

Knowledge Test based on Question Answering

Dhinesh.N, Pradeepan.K, Bharath.R
BE CSE, Semester 8

Ms. M. Saritha, B.E., M.E.,i
Supervisor

Project Review: Second Review (28 February 2013)
Department of Computer Science and Engineering
SSN College of Engineering

Abstract

Everybody always have an urge to test themselves on various domains, based on their interest. There are web products to serve this purpose, but only for certain fields and most of the questions are hard coded. There is no centralized system that judges the users knowledge on their interested field. We propose a system that gets the users interested field and frames a set of questions based on articles excerpted from certain trusted information providers over the Internet. Along with the queries, the answers to them are also retrieved. The user enters the answers and his answers are evaluated to give an aggregated score to judge his knowledge level on that particular domain.

Contents

1	Detailed Design	3
1.1	Architecture Diagram	3
2	Modules Identified and Algorithms	4
2.1	Key-phrase Input and Topic Selection	4
2.2	Parsing and Extraction of necessary data	5
2.3	Search for Keywords	5
2.4	Key Phrase Extraction Techniques	5
2.5	Selection of declarative sentences	6
2.6	Tokenization and Syntax analysis	6
2.7	Question and Answer framing	7
2.8	Answer evaluation	7
3	Implementation Details	7
3.1	Software Used	7
3.2	Corpus used	8
3.3	Module Implementation with Snapshots	8
	References	11

1 Detailed Design

1.1 Architecture Diagram

Figure 1. Question and Answer Generation

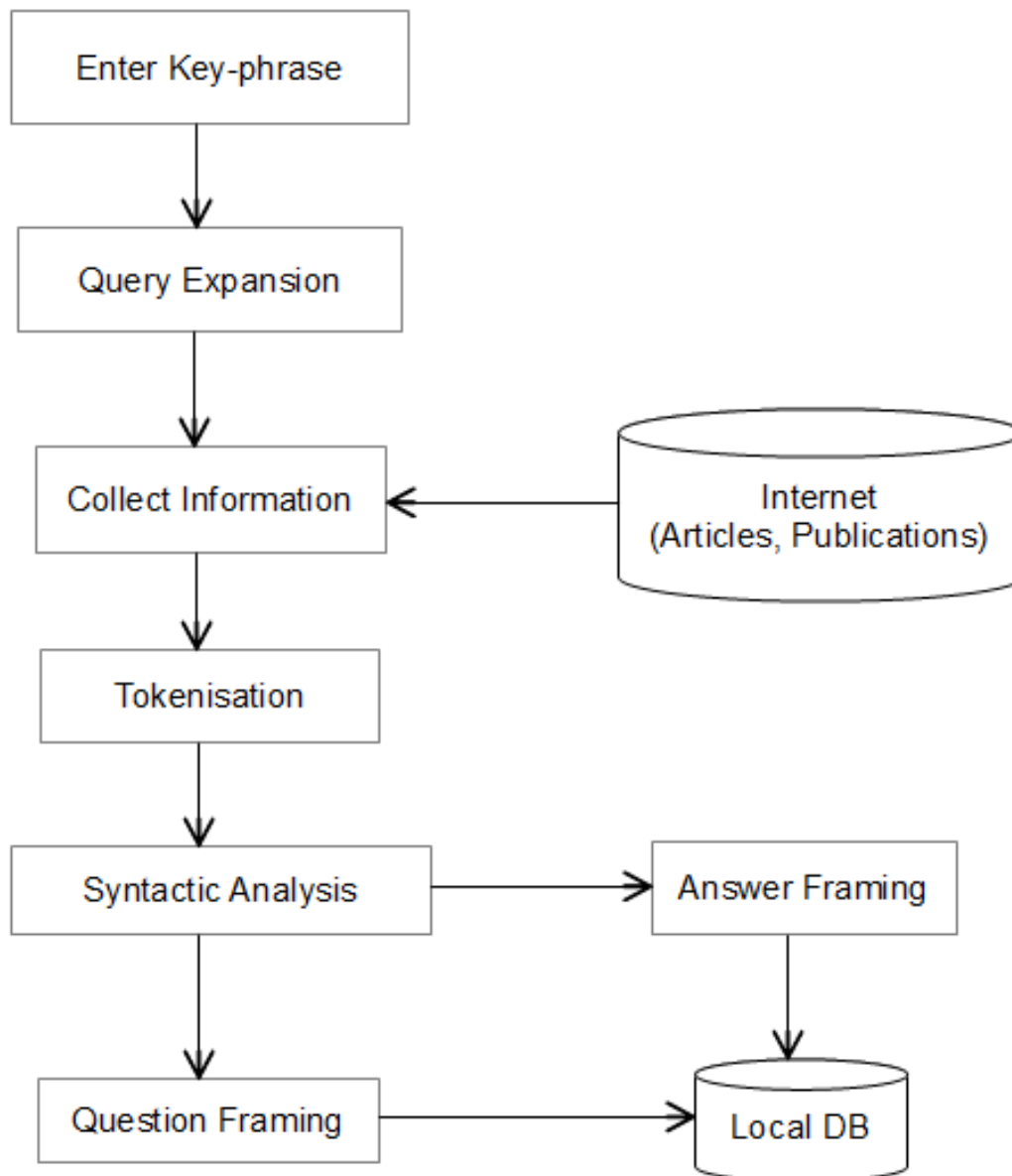
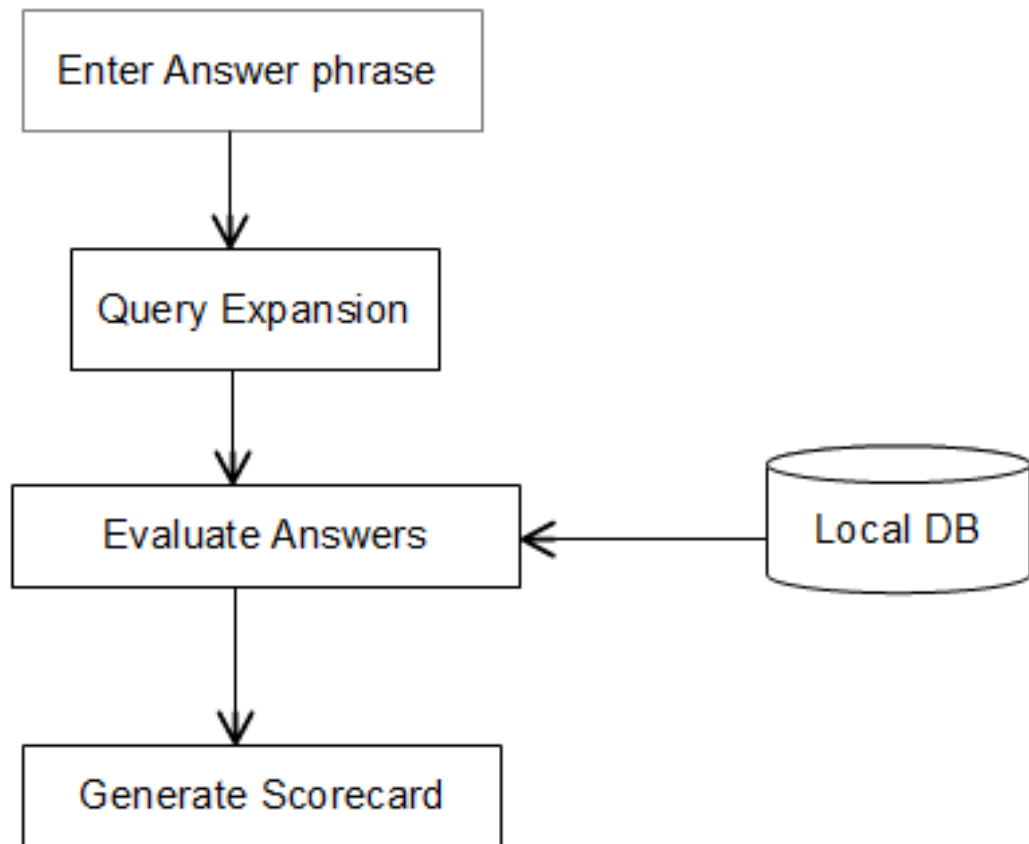


Figure 2. Answer Evaluation



2 Modules Identified and Algorithms

2.1 Key-pharse Input and Topic Selection

The user types the TOPIC of his interest in the Text box specified. The word entered is searched in the wikipedia using the Mediawiki API. In cases of disambiguation, the user is asked to select the exact topic of his need, provided by the Wikipedia disambiguation page.

2.2 Parsing and Extraction of necessary data

The API call returns a JSON file which includes the Title and Body of the article and other metadata. Using BeautifulSoup python library, the JSON file is parsed so that the body is extracted. The body, which contains the hyperlinks and other internal metadata in it, is parsed again to gather the text-only content.

2.3 Search for Keywords

This text-only body content is given as input to the carrot2 API, which finds the keywords of the articles, using LINGO algorithm. These keywords help in framing valid and meaningful questions. The list of keywords along with body of the article are used as input to the following algorithm. The keywords are located in the body of the article by parsing through it.

2.4 Key Phrase Extraction Techniques

Key phrases provide important information about the content of a document. Two approaches for the automatic extraction of key phrases have been studied. Supervised techniques require labeled data to train the system and tend to be more accurate but also more restricted. Unsupervised techniques do not require training sets and tend to be applicable to wider knowledge domains, but they are also less accurate. Turney introduced a system for key phrase extraction called GenEx, which is based on a set of parameterized heuristic rules tuned by a genetic algorithm. Frank et al. applied a Nave Bayes classifier for key phrase extraction on the same data used by Turney, which improved the results. Both GenEx and the Nave Bayes classifier are examples of supervised approaches to key phrase extraction. In general, supervised approaches require an annotated training set, which is often not practical. To eliminate the need for training data, several authors have developed unsupervised approaches to key phrase extraction. Barker and Cornacchia ranked noun phrases extracted from a document by using simple heuristics based on the length and the frequency of their head noun. Bracewell et al. clustered terms which share the same noun term from a list of extracted noun phrases. Another widely adopted unsupervised approach for key phrase extraction is to use graph-based ranking methods. Mihalcea and Tarau represented a document as a term graph based on term relatedness; a graph based ranking algorithm is then used to assign importance scores to each term. The Lingo algorithm, another unsupervised approach, is generally used for clustering web search results. It is based on singular value decomposition (SVD). The cluster-label induction

phrase in Lingo involves following steps. First, a term- document matrix A is built from the input documents. Second, the term-document matrix is broken into three matrix (U , S , and V) by performing SVD, such that $A \approx USVT$. Third, k column vectors of U are extracted. Each column vector refers to a cluster or latent concept. Fourth, the semantic similarities between latent concepts and single words phrases are calculated by using classic cosine distances, where each column vector of matrix P represents a single word or phrase. Last, we choose the most similar single word or phrase as the concept label by finding the largest value in each row of matrix M . Rows of the matrix M represent latent concepts, its columns represent phrases or single words, and individual values are the cosine similarities.

2.5 Selection of declarative sentences

The parsing process also picks the sentences which contain the keywords. Now with the handful of picked sentences and keywords, the next step of the algorithm starts.

2.6 Tokenization and Syntax analysis

Using NLTK for python, each of the sentences are tokenised and broken down into array of words. Tokenization is the process of splitting the given input into a stream of token. In this case, the input is the text of the received message, the token being an English word. This stage is simple in case of a whitespace-delimited language like English. It involves chopping the stream of text into words based on spaces, punctuation marks. After this stage, a bag of words is obtained. These array of words are fed as input to the Tagger inbuilt function, which categorises the words into syntactical units (i.e. nouns, verbs, pronouns, adjectives etc.) NLTK based analysis is used for tagging tokens as part of sentence(proper noun, adjective, verb etc.). The tags are used to identify type of sentences. Based on set of rules for different sentences, questions are framed simply by replacing one or more token, or by completely reframing the sentence. Example: Sentence: And now for something completely different Tokens:

- (And, CC)
- (now, RB)
- (for, IN)

- (something, NN)
- (completely, RB)
- (different, JJ)

2.7 Question and Answer framing

By knowing the syntax of the sentences, the assertive sentences are converted into interrogative sentences by following the rules of English grammar. Eg) INPUT: Rio is nicknamed the Cidade Maravilhosa or "Marvelous City". OUTPUT : What is nicknamed the Cidade Maravilhosa or "Marvelous City"? All the sentences are changed into interrogative questions. In the meantime, the answers to the questions, which are extracted from the sentences are also stored for future use. Both the questions and the answers are stored in the local database.

2.8 Answer evaluation

A php page, retrieves all the questions and presents them to the user. The user then answers all the questions specified. When the submit button is clicked upon, the user's answers are then stored in the database. Both the user given answers and the actual answers are checked for equality. If the answers match then it is graded with +1 points, if it is wrong no points are added. By calculation the percentage for correct answers to the total questions, the user is given appropriate marks.

3 Implementation Details

3.1 Software Used

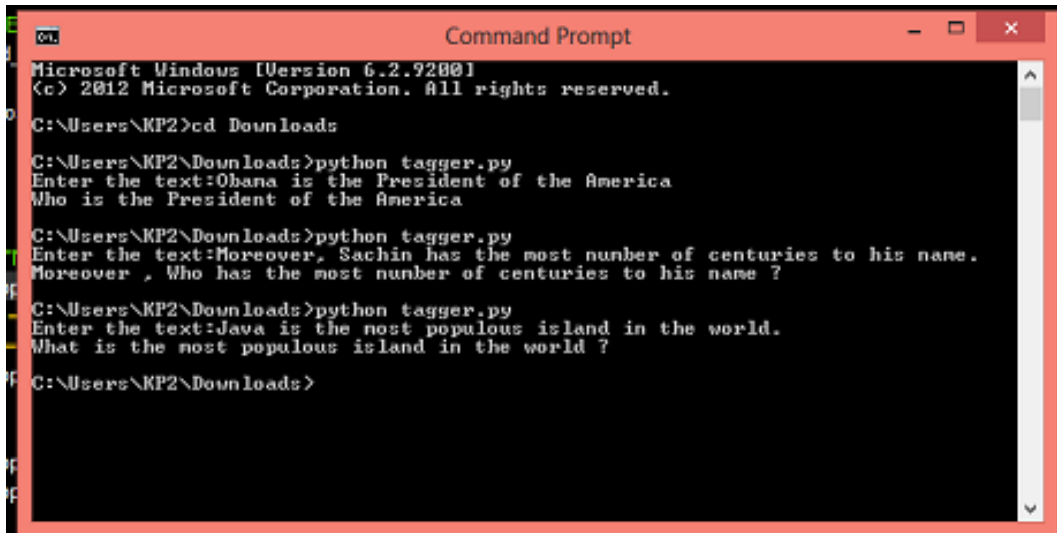
For the text processing Natural Language Tool Kit (NLTK) in python environment is used. Various libraries including Tagger library are being used. For information retrieval from Wikipedia, Mediawiki API python environment is used. For parsing of HTML content, BeautifulSoup (BS4) library, which deals with HTML tags, is used. For the utilisation of lingo algorithm, carrot2 API is used, in python environment.

3.2 Corpus used

For the purpose of Disambiguation, wikipedia disambiguation pages are used. For the purpose of tagging, builtin NLTK dictionary is used.

3.3 Module Implementation with Snapshots

Fig 3: Question Framing



```
Microsoft Windows [Version 6.2.9200]
(c) 2012 Microsoft Corporation. All rights reserved.

C:\Users\KP2>cd Downloads

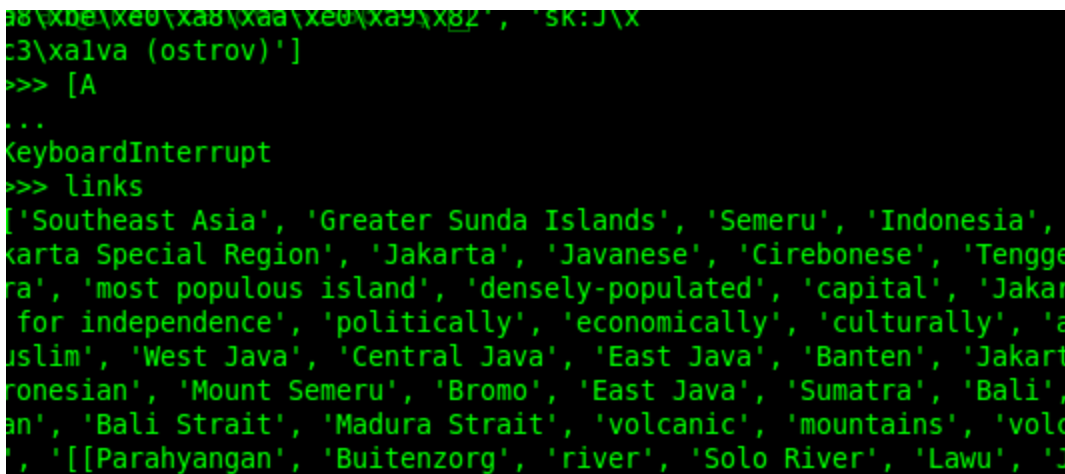
C:\Users\KP2\Downloads>python tagger.py
Enter the text:Obama is the President of the America
Who is the President of the America

C:\Users\KP2\Downloads>python tagger.py
Enter the text:Moreover, Sachin has the most number of centuries to his name.
Moreover , Who has the most number of centuries to his name ?

C:\Users\KP2\Downloads>python tagger.py
Enter the text:Java is the most populous island in the world.
What is the most populous island in the world ?

C:\Users\KP2\Downloads>
```

Fig 4: List of key-words



```
['Southeast Asia', 'Greater Sunda Islands', 'Semeru', 'Indonesia', 'Jakarta Special Region', 'Jakarta', 'Javanese', 'Cirebonese', 'Tenggera', 'most populous island', 'densely-populated', 'capital', 'Jakarta for independence', 'politically', 'economically', 'culturally', 'Muslim', 'West Java', 'Central Java', 'East Java', 'Banten', 'Jakarta', 'Indonesian', 'Mount Semeru', 'Bromo', 'East Java', 'Sumatra', 'Bali', 'Bali Strait', 'Madura Strait', 'volcanic', 'mountains', 'volcano', 'Parahyangan', 'Buitenzorg', 'river', 'Solo River', 'Lawu']
```


Fig 5: JSON output with metadata

```

- pages: {
  - 69336: {
    pageid: 69336,
    ns: 0,
    title: "Java",
    - revisions: [
      - {
        contentformat: "text/x-wiki",
        contentmodel: "wikitext",
        *: "{{{about|the Indonesian island}}} {{Infobox islands |name = J
name = Java |native name link = Indonesian language |location =
|archipelago = [[Greater Sunda Islands]] |area km2 = 138794 |ra
|country admin divisions = [[Banten]],<br />[[Jakarta|Jakarta&n
[[Yogyakarta|Yogyakarta Special Region]] |country largest city
1064 |ethnic groups = [[Javanese people|Javanese]] (inc. [[Cire
people|Betawi]], [[Madurese people|Madurese]] }} ''Java'' ({{
island of [[Madura]] which is administered as part of the provi

```

Fig 6: Metadata with tables

Since its original release, ''DotA'' has become a feature at several wor as the Cyberathlete Amateur and CyberEvolution leagues; in a 2008 articl popular and most-discussed free, non-supported game mod in the world".<| Dota2></ref>"

```

c
"{{{For}}}
{{{pp-semi|small=yes}}}
{{{Taxobox
| name           = Domestic dog
| fossil_range   = {{{Fossil range|0.015|0}}}<small>[[Pleistocene]]&nbsp;&
| status         = DOM
| image          = YellowLabradorLooking new.jpg <!-- Please do not chang
| image_width    = 260px
| image_caption  = Yellow [[Labrador Retriever]], the most registered bre
| regnum         = [[Animal]]ia
| phylum       = [[Chordate|Chordata]]
| classis       = [[Mammal]]ia
| ordo          = [[Carnivora]]
| familia       = [[Canidae]]
| genus         = ''[[Canis]]''
| species       = ''[[Gray Wolf|C. lupus]]''
| subspecies    = ''''C. l. familiaris''''<ref name = "MSW3 Canis lupu
Browse: Canis lupus familiaris|publisher=Bucknell.edu |year=2005 |access
| trinomial     = ''Canis lupus familiaris''<ref name="MSW3 Lupus"/>

```

Fig 7: Text only content after Parsing

"Java () is an island of Indonesia. With a population of 13 world's most populous island, and one of the most densely-p is located on western Java. Much of Indonesian history took East Indies. Java was also the center of the Indonesian str

Formed mostly as the result of volcanic eruptions, Java is spine along the island. It has three main languages, though its residents are bilingual, with Indonesian as their first ethnicities, and cultures.

Java is divided into four provinces, West Java, Central Jav

==Etymology==

The origins of the name "Java" are not clear. One possibili prior to Indianization the island had different names. Ther plant for which the island was famous. "Yawadvipa" is menti Java, in search of Sita. It was hence referred to in Indian word is derived from a Proto-Austronesian root word, meanin

==Geography==

Java lies between Sumatra to the west and Bali to the east.

Fig 8: Sample I/O for Parsing

```
>>> input_data = " '''Java''' is an [[island]] near [[Indonesia | Indonesian islands]]  
>>> url='http://newsbbc.com/island_news/' </ref>}}."  
>>> parsedData = json.loads(requests.post("http://localhost:8888", data=input_data, pr  
or "{ 'data': 'No Data' }")['data'].replace('\n',''))  
>>>  
>>>  
>>> print parsedData  
Java is an island near Indonesian islands .  
>>>  
>>>
```

Fig 9: Sentence Split-up

```
>>>
>>> for h in range(8):
...     print sentences[h], '\n'
...

Java () is an island of Indonesia.

With a population of 135 million (excluding the 3.6 million on the island of Java), Java is the most populous island, and one of the most densely-populated places on the globe.

Java is the home of 60 percent of the Indonesian population.

The Indonesian capital city, Jakarta, is located on western Java.

Much of Indonesian history took place on Java.

It was the center of powerful Hindu-Buddhist empires, the Islamic sultanate of Mataram.

Java was also the center of the Indonesian struggle for independence during the 1940s.
```

References

- [1] Liu, Ming *Using Wikipedia and Conceptual Graph Structures to Generate Questions for Academic Writing Support.*
<http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6165258&contentType=Journals+%26+Magazines&queryText%3Dquestion+generation+wikipedia>
- [2] Yajie Miaos . *Improving Question Answering Based on Query Expansion with Wikipedia.*
<http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=5671408&contentType=Conference+Publications&queryText%3Danswers+wikipedia>
- [3] Figueroa, A. *Mining Wikipedia Resources for Discovering Answers to List Questions in Web Snippets.*
<http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=4725906&contentType=Conference+Publications&queryText%3Danswers+wikipedia>

- [4] Shaidah Jusoh, Hejab M. Al Fawareh , *Semantic Extraction from Texts* published 2009 International Conference on Computer Engineering and Applications, Singapore
- [5] Silvia Calegari and Davide Ciucci, *Fuzzy Ontology, Fuzzy Description Logics and Fuzzy-OWL*, ACM 7th international workshop on Fuzzy Logic and Applications
- [6] [*English WordList with Meanings*](#)
- [7] [*WordNet Database, Princeton University*](#)