

UNIVERSITY OF NORTH TEXAS
COMPUTER SCIENCE DEPARTMENT
CSCE 5200 Information Retrieval and Web Search (Fall 2022)
Project Group Number – **25**
UNT – Search Engine

Team Members:

- 1) Sai Manoj Malepati
- 2) Sai Meghana Gushidi
- 3) Goutham Pallapothu
- 4) Abhay Arora
- 5) Sai Anjani Pinreddy
- 6) Sri Sai Charan Yerramsetty

Source Code

- .Zip Submitted in Canvas
- Git Repo: https://github.com/abhayarora23UNT/UNT_IRS_Project

Steps to Run the Project :

1. Download the zip file of the project.
2. Extract all files from the zip.
3. Open the command prompt and change the directory to the project file.
4. Run the command `flask run`.
 - a. To change the default IP, run this command `flask run --host=0.0.0.0`
5. Open this URL in the browser: <http://127.0.0.1:5000/>

Technologies/Framework/Language

- Languages: Python, HTML
- Libraries: Beautiful soup, urllib, pandas, nltk, sklearn

Implementation Details

- 1) In the first phase, the application crawls the unt.edu website using python libraries: Beautiful soup, requests, urllib, and collections.
- 2) The crawler returned more than 2000 URLs.
- 3) After getting all the links, we parsed the HTML content, removed all tags, cleaned the text,
- 4) Then we removed all the characters from the text, saved the link, and cleaned the text in the corpus file.
- 5) After that, we query a search keyword using a vector space model, check document similarity using cosine similarity, and finally sort the result set.
- 6) The top 11 search results will be displayed on the UI
- 7) For routing, flask decorators are used.

Crawling via BeautifulSoup

```
soup = BeautifulSoup(response.text, "lxml")

for link in soup.find_all('a'):

    anchor = link.attrs["href"] if "href" in link.attrs else ''

    if anchor.startswith('/'):

        local_link = base_url + anchor

        local_urls.add(local_link)

    elif strip_base in anchor:

        local_urls.add(anchor)

    elif not anchor.startswith('http'):

        local_link = path + anchor

        local_urls.add(local_link)

    else:

        foreign_urls.add(anchor)

for i in local_urls:

    if not i in new_urls and not i in processed_urls:

        new_urls.append(i)
```

VectorSpace Model using TfidfVectorizer

```
for data in corpusData:
    tokenData = toGenerateTokens(data)
    LinkData = toReadStopWords(tokenData)
    LinkData = wordStemmer(LinkData)
    LinkData = ' '.join(LinkData)
    ModifiedCorpus.append(LinkData) # This contains data after preprocessing
stage

vectorizerX = TfidfVectorizer(stop_words='english')
documentVector = vectorizerX.fit_transform(ModifiedCorpus)
dataFrame = pd.DataFrame(documentVector.toarray(),
                          columns=vectorizerX.get_feature_names_out())

query = toGenerateTokens(query)
query = toReadStopWords(query)
queryData = []
for word in wordStemmer(query):
    queryData.append(word)
queryData = ' '.join(queryData)
queryVector = vectorizerX.transform([queryData])
#Using cosineSimilarities to find similarity between query and existing link
and content
cosineSimilarityValues = cosine_similarity(documentVector,
queryVector).flatten()
#Top 11 documents will be displayed on the top
topScoredDocuments = cosineSimilarityValues.argsort()[::-12:-1]
```

Outputs

- The home screen will show a search box.
- The user will enter some word in the search box and click the Search Button.
- The system will retrieve the top results 11 using the vector space model

Sample Queries

Input 1: Computer

➤ Results from our application

The screenshot shows a web browser window with the address bar displaying '127.0.0.1:5000/computer'. The page features a green header with the UNT logo and the text 'University Of North Texas Search'. Below the header is a search bar with the placeholder text 'Enter keyword to search' and a green 'Search' button. The search results are titled 'Search Results for computer' and list three items:

- [Computer Science and Engineering](https://www.unt.edu/academics/undergrad/computer-science-and-engin)
<https://www.unt.edu/academics/undergrad/computer-science-and-engin>
HomeAdmissionsAcademicsStudent LifeAbout UNTResearchLocationsAthleticsGiving Computer Science and Engineering Home ...
- [Information Technology](https://www.unt.edu/academics/undergrad/information-technolog)
<https://www.unt.edu/academics/undergrad/information-technolog>
HomeAdmissionsAcademicsStudent LifeAbout UNTResearchLocationsAthleticsGiving Information Technology Home ...
- [Computer Science \(B.S.\)](http://frisco.unt.edu/programs/computer-sci)
<http://frisco.unt.edu/programs/computer-sci>
ereHomeProgramsProject Design and AnalysisComputer Science (B.S.) Be prepared for a high-paying job in any industry If you're good at multi-

➤ Results from original Unt.edu :

The screenshot shows a web browser window with the address bar displaying 'unt.edu/search-results?search=computer&sa=Search'. The page features a green header with the UNT logo and the text 'UNIVERSITY OF NORTH TEXAS'. Below the header is a navigation bar with links to 'Admissions', 'Academics', 'Student Life', 'About UNT', 'Research', 'Locations', 'Athletics', and 'Giving'. The search results are titled 'Search Results' and show a search bar with the query 'computer' and a 'Search' button. The results indicate 'About 12,380,000,000 results (0.35 seconds)' and are sorted by 'Relevance'. The results list three items:

- [Computer Science and Engineering: Home](https://computerscience.engineering.unt.edu)
computerscience.engineering.unt.edu
I would like to extend my warmest welcome to you as the Chair to the Department of **Computer** Science and Engineering at the University of North Texas. Computing ...
- [Computer Labs - University Libraries - UNT](#)
[University Libraries > services > computer-labs](#)
Willis Library is a Student **Computer** Lab that offers Dell and Mac **computer** workstations, laptop checkouts, printing and scanning services. Sycamore Library - ...
- [Computer Science \(B.S.\) | UNT at Frisco](#)

Input 2: Student Life

➤ Results from our application

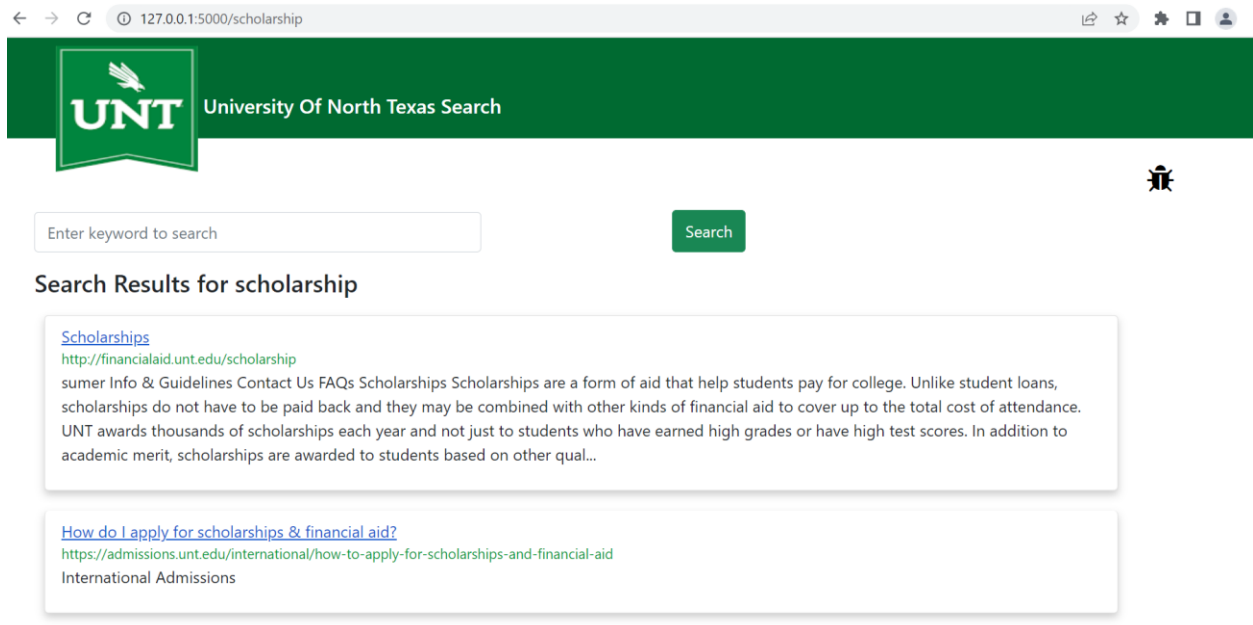
The screenshot shows a web browser with the address bar displaying '127.0.0.1:5000/student%20life'. The page features a green header with the UNT logo and the text 'University Of North Texas Search'. Below the header is a search bar with the placeholder text 'Enter keyword to search' and a green 'Search' button. The search results are titled 'Search Results for student life' and contain two identical entries. Each entry has a blue link 'Spiritual Life', a green URL 'http://studentaffairs.unt.edu/office-of-spiritual-lif', and a snippet of text: 'matter what your beliefs are. You are hereHomeDepartmentsSpiritual Life Spiritual Life Virtual Resources Remote Meditation Sessions Get Connected Upcoming Events by Spiritual Life More Events Programs by Spiritual Life Spiritual Life The University of North Texas recognizes the importance of a healthy and diverse spiritual life on a... By: Spiritual Life Spiritual Life Gardening Learn about gardening, spend time outside, get your hands in the dirt, and grow food to donate to our local... By: Sp...'. A small star icon is visible in the top right corner of the page.

➤ Results from original Unt.edu :

The screenshot shows the original UNT.edu search results page. The browser address bar displays 'unt.edu/search-results?search=student+life&sa=Search'. The page has a green header with the UNT logo and a navigation menu including 'Admissions', 'Academics', 'Student Life', 'About UNT', 'Research', 'Locations', 'Athletics', and 'Giving'. Below the header is a search bar with the placeholder text 'Search' and a green 'Search' button. The search results are titled 'Search Results' and show 'About 6,380,000,000 results (0.48 seconds)'. The results are sorted by 'Relevance'. The first result is 'Student Life | University of North Texas' with a green URL 'www.unt.edu > student-life' and a snippet: 'UNT is a student-focused, public, research university located in Denton, Texas. As one of Texas' largest universities, we offer 109 bachelor's, ...'. The second result is 'Student Activities | Division of Student Affairs' with a green URL 'studentaffairs.unt.edu > student-activities-center' and a snippet: 'We also offer programs and services specifically designed to help off-campus, commuter, graduate, and non-traditional students connect to campus life. UNT ...'. The third result is 'Center for Fraternity and Sorority Life | Division of Student Affairs' with a green URL 'studentaffairs.unt.edu > cfsf' and a snippet: 'UNT CFSL Diversity Statement. *On behalf of all the Greek-lettered organizations at the University of North Texas, the Interfraternity Council, National Pan- ...'. A small star icon is visible in the top right corner of the page.

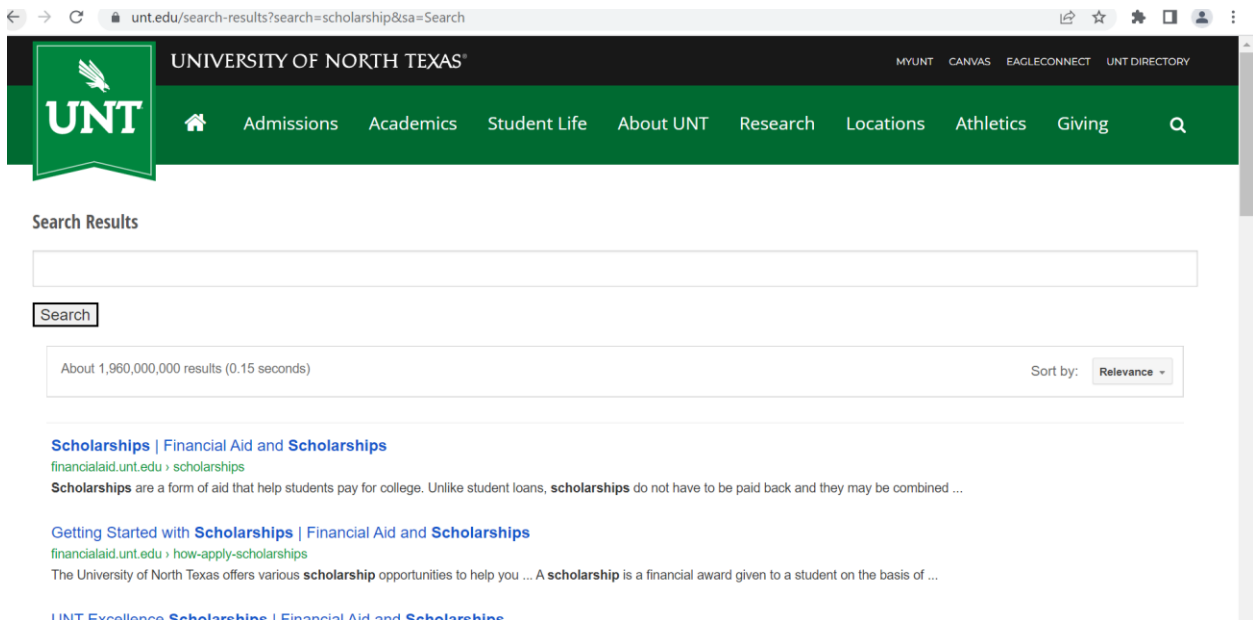
Input3: Scholarship

➤ Results from our application



A screenshot of a web browser showing the University of North Texas Search page. The browser's address bar displays "127.0.0.1:5000/scholarship". The page features a green header with the UNT logo and the text "University Of North Texas Search". Below the header is a search bar with the placeholder text "Enter keyword to search" and a green "Search" button. The search results are titled "Search Results for scholarship". The first result is "Scholarships" with a link to "http://financialaid.unt.edu/scholarship". The text of the result states: "sumer Info & Guidelines Contact Us FAQs Scholarships Scholarships are a form of aid that help students pay for college. Unlike student loans, scholarships do not have to be paid back and they may be combined with other kinds of financial aid to cover up to the total cost of attendance. UNT awards thousands of scholarships each year and not just to students who have earned high grades or have high test scores. In addition to academic merit, scholarships are awarded to students based on other qual...". The second result is "How do I apply for scholarships & financial aid?" with a link to "https://admissions.unt.edu/international/how-to-apply-for-scholarships-and-financial-aid" and the text "International Admissions".

➤ Results from original Unt.edu :



A screenshot of the University of North Texas search results page. The browser's address bar shows "unt.edu/search-results?search=scholarship&sa=Search". The page has a green header with the UNT logo and navigation links: "Admissions", "Academics", "Student Life", "About UNT", "Research", "Locations", "Athletics", "Giving", and a search icon. Below the header is a search bar with the placeholder text "Search" and a green "Search" button. The search results are titled "Search Results". Below the search bar, it says "About 1,960,000,000 results (0.15 seconds)" and "Sort by: Relevance". The first result is "Scholarships | Financial Aid and Scholarships" with a link to "financialaid.unt.edu > scholarships". The text of the result states: "Scholarships are a form of aid that help students pay for college. Unlike student loans, scholarships do not have to be paid back and they may be combined ...". The second result is "Getting Started with Scholarships | Financial Aid and Scholarships" with a link to "financialaid.unt.edu > how-apply-scholarships". The text of the result states: "The University of North Texas offers various scholarship opportunities to help you ... A scholarship is a financial award given to a student on the basis of ...". The third result is "UNT Excellence Scholarships | Financial Aid and Scholarships".

References

1. <http://www.learningaboutelectronics.com/Articles/How-to-find-all-hyperlinks-on-a-web-page-in-Python-using-BeautifulSoup.php>
2. <https://www.kaggle.com/code/adeptnugopal/nlp-text-similarity-using-cosine-count-vectorizer>
3. <https://towardsdatascience.com/vector-space-models-48b42a15d86d>
4. <https://www.python.org/>
5. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
6. <https://www.w3schools.com/html/>