

# Comic Book Recommendation System

By: Kelly Dong

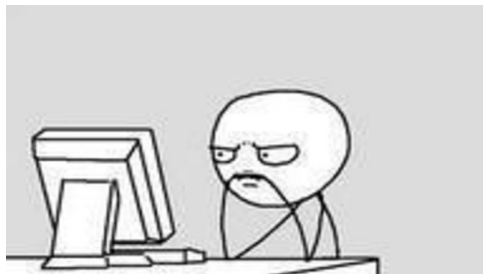


# Project Overview

The comic book industry is a behemoth mess with an enormous amount of volumes and issues. It is a constant challenge to search through the large backlog for a story that will interest you.

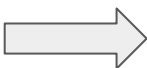
Using natural language processing and feature engineering, we will create a model that takes in an issue name that you have read before and output recommendations that are similar to your input.

The resulting product will save the user time and energy as they browse for their next comic book.



# Dataset Overview and Preprocessing

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7061 entries, 0 to 7060
Data columns (total 32 columns):
 #   Column              Non-Null Count  Dtype
---  -
0   Unnamed: 0          6295 non-null  float64
1   aliases             0 non-null     float64
2   api_detail_url      7061 non-null  object
3   associated_images   7061 non-null  object
4   character_credits   7061 non-null  object
5   character_died_in   7061 non-null  object
6   concept_credits     7061 non-null  object
7   cover_date          7061 non-null  object
8   date_added          7061 non-null  object
9   date_last_updated   7061 non-null  object
10  deck                475 non-null   object
11  description          6982 non-null  object
12  first_appearance_characters  0 non-null     float64
13  first_appearance_concepts  0 non-null     float64
14  first_appearance_locations  0 non-null     float64
15  first_appearance_objects  0 non-null     float64
16  first_appearance_storyarcs  0 non-null     float64
17  first_appearance_teams    0 non-null     float64
18  has_staff_review      7061 non-null  object
19  id                   7061 non-null  int64
20  image                7061 non-null  object
21  issue_number          7061 non-null  object
22  location_credits      7061 non-null  object
23  name                 6767 non-null  object
24  object_credits        7061 non-null  object
25  person_credits        7061 non-null  object
26  site_detail_url       7061 non-null  object
27  store_date            3003 non-null  object
28  story_arc_credits     7061 non-null  object
29  team_credits          7061 non-null  object
30  team_disbanded_in     7061 non-null  object
31  volume                7061 non-null  object
dtypes: float64(8), int64(1), object(23)
memory usage: 1.7+ MB
```



```
<class 'pandas.core.frame.DataFrame'>
Index: 6767 entries, 0 to 7060
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  -
0   character_credits    6767 non-null  object
1   character_died_in    6767 non-null  object
2   concept_credits      6767 non-null  object
3   cover_date           6767 non-null  datetime64[ns]
4   description          6767 non-null  object
5   has_staff_review      6767 non-null  object
6   id                   6767 non-null  int64
7   issue_number          6767 non-null  object
8   location_credits      6767 non-null  object
9   name                 6767 non-null  object
10  object_credits        6767 non-null  object
11  person_credits        6767 non-null  object
12  story_arc_credits     6767 non-null  object
13  team_credits          6767 non-null  object
14  team_disbanded_in     6767 non-null  object
15  volume                6767 non-null  object
16  combined_description  6767 non-null  object
dtypes: datetime64[ns](1), int64(1), object(15)
memory usage: 1.2+ MB
```

```
def clean_description(html_text):
    soup = BeautifulSoup(html_text, 'html.parser')
    cleaned_text = soup.get_text(separator='\n')
    cleaned_text = cleaned_text.replace('\n', '').replace('\n', '')
    return cleaned_text
```

'<p><em>"PATH TO DOOM" Chapter

One</em></p><p><em>Superman returns to Metropolis just in  
time to meet the city of tomorrow's newest protector: Lex Luthor.

But it's not long before these dueling titans meet someone

unexpected — the new Clark Kent!</em></p><p><em>DON'T

MISS: ACTION COMICS returns to its original numbering with

this issue!</em></p><p><em>NOW SHIPPING TWICE

MONTHLY!</em></p><h4>List of covers and their

creators:</h4><table data-max-width="true"><thead><tr><th

scope="col">Cover</th><th scope="col">Name</th><th

scope="col">Creator(s)</th><th scope="col">Sidebar

Location</th></tr></thead><tbody><tr><td>Reg</td><td>Regul

ar Cover</td><td>Ivan Reis, Joe Prado & Sonia

Oback</td><td>1</td></tr><tr><td>Var</td><td>Variant

Cover</td><td>Ryan

Sook</td><td>3</td></tr><tr><td>2nd</td><td>Second Printing

Cover</td><td>Ivan Reis, Joe Prado & Sonia

Oback</td><td>2</td></tr></tbody></table>'

"PATH TO DOOM" Chapter One  
Superman returns to Metropolis  
just in time to meet the city of tomorrow's newest protector: Lex  
Luthor. But it's not long before these dueling titans meet someone  
unexpected — the new Clark Kent!  
DON'T MISS: ACTION  
COMICS returns to its original numbering with this issue!  
NOW SHIPPING TWICE MONTHLY!  
List of covers and their  
creators:  
CoverNameCreator(s)Sidebar LocationRegRegular  
CoverIvan Reis, Joe Prado & Sonia Oback1VarVariant CoverRyan  
Sook32ndSecond Printing CoverIvan Reis, Joe Prado & Sonia  
Oback2'

```

7056    [{'api_detail_url': 'https://comicvine.gamespo...
7057    [{'api_detail_url': 'https://comicvine.gamespo...
7058    [{'api_detail_url': 'https://comicvine.gamespo...
7059    [{'api_detail_url': 'https://comicvine.gamespo...
7060    [{'api_detail_url': 'https://comicvine.gamespo...
Name: character_credits, Length: 6767, dtype: object

```

```

def get_names(json_str, index):
    json_str = json_str.replace("'", '"')
    try:
        lst = json.loads(json_str)
    except json.JSONDecodeError as e:
        #print(f"JSONDecodeError at row {index}: {e}")
        return []
    names = [item["name"] for item in lst]
    return names

```

```

7056    [Empress, Hal Jordan, Kon-El, Secret, Slobo, S...
7057    [Buzz, Hal Jordan, Harm, Secret, Spectre]
7058    [Arrowette, Colonel Rajak, Kon-El, Red Tornado...
7059    [Arrowette, Empress, Kon-El, Secret, Slobo, Sn...
7060    [Agent Donald Fite, Agent Ishido Maad, Arrowet...
Name: character_credits, Length: 6767, dtype: object

```

```

7056    [Empress, Hal_Jordan, Kon-El, Secret, Slobo, S...
7057    [Buzz, Hal_Jordan, Harm, Secret, Spectre]
7058    [Arrowette, Colonel_Rajak, Kon-El, Red_Tornado...
7059    [Arrowette, Empress, Kon-El, Secret, Slobo, Sn...
7060    [Agent_Donald_Fite, Agent_Ishido_Maad, Arrowet...
Name: character_credits, Length: 6767, dtype: object

```

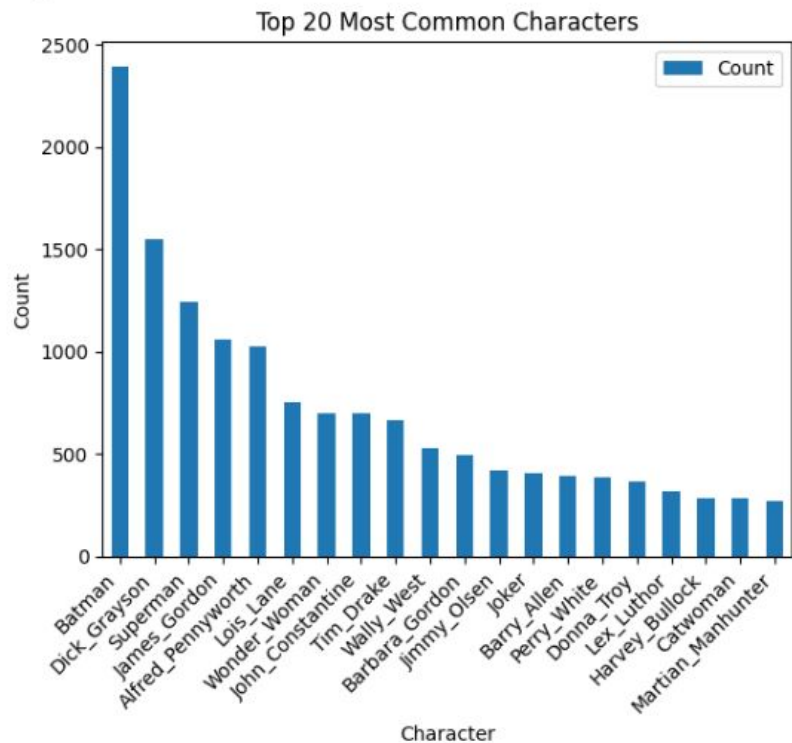
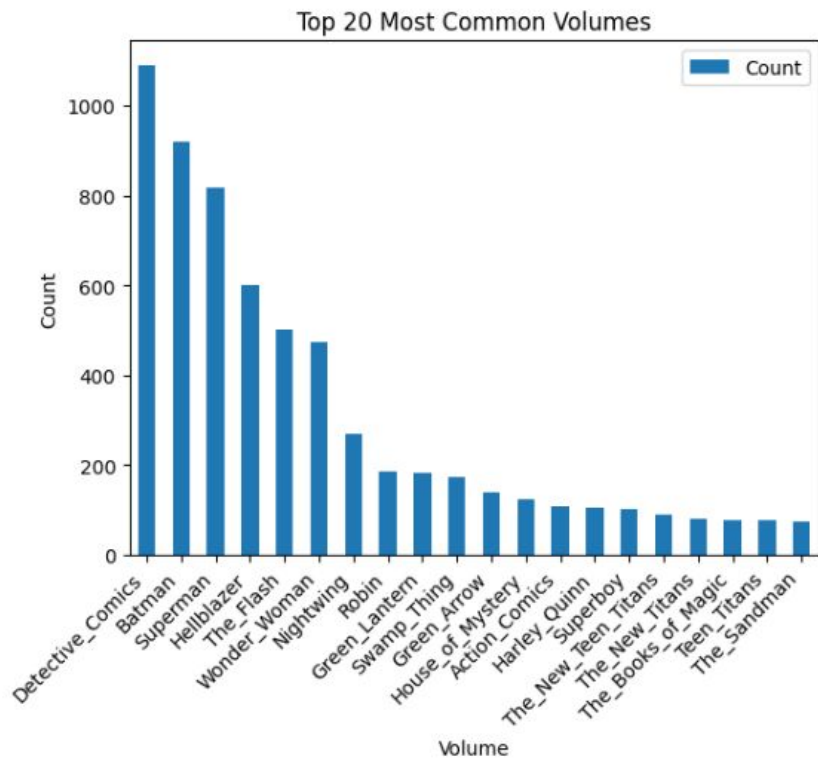
Adding underscores was important to having each character/element register as one feature when we tokenize the columns later.

```

df['character_credits'].apply(lambda x:[name.replace(' ', '_') for name in x])

```

# EDA



```
tfidf_des = TfidfVectorizer(stop_words='english')
tfidf_des_matrix = tfidf_des.fit_transform(df['description'])
```

```
tfidf_char = TfidfVectorizer(stop_words='english')
tfidf_char_matrix = tfidf_char.fit_transform(df['character_credits'])
```

```
tfidf_concept = TfidfVectorizer(stop_words='english')
tfidf_concept_matrix = tfidf_concept.fit_transform(df['concept_credits'])
```

```
tfidf_location = TfidfVectorizer(stop_words='english')
tfidf_location_matrix = tfidf_location.fit_transform(df['location_credits'])
```

```
tfidf_object = TfidfVectorizer(stop_words='english')
tfidf_object_matrix = tfidf_object.fit_transform(df['object_credits'])
```

```
tfidf_person = TfidfVectorizer(stop_words='english')
tfidf_person_matrix = tfidf_person.fit_transform(df['person_credits'])
```

```
tfidf_arc = TfidfVectorizer(stop_words='english')
tfidf_arc_matrix = tfidf_arc.fit_transform(df['story_arc_credits'])
```

```
tfidf_team = TfidfVectorizer(stop_words='english')
tfidf_team_matrix = tfidf_team.fit_transform(df['team_credits'])
```

```
tfidf_vol = TfidfVectorizer(stop_words='english')
tfidf_vol_matrix = tfidf_vol.fit_transform(df['volume'])
```

```
Description Length <= 20: 107
Number of entries with 0 volume: 0
Number of entries with 0 characters: 322
Number of entries with 0 persons: 630
Number of entries with 0 locations: 1527
Number of entries with 0 concepts: 1923
Number of entries with 0 teams: 2300
Number of entries with 0 objects: 4500
Number of entries with 0 story arcs: 6354
```

```
all_matrices = [tfidf_des_matrix, tfidf_char_matrix, tfidf_concept_matrix,
                 tfidf_location_matrix, tfidf_object_matrix, tfidf_person_matrix,
                 tfidf_arc_matrix, tfidf_team_matrix, tfidf_vol_matrix]
```

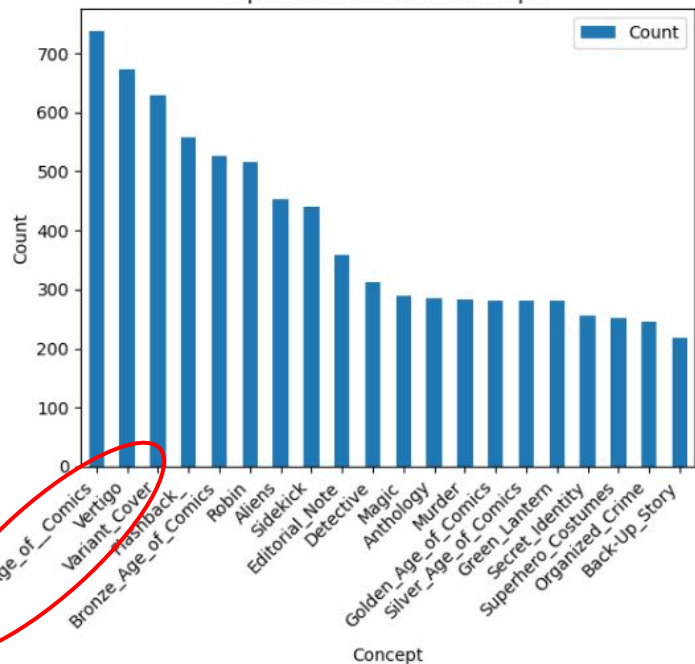
```
weights = [3, 2, 1,
            1, 0, 1,
            0, 1, 0]
```

```
# Scale each TF-IDF matrix by its corresponding weight
weighted_tfidf_matrices = [matrix * weight for matrix, weight in zip(all_matrices, weights)]
```

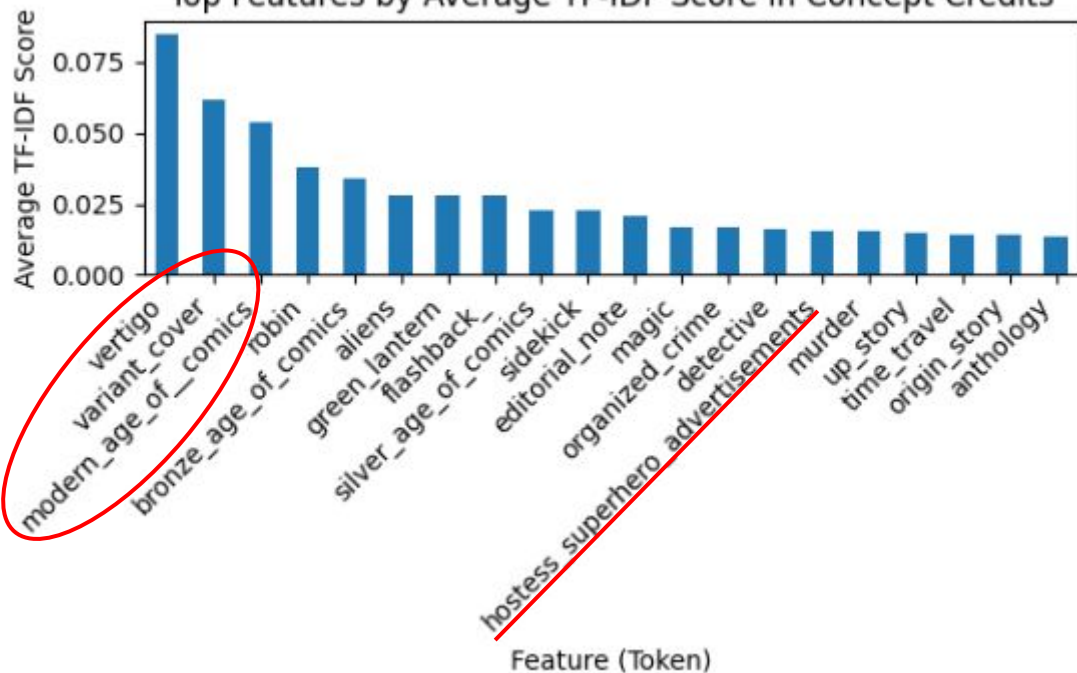
```
combined_matrix = hstack(weighted_tfidf_matrices)
```



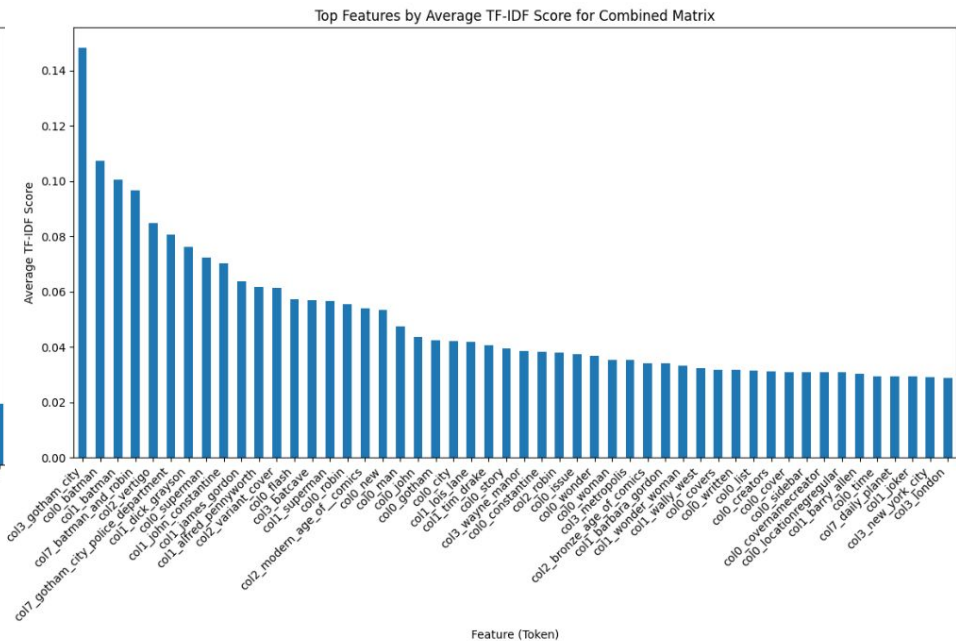
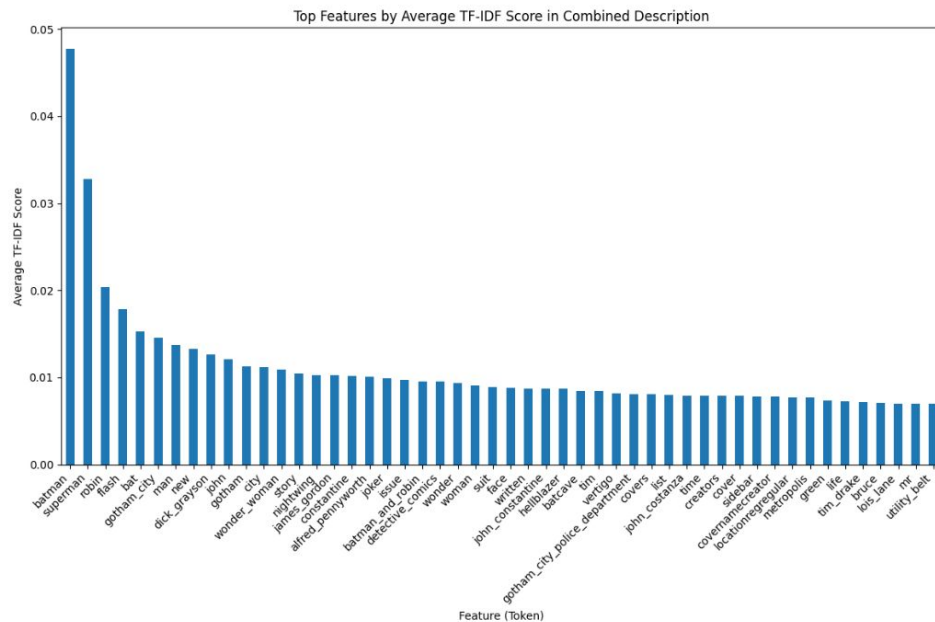
Top 20 Most Common Concepts



Top Features by Average TF-IDF Score in Concept Credits





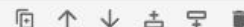


# The Model

```
cosine_sim = linear_kernel(combined_matrix, combined_matrix)
```

```
def get_recommendations(df, title, cosine_sim=cosine_sim):  
    # Get the index of the issue that matches the title  
    idx = df[df['name'] == title].index[0]  
    # Get the pairwise similarity scores of all issues with that issue  
    sim_scores = list(enumerate(cosine_sim[idx]))  
    # Sort the issues based on the similarity scores  
    sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)  
    # Get the scores of the 10 most similar issues  
    sim_scores = sim_scores[0:11]  
    # Get the issue indices  
    issue_indices = [i[0] for i in sim_scores]  
    # Return the top 10 most similar issues  
    return df[['issue_number', 'name', 'description', 'character_credits', 'concept_credits', 'volume']].iloc[issue_indices]
```

```
test1 = get_recommendations(df, "Robins vs. Zombies: Robin War", cosine_sim)
test1
```



issue_number		name	description	character_credits	concept_credits	volume
2373	13	Robins vs. Zombies: Robin War	A "Robin War" tie-in! With Robins fighting cop...	['Colton_Rivera', 'Damian_Wayne', 'Efrem', 'He...	[]	['Gotham_Academy']
2374	14	Yearbook Part One; Animal Science 101; Queen G...	An all-new era of GOTHAM ACADEMY begins here w...	['Clayface_(Karlo)', 'Colton_Rivera', 'Dillyn'...	[]	['Gotham_Academy']
2378	18	Yearbook Part Five; Whatever Happened to Profe...	As the "Gotham Academy Yearbook" storyline com...	['Coach_Humphreys', 'Colton_Rivera', 'Damian_W...	[]	['Gotham_Academy']
2368	7	Curse of the Inishtree Quill	Special guest student Damian Wayne drops by th...	['Batman', 'Bookworm', 'Colton_Rivera', 'Damia...	['Joker_75th_Anniversary_Variant']	['Gotham_Academy']
2362	1	Welcome to Gotham Academy	WELCOME TO GOTHAM ACADEMY! Gotham City's most ...	['Aunt_Harriet', 'Batman', 'Calamity', 'Colton...	['Batman_Villains', 'The_New_52']	['Gotham_Academy']
2376	16	Yearbook Part Three; Maps' Day Out; Boring Sun...	It's "Yearbook" part 3! As Olive and the gang ...	['Barbara_Gordon', 'Batman', 'Colton_Rivera', ...	['Robin']	['Gotham_Academy']
2365	4	The Secret of the Symbol	The hunt for the Ghost of Gotham Academy begins!	['Batman', 'Calamity', 'Coach_Humphreys', 'Gra...	['Batman_Villains']	['Gotham_Academy']
2370	9	Calamity	If the gang thought it was hard to keep up wit...	['Calamity', 'Clayface_(Karlo)', 'Coach_Humphr...	[]	['Gotham_Academy']
2375	15	Yearbook Part Two; Staff Party; Serpents & Sec...	It's part two of "Gotham Academy Yearbook"! Th...	['Batman', 'Bookworm', 'Clayface_(Karlo)', 'Co...	['Anthology']	['Gotham_Academy']
2366	5	Save The Last Dance	This month's assignment: Uncover the hideous s...	['Batman', 'Calamity', 'Colton_Rivera', 'Headm...	[]	['Gotham_Academy']
2371	10	The Cursed Play	"Bubble, bubble, toil and trouble!" To investi...	['Calamity', 'Clayface_(Karlo)', 'Colton_River...	[]	['Gotham_Academy']

```
test2 = get_recommendations(df, "A Death in the Family Chapter 1 and 2", cosine_sim)
test2
```

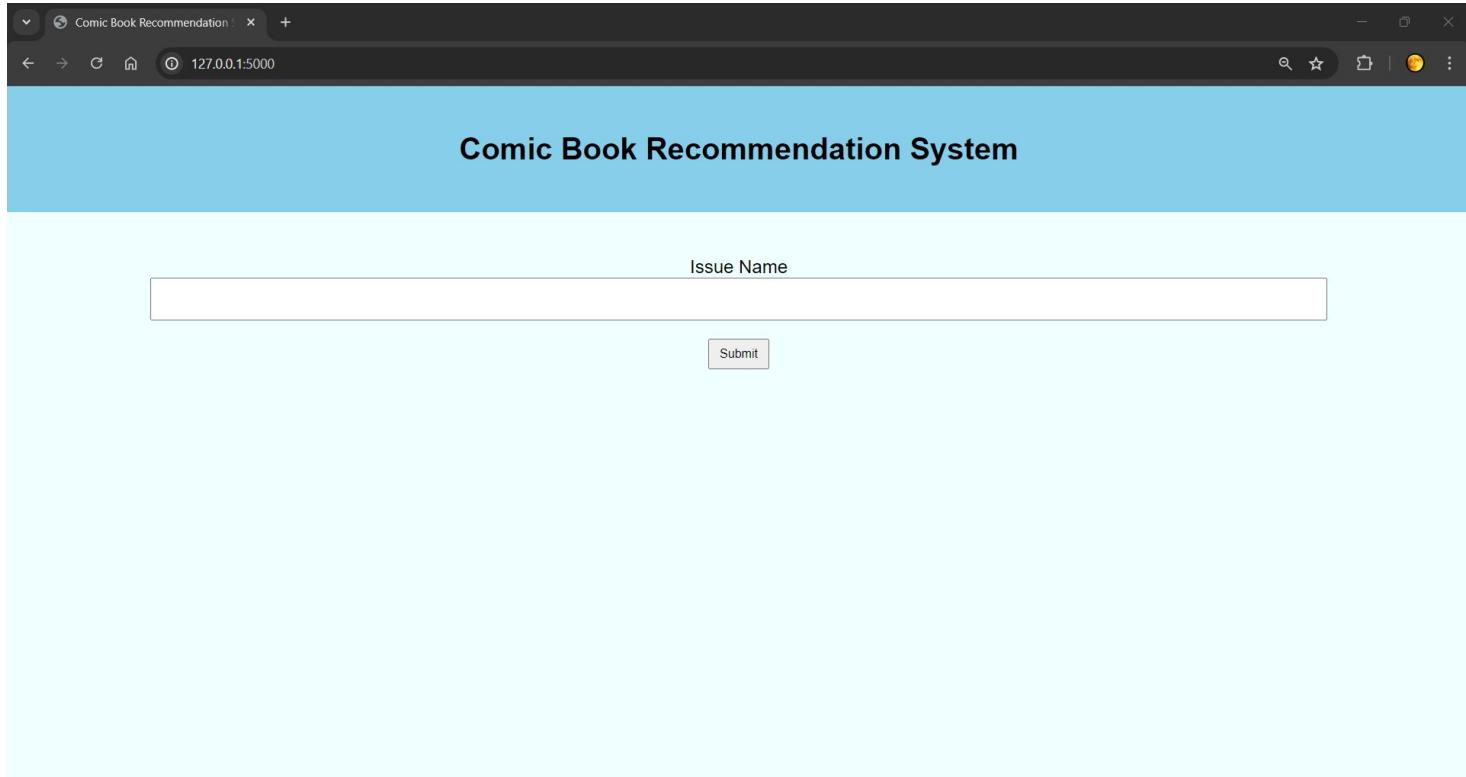
issue_number		name	description	character_credits	concept_credits	volume
611	426	A Death in the Family Chapter 1 and 2	Jason Todd's been getting out of control in hi...	['Alfred_Pennyworth', 'Batman', 'Catherine_Tod...]	['Batman_Villains', 'Editorial_Note', 'Flashba...]	['Batman']
612	427	A Death In the Family Chapter 3 and 4	Finally discovering his mother isn't the happy...	['Alfred_Pennyworth', 'Batman', 'Jason_Todd', ...]	['Assassination', 'Batman_Villains', 'Cliffhan...]	['Batman']
595	410	Two Of A Kind	The training of Jason Todd. Batman tries to ke...	['Alfred_Pennyworth', 'Batman', 'Dick_Grayson'...]	['Martial_Arts', 'Origin_Story']	['Batman']
596	411	Second Chance	Jason Todd's debut as Robin continues now that...	['Alfred_Pennyworth', 'Batman', 'James_Gordon'...]	['Baseball', 'Henchmen']	['Batman']
613	428	A Death in the Family Chapter 5	Picking up directly where last issue left off,...	['Alfred_Pennyworth', 'Ayatollah_Khomeini', 'B...]	['Batman_Villains', 'Flashback_', 'Martial_Art...]	['Batman']
593	408	"Did Robin Die Tonight?"	When Dick Grayson is injured by the Joker, Bat...	['Alfred_Pennyworth', 'Batman', 'Catherine_Tod...]	['Flashback_', 'Origin_Story']	['Batman']
1315	580	Double Image	Batman has reasons to believe that the "Harvey...	['Alfred_Pennyworth', 'Batman', 'Charlatan', '...]	['Flashback_', 'Origin_Story']	['Detective_Comics']
720	534	A Wound on the Heart of Heaven	Legacy' part 5, continued from BATMAN: SHADOW ...	['Barbara_Gordon', 'Batman', 'Lady_Shiva']	[]	['Batman']
915	645	Show Me Yesterday, For I Can't Find Today : A...	With the Red Hood's identity revealed to be Ja...	['Alfred_Pennyworth', 'Batman', 'Black_Mask', ...]	[]	['Batman']
3961	8	Death's Door	Batman returns! Bruce Wayne continues his trai...	['Bane', 'Batman', 'Dick_Grayson', 'Harold', '...]	['Ninjas', 'Robin', 'Secret_Identity']	['Robin']
594	409	Just Another Kid On Crime Alley!	Batman inadvertently enrolls street kid Jason ...	['Batman', 'James_Gordon', 'Jason_Todd', 'Ma_G...]	['Drug_Trafficking', 'Origin_Story']	['Batman']

```
test3 = get_recommendations(df, "War for the Books of Magic, Part 1", cosine_sim)
test3
```



issue_number		name	description	character_credits	concept_credits	volume
3601	12.0	War for the Books of Magic, Part 1	FAUST'S true master is revealed!The divided te...	['Black_Boris', 'Black_Orchid_(Garcia)', 'Blac...	['Magick']	['Justice_League_Dark']
3598	9.0	The Black Room	New series writer JEFF LEMIRE introduces a new...	['Andrew_Bennett', 'Black_Orchid_(Garcia)', 'D...	['Magick']	['Justice_League_Dark']
3600	11.0	The Black Room, Part Three	The secrets of the BLACK ROOM revealed!The tea...	['Abnegazar', 'Black_Orchid_(Garcia)', 'Deadma...	['Magick']	['Justice_League_Dark']
3605	15.0	The Death of Magic, Part 1: Up is Down	A new story arc starts here!Trapped in a techn...	['Alkion', 'Black_Orchid_(Garcia)', 'Deadman',...]	['Magick']	['Justice_League_Dark']
3607	17.0	The Death of Magic, Part 3: Prisoners of Epoch	Constantine and the others are trapped on a wo...	['Alkion', 'Black_Orchid_(Garcia)', 'Deadman',...]	['Magick']	['Justice_League_Dark']
3606	16.0	The Death of Magic, Part 2: Night of the Hunter	The team is trapped on a magic-less planet tha...	['Alkion', 'Black_Orchid_(Garcia)', 'Deadman',...]	['Magick']	['Justice_League_Dark']
3608	18.0	The Death of Magic, Part 4: The Last Stand	The conclusion to the "DEATH OF MAGIC" comes c...	['Alkion', 'Black_Orchid_(Garcia)', 'Deadman',...]	['Magick']	['Justice_League_Dark']
3604	14.0	Enter the House of Mystery	A major new storyline starts here!In the after...	['Alec_Holland', 'Amethyst', 'Animal_Man', 'Ap...	['Magick']	['Justice_League_Dark']
3603	13.0	War for the Books of Magic, Part 2: Revelations	House of Mystery vs. House of Secrets!The team...	[]	['Magick']	['Justice_League_Dark']
3610	20.0	Horror City, Part 2: The Nightmare Gospel	The Flash and Swamp Thing guest-star as the te...	['Alec_Holland', 'Barry_Allen', 'Deadman', 'Do...	['Magick']	['Justice_League_Dark']
3611	21.0	Horror City, Conclusion: Die, Die, Die My Darling	Want to know how each team member is going to ...	['Alec_Holland', 'Amethyst', 'Andrew_Bennett',...]	['Magick']	['Justice_League_Dark']

# My Web App Demo



The screenshot shows a web browser window with a single tab titled 'Comic Book Recommendation'. The address bar displays '127.0.0.1:5000'. The page has a light blue header with the text 'Comic Book Recommendation System'. Below the header, on a light cyan background, is a form. The form consists of a text input field with the placeholder text 'Issue Name' and a 'Submit' button centered below it.

Comic Book Recommendation System

Issue Name

Submit

# Next Steps...

- Continue Collecting Data
- Fine Tune Recommendation Model
  - allow user to fiddle with feature weights
- Make Web App easier to use
  - Issue Search Bar
    - make more forgiving towards typos
    - drop down menu with options
  - Search Other Categories (ex. description, character\_credits)



# Thank You!

