

Cady Baltz
CS 4395.001 Human Language Technologies
Portfolio Assignment: Finding or Building a Corpus
3/11/2023

Introduction:

In this project, I designed a web crawler to find websites related to tourism in Japan. The goal of this program was to create a knowledge base of information that could be used to create a travel agent chatbot that can help answer questions about visiting Japan.

Knowledge Base Creation:

1) Finding 15 Relevant URLs

First, I found a website related to traveling in Japan that covered a wide range of information, from Japan's history and culture to its most popular tourist sites. Most importantly, this website contained external links that would allow me to further traverse the Internet for more information about this topic. The URL of this website is: <https://en.m.wikivoyage.org/wiki/Japan>.

Then, I created a `web_crawler()` function to find fourteen additional relevant links branching off of my initial starter URL. Writing this function required a lot of trial and error in order to ensure that the websites I found contained useful information. To find links, I created a 'link_queue' that initially only contained my starter URL. Then, I scraped the text off of this URL using the Python BeautifulSoup library, including all of its links. I added all of these links to the queue to continue searching indefinitely (although I added a cap of 1000 iterations to prevent infinite loops).

In order for a link to be added to my final 'related_links' list, it had to satisfy the following requirements:

1. The URL could not be in my list of "blocked_hosts".
 - a. After observing the output of my function, I noticed some websites that would not allow me to scrape their text. For example, certain websites included a "Prove you are not a robot" message. To handle these edge cases, I added the website's domain name to a manual list in my program each time I encountered one.
2. It had to have a different hostname than all of the existing links in the array
 - a. This requirement helped ensure diversity in my webpage selection.
3. It had to be successfully scraped by BeautifulSoup.

- a. I used a try/except block to try scraping the website and check whether it was successful.
4. It had to contain at least 500 words of paragraph text (text within <p> tags).
 - a. I specifically searched within the <p> tags using BeautifulSoup to exclude webpage text that was not useful, like text from navigation bars and footers.
 - b. I included the 500 words requirement to ensure that I was scraping enough text to build a sizable knowledge base.
5. The text had to use ASCII characters.
 - a. I added this requirement after I found that a lot of the webpages being returned were written in Japanese.

I continued checking links from my queue using these requirements until I had fifteen valid links.

2) Cleaning the Raw Webpage Text

After extracting the text within the <p> tags using BeautifulSoup, I saved this raw text to its own file in my 'raw_page_texts' directory. Then, I iterated through all of these raw text files to clean them in my clean_text() function. In this function, I clean the text by removing extra whitespace, including new lines and tabs. Then, I used NLTK's sent_tokenize() method to separate the raw text into individual sentences. Then, I iterated through each tokenized sentence and printed it on a line in a new file in my 'clean_page_texts' directory.

3) Determining the Top 25 Terms

To determine the top 25 terms, I applied simple term frequency when analyzing my documents. In my get_top_25_terms() method, I iterated through each cleaned text file from the previous step. Then, I tokenized each individual sentence using the NLTK word_tokenize() method. Finally, I iterated through each token and incremented its count in a dictionary I created.

To ensure my top 25 terms were useful, I applied the following requirements to them:

1. They are alphabetic
2. They are not in the English stop words list from NLTK
3. They have a length greater than five characters

Then, I simply sorted my count dictionary and returned the 25 terms with the greatest counts across all of my documents.

4) Selecting the Top 10 Terms

After finding my top 25 terms, I printed them all to console and manually observed them. I selected ten terms that I thought would be useful for a chatbot travel agent to recognize, and that I thought would result in substantial information in my knowledge base.

Selected terms:

1. japanese
2. country
3. english
4. popular
5. cities
6. hotels
7. restaurants
8. travel
9. international
10. stores

5) Creating the Knowledge Base

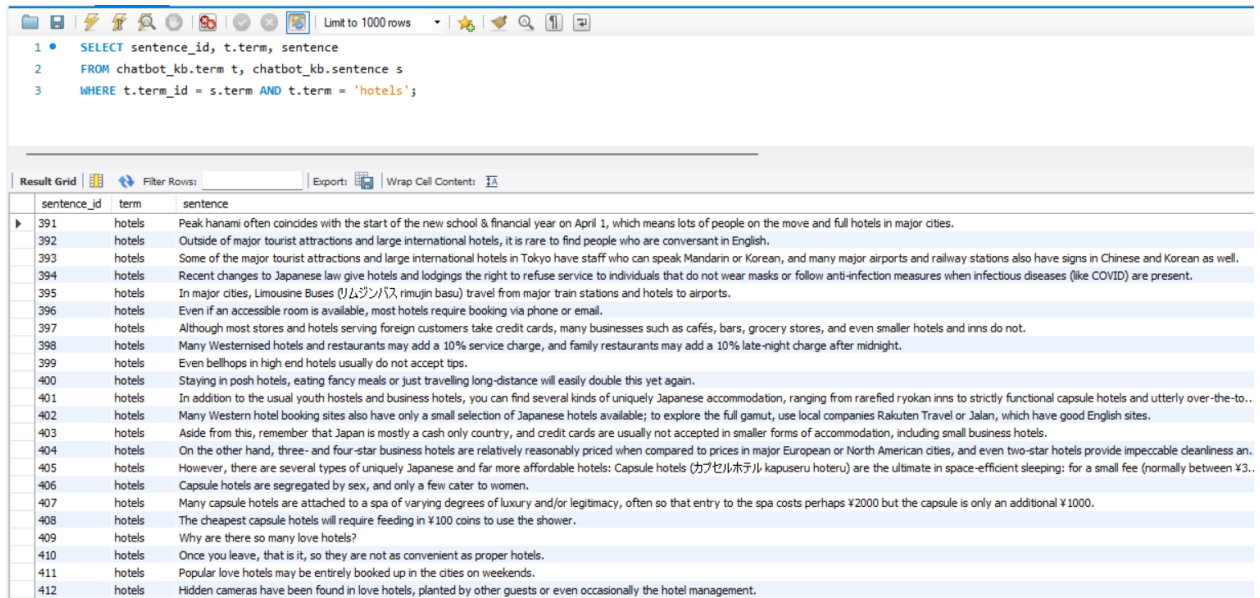
Once I had my top 10 terms, I searched for information related to these terms in my cleaned text files. I created a new dictionary with 10 keys – one for each term listed above. Then, I iterated through each sentence of each cleaned text file. If the sentence contained one of the terms listed above, I included it in the corresponding entry of the dictionary. Then, I dumped this new dictionary into a pickle file named “chatbot_kb.p”.

5) Creating a SQL Database

To further improve the usability of my knowledge base, I created a SQL database for it. The database consists of two tables: term and sentence. The ‘term’ table contains the 10 terms listed above, as well as a unique ID for each. The ‘sentence’ table contains pieces of knowledge that map to these terms. Each entry of the ‘sentence’ table includes a unique ID, a sentence of information (with a maximum length of 1000 characters for handling storage concerns), and a foreign key that maps to the ‘term’ table.

By arranging my data in a SQL database, I can easily query it and view the information. In total, my database contains 871 entries. On the next page, you can find screenshots of my SQL knowledge base showing the execution of a few different queries.

Query for 'hotels':



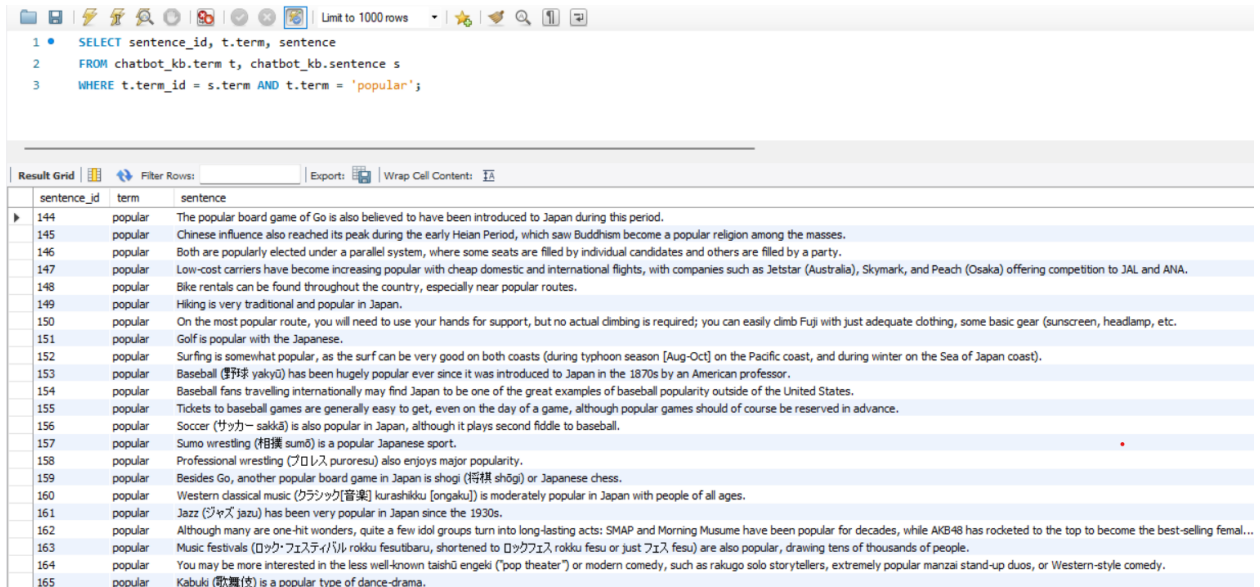
Limit to 1000 rows

```
1 • SELECT sentence_id, t.term, sentence
2 FROM chatbot_kb.term t, chatbot_kb.sentence s
3 WHERE t.term_id = s.term AND t.term = 'hotels';
```

Result Grid | Filter Rows: | Exports: | Wrap Cell Contents: [↗](#)

	sentence_id	term	sentence
▶	391	hotels	Peak hanami often coincides with the start of the new school & financial year on April 1, which means lots of people on the move and full hotels in major cities.
	392	hotels	Outside of major tourist attractions and large international hotels, it is rare to find people who are conversant in English.
	393	hotels	Some of the major tourist attractions and large international hotels in Tokyo have staff who can speak Mandarin or Korean, and many major airports and railway stations also have signs in Chinese and Korean as well.
	394	hotels	Recent changes to Japanese law give hotels and lodgings the right to refuse service to individuals that do not wear masks or follow anti-infection measures when infectious diseases (like COVID) are present.
	395	hotels	In major cities, Limousine Buses (リムジンバス rimujin basu) travel from major train stations and hotels to airports.
	396	hotels	Even if an accessible room is available, most hotels require booking via phone or email.
	397	hotels	Although most stores and hotels serving foreign customers take credit cards, many businesses such as cafés, bars, grocery stores, and even smaller hotels and inns do not.
	398	hotels	Many Westernised hotels and restaurants may add a 10% service charge, and family restaurants may add a 10% late-night charge after midnight.
	399	hotels	Even bellhops in high end hotels usually do not accept tips.
	400	hotels	Staying in posh hotels, eating fancy meals or just travelling long-distance will easily double this yet again.
	401	hotels	In addition to the usual youth hostels and business hotels, you can find several kinds of uniquely Japanese accommodation, ranging from rarefied ryokan inns to strictly functional capsule hotels and utterly over-the-to...
	402	hotels	Many Western hotel booking sites also have only a small selection of Japanese hotels available; to explore the full gamut, use local companies Rakuten Travel or Jalan, which have good English sites.
	403	hotels	Aside from this, remember that Japan is mostly a cash only country, and credit cards are usually not accepted in smaller forms of accommodation, including small business hotels.
	404	hotels	On the other hand, three- and four-star business hotels are relatively reasonably priced when compared to prices in major European or North American cities, and even two-star hotels provide impeccable cleanliness an...
	405	hotels	However, there are several types of uniquely Japanese and far more affordable hotels: Capsule hotels (カプセルホテル kapuseru hoteeru) are the ultimate in space-efficient sleeping: for a small fee (normally between ¥3...
	406	hotels	Capsule hotels are segregated by sex, and only a few cater to women.
	407	hotels	Many capsule hotels are attached to a spa of varying degrees of luxury and/or legitimacy, often so that entry to the spa costs perhaps ¥2000 but the capsule is only an additional ¥1000.
	408	hotels	The cheapest capsule hotels will require feeding in ¥100 coins to use the shower.
	409	hotels	Why are there so many love hotels?
	410	hotels	Once you leave, that is it, so they are not as convenient as proper hotels.
	411	hotels	Popular love hotels may be entirely booked up in the cities on weekends.
	412	hotels	Hidden cameras have been found in love hotels, planted by other guests or even occasionally the hotel management.

Query for 'popular':



Limit to 1000 rows

```
1 • SELECT sentence_id, t.term, sentence
2 FROM chatbot_kb.term t, chatbot_kb.sentence s
3 WHERE t.term_id = s.term AND t.term = 'popular';
```

Result Grid | Filter Rows: | Exports: | Wrap Cell Contents: [↗](#)

	sentence_id	term	sentence
▶	144	popular	The popular board game of Go is also believed to have been introduced to Japan during this period.
	145	popular	Chinese influence also reached its peak during the early Heian Period, which saw Buddhism become a popular religion among the masses.
	146	popular	Both are popularly elected under a parallel system, where some seats are filled by individual candidates and others are filled by a party.
	147	popular	Low-cost carriers have become increasingly popular with cheap domestic and international flights, with companies such as Jetstar (Australia), Skymark, and Peach (Osaka) offering competition to JAL and ANA.
	148	popular	Bike rentals can be found throughout the country, especially near popular routes.
	149	popular	Hiking is very traditional and popular in Japan.
	150	popular	On the most popular route, you will need to use your hands for support, but no actual climbing is required; you can easily climb Fuji with just adequate clothing, some basic gear (sunscreen, headlamp, etc.
	151	popular	Golf is popular with the Japanese.
	152	popular	Surfing is somewhat popular, as the surf can be very good on both coasts (during typhoon season [Aug-Oct] on the Pacific coast, and during winter on the Sea of Japan coast).
	153	popular	Baseball (野球 yakyū) has been hugely popular ever since it was introduced to Japan in the 1870s by an American professor.
	154	popular	Baseball fans travelling internationally may find Japan to be one of the great examples of baseball popularity outside of the United States.
	155	popular	Tickets to baseball games are generally easy to get, even on the day of a game, although popular games should of course be reserved in advance.
	156	popular	Soccer (サッカー sakka) is also popular in Japan, although it plays second fiddle to baseball.
	157	popular	Sumo wrestling (相撲 sumō) is a popular Japanese sport.
	158	popular	Professional wrestling (プロレス puroresu) also enjoys major popularity.
	159	popular	Besides Go, another popular board game in Japan is shogi (将棋 shōgi) or Japanese chess.
	160	popular	Western classical music (クラシック音楽 kurashikku ongaku) is moderately popular in Japan with people of all ages.
	161	popular	Jazz (ジャズ jazz) has been very popular in Japan since the 1930s.
	162	popular	Although many are one-hit wonders, quite a few idol groups turn into long-lasting acts: SMAP and Morning Musume have been popular for decades, while AKB48 has rocketed to the top to become the best-selling femal...
	163	popular	Music festivals (ロックフェスティバル roku fesutobaru, shortened to ロックフェス roku fesu or just フェス fesu) are also popular, drawing tens of thousands of people.
	164	popular	You may be more interested in the less well-known taishū engaku ("pop theater") or modern comedy, such as rakugo solo storytellers, extremely popular manzai stand-up duos, or Western-style comedy.
	165	popular	Kabuki (歌舞伎) is a popular type of dance-drama.

Query for 'restaurants':

Limit to 1000 rows

```

1 • SELECT sentence_id, t.term, sentence
2 FROM chatbot_kb.term t, chatbot_kb.sentence s
3 WHERE t.term_id = s.term AND t.term = 'restaurants';

```

Result Grid

Filter Rows:

Exports

Wrap Cell Contents:

	sentence_id	term	sentence
509		restaurants	Though their food is relatively uninteresting, these restaurants usually have illustrated menus, so travellers who cannot read Japanese can use the photos to choose and communicate their orders.
510		restaurants	Vegetarians may want to seek out Indian or Italian restaurants in larger cities.
511		restaurants	However, ID verification is almost never requested at restaurants, bars, or convenience stores, so long as the purchaser does not appear obviously underage.
512		restaurants	In Japanese restaurants, beer is typically served in various sizes of bottles (瓶 bin), or draft (生 nana meaning "fresh").
513		restaurants	Most restaurants serve filtered tap water for free.
514		restaurants	Since 2020, even restaurants in Tokyo only allow smoking in dedicated, separately ventilated smoking sections.
515		restaurants	Amid the COVID-19 outbreak, there has been a perceived spike in xenophobia, with some shops and restaurants having refused service to foreigners, especially Chinese people.
516		restaurants	Christmas Eve is considered to be one of the most romantic days of the year in Japan, and restaurants will be fully booked by young couples looking to have a romantic night out, so be sure to make your dinner reservations well in advance.
517		restaurants	In addition to public transport, smart cards are used for all sorts of electronic payments, so they can be used at vending machines, convenience stores, fast food restaurants, etc.
518		restaurants	Hostesses work in bars and sing karaoke to entertain, compared to geisha coming to tea houses and restaurants to perform traditional Japanese arts.
519		restaurants	Maid cafés and other cosplay restaurants have employees dressed as French maids pamper their clients while serving them beverages and food.
520		restaurants	This includes vending machines, convenience stores, fast food restaurants, etc.
521		restaurants	Many Westernised hotels and restaurants may add a 10% service charge, and family restaurants may add a 10% late-night charge after midnight.
522		restaurants	The Michelin Guide is considered by many Western visitors to be the benchmark of good restaurants in Japan.
523		restaurants	But many top fine dining restaurants are not listed in it by choice.
524		restaurants	Many restaurants give you a hot towel (o-shibori) to wipe your hands with (not your face) as soon as you sit down.
525		restaurants	In all types of Japanese restaurants, staff generally ignore you until you ask for something.
526		restaurants	Tipping is not customary in Japan, although many sit-down restaurants apply 10% service charges and 24-hour "family restaurants" usually have a 10% late-night surcharge.
527		restaurants	The number of restaurants (レストラン resutoran) in Japan is stupendous, and you will never run out of places to go.
528		restaurants	Japan is tied with France for first place as the country with the most Michelin-star restaurants.
529		restaurants	Most sushi restaurants keep a handy multilingual decoding key on hand or on the wall.
530		restaurants	At the finest sushi restaurants, the chef puts a dab of fiery wasabi radish into the sushi, and glazes the fish with soy sauce for you.

Query for 'hotels':

</

6) Sample Dialog

Using pieces of knowledge I found in my database, here is what I envision could be a chatbot conversation. All of the Chatbot responses come directly from my knowledge base with no modification of the text. I also highlighted the keywords that the user said that would help the chatbot form its answer using my knowledge base.

Chatbot: Hello, how can I help you today?

User: I want to plan a trip to Japan! What is the best way to **travel** there?

Chatbot: Delta Air Lines, United Airlines, and American Airlines also operate flights from numerous U.S. cities into Narita or Haneda, as does Air Canada from several Canadian cities.

User: What is the best way to **travel** around Japan?

Chatbot: A tourist who plans to travel a great deal around the country should consider investing in a Japan Rail Pass, which offers — with a few exceptions — unlimited travel on all Japan Railway (JR) services including bullet trains, limited express and regular commuter trains.

User: What are the most **popular** things to do in Japan?

Chatbot: Hiking is very traditional and popular in Japan.

User: Hiking sounds fun! Where should I hike?

Chatbot: On the most popular route, you will need to use your hands for support, but no actual climbing is required; you can easily climb Fuji with just adequate clothing, some basic gear (sunscreen, headlamp, etc.

User: What about food? Does Japan have good **restaurants**?

Chatbot: Japan is tied with France for first place as the country with the most Michelin-star restaurants.

User: That sounds great! I want to go to Japan now. Can you help me book a **hotel**?

Chatbot: Some major chains of cheaper business hotels include Tokyu REI Hotels, known for its generously sized rooms, Sunroute Hotels and Toyoko Inn. Even if an accessible room is available, most hotels require booking via phone or email.