# Homework 6

1. Evaluate text similarity of Amazon book search results by doing the following:

   > a. Do a book search on Amazon via the search box. Manually copy the full book title (including subtitle) of each of the top 24 books listed in the first two pages of search results.
   >
   > b. In Python, run one of the text-similarity measures covered in this course, e.g., cosine similarity. Compare each of the book titles, pairwise, to every other one.
   >
   > c. Which two titles are the most similar to each other? Which are the most dissimilar? Where do they rank, among the first 24 results?

**Titles 7 and 10 are the most similar with a score of 0.91287. Both are titled "The Great Gatsby" so this was expected.**

**Titles 15 and 14 are the most dissimilar due to the length of the titles being vastly different. They scored 0.26111.**

In [ ]:
```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import numpy as np
```

In [ ]:
```
amazon_books = [
                "The Great Gatsby: The Original 1925 Edition (A F. Scott Fitzgerald Classic Novel)",
                "The Great Gatsby: The Original 1925 Edition (F. Scott Fitzgerald Classics)",
                "The Great Gatsby: A Classic 1925 Jazz Age Novel",
                "The Great Gatsby: The Only Authorized Edition",
                "The Great Gatsby by F. Scott Fitzgerald",
                "The Great Gatsby (Wordsworth Collector's Editions)",
                "The Great Gatsby",
                "The Great Gatsby: Illustrated Edition",
                "THE GREAT GATSBY: An Illuminated Edition",
                "The Great Gatsby",
                "The Great Gatsby: A Novel: Illustrated Edition",
                "The Great Gatsby: One of the greatest novels of American Literature, a Masterpiece (Annotated)",
                "The Great Gatsby: A Graphic Novel Adaptation",
                "The Great Gatsby",
                "The Great Gatsby (Gold Edition): Original version (1925) of F. Scott Fitzgerald's masterpiece, enriched with historical introduction and original period photos.",
                "The Great Gatsby: Original 1925 Edition",
                "The Great Gatsby (Annotated): Large-Print Edition",
                "The Great Gatsby and Other Works (Leather-bound Classics)",
                "The Great Gatsby (SeaWolf Press Classic)",
                "The Great Gatsby",
                "The Great Gatsby: With a New Historical Introduction for the Classroom",
                "The Great Gatsby (Annotated)",
                "The Great Gatsby: A F. Scott Fitzgerald Classics (The Original 1925 Edition)",
                "El gran Gatsby / The Great Gatsby (Spanish Edition)"
                ]
```

In [ ]:
```
count_vectorizer = CountVectorizer(stop_words='english')
count_vectorizer = CountVectorizer()
amazon_matrix = count_vectorizer.fit_transform(amazon_books)
```

In [ ]:
```
amazon_max = cosine_similarity(amazon_matrix)
amazon_min = cosine_similarity(amazon_matrix)
amazon_max[amazon_max > 0.99999999999] = 0
```

In [ ]:
```
pd.set_option('display.max_columns', None)
amazon_df = pd.DataFrame(amazon_max,columns = amazon_books,index=amazon_books)
amazon_df
```

Out[ ]:

| | The Great Gatsby: The Original 1925 Edition (A F. Scott Fitzgerald Classic Novel) | The Great Gatsby: The Original 1925 Edition (F. Scott Fitzgerald Classics) | The Great Gatsby: A Classic 1925 Jazz Age Novel | The Great Gatsby: The Only Authorized Edition | The Great Gatsby by F. Scott Fitzgerald | The Great Gatsby (Wordsworth Collector's Editions) | The Great Gatsby | The Great Gatsby: Illustrated Edition | THE GREAT GATSBY: An Illuminated Edition | The Great Gatsby | The Great Gatsby: A Novel: Illustrated Edition | The Great Gatsby: One of the greatest novels of American Literature, a Masterpiece (Annotated) | The Great Gatsby: A Graphic Novel Adaptation | The Great Gatsby | The Great Gatsby (Gold Edition): Original version (1925) of F. Scott Fitzgerald's masterpiece, enriched with historical introduction and original period photos. | The Great Gatsby: Original 1925 Edition | The Great Gatsby (Annotated): Large-Print Edition |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **The Great Gatsby: The Original 1925 Edition (A F. Scott Fitzgerald Classic Novel)** | 0.000000 | 0.880705 | 0.686406 | 0.647150 | 0.679366 | 0.452911 | 0.640513 | 0.620174 | 0.566139 | 0.640513 | 0.679366 | 0.403604 | 0.566139 | 0.640513 | 0.591312 | 0.792594 | 0.524142 |
| **The Great Gatsby: The Original 1925 Edition (F. Scott Fitzgerald Classics)** | 0.880705 | 0.000000 | 0.510310 | 0.673575 | 0.707107 | 0.471405 | 0.666667 | 0.645497 | 0.589256 | 0.666667 | 0.589256 | 0.420084 | 0.471405 | 0.666667 | 0.615457 | 0.824958 | 0.545545 |
| **The Great Gatsby: A Classic 1925 Jazz Age Novel** | 0.686406 | 0.510310 | 0.000000 | 0.471405 | 0.433013 | 0.433013 | 0.612372 | 0.474342 | 0.433013 | 0.612372 | 0.577350 | 0.342997 | 0.577350 | 0.612372 | 0.301511 | 0.577350 | 0.400892 |
| **The Great Gatsby: The Only Authorized Edition** | 0.647150 | 0.673575 | 0.471405 | 0.000000 | 0.544331 | 0.544331 | 0.769800 | 0.745356 | 0.680414 | 0.769800 | 0.680414 | 0.485071 | 0.544331 | 0.769800 | 0.355335 | 0.680414 | 0.629941 |
| **The Great Gatsby by F. Scott Fitzgerald** | 0.679366 | 0.707107 | 0.433013 | 0.544331 | 0.000000 | 0.500000 | 0.707107 | 0.547723 | 0.500000 | 0.707107 | 0.500000 | 0.396059 | 0.500000 | 0.707107 | 0.435194 | 0.500000 | 0.462910 |
| **The Great Gatsby (Wordsworth Collector's Editions)** | 0.452911 | 0.471405 | 0.433013 | 0.544331 | 0.500000 | 0.000000 | 0.707107 | 0.547723 | 0.500000 | 0.707107 | 0.500000 | 0.396059 | 0.500000 | 0.707107 | 0.261116 | 0.500000 | 0.462910 |
| **The Great Gatsby** | 0.640513 | 0.666667 | 0.612372 | 0.769800 | 0.707107 | 0.707107 | 0.000000 | 0.774597 | 0.707107 | 0.000000 | 0.707107 | 0.560112 | 0.707107 | 0.000000 | 0.369274 | 0.707107 | 0.654654 |
| **The Great Gatsby: Illustrated Edition** | 0.620174 | 0.645497 | 0.474342 | 0.745356 | 0.547723 | 0.547723 | 0.774597 | 0.000000 | 0.730297 | 0.774597 | 0.912871 | 0.433861 | 0.547723 | 0.774597 | 0.381385 | 0.730297 | 0.676123 |
| **THE GREAT GATSBY: An Illuminated Edition** | 0.566139 | 0.589256 | 0.433013 | 0.680414 | 0.500000 | 0.500000 | 0.707107 | 0.730297 | 0.000000 | 0.707107 | 0.666667 | 0.396059 | 0.500000 | 0.707107 | 0.348155 | 0.666667 | 0.617213 |
| **The Great Gatsby** | 0.640513 | 0.666667 | 0.612372 | 0.769800 | 0.707107 | 0.707107 | 0.000000 | 0.774597 | 0.707107 | 0.000000 | 0.707107 | 0.560112 | 0.707107 | 0.000000 | 0.369274 | 0.707107 | 0.654654 |
| **The Great Gatsby: A Novel: Illustrated Edition** | 0.679366 | 0.589256 | 0.577350 | 0.680414 | 0.500000 | 0.500000 | 0.707107 | 0.912871 | 0.666667 | 0.707107 | 0.000000 | 0.396059 | 0.666667 | 0.707107 | 0.348155 | 0.666667 | 0.617213 |
| **The Great Gatsby: One of the greatest novels of American Literature, a Masterpiece (Annotated)** | 0.403604 | 0.420084 | 0.342997 | 0.485071 | 0.396059 | 0.396059 | 0.560112 | 0.433861 | 0.396059 | 0.560112 | 0.396059 | 0.000000 | 0.396059 | 0.560112 | 0.361961 | 0.396059 | 0.458349 |
| **The Great Gatsby: A Graphic Novel Adaptation** | 0.566139 | 0.471405 | 0.577350 | 0.544331 | 0.500000 | 0.500000 | 0.707107 | 0.547723 | 0.500000 | 0.707107 | 0.666667 | 0.396059 | 0.000000 | 0.707107 | 0.261116 | 0.500000 | 0.462910 |
| **The Great Gatsby** | 0.640513 | 0.666667 | 0.612372 | 0.769800 | 0.707107 | 0.707107 | 0.000000 | 0.774597 | 0.707107 | 0.000000 | 0.707107 | 0.560112 | 0.707107 | 0.000000 | 0.369274 | 0.707107 | 0.654654 |
| **The Great Gatsby (Gold Edition): Original version (1925) of F. Scott Fitzgerald's masterpiece, enriched with historical introduction and original period photos.** | 0.591312 | 0.615457 | 0.301511 | 0.355335 | 0.435194 | 0.261116 | 0.369274 | 0.381385 | 0.348155 | 0.369274 | 0.348155 | 0.361961 | 0.261116 | 0.369274 | 0.000000 | 0.609272 | 0.322329 |
| **The Great Gatsby: Original 1925 Edition** | 0.792594 | 0.824958 | 0.577350 | 0.680414 | 0.500000 | 0.500000 | 0.707107 | 0.730297 | 0.666667 | 0.707107 | 0.666667 | 0.396059 | 0.500000 | 0.707107 | 0.609272 | 0.000000 | 0.617213 |
| **The Great Gatsby (Annotated): Large-Print Edition** | 0.524142 | 0.545545 | 0.400892 | 0.629941 | 0.462910 | 0.462910 | 0.654654 | 0.676123 | 0.617213 | 0.654654 | 0.617213 | 0.458349 | 0.462910 | 0.654654 | 0.322329 | 0.617213 | 0.000000 |
| **The Great Gatsby and Other Works (Leather-bound Classics)** | 0.369800 | 0.481125 | 0.353553 | 0.444444 | 0.408248 | 0.408248 | 0.577350 | 0.447214 | 0.408248 | 0.577350 | 0.408248 | 0.323381 | 0.408248 | 0.577350 | 0.284268 | 0.408248 | 0.377964 |
| **The Great Gatsby (SeaWolf Press Classic)** | 0.566139 | 0.471405 | 0.577350 | 0.544331 | 0.500000 | 0.500000 | 0.707107 | 0.547723 | 0.500000 | 0.707107 | 0.500000 | 0.396059 | 0.500000 | 0.707107 | 0.261116 | 0.500000 | 0.462910 |
| **The Great Gatsby** | 0.640513 | 0.666667 | 0.612372 | 0.769800 | 0.707107 | 0.707107 | 0.000000 | 0.774597 | 0.707107 | 0.000000 | 0.707107 | 0.560112 | 0.707107 | 0.000000 | 0.369274 | 0.707107 | 0.654654 |
| **The Great Gatsby: With a New Historical Introduction for the Classroom** | 0.480384 | 0.500000 | 0.408248 | 0.577350 | 0.471405 | 0.471405 | 0.666667 | 0.516398 | 0.471405 | 0.666667 | 0.471405 | 0.420084 | 0.471405 | 0.666667 | 0.430820 | 0.471405 | 0.436436 |
| **The Great Gatsby (Annotated)** | 0.554700 | 0.577350 | 0.530330 | 0.666667 | 0.612372 | 0.612372 | 0.866025 | 0.670820 | 0.612372 | 0.866025 | 0.612372 | 0.606339 | 0.612372 | 0.866025 | 0.319801 | 0.612372 | 0.755929 |
| **The Great Gatsby: A F. Scott Fitzgerald Classics (The Original 1925 Edition)** | 0.880705 | 0.000000 | 0.510310 | 0.673575 | 0.707107 | 0.471405 | 0.666667 | 0.645497 | 0.589256 | 0.666667 | 0.589256 | 0.420084 | 0.471405 | 0.666667 | 0.615457 | 0.824958 | 0.545545 |
| **El gran Gatsby / The Great Gatsby (Spanish Edition)** | 0.526235 | 0.547723 | 0.447214 | 0.632456 | 0.516398 | 0.516398 | 0.730297 | 0.707107 | 0.645497 | 0.730297 | 0.645497 | 0.383482 | 0.516398 | 0.730297 | 0.337100 | 0.645497 | 0.597614 |

In [ ]:
```
max_values = []
for l in range(len(amazon_max)):
    val = np.argmax(amazon_max[l])
    max_values.append(val)

min_values = []
for l in range(len(amazon_min)):
    val = np.argmin(amazon_min[l])
    min_values.append(val)
```

In [ ]:
```
max_nums = []
index=0
for idx,arr in enumerate(max_values):
    num = amazon_max[idx][arr]
    max_nums.append(num)
max(max_nums)
```

Out[ ]:
```
0.9128709291752769
```

In [ ]:
```
min_nums = []
for idx,arr in enumerate(min_values):
    num = amazon_min[idx][arr]
    min_nums.append(num)

min(min_nums)
```

Out[ ]:
```
0.2611164839335468
```

---

1. Now evaluate using a major search engine.

   > a. Enter one of the book titles from question 1a into Google, Bing, or Yahoo!. Copy the capsule of the first organic result and the 20th organic result. Take web results only (i.e., not video results), and skip sponsored results.
   >
   > b. Run the same text similarity calculation that you used for question 1b on each of these capsules in comparison to the original query (book title).
   >
   > c. Which one has the highest similarity measure?

**The book title to the first organic search result has the highest similarity measure with a score of 0.866025. While comparing Originally searched items to the 20th value returns a close score of 0.866025. Scores between the returned google results posted a similarity score of 0.693375**

Submit all of your inputs and outputs and your code for this assignment, along with a brief written explanation of your findings.

In [ ]:
```
google_result = [
                 "The Great Gatsby",
                 "The Great Gatsby - Wikipedia",
                 "What makes The Great Gatsby great? | Books | The Guardian"
                ]
```

In [ ]:
```
count_vectorizer = CountVectorizer(stop_words='english')
count_vectorizer = CountVectorizer()
google_matrix = count_vectorizer.fit_transform(google_result)
```

In [ ]:
```
pd.DataFrame(cosine_similarity(google_matrix),columns=google_result,index=google_result)
```

Out[ ]:

| | The Great Gatsby | The Great Gatsby - Wikipedia | What makes The Great Gatsby great? \| Books \| The Guardian |
|---|---|---|---|
| **The Great Gatsby** | 1.000000 | 0.866025 | 0.800641 |
| **The Great Gatsby - Wikipedia** | 0.866025 | 1.000000 | 0.693375 |
| **What makes The Great Gatsby great? \| Books \| The Guardian** | 0.800641 | 0.693375 | 1.000000 |