Ενδεικτικά Τρεξίματα Κώδικα :

Η συνάρτηση scrape_wikipedia_articles αντλεί δεδομένα από τις παραγράφους και τους πίνακες των wiki ιστοσελίδων.

Κλήση συνάρτησης : scrape_wikipedia_articles(start_urls)

Αποτέλεσμα:

```
In [30]: runfile('C:/Users/xristos/Desktop/ΠΛΗΡΟΦΟΡΙΚΗ/ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ/main ex.py',
wdir='C:/Users/xristos/Desktop/ΠΛΗΡΟΦΟΡΙΚΗ/ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ')
Scraping https://en.wikipedia.org/wiki/Dot-com_bubble
Scraping https://en.wikipedia.org/wiki/Real_estate_investment_trust
Scraping https://en.wikipedia.org/wiki/2007%E2%80%932008_financial_crisis
Scraping https://en.wikipedia.org/wiki/COVID-19_recession
Scraping https://en.wikipedia.org/wiki/Financial_crisis
Scraping https://en.wikipedia.org/wiki/S%26P_500
Scraping https://en.wikipedia.org/wiki/List_of_S%26P_500_companies
Scraping https://en.wikipedia.org/wiki/Nasdaq
Scraping https://en.wikipedia.org/wiki/Cryptocurrency
Scraping https://en.wikipedia.org/wiki/Energy_industry
Scraping https://en.wikipedia.org/wiki/Goldman_Sachs
Scraping https://en.wikipedia.org/wiki/Stock_market_crash
Scraping https://en.wikipedia.org/wiki/Category:Real_estate_companies_of_the_United_States
Scraping https://en.wikipedia.org/wiki/Blockchain
Scraping https://en.wikipedia.org/wiki/Bitcoin
Saved 15 articles to wikipedia_articles.json
```

Αν ανοίξουμε το αρχείο wikipedia_articles.json θα δούμε ότι είναι αποθηκευμένα όλα τα άρθρα.


```
"title": "Dot-com bubble",
"content": "\n The dot-com bubble (or dot-com boom) was
"url": "https://en.wikipedia.org/wiki/Dot-com bubble",
"tables": []
```


Η συνάρτηση preprocess_articles("wikipedia_articles.json") “καθαρίζει” όλους τους όρους από ειδικούς χαρακτήρες και δημιουργεί το αρχείο original_articles.json. Το τελευταίο περιέχει όλες τις λέξεις στην αρχική τους ρίζα όπως φαίνεται παρακάτω.

Εκτέλεση Κώδικα:

```
Processing article: Dot-com bubble
Processing article: Real estate investment trust
Processing article: 2007–2008 financial crisis
Processing article: COVID-19 recession
Processing article: Financial crisis
Processing article: S&P 500
Processing article: List of S&P 500 companies
Processing article: Nasdaq
Processing article: Cryptocurrency
Processing article: Energy industry
Processing article: Goldman Sachs
Processing article: Stock market crash
Processing article: Category:Real estate companies of the United States
Processing article: Blockchain
Processing article: Bitcoin
Saved preprocessed articles to original_articles.json
```

Αρχείο original_articles.json:


```
"original_words": "bubble boom stock market bubble ballooned peaked friday march period market growth coincided
```


Η συνάρτηση build_inverted_index() δημιουργεί το αρχείο inverted_index.json το οποίο περιέχει κάθε όρο και το id των άρθρων που παρουσιάζονται.


```
"case": [
    0,
    2,
    3,
    8,
    10,
    13,
    14
],
"More": [
    0,
    1,
    4,
    13
],
"there": [
    0,
    1,
    2,
    3,
    4,
    7,
    8,
    9,
    10,
    11,
    13,
    14
],
"20%": [
    0,
    2,
    3,
    8,
    10,
    11,
    14
],
"similar": [
    0,
    1,
    3,
    4,
    8,
    13,
    14
],
```


Οι συναρτήσεις calculate_tf και calculate_idf δημιουργούν τα αρχεία TF.json και IDF.json αντίστοιχα.

Περιεχόμενο TF.json:


```
"0": {
    "The": 0.006780870806566738,
    "dot-com": 0.009635974304068522,
    "bubble": 0.0035688793718772305,
    "(or": 0.00035688793718772306,
    "boom)": 0.00035688793718772306,
    "was": 0.006067094932191292,
    "a": 0.017844396859386154,
    "stock": 0.005710206995003569,
    "market": 0.0053533190578158455,
    "that": 0.007494646680942184,
    "ballooned": 0.00035688793718772306,
    "during": 0.0024982155603140615,
    "the": 0.053176302640970737,
    "late-1990s": 0.00035688793718772306,
```


Περιεχόμενο IDF.json:


```
"The": -0.06453852113757118,
"dot-com": 0.9162907318741551,
"bubble": 0.5108256237659907,
"(or": 0.9162907318741551,
"boom)": 2.0149030205422647,
"was": 0.06899287148695142,
"a": 0.0,
"stock": 0.06899287148695142,
"market": 0.0,
"that": 0.0,
"ballooned": 2.0149030205422647,
"during": 0.14310084364067324,
"the": -0.06453852113757118,
"late-1990s": 2.0149030205422647,
```


Ενδεικτικό τρέξιμο query όπως αναφέρεται και στο read.me :

```
#run a query and print results with tf-idf method
query = "what is bitcoin"
results = rank_query(query,method="TF-IDF")
print(f"Ranked results for query '{query}':")
for result in results:
    print(f"Document ID: {result['id']}, Title: {result['title']}, Score: {result['score']:.4f}")
```

Αποτελέσματα κώδικα:

```
Ranked results for query 'what is bitcoin':
Document ID: 14, Title: Bitcoin, Score: 0.0282
Document ID: 8, Title: Cryptocurrency, Score: 0.0085
Document ID: 13, Title: Blockchain, Score: 0.0049
Document ID: 0, Title: Dot-com bubble, Score: 0.0002
Document ID: 7, Title: Nasdaq, Score: 0.0002
Document ID: 4, Title: Financial crisis, Score: 0.0001
Document ID: 9, Title: Energy industry, Score: 0.0001
Document ID: 3, Title: COVID-19 recession, Score: 0.0001
Document ID: 11, Title: Stock market crash, Score: 0.0001
Document ID: 2, Title: 2007–2008 financial crisis, Score: 0.0001
Document ID: 10, Title: Goldman Sachs, Score: 0.0000
Document ID: 1, Title: Real estate investment trust, Score: 0.0000
Document ID: 5, Title: S&P 500, Score: 0.0000
Document ID: 6, Title: List of S&P 500 companies, Score: 0.0000
```

Παρατηρούμε ότι τα τρία πρώτα άρθρα είναι και τα πιο σχετικά. Τα υπόλοιπα άρθρα εμφανίζονται διότι το σύνολο των δεδομένων μας αποτελείται μόνο από 14 άρθρα.

Δημιουργία μερικών queries για την αξιολόγηση της μηχανής:

```
#evaluate engine
query_relevance = {
    "Financial Crisis": [0, 2, 3, 4, 11],
    "Covid Economy Consequences": [3],
    "What companies Are included to nasdaq": [5,6,7,9],
    "Banks of America": [6,7,10],
    "Real Estate Investments": [1,12],
    "Blockchain and Crypto": [8,13,14],
}
```

Οι μετρήσεις :

```
{'Precision': 0.22443482443482443, 'Recall': 0.9166666666666666, 'F1-Score':
0.34932725083653876, 'MAP': 0.6916812816812817}
```

Τέλος, το παρόν pdf έχει μόνο screenshot και επιστρεπτέες τιμές. Μεγαλύτερη και λεπτομερή ανάλυση έχει το pdf που ανέβηκε στο e-class της σχολής.