Farishah Nahrin
CS 6301.M02

Web Crawler: Report & Reflection

(1) Describe how you created your knowledge base, indicate your top 10 terms
(2) Write up a sample dialog you would like to create with a chatbot based on your knowledge base
(3) What was your starter site
(4) What kind of data did you get
(5) How did you clean up the data
(6) What were your top terms
(7) Show us your knowledge base, include screen shots of the knowledge base
(8) How might you use this data for a chatbot

Describe how you created your knowledge base, and indicate your top 10 terms
APIs used: Python Requests Library, and Scrapy (because instructions said that other APIs can be used for scraping).

The knowledge base was created with the following steps:

1. The entire text data was aggregated, with data from all the internal and external relevant links.
    a. The code uses the requests library to fetch the HTML content of the virtualbangladesh site.
    b. Once the HTML content is fetched, it can be passed to the TextResponse object of the Scrapy library for parsing.
    c. The TextResponse object allows for easy extraction of content using CSS selectors or XPath expressions. The code could then iterate through all the URLs on the site, and then it uses Requests and Scrapy again, to extract the relevant information.
    d. After aggregating all the scraped data, the code does the tokenization, which involves splitting the text into individual words or tokens. And this helps prepare the data for further processing, like filtering non-alphanumeric characters and removing stop words.
2. Then the entire data is cleaned up and the sentence gets tokenized.
    a. After aggregating the scraped data, I did the tokenization, which involved splitting the text into individuals words or tokens. And this helps prepare the data for further processing, like filtering non-alphanumeric characters and removing stop words. This is how the entire data was cleaned and how the sentences were tokenized.
    b. sent_tokenize from the NLTK library was used to do the sentence tokenization
    c. The reason why I am using Scrapy instead of BeautifulSoup, is because i have a bit more experience with Scrapy. And the instructions said, "be able to do web

scraping with Beautiful Soup or other APIs" So this is the other API i have chosen to scrape with. After extracting the text with Scrapy, the code could pass the text to sent_tokenize to split it into individual sentences. This allows for the processing on a sentence-level, which was used for the knowledge base.

3. Then the top five sentences, for each keyword, contained this said keyword, and then it was combined with ‘,’ as a delimiter.

The Top 10 Terms I picked was: ['bangladesh', 'pakistan', 'east', 'awami', 'muslim', 'government', 'india', 'military', 'state', 'political']

Write up a sample dialog you would like to create with a chatbot based on your knowledge base

Note: All chatbot answers are copied, directly from my knowledge base.

Chatbot: Welcome to Virtual Bangladesh! What would you like to learn about Bangladesh
User Question: When did Bangladesh gain independence and political power?
Chatbot Answer: Pakistan’s forces surrendered on December 16, 1971. india had taken numerous prisoners and gained control of a large area of East Pakistan, which is now Bangladesh.

User Question: What was Bangladesh called before it gained independence?
Chatbot Answer: The war pitted bangladesh (then east pakistan) and later helped by india against the west pakistanis and lasted over a duration of nine months.

User Question: Why was there a war between Pakistan and Bangladesh?
Chatbot Answer: Muhammad Ali Jinnah, leader of the muslim league, publicly endorsed the “pakistan resolution” that called for the creation of an independent state in regions where muslims were a majority. The war pitted Bangladesh against the West Pakistanis and lasted over a duration of nine months.

What was your starter site?
This was the starter site: https://www.virtualbangladesh.com/

What kind of data did you get?
“Virtual Bangladesh is a comprehensive website about the country of Bangladesh. It provides a wealth of information about the history, culture, geography, people, and economy of Bangladesh. The website also includes a wide range of articles, maps, and photos about the country, as well as a directory of resources for travelers and researchers. The aim of the website is to provide a virtual representation of Bangladesh and to promote the country's rich cultural heritage, as well as to provide information and resources to those interested in learning more about Bangladesh.” Personally, I chose this as my starter site, because my motherland is Bangladesh. And although I was not born nor raised in Bangladesh, I wanted to use a starter site

that can help me learn more about my country, and also potentially help others learn about the history of Bangladesh, and how Bangladesh came to be. The script aims to store all the text data from all the links present on the website.

How did you clean up the data?
A simple regular expression was applied on the entire text data. The regex expression that was used was "[^\S\n]+\|*". The regex expression "[^\S\n]+|*" is a pattern used to match and capture a specific type of string in a text.

Here's a detailed explanation of each part of the expression:

1. "[^\S\n]" - This part of the expression uses a negated character class to match any whitespace characters except newline characters. The "\S" negates all non-whitespace characters and "\n" negates newline characters.
2. "+" - The plus sign means that one or more consecutive whitespace characters (excluding newline characters) must be matched.
3. "|*" - The pipe symbol "|" matches either a pipe symbol or nothing. The asterisk after the pipe symbol means that zero or more consecutive pipe symbols can be matched.

Basically this entire expression matches one or more consecutive whitespace characters (excluding newline characters) followed by zero or more consecutive pipe symbols.
My further processing steps include:

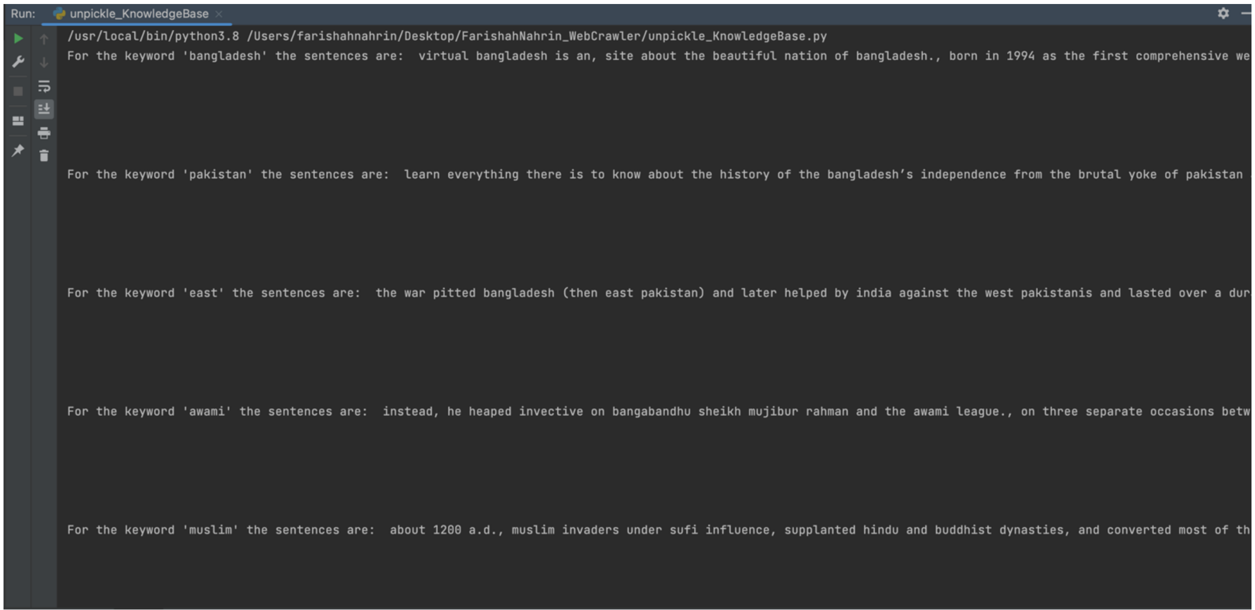
1. Tokenization: The text wsa tokenized using a regular expression tokenizer (RegexpTokenizer) which is configured to only keep word characters (\w+). The tokenizer splits the text into individual tokens (also known as words) based on the provided pattern, and converted to lower-case.
2. Stopword removal: The code removes stopwords from the list of tokens. Stopwords are common words such as "the", "is", "a", etc. which don't carry much meaning and are usually removed from the text to improve efficiency. The code uses the stopwords module from NLTK to get a list of stopwords in English.
3. Alpha tokens: The code removes non-alphabetic tokens by keeping the tokens that contain only alphabetic characters (using the isalpha() method).

What were your top terms?
The top 40 words that were determined with tf-df were:
['bangladesh', 'pakistan', 'র', 'east', 'league', 'bengal', 'government', 'west', 'awami', 'pakistani', 'national', 'people', 'political', 'new', 'state', 'one', 'ন', 'military', 'ব', 'ত', 'ক', 'bangla', 'also', 'dhaka', 'india', 'দ', 'first', 'would', 'ম', 'bengali', 'bangalis', 'two', 'bangali', 'স', 'ল', 'য', 'march', 'country', 'muslim', 'war']

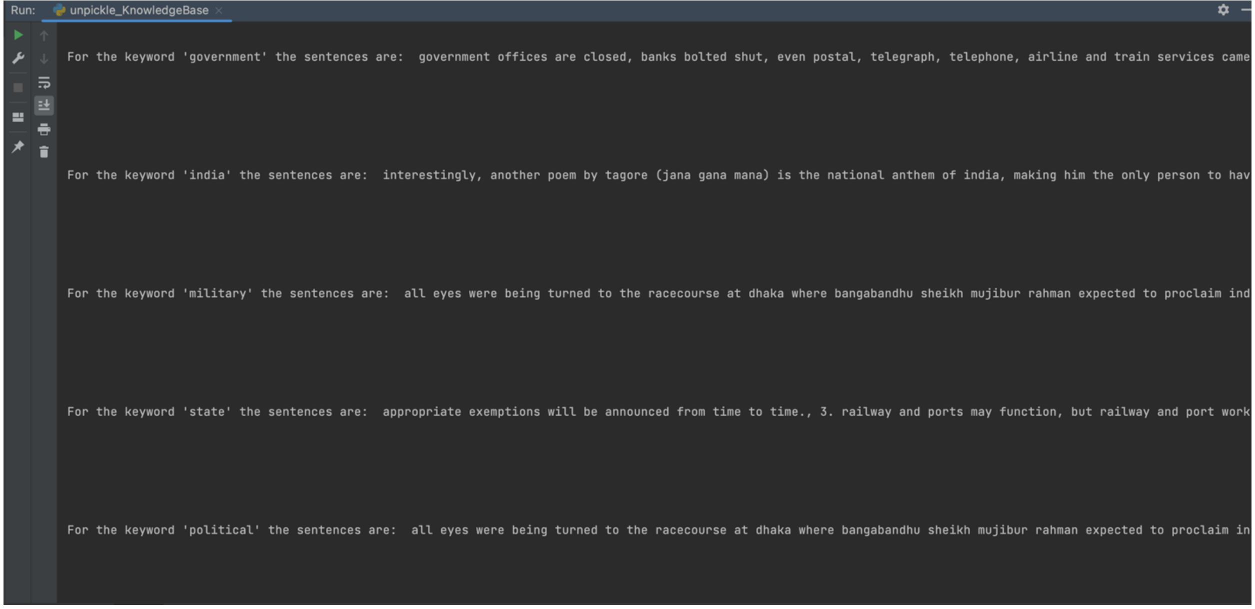
Note: Yes, there were a 10 Bengali words/letters that slipped in the top 40 because in TD-IDF, across the corpus, the relevance in calculated, therefore it can't identify if a Bengali character is in it. So, the Bengali token is treated as an English token. And NLTK does not support Bengali, so it was not possible to filter out Bangla stop words, or Bangla words in general. Nevertheless, the website I chose was mostly in English.

My top 10 manually selected words are:
['bangladesh', 'pakistan', 'east', 'league', 'muslim', 'government', 'india', 'military', 'state', 'political']

Show us your knowledge base, include screen shots of the knowledge base
I had to un-pickle the file using a separate .py file I created (unpickle_KnowledgeBase.py), so based on that, this is the output. (As shown in the next page).
Sentences are displayed one line, so I have done Ctrl + A, copied my un-pickled knowledge base, and then just pasted it right under the screenshots, so it is easier to see.



```
/usr/local/bin/python3.8 /Users/farishahnahrin/Desktop/FarishahNahrin_WebCrawler/unpickle_KnowledgeBase.py
For the keyword 'bangladesh' the sentences are:  virtual bangladesh is an, site about the beautiful nation of bangladesh., born in 1994 as the first comprehensive we




For the keyword 'pakistan' the sentences are:  learn everything there is to know about the history of the bangladesh’s independence from the brutal yoke of pakistan




For the keyword 'east' the sentences are:  the war pitted bangladesh (then east pakistan) and later helped by india against the west pakistanis and lasted over a dur




For the keyword 'awami' the sentences are:  instead, he heaped invective on bangabandhu sheikh mujibur rahman and the awami league., on three separate occasions betw




For the keyword 'muslim' the sentences are:  about 1200 a.d., muslim invaders under sufi influence, supplanted hindu and buddhist dynasties, and converted most of th
```



```
For the keyword 'government' the sentences are:  government offices are closed, banks bolted shut, even postal, telegraph, telephone, airline and train services came




For the keyword 'india' the sentences are:  interestingly, another poem by tagore (jana gana mana) is the national anthem of india, making him the only person to hav




For the keyword 'military' the sentences are:  all eyes were being turned to the racecourse at dhaka where bangabandhu sheikh mujibur rahman expected to proclaim ind




For the keyword 'state' the sentences are:  appropriate exemptions will be announced from time to time., 3. railway and ports may function, but railway and port work




For the keyword 'political' the sentences are:  all eyes were being turned to the racecourse at dhaka where bangabandhu sheikh mujibur rahman expected to proclaim in
```

The Full Knowledge Base:

1. For the keyword **'bangladesh'** the sentences are: virtual bangladesh is an, site about the beautiful nation of bangladesh., born in 1994 as the first comprehensive web-site on bangladesh, one will find here almost everything there is to know about bangladesh.from images and sounds that describe bangladesh's charm to little known esoteric facts.welcome and be enthralled by the beauty that is bangladesh.virtual bangladesh is an, site about the beautiful nation of bangladesh., born in 1994 as the first comprehensive web-site on bangladesh, one will find here almost everything there is to know about bangladesh.from images and sounds that describe bangladesh's charm to little known esoteric facts.
2. For the keyword **'pakistan'** the sentences are: learn everything there is to know about the history of the bangladesh's independence from the brutal yoke of pakistan after a nine month long liberation war (bengali:,).the war pitted bangladesh (then east pakistan) and later helped by india against the west pakistanis and lasted over a duration of nine months.many were killed by pakistani troops.soon thousands of people were locked in battle with west pakistan soldiers and sailors.the trouble gained a new dimension when a unit of the east pakistan rifles refused to fire on bangali demonstrators.
3. For the keyword **'east'** the sentences are: the war pitted bangladesh (then east pakistan) and later helped by india against the west pakistanis and lasted over a duration of nine months.the trouble gained a new dimension when a unit of the east pakistan rifles refused to fire on bangali demonstrators.general tikka khan takes over as governor east pakistan., yahya khan was clear in his mind about what he should do.those were :, 1. no tax campaign will continue., 2. the secretariat, government and semi-government offices, high court and other courts throughout east bangla will observe hartals.appropriate exemptions will be announced from time to time., 3. railway and ports may function, but railway and port workers will not cooperate if railway or ports are used for mobilizing of forces for the purpose of repression against the people of east bangla., 4. radio, television and newspapers shall give complete versions of bangabandhu's statement and shall not suppress news about the people's movement, otherwise bangali worker in these establishment shall cooperate., 5. only local and inter-district telephone communication shall function., 6. all educational institution shall remain closed., 7. banks shall not effect remittances to the western wing either through the state bank or otherwise., 8. black flags shall be hoisted on all buildings everyday., 9. hartal (strike) is withdrawn in all other spheres but complete hartal may be declared at any moment depending on the situation., 10. a 'sangram parishad' should be organized in each union, mohallah, thana, sub-division – under the leadership of the local awami league units., bangabandhu sheikh mujibur rahman exhorted his people to turn every house in east bangla into a fortress., finally, raising his fist bangabandhu cried out at the top of his voice : "our struggle this time is a struggle for freedom, our struggle this time is a struggle for independence.
4. For the keyword **'awami'** the sentences are: instead, he heaped invective on bangabandhu sheikh mujibur rahman and the awami league., on three separate

occasions between march 3 and march 24 bangali members of armed forces approached sheikh mujibur rahman for guidance because they had no illusions about what was coming., in the evening bangabandhu was engaged in an emergency meeting of the party’s working committee to consider the president’s new date for the national assembly meeting.the awami leaguers also had to decide whether or not to make the declaration of independence that the people were clamoring for.on the one side were the powerful student groups insisting to announce the break with the west pakistan, with them also were the street crowds., the discussions had taken up the whole night but the awami league was still undecided.appropriate exemptions will be announced from time to time., 3. railway and ports may function, but railway and port workers will not cooperate if railway or ports are used for mobilizing of forces for the purpose of repression against the people of east bangla., 4. radio, television and newspapers shall give complete versions of bangabandhu’s statement and shall not suppress news about the people’s movement, otherwise bangali worker in these establishment shall cooperate., 5. only local and inter-district telephone communication shall function., 6. all educational institution shall remain closed., 7. banks shall not effect remittances to the western wing either through the state bank or otherwise., 8. black flags shall be hoisted on all buildings everyday., 9. hartal (strike) is withdrawn in all other spheres but complete hartal may be declared at any moment depending on the situation., 10. a ‘sangram parishad’ should be organized in each union, mohallah, thana, sub-division – under the leadership of the local awami league units., bangabandhu sheikh mujibur rahman exhorted his people to turn every house in east bangla into a fortress., finally, raising his fist bangabandhu cried out at the top of his voice : “our struggle this time is a struggle for freedom, our struggle this time is a struggle for independence.as contact could not be established between the leaders of the awami league, major zia was requested by the rebel station to broadcast a message of independence to the people of bangladesh., at 7:45 pm on 26th march 1971, major zia broadcast the message which became historic in the struggle for independence., “this is swadhin bangla betar kendra.

5. For the keyword **'muslim'** the sentences are: about 1200 a.d., muslim invaders under sufi influence, supplanted hindu and buddhist dynasties, and converted most of the population of the eastern areas of bengal to islam.in 1859, the british crown replaced the east india company, extending british dominion from bengal in the east to the indus river in the west., in the late 19th and early 20th centuries, muslim and hindu leaders began to press for a greater degree of independence.growing concern about hindu domination of the movement led muslim leaders to form the all-india muslim league in 1906. in 1913, the league formally adopted the same goal as the indian national congress: self-government for india within the british empire.the congress and the league were unable, however, to agree on a formula to ensure the protection of muslim religious, economic, and political rights.over the next 2 decades, mounting tension between hindus and muslims led to a series of bitter intercommunal conflicts., the idea of a separate muslim state emerged in the 1930s.
6. For the keyword 'government' the sentences are: government offices are closed, banks bolted shut, even postal, telegraph, telephone, airline and train services came to a

standstill.those were :, 1. no tax campaign will continue., 2. the secretariat, government and semi-government offices, high court and other courts throughout east bangla will observe hartals.this proved that bangabandhu’s directives were being obeyed even at that top level., pakistan international airlines canceling most of it’s international services, concentrated all available aircraft of ferrying “government passengers” to dhaka.growing concern about hindu domination of the movement led muslim leaders to form the all-india muslim league in 1906. in 1913, the league formally adopted the same goal as the indian national congress: self-government for india within the british empire.the congress party and the muslim league could not, however, agree on the terms for drafting a constitution or establishing an interim government.

7. For the keyword **'india'** the sentences are: interestingly, another poem by tagore (jana gana mana) is the national anthem of india, making him the only person to have penned national anthems of two nations.the war pitted bangladesh (then east pakistan) and later helped by india against the west pakistanis and lasted over a duration of nine months.they were followed by representatives of the dutch, the french, and the british east india companies.in 1859, the british crown replaced the east india company, extending british dominion from bengal in the east to the indus river in the west., in the late 19th and early 20th centuries, muslim and hindu leaders began to press for a greater degree of independence.at the movement’s forefront was the largely hindu indian national congress.
8. For the keyword **'military'** the sentences are: all eyes were being turned to the racecourse at dhaka where bangabandhu sheikh mujibur rahman expected to proclaim independence on march 7, 1971., on the other side yahya khan saw the remedy only in terms of applying greater force – a military solution for a political problem., non-violent non-cooperation movement and daily shut-down from 7 am to 2 pm continue.all eyes were centered on the dais where bangabandhu sheikh mujibur rahman was expected any moment., in his speech bangabandhu declared a four-point demand to consider the national assembly meeting on march 25, 1971. they were :, 1. the immediate withdrawal of the martial law., 2. immediate withdrawal of all military personnel to their barracks., 3. an inquiry into the loss of life., 4. immediate transfer of power to the elected representative of the people before the assembly meeting march 25., bangabandhu also unfolded a program of several directives that was the extent of the civil disobedience movement.kazi husne ara rushed out and brought with her mahbub hassan, belal ahmed and abul kashem sandwipi., making hurried trips between agrabad broadcasting station and its transmission center at kalurghat they failed to secure permission from higher authorities to run the station…, it was decided that they should go back to the other side of kalurghat bridge where rations had just been delivered to the jawans of east bengal regiment under the command of major ziaur rahman and plead with them for assistance to run the kalurghat transmitter as a broadcasting station…, inside the kalurghat station, …the engineer [ashikul islam] had an interview with the commandant who agreed to send some military guards to protect the kalurghat transmitter…, ..as the bengali soldiers took positions to guard the transmission centre, the rebels put their heads together and secured the help of a few

engineers of the kalurghat industrial complex to convert it into a broadcasting station., as kalurghat was getting organized into a nerve-centre for coordinating the liberation struggle, baluch troops had invaded the ebr barracks where under the command of major zia a bloody battle raged.dominion status was rejected in 1956 in favor of an “islamic republic within the commonwealth.” attempts at civilian political rule failed, and the government imposed martial law between 1958 and 1962 and 1969 and 1972. the government was dominated by military and oligarchies all rooted in the west.successive military coups occurred on november 3 and 7, resulting in the emergence of army chief of staff gen. ziaur rahman (zia), as strongman.

9. For the keyword **'state'** the sentences are: appropriate exemptions will be announced from time to time., 3. railway and ports may function, but railway and port workers will not cooperate if railway or ports are used for mobilizing of forces for the purpose of repression against the people of east bangla., 4. radio, television and newspapers shall give complete versions of bangabandhu’s statement and shall not suppress news about the people’s movement, otherwise bangali worker in these establishment shall cooperate., 5. only local and inter-district telephone communication shall function., 6. all educational institution shall remain closed., 7. banks shall not effect remittances to the western wing either through the state bank or otherwise., 8. black flags shall be hoisted on all buildings everyday., 9. hartal (strike) is withdrawn in all other spheres but complete hartal may be declared at any moment depending on the situation., 10. a ‘sangram parishad’ should be organized in each union, mohallah, thana, sub-division – under the leadership of the local awami league units., bangabandhu sheikh mujibur rahman exhorted his people to turn every house in east bangla into a fortress., finally, raising his fist bangabandhu cried out at the top of his voice : “our struggle this time is a struggle for freedom, our struggle this time is a struggle for independence.over the next 2 decades, mounting tension between hindus and muslims led to a series of bitter intercommunal conflicts., the idea of a separate muslim state emerged in the 1930s.on march 23, 1940, muhammad ali jinnah, leader of the muslim league, publicly endorsed the “pakistan resolution” that called for the creation of an independent state in regions where muslims were a majority., at the end of world war ii, the united kingdom, under considerable international pressure to reduce the size of its overseas empire, moved with increasing urgency to grant india independence.in june 1947, the uk declared it would grant full dominion status to two successor states–india and pakistan.the various princely states could freely join either india or pakistan.
10. For the keyword **'political'** the sentences are: all eyes were being turned to the racecourse at dhaka where bangabandhu sheikh mujibur rahman expected to proclaim independence on march 7, 1971., on the other side yahya khan saw the remedy only in terms of applying greater force – a military solution for a political problem., non-violent non-cooperation movement and daily shut-down from 7 am to 2 pm continue.the congress and the league were unable, however, to agree on a formula to ensure the protection of muslim religious, economic, and political rights.the capital of federal pakistan was at islamabad., bangabandhu shekih mujibur rahman, pakistan’s history for the next 26 years was marked by political instability and economic difficulties.dominion status was rejected in 1956 in favor of an “islamic republic within the commonwealth.”

> attempts at civilian political rule failed, and the government imposed martial law between 1958 and 1962 and 1969 and 1972. the government was dominated by military and oligarchies all rooted in the west.by the time pakistan's forces surrendered on december 16, 1971, india had taken numerous prisoners and gained control of a large area of east pakistan, which is now bangladesh., mujibur rahman came to office with immense personal popularity but had difficulty quickly transforming this support into political legitimacy.

How might you use this data for a chatbot?

Personally, I chose this as my starter site, because my motherland is Bangladesh. And although I was not born nor raised in Bangladesh, I wanted to use a starter site that can help me learn more about how this young country came to be, and also potentially help others learn about the history of Bangladesh . It would be really cool if the starter site I chose, could also use this knowledge base, for a chatbot on their site, so that people can easily ask questions and receive answers about the history of Bangladesh, without having to scour through the site. (But I would have to sell them this product, first!) Also, if any travel blogs, agencies, or sites, wanted to use this knowledge base, for a chatbot, they could, since this information can help tourists learn more about the arduous journey of Bangladesh's independence. The Bangladesh Embassy includes a page that details the history of Bangladesh, so the information from this chatbot could also be useful to their page visitors. Lastly, the knowledge base could be useful for a chatbot at the main Bangladesh airport. At the DFW Airport, there are large screens, where users can tap through and find more information about Dallas or its Airport facilities. So, if there was a chatbot at the main Bangladesh airport, the knowledge base could be used to help tourists, younger folks, or locals, to learn more about how Bangladesh came to be.