

Knowledge Test based on Question Answering

Dhinesh.N, Pradeepan.K, Bharath.R
BE CSE, Semester 7

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

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

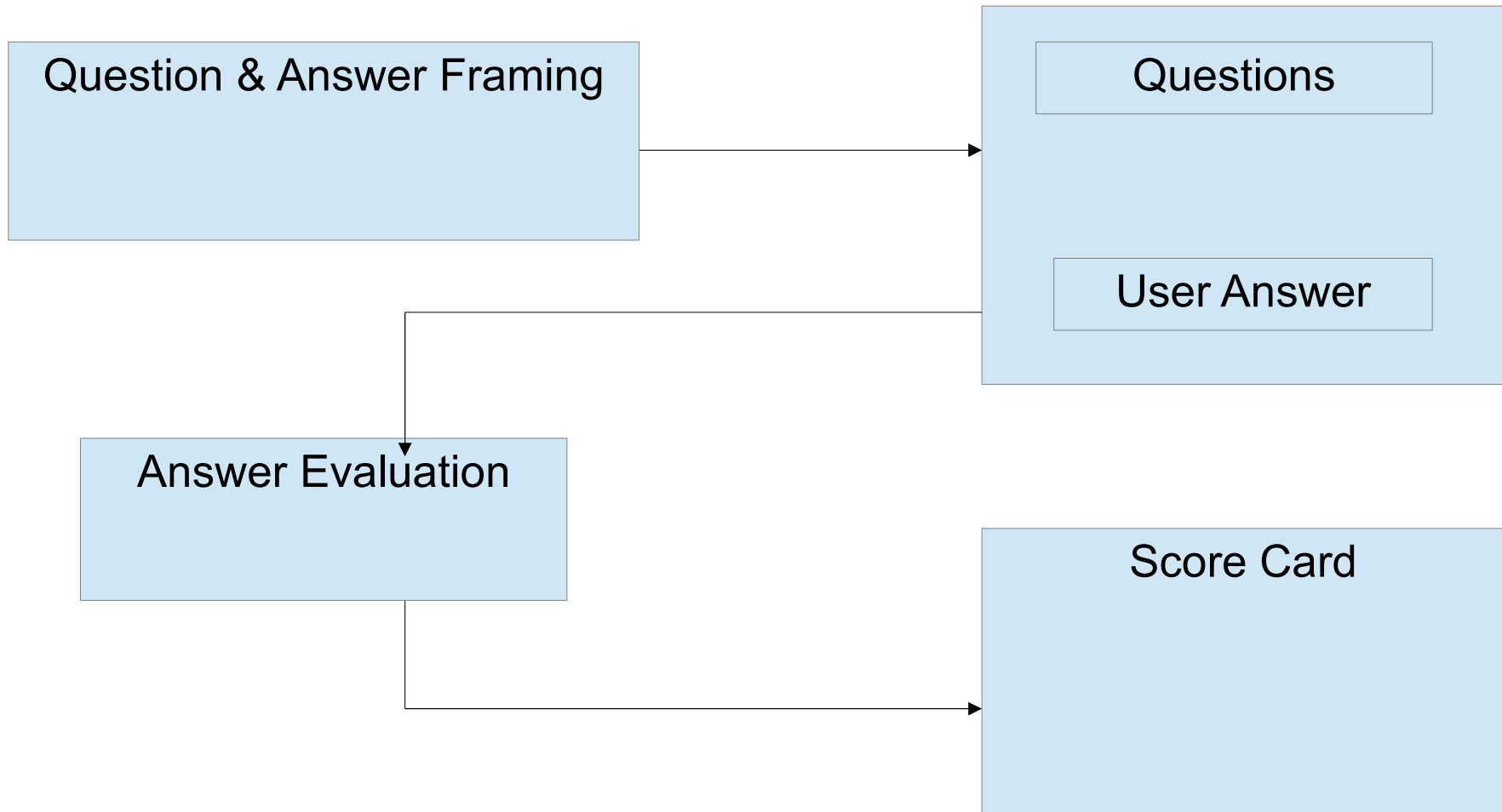
Outline

- Design - Outline
- Design - Detailed
- Module Split-up
- Implementation Details
 - Previous Results
 - Module Output/Screen-shot

Design Outline

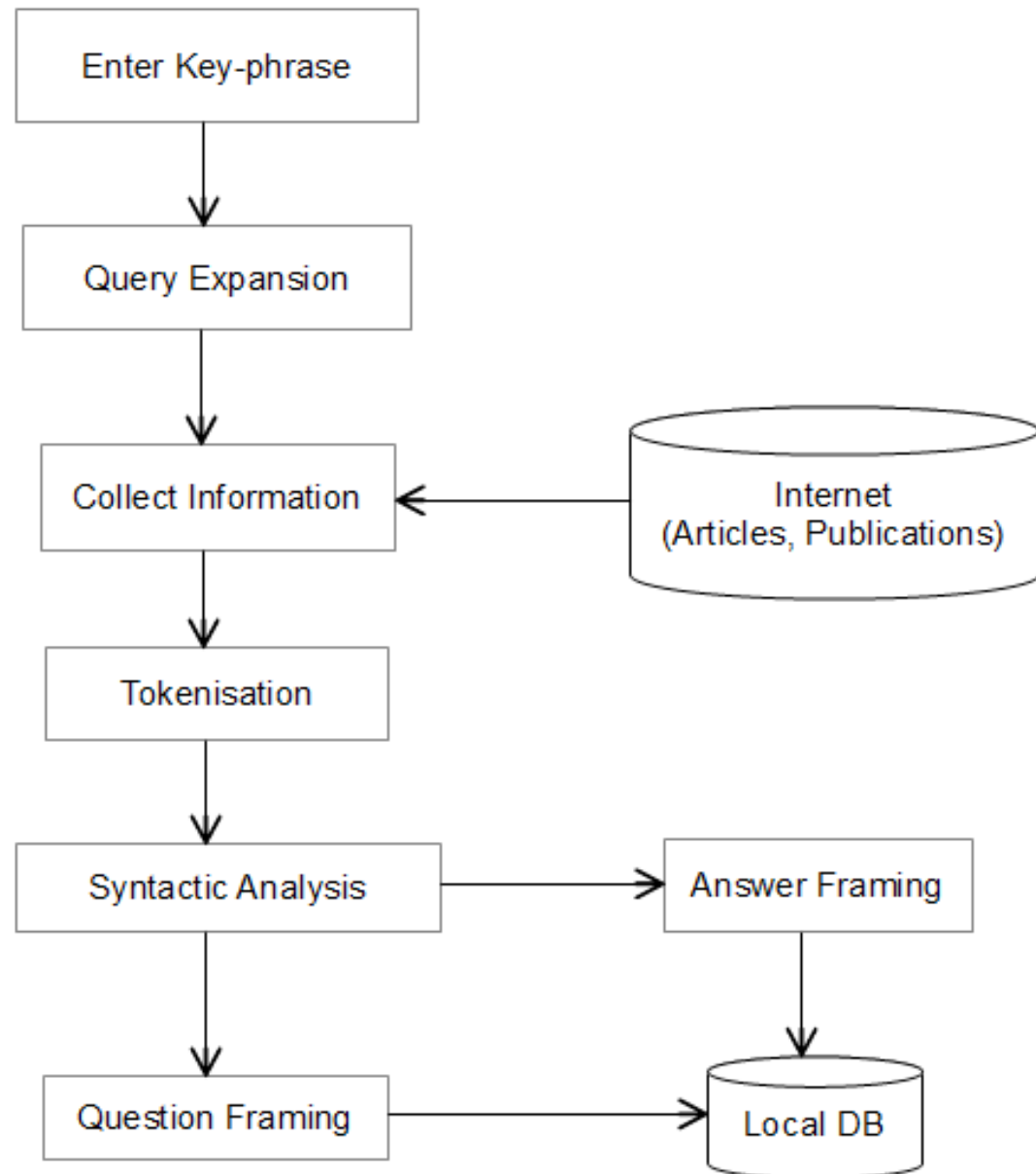
SYSTEM

USER



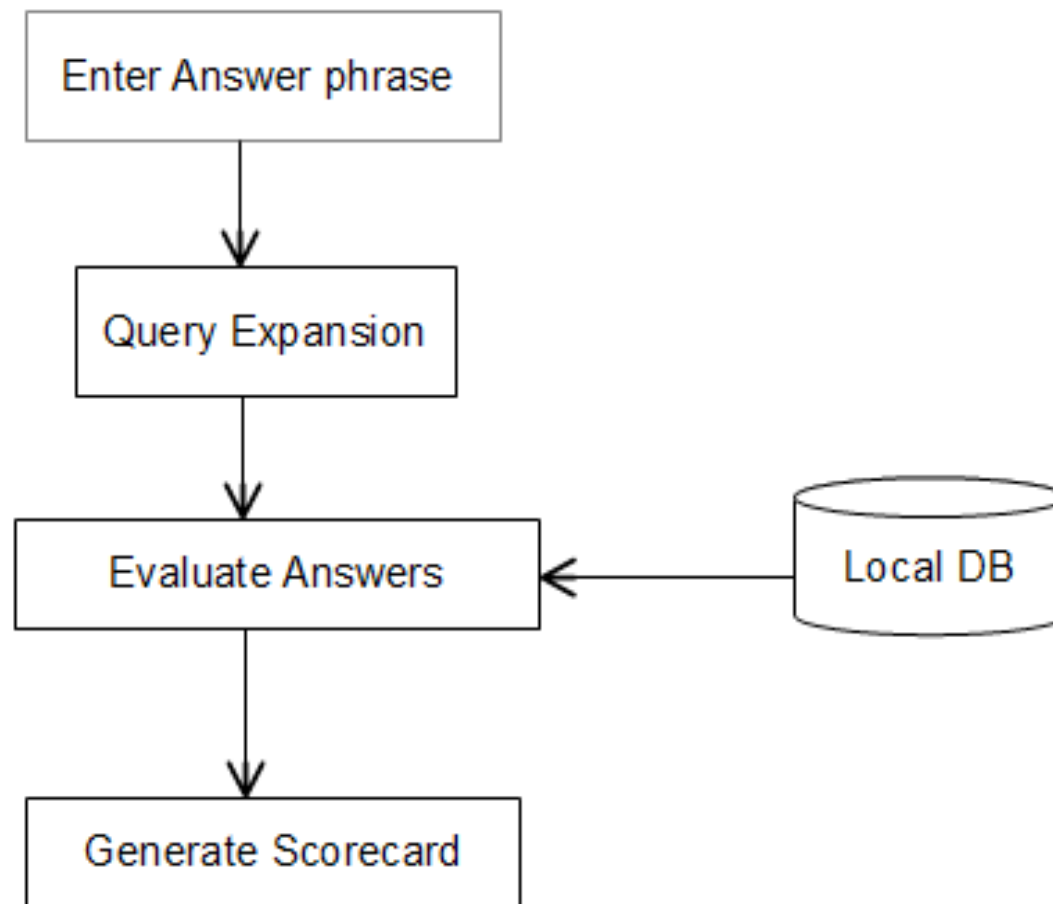
Detailed Design

Figure 1. Question and Answer Generation



Detailed Design

Figure 2. Answer Evaluation



Module Split-up

I. Key-phrase Input and Topic Selection

- User inputs the topic of his interest.
- The input is searched for in Wikipedia using Mediawiki API.
- In case of disambiguation, wikipedia disambiguations for the word is displayed for user to select.

II. Parsing and Extraction of necessary data

- The API returns a JSON file with head and body of article.
- Body of file is extracted and parsed using BeautifulSoup python library.

III. Search for Keywords

- The text-only content is given to Carot2 API, which determines the keywords of the article.
- The Keywords aid in extracting meaningful sentences.

IV. Selection of Declarative sentences

- Sentences containing the keywords are parsed.
- The sentences are used to frame questions.

V. Tokenisation and Syntax Analysis

- Sentences are broken into array of words using NLTK for python.
- These array of words are fed as input to the Tagger inbuilt function, which categorizes the words into syntactical units (i.e. nouns, verbs, pronouns, adjectives etc.)
- The tags are used to identify type of sentences.

VI. Question and Answer framing

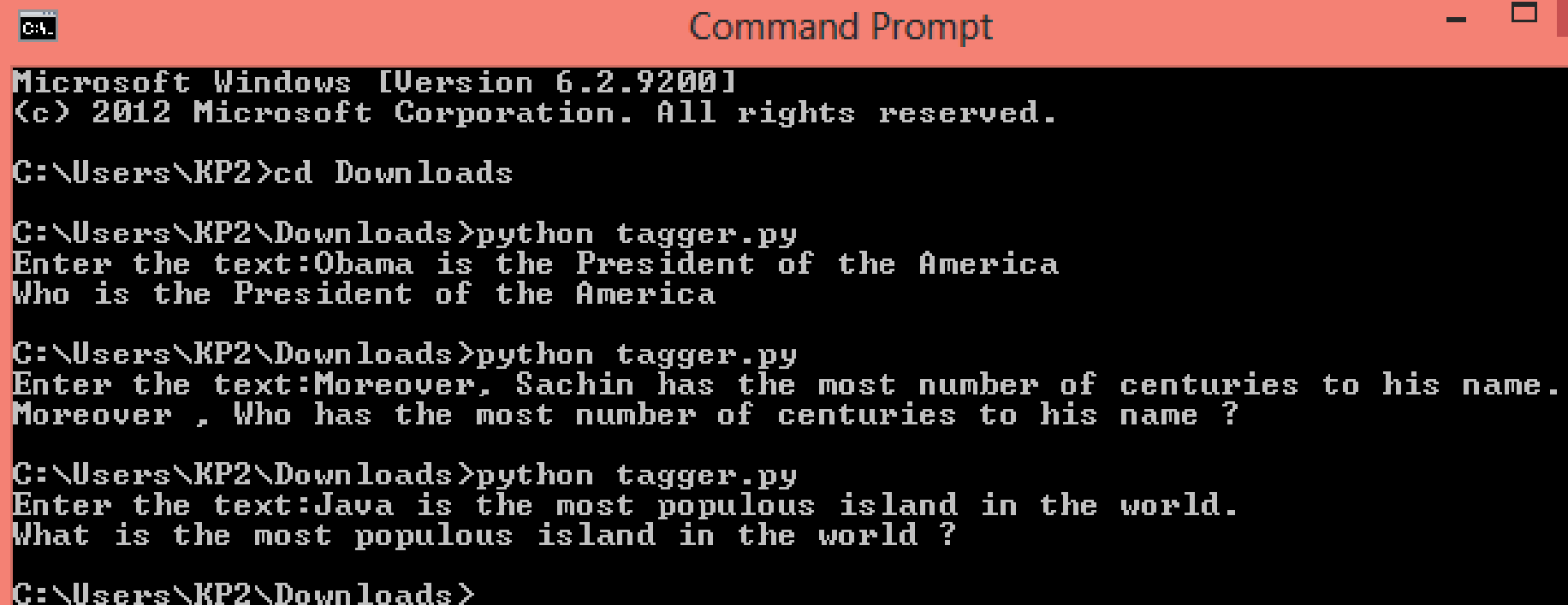
- Assertive sentences are converted to interrogative sentences using the rules of English grammar.
- Similarly, all converted sentences are stored locally.
- Answer to questions, known through the sentences are also stored.
- Questions are retrieved from this database as and when required by user.
- Answers, so stored, are used for query expansion and evaluation of user answers to questions

VII. Answer Evaluation

- A PHP file retrieves and displays questions to user.
- User enters answers to these questions.
- User answers are evaluated for equality with answers stored in the database.
- Scores are awarded based on the correctness of the answers.

Implementation Details

Previous Results



```
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>
```

Module Output/Screenshot

List of Key-Words

```
a8\xde\xed\xab\xad\xed\xay\x82', 'SK:J\x
c3\xalva (ostrov)']
>>> [A
...
KeyboardInterrupt
>>> links
['Southeast Asia', 'Greater Sunda Islands', 'Semeru', 'Indonesia',
'Jakarta Special Region', 'Jakarta', 'Javanese', 'Cirebonese', 'Tengge
ra', 'most populous island', 'densely-populated', 'capital', 'Jakarta
for independence', 'politically', 'economically', 'culturally', 'a
uslim', 'West Java', 'Central Java', 'East Java', 'Banten', 'Jakarta
onesian', 'Mount Semeru', 'Bromo', 'East Java', 'Sumatra', 'Bali',
an', 'Bali Strait', 'Madura Strait', 'volcanic', 'mountains', 'volc
', '[[Parahyangan', 'Buitenzorg', 'river', 'Solo River', 'Lawu', 'J
```

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 = Jawa |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
```

Metadata with tables

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



```
"{{For}}
{{pp-semi|small=yes}}
{{Taxobox
| name           = Domestic dog
| fossil_range   = {{Fossil range|0.015|0}}<small>[[Pleistocene]]&nbsp;&#x2013;&#x2013;Present
| status         = DOM
| image          = YellowLabradorLooking new.jpg <!-- Please do not change
| image_width    = 260px
| image_caption  = Yellow [[Labrador Retriever]], the most registered breed
| 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 lupus familiaris" />
Browse: Canis lupus familiaris|publisher=Bucknell.edu |year=2005 |accessdate=
| trinomial      = ''Canis lupus familiaris''<ref name="MSW3 Lupus"/>
```

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.

Example 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 .
>>>
>>>
>>>
```

Sentences 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 Sumatra), 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, and the Dutch colonial capital.

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

Thank You