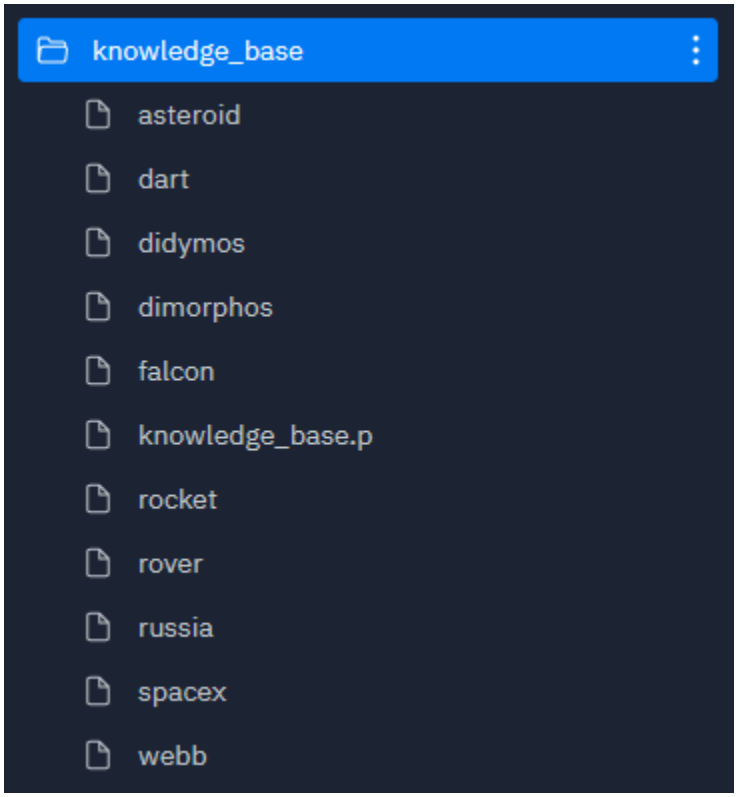
Ryan Dimaranan and Hannah Valena
CS 4395.001

Portfolio: Finding or Building a Corpus

How we Created Our Knowledge Base

To create our knowledge base, we started out with a website that was relevant to our topic and had sites from outside its domain. Using Python, we built a web crawler function that recursively found relevant websites to our topic by crawling sites that had our topic in its URL. Our function took the first two sites from the starter site relevant to our topic, added it to our list of websites, then crawled those two sites and repeated until we had 15 relevant websites. Then we scraped the data off of these sites, removing unrelated text from typical HTML tags that include no valuable text data, such as stylesheets and scripts. With our new corpus of documents we removed non alphanumeric characters and stopwords and used this filter data to create a TF-IDF representation of the terms in each document. From this data representation we retrieved the top 40 terms and selected 10 words we thought were most relevant to our topic. To build the knowledge base itself, we took sentences that had the specific term, cleaned them up and added it to a Python dictionary and also saved the data for each term in its own file.

Knowledge Base



```
Extracting 40 key terms...

Top 40 key terms: ['asteroid', 'dart', 'webb', 'hera', 'falcon', 'dimorphos', 'ball', 'basketball', 'esa', 'planet
ary', 'moon', 'telescope', 'didymos', 'double', 'redirection', 'tennis', 'apart', 'rover', 'spacex', 'payloads', '
light', 'year', 'pst', 'away', 'russia', 'far', 'monisha', 'ravisetti', 'launches', 'merlin', 'jwst', 'pixel', 'ja
vascript', 'enable', 'debris', 'dataset', 'dragon', 'december', 'spent', 'impact']

The top 10 key terms are: ['webb', 'dart', 'falcon', 'asteroid', 'didymos', 'dimorphos', 'spacex', 'rover', 'russi
a', 'rocket']

Building a knowledge base...

Dumping knowledge base dict to pickle file...

Creating knowledge base text files for each term...

Knowledge base:

{'webb': ["nasa's james webb space telescope just launched: what happens next - cnet amazon prime day pixel 7 phon
e pixel 7 vs. rivals target deal days pixel 7 pro cameras pixel watch google pixel event mcdonald's boo buckets yo
ur guide to a better future want cnet to notify you of price drops and the latest stories?", "no, thank you accept
 science follow nasa's james webb space telescope just launched: what happens next and why this is such a monument
al event for the entire field of astronomy.", "see full bio 4 min read this false color infrared exposure captures
 the christmas morning liftoff from south america of nasa's james webb space telescope.", 'the james webb space te
lescope launched, embarking on a journey to recalibrate how we view the universe.', "not only will webb teach us a
bout hidden regions of space, it has the power to prove whether we've correctly documented the events that took pl
ace immediately after the big bang.", "after a seamless deployment of webb's solar array about 30 minutes later, t
he telescope began charging up for the rest of its cosmic expedition.", 'webb is set to travel 1 million miles (1.
', 'webb will offer us a new story of the cosmos, completely unfiltered.', "this will be a giant step forward from
 the hubble telescope, which launched along with the discovery space shuttle in 1990. but before we get into the i
ncredible data that webb promises to reveal, here's the context of what just blasted into outer space after two de
cades of work and about $10 billion.", 'you can also take a deeper dive into the technical aspects of webb here.',
 "webb's impressive specsthis is a 3d rendering of how james webb will look in space once fully deployed.", "near-
infrared camera (nircam): webb's primary imager will detect the earliest stars and galaxies that formed.", "why we
bb is a very, very big dealwebb's promise rests on its unprecedented infrared imaging capabilities, especially wit
h nircam.", "a quick physics recap: to get to webb's promise, we have to talk about the electromagnetic spectrum."
, 'webb, by contrast, is designed for the job.', 'further, when you take into consideration exactly where that inf
rared light is coming from, in a sense webb has a time machine onboard.', "lockheed martin engineer alison nordt w
orks on webb's nircam.", 'webb can look much farther into deep space, about 13.', 'as per predictions of the exper
ts behind webb, the telescope could reveal habitable exoplanets, secrets of black holes and perhaps even evidence
of life beyond earth.', "after sighing a breath of relief, astronomers will sit tight for the next six months, awa
iting webb's command on how to alter, amend and footnote the entire field of astronomy.", 'one of the last views w
e will ever have of the webb space telescope as it starts a journey of a million miles on dec. 25. nasa/screenshot
 by cnet get the cnet science newsletter unlock the biggest mysteries of our planet and beyond with the cnet scien
ce newsletter.', "nasa's james webb space telescope launches: 'milestone achieved' - cnet x science nasa's james w
ebb space telescope launches: 'milestone achieved' the agency's most powerful telescope ever just blasted off.", '
the james webb space telescope -- a multi-billion dollar, gold-plated and unimaginably precise piece of machinery
-- successfully blasted off from south america, beginning its legendary trip among the stars.', "at 4: 20 a. m. pt
 (9: 20 a. m. local time in french guiana), webb's launch window opened.", "keep up to date with the launch info a
nd updates below, on nasa tv or on nasa's webb stream on youtube.", 'december 25, 2021, 5: 03 am pst ground contro
l to webb, we have communication by monisha ravisetti this outstanding, meticulously crafted and trailblazing piec
e of machinery is now on its way to show us the first light in the universe -- from 13.', "we'll start hearing bac
k from webb six months from now.", '@nasawebb is safely in space, powered on, and communicating with ground contro
llers.', "webb's solar panels have deployed and it's charging up for the cosmic adventure of a lifetime.", '#nasaw
ebb’s solar array has successfully deployed, and webb’s batteries are charging up ⚡ #unfoldtheuniverse pic.', 'co
m/8ozjigro6p— nasa webb telescope (@nasawebb) december 25, 2021 december 25, 2021, 4: 51 am pst our first and last
 view of webb by monisha ravisetti amid applause, the upper stage camera offers us the very first views of webb, a
nd soon after, our very last views as the telescope moves toward its work station.', 'nasa december 25, 2021, 4: 4
7 am pst soon to enter critical unfurling procedures by monisha ravisetti main engines are off "exactly as expecte
d" and webb is in its coasting phase.', "december 25, 2021, 4: 36 am pst awaiting solar deployment by monisha ravi
setti after webb's thrusters are working, its solar arrays are scheduled to be deployed at the 30 minute mark post
-launch.", '"within the first day, we expect to get signals back that it has deployed the solar panels and that it
 is power positive, " according to webb engineer alison nordt, the space science and instrumentation director at a
erospace giant lockheed martin.', "once that's done, ground control will be waiting for its very first communicati
on with webb.", "everything's still looking nominal and the section of the rocket containing webb is now traveling
 alone at 220 kilometers in altitude and a bit over 7 kilometers per second.", '#nasawebb is now on its way to its
 new home a million miles from earth, where it will #unfoldtheuniverse in ways we’ve never seen before.', 'co/td5j
9eusnh— nasa webb telescope (@nasawebb) december 25, 2021 december 25, 2021, 4: 20 am pst 30 seconds.', "get ready
 to say 3. . . 2. . . 1. . . by monisha ravisetti nasa december 25, 2021, 4: 18 am pst approaching the final co
down by monisha ravisetti nasawebb's right up there, experiencing its last days on earth in the tip of the ariane
5 rocket.", "december 25, 2021, 4: 13 am pst confirmed, webb will launch in 7 minutes by monisha ravisetti it's a
```

Sample Dialog

Me: When did the James Webb telescope launch?
Bot: The James Webb telescope launched on 25 December 2021
Me: What are some recent developments in space from Russia?
Bot: Next Mars rover likely won't launch in 2022 due to Russia sanctions
Me: Will the asteroid Dimorphos impact the Earth?
Bot: Dimorphos poses no threat to earth.
Me: What is the size of Dimorphos?
Bot: Dimorphos measures only 160 meters (525 feet) across.
Me: What is the size of Didymos?
Bot: Didymos measures just 780 meters (a half-mile) across
Me: Who was on the most recent SpaceX launch?
Bot: Commander Jared Isaacman, Hayley Arceneaux, Sian Proctor and Chris Sembroski blasted into space aboard a SpaceX crew dragon as part of the Inspiration4 mission.