IBM Bluemix - Developing a Travel guide using Watson services

by

Mahalakshmy Krishnamoorthy

A Project Report Submitted

in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in

Computer Science

Supervised by

Dr. Carol Romanowski

Department of Computer Science

B. Thomas Golisano College of Computing and Information Sciences

Rochester Institute of Technology

Rochester, New York

May 2016
Abstract

The amount of information that is available on the internet is enormous and also most of this information is in the form of unstructured data. This unstructured data can be efficiently put to use to make applications using the cognition feature. The cognitive applications are such that they follow a human like thinking approach. IBM Watson is one of the leading systems that helps to achieve build a cognitive application. In general, to obtain information about things, places, people, institutions, etc. or in general everything usually the first approach one takes are to make use of search engine. Developing a domain specific tool will provide better and efficient results. In general, every human is passionate and for a break loves to travel and there is an abundant amount of data about places, tourist spots and things to explore information available on the internet. There are millions of people who travel to New York City every year and considering the vastness and the enormous number of things one can do the most helpful thing is to have a travel guide or an adviser. It is with all this data in the unstructured format and with the use of Watson services that the travel application is created.
Table of Contents

Abstract ................................................................................................................................. 2

1 Introduction ...................................................................................................................... 3
  1.1 Cognitive Computing & Existing systems .................................................................. 3
  1.2 IBM Watson ............................................................................................................... 5
  1.3 IBM Bluemix ............................................................................................................ 5
  1.4 Objective .................................................................................................................. 5
  1.5 Motivation ................................................................................................................ 6

2 Design & Approach ........................................................................................................... 7

3 Implementation ................................................................................................................ 9
  3.1 Data Collection and Cleaning ..................................................................................... 9
  3.2 Application Creation on Bluemix ............................................................................... 12
  3.3 Retrieve and Rank ..................................................................................................... 12
     3.3.1 Retrieve Component based on Apache Solr ....................................................... 13
     3.3.2 Rank Component based on Trained Ranker ....................................................... 15
  3.4 User Interface Implementation .................................................................................. 16
  3.5 Testing ..................................................................................................................... 16
     3.5.1 Testing of Retrieve Component based on Apache Solr .................................... 16
     3.5.2 Testing of Rank Component based on Trained Ranker ..................................... 19
  3.6 Results ...................................................................................................................... 22

4 Conclusion ...................................................................................................................... 24
  4.1 Current Status .......................................................................................................... 24
  4.2 Future Work ............................................................................................................. 24
  4.3 Lessons Learned ....................................................................................................... 24

5 References ..................................................................................................................... 25
Chapter 1

Introduction

1.1 Cognitive Computing & Existing systems

The cognitive systems that are developed are modeled on the basic concepts of artificial intelligence and neural networks and other such basic advancements. The systems that were initially developed could be programmed such as to output the results based on the if-else or other similar set of parameters. With development in each of the fields the systems became more intelligent and had the capability to analyze the situations and make decisions to problems with better understanding. Cognitive systems are such that they for making decisions or finding solutions to problems will use an approach similar to humans and thus are defined as systems with human like thinking capability. When a human has to find answer to a question or problem he will think about all the situations for the question gather all the data or think about all the things he knows about that topic and then find the solution to it. This is the kind of solution the cognitive systems try to achieve [1].

The most important aspect of cognitive system is the type of data that it can deal with. Cognitive systems can make use of unstructured data as well to analyze the problem and find results. Unstructured data consists of almost 80% of the data that is present these days. What this unstructured data mainly constitutes are emails, blogs, reports, social media data, etc. [2]. They are such that they have no defined fields, entities and relations amongst them so it becomes very difficult to map them into a relational database and then use them for analyzing the data. Thus for analyzing this data, finding hidden patterns or developing recommendation, prediction from them there cannot be any pre-defined type of database that can be used. For example, consider an organization that wants to track the writing and presentation skills of the employees for helping them improve or better organize their work in this case the main source of data would be the email conversations, presentation documents and other such documents that will be used as the data source. It would be a difficult task to model a database for such documents since the data that is present in each of them is not going to be uniform and will vary with each employee. In such situations developing a cognitive application that can handle unstructured data would be very beneficial. Some of the leading examples of Cognitive systems that have been developed include:

- Watson which has been developed by IBM. It has been evolving ever since it has been developed [2].
- DARPA DeepDive is an open source project developed by professor Christopher Re of the University of Wisconsin [3].
- Apache UIMA which is a part of the Apache Foundation and makes use of the Unstructured Information Management(UIMA) [3].

All of these systems are developed with the UIMA framework in view and the process is similar in various stages. They all have Entity Relationship built for all of the data with various linking. The prominent differences include on how they process the input and form the Lexical Answer Type for each of them.
1.2 IBM Watson

IBM Watson is one of the leading cognitive systems that has been built and is basically a Question Answering system. Watson can be trained and tested for any particular domain and will work with the data that is provided to it. The quiz show Jeopardy in which Watson has been a part and won against participants is one of the highlights of the potential of Watson. Watson can process a large amount of unstructured data which can be in the form of word files, PDFs or web pages. It makes use of scoring algorithm to find the best possible answer to the question and ranks all its findings. All of this is supported by the natural language processing, machine learning and artificial intelligence techniques. The main challenge for the system is to make human like decisions to difficult problems.

Watson follows the standard processing to find answer to a question and is built with answering Jeopardy style question as foundation. The process begins with analyzing the question and forming answers step by step from the question analysis. The question is first transformed where the entity extraction is the main focus and help in forming the Lexical Answer Types(LAT) [4]. There are several hypothesis generations, filtering and other such techniques followed for answer generation. For all the generated answers there is a confidence and evidence result that is generated. Based on this score the most suitable answer is retrieved [5, 6].

1.3 IBM Bluemix

IBM Bluemix is used in development of various applications and is like a gateway to the Watson services and thus it serves as a Platform as a Service [7]. Bluemix platform comprises of several services to help in developing applications some of them include Cloud Foundry, various runtimes, support for various algorithms, support for several databases and analysis. It is with the help of the Watson services that are present in this platform that the application has been developed [8]. The two main Watson services implemented and explored in this application are the Document Conversion service and the Retrieve and Rank service. Several other Watson services such as Natural Language Classifier, Dialog, Twitter services have also been experimented and studied upon.

1.4 Objective

The main objective of the project is to develop a travel guide to New York City as a Cognitive application using Watson services on the IBM Bluemix platform. The application is developed by exploring the advantage of Watson on being able to handle unstructured data and also to be able to answer questions efficiently with a human like thinking nature rather than just talking about the information that is present in the data.

1.5 Motivation

Ever since the evolution of the internet and advanced technologies and system to use the information available there is an enormous change taking place every day leading to better and efficient applications that can be used in day to day life. One of the main motivation behind developing on a cognitive application is exploring the key features and functionalities that are available for a particular domain. Also, this gives an opportunity to explore the differences that exist between similar platform based applications that exist such as Siri, Google Now, Cortana, etc. One of the main advantage of using
Watson services is to make use of unstructured information that is data present as emails, messages, blogs without proper structure. Also, since the application is trained, tested and developed for a particular domain it has more control over the domain and the tuning of the system to make it more efficient. Some of the domains in which Watson applications have been successfully developed and are being developed include healthcare, financial market, businesses, travel, etc. In the healthcare domain it has helped in developing in applications on treating cancer and research into oncology at various organizations and research institutes such as University of Texas MD Anderson Cancer Center, Mayo Clinic, etc. [9].

To obtain information or learn about a new thing the medium that is used these days and is a wealth of information is the internet. According to statistics and facts about searching the internet there are billions of searches that are done every day. The efficiency of the results to these searches depends about the topic and also on how the query is framed and what kind of output is expected. For these searches traditionally in the initial days google, yahoo search and other search techniques have been used. With advancement in development of technology and devices there are a lot of mobile based search tools and personal agents that are present these days. Some of the most used platforms are Siri, Cortana and Google Now.

The main area this project focuses is developing a cognitive application for travel domain. In general, every human being has a passion to travel, explore, see new places and record memories. To obtain information about travelling one would make use of travel agents, travel websites for bookings and will also have to search for a lot of information on the internet. There are a lot of travel planning help websites, mobile based applications that have been developed for this purpose. New York city has a lot of places to visit and more than millions of visitors get attracted to this place every year.

The main users of this application will be individuals who plan to visit the city and in order to visit would want information and learn more about the place. It can also be used by travel agents to find more efficient information or to train the system with more personalized information and this will help in improving the standards of their services. It can also be used by agencies for better planning and development that would help the analyze the questions, learn more about the queries, difficulties and obtain constructive feedback for improvement. The application can also be installed at prominent locations such as famous landmarks, museums, airports, train stations where people could get information about things to do, buy or eat.
Chapter 2

Design & Approach

The above diagram describes the basic steps and the flow of the application development. They can be summarized as below:

- **Data Collection & Cleaning:**

  This is the initial step comprises of collecting data and parsing it to form the data for Watson to efficiently answer questions. The data has to be cleaned and formatted in such a manner that the query search can lead to best results. By using Java and Scripting the data is cleaned on implementing the Document Conversion service.

- **Retrieve and Rank Implementation:**

  The main question answering component is implemented in this stage. It is implemented in two ways one using the Apache Solr system and the second by using the Trained Ranker model. This component is implemented the services using Java, Scripting and Node.js.
● Application creation:

This step involves creating an application on IBM Bluemix to bind all the Watson services. The User interface is built using Node.js with input box to enter the question and the output appears in the next step.

● Testing & Result Analysis:

The application has been tested on several questions and an analysis of how the answers differ between the Retrieve and Rank component have been analyzed.
Chapter 3

Implementation

3.1 Data Collection & Cleaning

The corpus consists of 348 documents collected from sources such as Wikipedia, Wikivoyage, wikitravel and other New York city travel information related websites. These documents have been at the first step stored as html files, then processed by implementing the Document Conversion Watson service [10] and stored in Json files to be given to the retrieve and rank service for creating the corpus. The main reason for implementing this service and forming the Json answer units is the compatibility of this kind of data as input as corpus for the retrieve and rank service.

All of the html files contain images, tables, figures, graphs which needs to be removed and just the text component should be present. This is done by defining a config unit for the conversion specifying which tags need to be removed and which tags need to be given more emphasis on for the answer units. The service implementation is written in Java which scans through all the files in the data collected and calls the Document Conversion Service to form answer-units. This generates a Json file with id to the answer-unit and text associated with it. Each id generated is unique and the text is based on the parameters mentioned in the custom configuration file.
Below is an example of the html file and its equivalent Answer-unit file created using the Document Conversion service with the defined custom configuration.

The Json answer units created as Json documents contain additional information and are parsed to create answer units with just the id and the text associated with it. All the other tags such as the parent_id for the document, direction are all removed. This is done using Java code which parses all the converted Json files and rewrites them with the required format.
Below is an example of the output answer unit created for the retrieve and rank service after removing all the unwanted tags.
3.2 Application creation on Bluemix

On the IBM Bluemix platform an application named travelappmk is created and the various Watson services that have to be used. Each service is then authorized using service credentials that are created and are bound to the application.

The travelappmk has the following Watson services:

- Document Conversion [10]
- Retrieve and Rank [11]

While creating the application the SDK for Java has been used but in the later stage to make it compatible to a web application the SDK for Node.js is also incorporated. The first step to use any of these services is to download the desired SDK and import it into the project. This helps to make use of the desired Watson service. In the next step the authorization for the application with the service username and password has to be specified. Once the authorization has been created as per the required function of the service the code can be implemented.

3.3 Retrieve and Rank

This is the main question-answering component of the application. This service contains a lot of input parameters that are submitted to it in addition to the question to be answered. This service requires two implementations one involving Apache Solr [12] and the second involving Trained rank model. Apache Solr is an index based ranking system used to implement a search engine and is from the Apache Lucene project and the Retrieve component is based on it. The Rank component uses machine learning features and has a trained model that is given to it along with the data. To implement this service Java, Scripting and Node.js for the user interface has been used. The main corpus that is loaded into this system is the Json files from the Document conversion step.
3.3.1 Retrieve component based on Apache Solr

The steps involved in building the Retrieve component are as follows:

1. Importing the Java SDK/NodeJS SDK into the project and creating the service and providing the authentication credentials.

2. Creating the Solr cluster which will have all the corpus data and files telling the system how to associate the data in the corpus and the various relationships. On completion of this, the Solr cluster ID is obtained.

   C:\Users\Mahalakshmy> curl -X POST -u "cZbb543-f132-40e3-99bc-0b3e163dea06" "https://gateway.watsonplatform.net/retrieve-and-rank/api/v1/solr_clusters" -d "
   {"solr_cluster_id":"scb41bd9c_db7d_4a44_9b1c_1309304b6415","cluster_name":null,"cluster_size":null,"solr_cluster_status":"NOT_AVAILABLE"}
   C:\Users\Mahalakshmy> curl -u "cZbb543-f132-40e3-99bc-0b3e163dea06" "https://gateway.watsonplatform.net/retrieve-and-rank/api/v1/solr_clusters/scb41bd9c_db7d_4a44_9b1c_1309304b6415" 
   {"solr_cluster_id":"scb41bd9c_db7d_4a44_9b1c_1309304b6415","cluster_name":null,"cluster_size":null,"solr_cluster_status":"READY"}

3. Adding all the answer-units documents and creating the config file with definitions on what are the attributes(id, text) present in the answer units and specifying which is the unique value(id) to efficiently search through the Solr collection. For adding all the files, a script has been written to parse through all files in the corpus and add them.
4. To test the working of this Retrieve component the URL can be tested on the browser and returns an answer to the question submitted. The username, password and solr_cluster_id generated have been specified while making the call.

**URL:**

https://[username]:[password]@gateway.watsonplatform.net/retrieve-and-rank/api/v1/solr_clusters/[solr_cluster_id]/solr/example_collection/select?q=where to eat in nyc&amp;wt=json&amp;fl=id,txt

The following output is obtained:
3.3.2 Rank component based on the Trained Ranker

The initial steps for this component are same as the Retrieve component. After adding the documents, a training set is created for training the rank service. The steps followed after that are as follows:

1. The training set is created such that it has various questions and for each question the various possibilities of the answer are specified with the id and rating to the relevance of the answer. The training set implemented for this service includes 100 training questions and various possibilities of the answer with its rating and is stored in csv format. Below is an example from the training set:

2. This data is given to the train.py code present along with the Retrieve and Rank service [11] of Watson which generates training data for the service.

3. Make use of SolrClient functions to create a SolrClient so that the application can be called using Java or Node.js

4. The application can be accessed by the following URL to find the relevant answer based on the trained model. The username, password and solr_cluster_id generated have been specified while making the call.

URL:

https://[username]:[password]@gateway.watsonplatform.net/retrieve-and-rank/api/v1/solr_clusters/{solr_cluster_id}/solr/example_collection/fcselect?ranker_id=3b140ax14-rank-1674&q=where%20to%20eat%20in%20nyc&wt=json&fl=id,text
The answers can be improved by re-ranking the system with better advanced training questions and providing more variations to the questions.

3.4 User Interface Implementation

The User Interface has been created using Node.js application and has a Textbox to enter a question and the output appears in the next step.

3.5 Testing

Testing has been done using several questions that are not present in the training set or by coming up with variations of the training set questions.

3.5.1 Testing of the Retrieve component based on Apache Solr

The Retrieve component has been tested on 100 questions by storing the questions in a file and calling the service for each of the questions using java code and storing the results in a file for analysis.
Some of the testing results are as follows:

**Question: Where is Little Italy?**

```
{responseHeader={status=0,QTime=1.params={q=where is little italy,wt=javabin,version=2}},
 response={numFound=691,start=0,docs=[SolrDocument{id=d998e360-30b-4abc-9047-93f9e470ed,
 text=[Little Italy on Mulberry Street used to extend as far south as Worth Street, as far north as Houston Street, as far west as Lafayette Street, and as far east as Bowery. It is now only three blocks on Mulberry Street. Little Italy originated as Mulberry Bend. Jacob Riis described Mulberry Bend as “the foul core of New York’s slums.” Bill Tonelli from New York magazine said, “Once, Little Italy was like an insular Neapolitan village created on these shores, with its own language, customs, and financial and cultural institutions.” Little Italy was not the largest Italian neighborhood in New York City, and East Harlem (as Italian Harlem) had a larger Italian population. Tonelli said that Little Italy “was perhaps the city’s poorest Italian neighborhood”. In 1910 Little Italy had almost 10,000 Italians; that was the peak of the community’s Italian population. At the turn of the 20th century over 90% of the residents of the Fourteenth Ward were of Italian birth or origins. Tonnelli said that it meant “that residents began moving out to more spacious digs almost as soon as they arrived.” Ferrara Cafe (est. 1892) In Little Italy After World War II, many residents of the Lower East Side began moving to Brooklyn, Staten Island, eastern Long Island, and New Jersey. Chinese immigrants became an increased presence after the U.S. Immigration Act of 1965 removed immigration restrictions, and the Manhattan Chinatown to Little Italy’s south expanded. In 2004, Tonelli said, “You can go back 30 years and find newspaper clips chronicling the expansion of Chinatown and mourning the loss of Little Italy.” Prior to 2004, several upscale businesses entered the northern portion of the area between Houston and Keenway Street.]}}
```

**Question: Where can I find cherry blossom?**

```
{responseHeader={status=0,QTime=1.params={q=where can I find cherry blossom,wt=javabin,version=2}},
 response={numFound=146,start=0,docs=[SolrDocument{id=39e3c3da-d0d-4f1-aa8f-127c3348e771, text=[The Garden has more than 200 cherry trees of forty-two Asian species and cultivated varieties, making it one of the foremost cherry-viewing sites outside Japan. The first cherries were planted at the garden after World War I, a gift from the Japanese government. Each spring at BBG, when the trees are in bloom, a month-long cherry blossom viewing festival called Hanami is held at the Cherry Esplanade, culminating in a weekend celebration called Sakura Matsuri. The Esplanade features two rows of cherry trees with trails and sitting areas on the side. Visitors may also sit on the field of grass between the rows of cherry trees. Cherry trees are found on the Cherry Esplanade and Cherry Walk, in the Japanese Hill-and-Pond Garden, and in many other locations in the Garden. Depending on weather conditions, the Asian flowering cherries]]}}
```
Question: How tall is 432 park avenue?
3.5.2 Testing of Rank component based on the Trained Ranker

The rank component has been tested on questions based on training set questions, variations in the formation of questions in the training set and some new questions. In the trained model relevant answers appear higher the order and provide better answers to questions. Below are results to some of the questions not present in the training set.

**Question:** Where can I find Halal guys?

**Answer 1:**

Nothing differentiates New York more from other American cities than the astonishing amount of food cooked and served on the streets. Starting with the thousands of hot dog stands on almost every street corner (try Hallo Berlin on 54th and Fifth for the best rated sausages), the possibilities are endless. People trek to Jackson Heights in Queens for a nibble of the famous arepas of the Arepa Lady. Freshly cooked Indian dosas are served up at a pittance at the NV Dosas stand in Washington Square Park. The Trinidadian/Pakistani Trinipak cart on 43rd and Sixth. Danny Meyer, the famous restaurateur, has a burger stand ("Shake Shack") in Madison Square Park as well as a new location on the upper west side. The halal offerings in midtown are legendary ( Kwik-Meal on 45th and Sixth; Chicken Guy/Halal Chicken on 53rd and Sixth and many others). Most carts serve lunch (from about eleven in the morning to five or six in the evening) and disappear after dark, so look for a cart near you, smell what's cooking, and enjoy a hot and often tasty lunch for a few dollars (a meal costs anywhere from about $2-8). Mornings, from about 8AM-10AM, the streets are dotted with coffee carts that sell coffee, croissants, bagels, and danish pastries and are good for a cheap breakfast: small coffee and bagel for a dollar or so. From 10AM to 7PM many vendors sell lunch and dinner choices, including hot dogs, hamburgers, gyros, and halal. Other street vendors sell italian ices, pretzels, ice cream, and roasted peanuts. Also, look around for the coffee truck (often found in Union Square), dessert truck, as well as Belgian waffle truck that roam around the city.

**Answer 2:**

Cranberry Cafe, 115 W 45th St (1 block from Times Sq). 24/7. A huge variety of sandwiches, pizza, sushi, and udon. Hot and cold salad bars. Sandwiches are $4-$6. Very popular with local workers, so it gets very busy at breakfast and lunch time. There is ample seating upstairs. A edit The Halal Guys, 53rd St and 6th Ave (carts on southeast and southwest corners), [18]. Weekdays: 10am-4am, weekends: 10am-5am. The Halal Guys are popular food carts that serve chicken, beef gyro, and falafel wraps and platters with rice, lettuce, and sauce. There are lines of customers waiting day and night. Eat outside standing or on nearby benches, or take it to go. These original carts have been so successful, they now have permanent restaurants and international franchises. Wrap $5, platter $7. A edit Poseidon Bakery, 629 9th Ave (btwn 44th and 45th Sts), 212-757-6173. Serves wonderful Greek pastries, savory and sweet alike. There is no place to sit, but the staff are very friendly and helpful, and you can have a delicious, cheap meal for takeout. A
Question: How much does hotel accommodation cost?

Answer 1:

The costs of hotel accommodation in New York City is generally higher than the American average, and Manhattan (where most visitors will want to base themselves) in particular has some of the most expensive accommodation in the world. Expect to pay up to $50 for a hostel, $100-200 for a budget room with shared bath, $250-350 for a mid-range hotel with a decent room and a restaurant and/or room service, right up to world famous luxury hotels such as the Waldorf Astoria or The Plaza, where a stay in the top suites can run into thousands of dollars a night. There is no shortage of choice however as all of the major international hotel chains such as Hilton, Hyatt, Marriott and Holiday Inn each have multiple properties in Manhattan. Most rooms below $200 in Manhattan are small with room for a bed, a tv, and little else, and may be located in less attractive areas of the island - for instance along the West Side Highway, or on the northern reaches beyond Central Park. You can reduce the costs by basing yourself in the outer boroughs, of course but beware of cheap hotels in Brooklyn or Queens where you may be sharing the premises with hourly customers! A smarter move is to stay along the New Jersey shore - Hoboken, Jersey City and Newark for instance all have major chain hotels which can be much cheaper than their Manhattan equivalents, and are connected across the Hudson via the PATH system. That said however, there are ways of finding accommodation in the city and producers have to be aware of their location, traffic and the closer time of your stay.

Answer 2:

New York City can be expensive for the budget traveller. Hostels generally cost about $25 to $35 per night for shared, dorm accommodations. Hotels start at $70 and run upwards, though location, time of year and day of week will effect the rates. A unique compromise between hotel and dorm hostel is Bowery's Whitehouse of New York, providing single or double accommodations in miniscule rooms with open, lattice ceilings, offering the security of an enclosed, locked room with the space of a dorm room. Rates are around $30 per person, plus taxes, depending on time of year. Reservations are highly recommended as far in advance as possible, as the Whitehouse fills up quickly. Reservations and rates can be found at the Bowery's Whitehouse of New York website. Brooklyn and Queens, both well serviced by the subway system are good alternatives to staying in Manhattan. As well, the PATH system running 24 hours means that travellers could also look at staying in Jersey City, Newark or Hoboken and still have great access to Manhattan.
Question: What are driving rules in NYC?

The citywide speed limit is 25mph. Unlike other places in the United States, New York City is a citywide No-Turn-On-Red zone. [10] Be careful when driving as some (but not all) entrances to New York City have signs alerting motorists that it is illegal to turn on red signal in NYC, and other drivers from out of town may not know this rule. As in the rest of New York State, talking on a cell-phone (without a hands-free device) or texting while driving is illegal. Even if you do have a hands-free device, minimize your talking and prioritize driving. There are red light cameras at 100 various intersections in New York City. [11] A camera will take a picture if you run a red light and a fine disputable on the web will be issued in 30 days. [12] However, since the camera does not identify who is driving the vehicle, no points will be issued against your drivers' license. Some bus lanes have video cameras. [13] A camera will take a video if you drive illegally in the bus lane other than to turn right and a fine disputable on the web will be issued in 30 days. [14] There are 20 to 40 speed limit cameras in the city located near schools. If there is an emergency vehicle trying to get through with its siren blaring, pull over to the side and move forward as necessary. Some avenues and many streets have only one-way traffic. Thankfully, one-way streets generally alternate direction, so if your destination is down a one-way street going in the wrong direction, go another block and double-back. A handy mnemonic is “Evans go East,” meaning that, for the most part, streets with even numbers will head east, and vice-versa (in Manhattan). The best gauge to determine a one way street’s direction is to check the direction parked cars face. Be wary of your surroundings when you park your car. While NYC is a safe city for its size, it’s not necessarily safe for your car as well. Make it as unworthy to steal as possible.
3.6 Results

The results of the application depend on what kind of training has been done to the system and how organized the corpus units are specified with their configurations. There are some variations that are noticed in the responses asked to both the components. It can be seen in some cases that more relevant and useful answers are found in the Rank component as compared to the Retrieve component. Even if the first answer found is not accurate the chances of the most relevant answer appearing higher up the order is more in case of the Rank component.

Example of the first answer found on two components for “Which is a good hotel to stay?” In this both the answers are relevant but the answer obtained from Rank model is more useful and provides the required details.

Answer - Rank Component

The costs of hotel accommodation in New York City is generally higher than the American average, and Manhattan (where most visitors will want to base themselves) in particular has some of the most expensive accommodation in the world. Expect to pay up to $50 for a hostel, $100-200 for a budget room with shared bath, $250-350 for a mid-range hotel with a decent room and a restaurant and/or room service; right up to world famous luxury hotels such as the Waldorf Astoria or The Plaza, where a stay in the top suites can run into thousands of dollars a night. There is no shortage of choice however as all of the major international hotel chains such as Hilton, Hyatt, Marriott and Holiday Inn each have multiple properties in Manhattan. Most rooms below $200 in Manhattan are small with room for a bed, a tv, and little else, and may be located in less attractive areas of the island. For instance along the West Side Highway, or on the northern reaches beyond Central Park. You can reduce the costs by basing yourself in the outer boroughs, of course but beware of cheap hotels in Brooklyn or Queens where you may be sharing the premises with hourly customers! A smarter move is to stay along the New Jersey shore - Hoboken, Jersey City and Newark for instance all have major hotel chains which can be much cheaper than their Manhattan equivalents, and are connected across the Hudson via the PATH system. That said however, there are ways of finding accommodation in a big name mainstream hotel in Manhattan at lower prices - it takes some determination - but it can be done using the following tips: January/February (but after the New Year period), are traditionally the "quiet" months in the city for tourism. If you can stomach the likelihood of heavy snow and bitterly cold weather (usually), New York has a character all of its own in times of heavy

Answer - Retrieve Component

Nice boutique hotel with good bar, Silverleaf Tavern, which serves a good 687. Lovely rooms including LCD TV's etc. Some rooms have a view of the Empire State Building. Edit the Wyndham Midtown 45, 205 E 45th St (btwn 2nd and 3rd Aves), 1-212-867-5100, [36]. 203 luxury living spaces, from studios to guest rooms to suites - all designed by residential architect Costa Kondylis, with custom-designed furniture by David Rockwell and state-of-the-art amenities. Edit Andaz 5th Avenue (Andaz 5th Avenue), 485 5th Ave, 1-212-601-1224, [37]. Enjoy a unique sense of welcome from this modern midtown Manhattan hotel. The hotel offers loft style Manhattan rooms and suites, meeting a event space, dining and free WIFI. Edit Redford Hotel - Murray Hill, 118 E 48th St (btwn Park and Lexington Aves), [38]. A small European style hotel located between the Empire State Building and the Chrysler Building. Edit Bryant Park Hotel, W 40th St (between 5th and 6th Aves, on Bryant Park), [39]. Distinctive black brick and gold trim building, amenities include deep soaking tubs, cashmere blankets, Filipino toiletries, Tibetan rugs in rooms. $2545. Edit Dylan Hotel, 52 E 41st St (between Madison and Park Aves), [40]. Irish boutique hotel with a 155 rooms and suites featuring a traditional Irish pub, outdoor patio for dining, WiFi access, and in-room safe. Edit Fitzpatrick Grand Central Hotel, 141 E 44th St, 1-212-351-6800, [41]. Irish boutique hotel with a 155 rooms and suites featuring a traditional Irish pub, outdoor patio for dining, WiFi access, and in-room safe. Edit Fitzpatrick Manhattan Hotel (Fitzpatrick Hotel), 607 Lexington Ave, 1-212-355-9100, [42]. Irish boutique hotel with a popular on-site Irish restaurant. Edit Four Seasons Hotel, 57 E 57th St (btwn Madison and Park Aves), 1-212-756-5700, [43]. Edit Grand Hyatt New York, Park Ave at Grand Central Terminal, 1-212-863-3234 (fax: 1-212-897-5772), [44]. Attached to Grand Central Station. Edit Hampton Inn
Another example of responses on both systems is for “Where is central park?” The left image is response from the Retrieve model and the right image is the response from Rank model. Both the answers are not wrong but the relevance is more with the answers obtained using the Rank component.
Chapter 4

Conclusion

4.1 Current status

The application has been implemented on a basic version making use of the Retrieve and Rank service as its main component. The document conversion was an important service which helped in building the suitable format for the corpus on which the question answering system is based. Currently the system is trained on 100 questions and has been tested on both the Retrieve and the Rank components. The answers obtained on both the components are efficient but in some cases the most relevant answer does not appear as the first answer but appears lower down in the order. The user interface is just a textbox to enter the question and submit button and displays only the answer in the first document found.

4.2 Future Work

The application currently developed is can be improvised by adding various features and also changing some of the existing modules. The system can be trained on more complex set of queries and the number of questions with their variations can be combined. One main component that can be added to improve the training of the system is incorporating the feedback unit which will take in user feedback on each question for a score and check if the question is present in the training set. If the question is present in the training set it would just add the unique id and the relevance rating specified by the user against it. In case the question is not present it can be added as a new question with the id and the relevance score. Other features that can be incorporated include developing a more user friendly User Interface with options to view all the 10 retrieved answers for a question.

4.3 Lessons Learned

Developing an application on cognitive platform helped explore how unstructured data can be processed and put to use. The IBM Bluemix platform and the Watson service implementation seemed very useful and by making use of the correct services there can be several useful applications that can be developed in a particular category. This also helps to understand a system better when it tries to form answers going beyond the stand searching or pre-defined rules that the traditional systems have been using.
References


