה	3	06/09 09:20	09:20	בזמן
blue bird airways	BZ 752	הרקליון	3	06/09 08:40	08:40	בזמן
blue bird airways	BZ 754	הרקליון	3	06/09 14:00	14:00	בזמן

2. המראות

Airline Name	Flight Num	Israel -> To	Terminal	schedule_time	updated_time	Status
blue air aviation	0B 169	בוקרשט	3	05/09 21:55	22:45	סופי
blue bird airways	BZ 111	לרנקה	3	06/09 13:30	13:30	לא סופי
blue bird airways	BZ 701	אתונה	3	06/09 15:30	15:30	לא סופי
blue bird airways	BZ 753	הרקליון	3	06/09 13:00	13:00	לא סופי
blue bird airways	BZ 777	סנטוריני	3	05/09 19:30	20:53	נחתה