CASEBOT: PROJECT1 **REPORT PART -1**

Name: Sahrudavi Caroline P. **NETID: SXP230055**

Web Crawling:

 Utilizing web crawling techniques, I collected 50 URLs. I used https://en.wikipedia.org/wiki/Christopher Nolan as the starter url. After collecting 25 wikipedia URLs, the crawler collects 25 more urls

```
https://www.indiewire.com/2020/08/tenet-critics-divided-christopher-nolan-1234581611/
https://www2.bfi.org.uk/news-opinion/sight-sound-magazine/features/christopher-nolan-escape-artist
https://pop.inquirer.net/37978/kenneth-branaph-must-evacuate-soldiers-trapped-dunkirk
https://www.empireonline.com/movies/features/empire-30-christopher-nolan-set-visits/
https://deadline.com/2024/01/christopher-nolan-france-honorary-cesa-r1235797659/
http://www.idvidordwell.net/blog/2017/08/09/dunkirk-part-2-the-art-film-as-event-movie/
https://www.indiewire.com/gallery/christopher-nolan-unmade-films/
https://www.latimes.com/entertainment/movies/la-ca-mm-dark-knight-10th-anniversary-20180822-story.html
https://indianexpress.com/article/entertainment/hollywood/happy-birthday-christopher-nolan-steven-spielberg-james-cameron-5860421/
https://www.indiewire.com/influencers/dunkirk-production-team-nolan/
https://www.telegraph.co.uk/films/2017/07/08/spitfires-flotillas-boats-rough-seas-1000-extras-christopher/
Starting URL:https://www.complex.com/pop-culture/tv-movie-creatives-that-defined-the-2010s/christopher-nolan
 Starting URL:http://time.com/3822892/christopher-nolan-2015-time-100
  Starting URL:http://filmlondon.org.uk/news/2010/july/nolans_mind_games
 Starting URL:https://variety.com/2015/film/news/christopher-nolans-syncopy-teaming-with-zeitgeist-on-blu-ray-releases-exclusive-1201437195/
 https://variety.com/2015/film/news/christopher-nolans-syncopy-teaming-with-zeitgeist-on-blu-ray-releases-exclusive-1201437195/#respond https://variety.com/2024/film/columns/christopher-nolan-tenet-flaws-oppenheimer-1235922910/ https://variety.com/t/christopher-nolan/
Starting URL:https://variety.com/2014/film/reviews/film-review-interstellar-1201338475/
  https://variety.com/2024/film/columns/christopher-nolan-tenet-flaws-oppenheimer-1235922910/
 https://variety.com/t/christopher-nolan/
Starting URL:https://www.hollywoodreporter.com/movies/movie-news/christopher-nolan-pitch-to-studios-including-apple-seeking-his-next-film-1235014132/
 https://www.hollywoodreporter.com/t/christopher-nolan/
https://www.hollywoodreporter.com/novies/movie-news/cillian-murphy-broke-his-head-oppenheimer-emily-blunt-1235834267/
https://www.hollywoodreporter.com/business/business-news/christopher-nolan-rips-hbo-max-as-worst-streaming-service-denounces-warner-bros-plan-4101408/
https://www.hollywoodreporter.com/movies/movie-news/christopher-nolan-pitch-to-studios-including-apple-seeking-his-next-film-1235014132/#respond
Starting URL:https://www.independent.ie/entertainment/movies/how-christopher-nolan-has-held-true-to-his-sweeping-vision-with-tenet-39466428.html
  Starting URL:https://www.nytimes.com/2014/11/02/magazine/the-exacting-expansive-mind-of-christopher-nolan.html
 Starting URL:https://variety.com/2017/film/news/christopher-nolan-dunkirk-oscars-movies-tv-spielberg-1202607836/
 https://variety.com/t/dunkirk/
https://variety.com/2024/film/columns/christopher-nolan-tenet-flaws-oppenheimer-1235922910/
https://variety.com/t/christopher-nolan/
Starting URL:http://filmcomment.com/article/cinematic-faith-christopher-nolan-scott-foundas
```

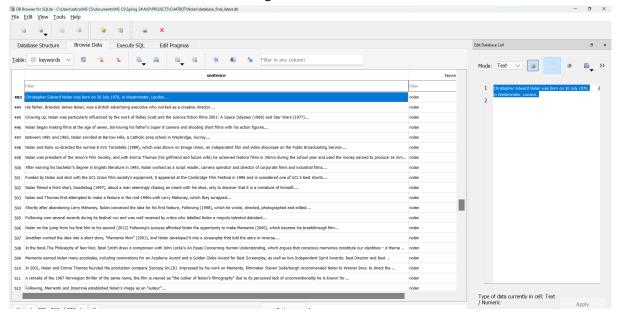
Snapshot of urls collected during crawling

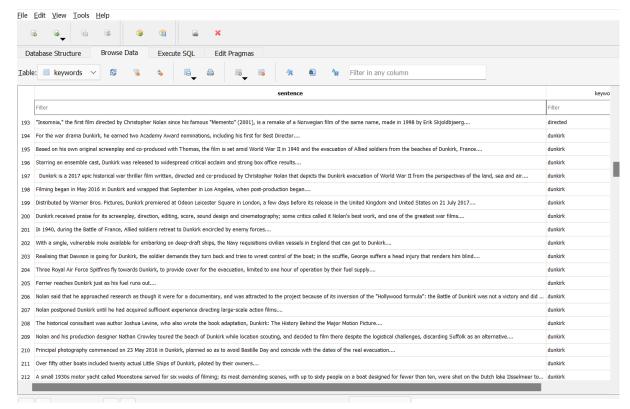
I then scraped all the collected 50 URLs for raw text. I got 34 urls after rejecting urls with request errors.

While scraping and cleaning - error requesting URLs

Clean files were then created after the gathered raw text underwent preprocessing steps to refine it for analysis. This included stripping HTML tags, removing citation number enclosed in square brackets, and tokenizing the text, removing the punctuation and stop words

- I employed Natural Language Processing (NLP) techniques to identify significant keywords within the processed text using tf-idf. These keywords served as essential markers for categorization and retrieval of relevant information.
- I manually picked 16 relevant keywords on which the knowledge base was populated
- I used an sqlite database to store the knowledge base
- My key terms are:
 - o nolan
 - academy
 - dunkirk
 - oppenheimer
 - o directed
 - o success
 - o awards
 - o critics
 - memento
 - o role
 - o knight
 - o cast
 - o inception
 - interstellar
 - o reviews
 - o tenet
- Screenshots of the knowledge base





Screenshots showing the 'keywords' table (sentence,keyword) which is the knowledge of my chatbotthe CaSebot (SQLite Database viewed in the DB Browser)