

Question Answering System for Yioop

Advisor

Dr. Chris Pollett

Committee Members

Dr. Thomas Austin

Dr. Robert Chun

By

Niravkumar Patel

Outline


- Problem Statement
- Question Answering System
- Yioop
- Proposed System
- Triplet Extraction Approach
- Integration with Yioop
- Observations
- Areas of Improvement
- Conclusion

Problem Statement

- Information Retrieval in Yioop doesn't provide a specific answer to the query entered in the form of a question
- Currently, Yioop treats both a query statement and a question in the same manner

Example 1

[Web](#) [Images](#) [Videos](#) [News](#)



what is http

0.03996 seconds. Showing 1 - 10 c

[SiteUptime - Website and Server Monitoring ServiceSiteUptime Frequently Asked Questions :: What is H](#)

[www.siteuptime.com/faq/questions...ionid=84](#) **Words:** plan knowledgebase authentication support
SiteUptime - Website and Server Monitoring Service. Login. Register. 1-866-744-6591. Overview. Example Reports. API. Register. Login. Compare. Demo. C
[Cached](#). [Similar](#). [Inlinks](#). IP:66.11.12.81.

[What Is 'HTTP' Protocol, and How Does It Affect Me?](#)

[netforbeginners.about.com/od/p/f...t-Me.htm](#) **Words:** internet protocol web policy transfer
Technology iPad. Internet Basics. About.com About Technology Internet Basics. Glossary of Internet Terminology and Acronyms P - Glossary of Internet T
[Cached](#). [Similar](#). [Inlinks](#). IP:208.185.127.82.

[HTTPS - What is Https? A simple explanation with video](#)

[www.jafaloo.com/https-what-is-ht...lanation](#) **Words:** data internet protocol server client
A simple explanation with video. Home How To Tech Blog Web Browsers. Google+. Tech Blog / Tutorial. A simple explanation. 4 Flares 4 Flares. While sur
[Cached](#). [Similar](#). [Inlinks](#). IP:192.254.225.237.


[What is HTTPS? - YouTube](#)



[www.youtube.com/watch?v=JCvPnwpWVUQ](#) **Words:** play sign loading video views
What is HTTPS. - YouTube. Upload. Sign in. Search. . Loading... This video is

Example 2

[Web](#) [Images](#) [Videos](#) [News](#)



who is the president of USA

0.04617 seconds. Showing 1 - 4 of 4

Search: [who is the president of us](#)

[Rajasthan Pradesh Congress Committee News Photos Videos - Rediff.com](#)
www.rediff.com/tags/-rajasthan-p...ommittee Words: congress rajasthan india news pradesh
Rajasthan Pradesh Congress Committee News Photos Videos - Rediff.com. Rediff News. Rediff News All
News. Rajasthan Pradesh Congress Committee Subscrib
[Cached](#). [Similar](#). [Inlinks](#). [IP:96.17.148.17](#).

[Centre keeps away from row over Delhi University's 4-year-course - Rediff.com India News](#)
www.rediff.com/news/report/admis...0623.htm Words: year delhi programme news ugc
Centre keeps away from row over Delhi University's 4-year-course - Rediff.com India News. News Centre
keeps away from row over Delhi University's 4-year
[Cached](#). [Similar](#). [Inlinks](#). [IP:0.0.0.0](#).

[Centre keeps away from row over Delhi University's 4-year-course - Rediff.com India News](#)
uswww.rediff.com/news/report/adm...0623.htm Words: year programme delhi news ugc
Centre keeps away from row over Delhi University's 4-year-course - Rediff.com India News. News Centre
keeps away from row over Delhi University's 4-year
[Cached](#). [Similar](#). [Inlinks](#). [IP:119.252.148.17](#).

1

Are questions really asked by users online?

- Search engines query log analysis shows that

Types of Query	Query log analysis
Informational	48%
Navigational	20%
Transactional	30%

- 27% of Informational queries are questions

Question Answering System



START
Natural Language Question Answering System

where is paris  [Ask Question >](#)

==> where is paris

The coordinates of Paris, France are 48.86 N, 2.33 E.

Paris is located in [France](#).

Source: START KB

[Paris, France](#) is located at 114.0 feet above sea level.

Source: [Global Gazetteer](#)



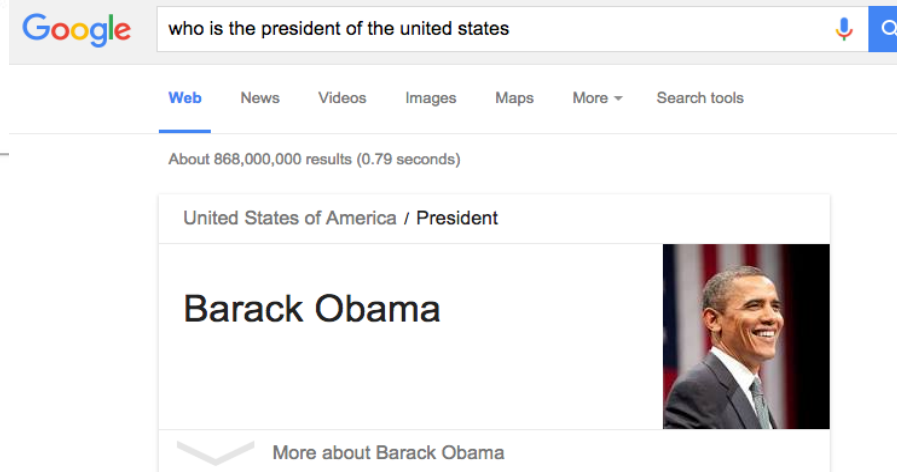
who is christopher columbus

[About](#) | [Images](#) [Videos](#) [News](#)

 **Christopher Columbus**

Christopher Columbus was an Italian explorer, navigator, colonizer and citizen of the Republic of Genoa. Under the auspices of the Catholic Monarchs of Spain, he completed four voyages across the Atlantic Ocean. Those voyag

[+ Show More](#) | [W More at Wikipedia](#)




Google who is the president of the united states

[Web](#) [News](#) [Videos](#) [Images](#) [Maps](#) [More](#) [Search tools](#)

About 868,000,000 results (0.79 seconds)

United States of America / President

Barack Obama 

[More about Barack Obama](#)

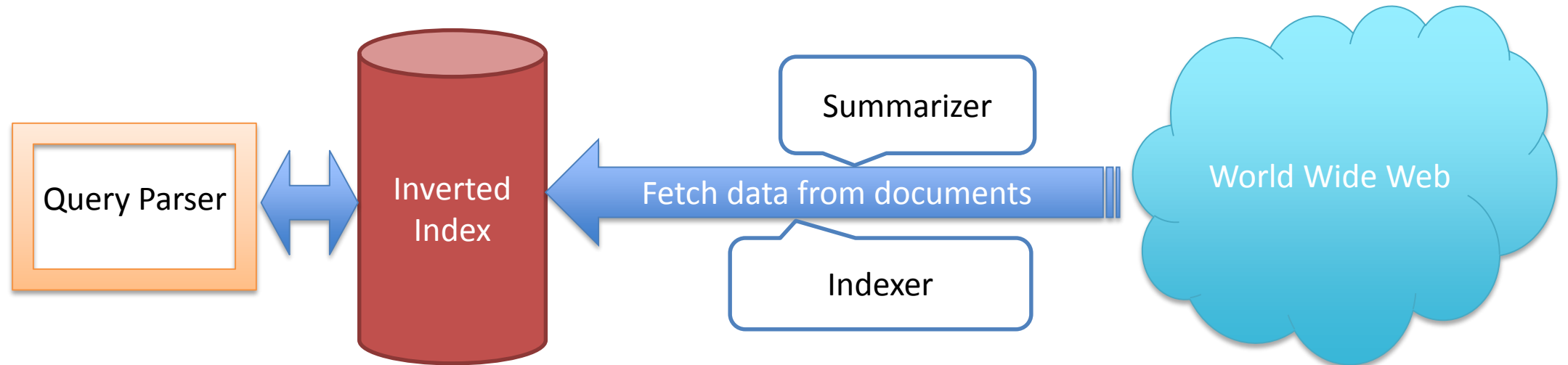
[President of the United States - Wikipedia, the free ...](#)
https://en.wikipedia.org/wiki/President_of_the_United_States [Wikipedia](#)
On January 20, 2009, **Barack Obama** became the 44th and current president. On November 6, 2012, he was re-elected and is currently serving the 57th term, which ends on January 20, 2017.

[Columbus](#) [Wikipedia, the free encyclopedia](#)

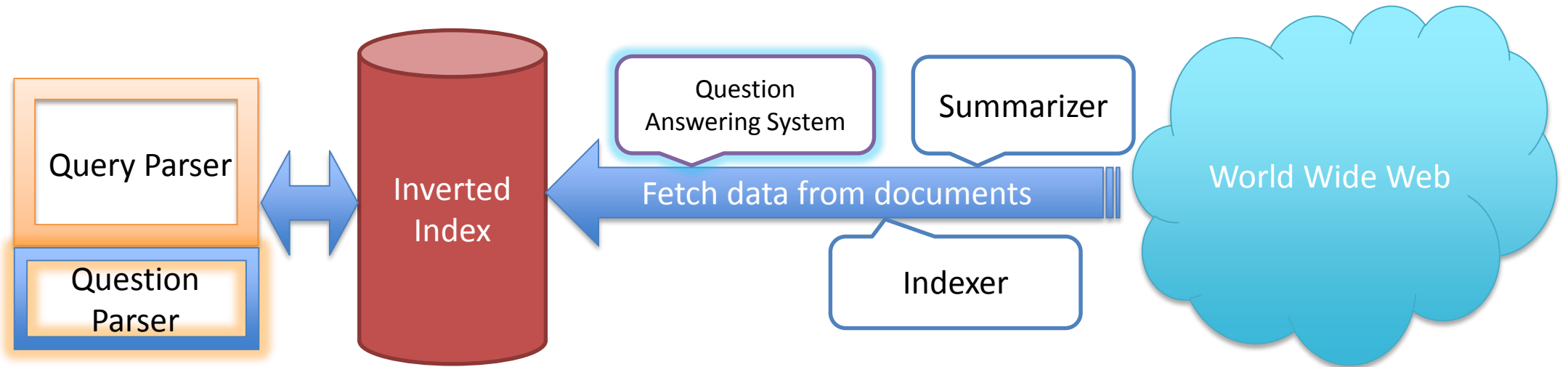
Yioop

- Open Source Search Engine
- Major processes
 - Crawler
 - Summarizer
 - Index builder
 - Query Parser

Information Flow in Yioop



Proposed Question Answering System for Yioop



Approaches for Question Answering System

- Named-entity recognition technology
- Knowledge based Approach
- Triplet Extraction Approach

Approach Chosen

- Triplet Extraction Algorithm
- Identifies extraction of relations between Subject and Object
- [SUBJECT – PREDICATE - OBJECT]

Prerequisite

- Structure of statement
- Part Of Speech Tagger
- Parse Tree Generation

Structure of Statement

- The phrase given to the system should be a complete statement.
- Ex.
 - “Barack Obama was born on August 4 1961” => [Barack Obama – born – on August 4 1961]
 - “Barack Obama Obama was born on August 4 1961” => [Barack Obama Obama – born – on August 4 1961]

Part of Speech Tagger

- Example
 - Statement: The grand jury commented on a number of other topics
- Output of statement
 - The[DT] grand[JJ] jury[NN] commented[VBD] on[IN] a[DT] number[NN] of[IN] other[JJ] topics[NNS]
- Tagging Technique used
 - Brill Tagger

Parse Tree Generation

- Grammar rules to interpret a common form of a statement

NP:	{<DT JJ NN.*>+}
PP:	{<IN><NP>}
VP:	{<VB.*><NP PP>+}
Statement:	{<NP><VP>}

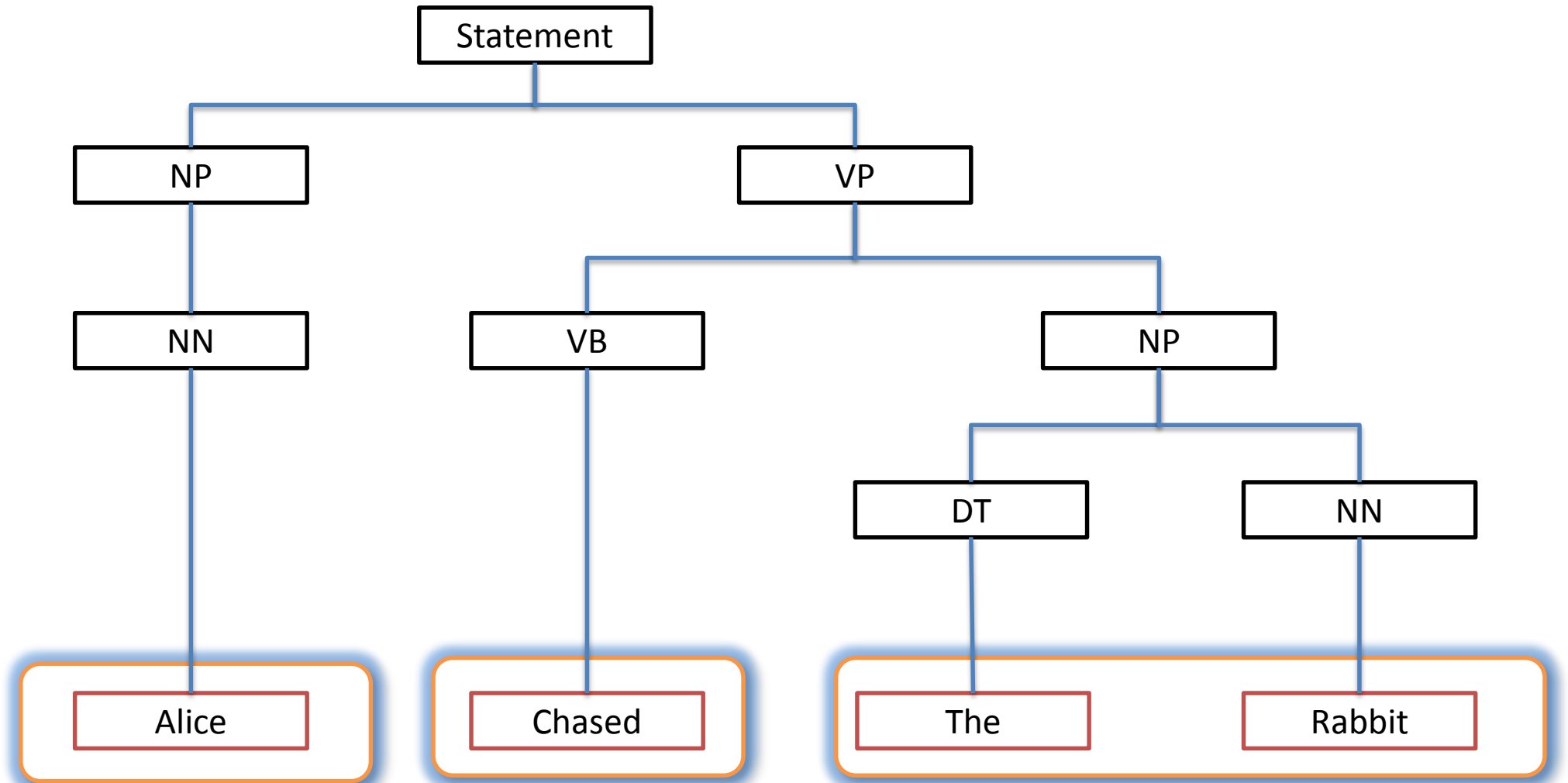
Triplet Extraction Algorithm

- Subject Extraction : Noun in the NP Sub-tree
- Predicate Extraction: Deepest Verb descendent of Verb Phrase
- Object Extraction : All siblings of VB in Verb Phrase Sub-tree

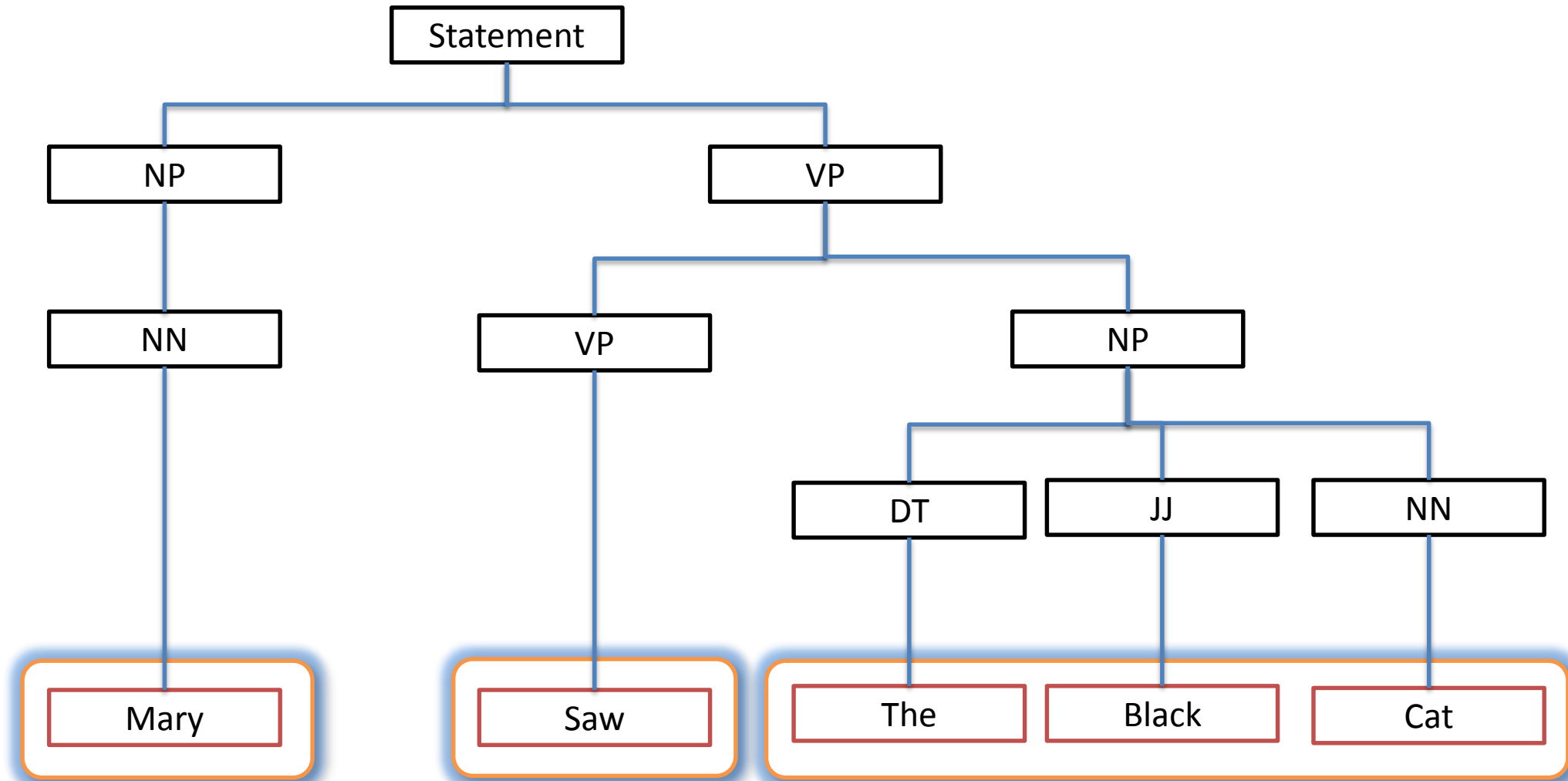
Triplet Extraction Algorithm

- Statement: Alice chased the rabbit
- POS output: Alice [NN] chased [VB] the [DT] rabbit [NN]
- Triplet Extracted: [ALICE – CHASED – THE RABBIT]

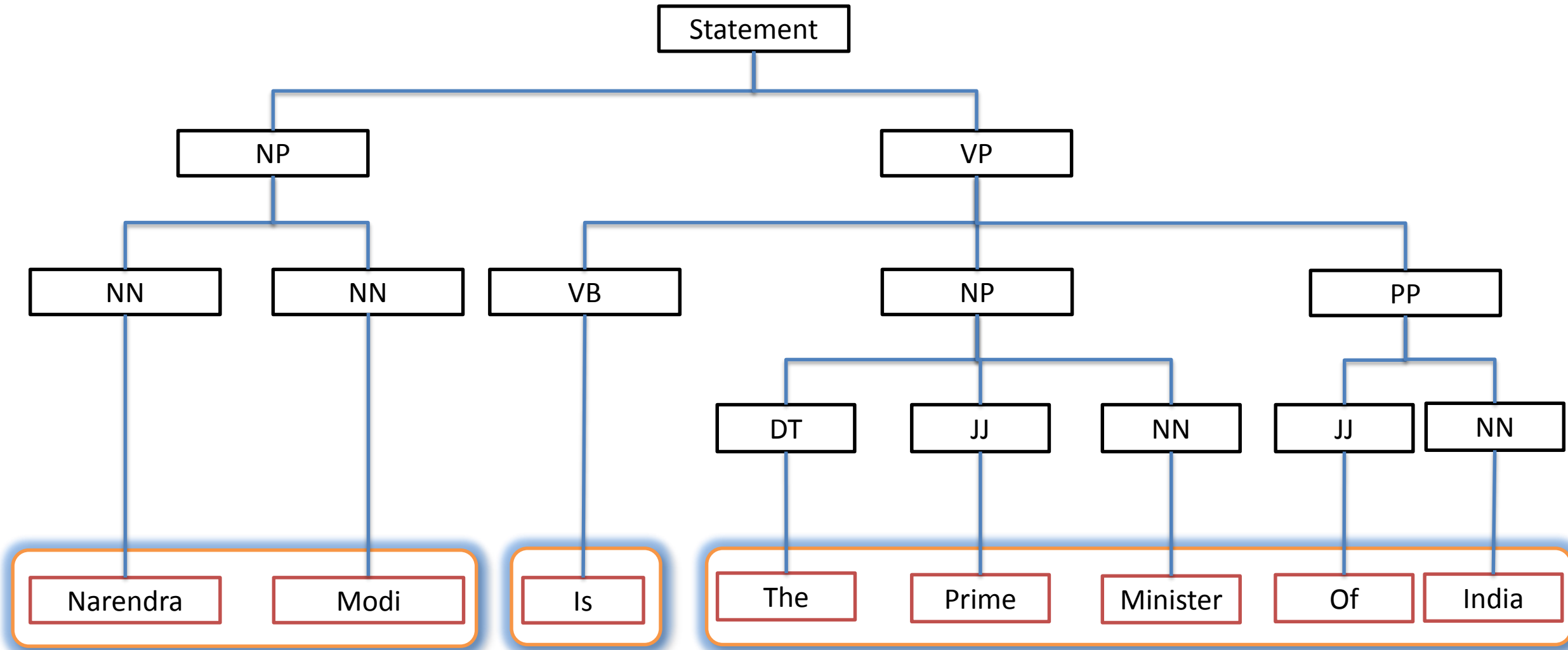
Ex. Alice chased the rabbit



Ex. Mary saw the black cat



Ex. Narendra Modi is the Prime Minister of India



Integration with Yioop

- At Crawl time
- At Query time

Integration at Crawl time

Page Extraction

Summary Generation

Phrase List Generation

Hypertext Transfer Protocol

From Wikipedia, the free encyclopedia

The **Hypertext Transfer Protocol (HTTP)** is an [application protocol](#) for distributed, collaborative, [hypermedia](#) information systems.^[1] HTTP is the foundation of data communication for the [World Wide Web](#).

[Hypertext](#) is structured text that uses logical links ([hyperlinks](#)) between nodes containing text. HTTP is the protocol to exchange or transfer hypertext.

The standards development of HTTP was coordinated by the [Internet Engineering Task Force \(IETF\)](#) and the [World Wide Web Consortium \(W3C\)](#), culminating in the publication of a series of [Requests for Comments \(RFCs\)](#). The first definition of HTTP/1.1, the version of HTTP in common use, occurred in [RFC 2068](#) in 1997, although this was obsoleted by [RFC 2616](#) in 1999.

Contents [\[hide\]](#)

- 1 Technical overview
- 2 History
- 3 HTTP session
- 4 HTTP Authentication
 - 4.1 Authentication Realms
- 5 Request methods
 - 5.1 Safe methods

After page rules applied

```
Array
(
    [ENCODING] => UTF-8
    [URL] => http://test-site.yioop.com/
    [IP_ADDRESSES] => Array
        (
            [0] => 1.1.1.1
        )
    [HTTP_CODE] => 200
    [MODIFIED] => 1448913409
    [TIMESTAMP] => 1448913409
    [TYPE] => text/html
    [HEADER] => page options test extractor
    [SERVER] => unknown
    [SERVER_VERSION] => unknown
    [OPERATING_SYSTEM] => unknown
    [LANG] => en
    [JUST_METAS] =>
    [ROBOT_METAS] => Array
        (
        )
    [TITLE] => Hypertext Transfer Protocol - Wikipedia, the free encyclopedia
    [DESCRIPTION] => 4 HTTP Authentication. 5 Request methods. 5.1 Safe methods. 5.2 Idempotent methods and web applications. 10 Request message. 11 Response message. 12.2 Server response. The server, which provides resources such as HTML files and other content, or performs other functions on behalf of the client, returns a response message to the client. HTTP provides multiple authentication schemes such as Basic access authentication and Digest access authentication which operate via a challenge-response mechanism whereby the server identifies and issues a challenge before serving the requested content. HTTP provides a general framework for access control and authentication, via an extensible set of challenge-response authentication schemes, which can be used by a server to challenge a client request and by a client to provide authentication information. For example, WebDAV defined 7 new methods and RFC 5789
```

Words and positions extracted to index from summary

```
Array
(
    [test site yioop hypertext transfer protocol wikipedia the free encyclopedia] => Array
        (
            [0] => 0
            [cond_max] => 5
        )
    [site yioop hypertext transfer protocol wikipedia the free encyclopedia 4] => Array
        (
            [0] => 1
            [cond_max] => 5
        )
    [yioop hypertext transfer protocol wikipedia the free encyclopedia 4 http] => Array
        (
            [0] => 2
            [cond_max] => 6
        )
    [hypertext transfer protocol wikipedia the free encyclopedia 4 http authent] => Array
        (
            [0] => 3
            [cond_max] => 10
        )
    [transfer protocol wikipedia the free encyclopedia 4 http authent 5] => Array
        (
            [0] => 4
            [cond_max] => 9
        )
)
```

Question Answer Triplets List

```
[qqque get option] => some of the method
[method qqque option] => get
[method get qqque] => head option
[the method for exampl head qqque option] => get
[the method for exampl head get qqque] => option
[action qqque _eg method] => take
[action take qqque] => _eg method
[the action the server qqque _eg method] => take
[the action the server take qqque] => _eg method
[qqque is protocol] => idempot
[http qqque protocol] => is
[http is qqque] => a stateless protocol
[qqque is a stateless protocol] => as http
[http qqque a stateless protocol] => is
[qqque be us] => challeng respons authent scheme
[authent scheme qqque us] => be
[authent scheme be qqque] => us by a server
[qqque be us by a server] => scheme
[authent scheme qqque us by a server] => be
[qqque digest access] => authent scheme such
[authent scheme qqque access] => digest
[authent scheme digest qqque] => access
[authent scheme such qqque access] => digest
[authent scheme such digest qqque] => access
[qqque digest access authent] => access authent
[authent qqque access authent] => digest
[authent digest qqque] => access authent
[qqque set of challeng respons authent scheme] => extens
```


Limitation on Triplet Extraction

- Number of words
- Fails to ignore triplets having unnecessary Subject/Object
 - [Nicholas Carsen – Wrote – In 2014]
 - [The university website – Made - Available]

Integration at Query Time



- Ex.
 - What is http? => HTTP is *qqque*
 - Who ate the dog? => *qqque* ate the dog
 - Where is SJSU University located? => [SJSU University - is located – *qqque*] & [SJSU University - located – *qqque*]

(*qqque* is the identifier for the question word)


Limitation on Question Parser

- Processes question that starts with Who, Where, What.
- Ex.
 - Why is the sky blue?
 - How many days in a year?
 - President of USA is?

Observation on results

Query results on Yioop without Question Answering System

Web Images Videos News


 what is http

0.03930 seconds. Showing 1 - 10 of 905

[SiteUptime - Website and Server Monitoring Service](#)
[SiteUptime Frequently Asked Questions](#)
[What is H](#)
www.siteuptime.com/faq/questions...ionid=84 Words: **plan knowledgebase authentication support**
SiteUptime - Website and Server Monitoring Service. Login. Register. 1-866-744-6591. Overview. Example Reports. API. Register. Login. Compare. Demo. C
[Cached](#) [Similar](#) [Inlinks](#) [IP:66.11.12.81](#)

[What is 'HTTP' Protocol, and How Does it Affect Me?](#)
netforbeginners.about.com/od/p/f...t-Me.htm Words: **internet protocol web policy transfer**
Technology iPad. Internet Basics. About.com About Technology Internet Basics. Glossary of Internet Terminology and Acronyms P - Glossary of Internet T
[Cached](#) [Similar](#) [Inlinks](#) [IP:208.185.127.82](#)

[HTTPS - What is Htpps? A simple explanation with video](#)
www.jafaloo.com/https-what-is-ht...lanation Words: **data internet protocol server client**
A simple explanation with video. Home How To Tech Blog Web Browsers. Google+. Tech Blog / Tutorial. A simple explanation. 4 Flares 4 Flares. While sur
[Cached](#) [Similar](#) [Inlinks](#) [IP:192.254.225.237](#)


[What is HTTPS? - YouTube](#)
 www.youtube.com/watch?v=JCvPnwpWVUQ Words: **play sign loading video views**
What is HTTPS. - YouTube. Upload. Sign in. Search. Loading... This video is unavailable. Watch Queue. TV Queue. Watch Queue TV Queue. Remove all Dis
[Cached](#) [Similar](#) [Inlinks](#) [IP:74.125.239.39](#)

[What is HTTP Method PROPFIND used for?](#)
nerdanswer.com/answer.php?q=489902&ref=q Words: **libreoffice images flag jpg propfind**
What is HTTP Method PROPFIND used for. New Questions. Recently Answered. Archives. Contact. What is HTTP Method PROPFIND used for. Date Published: 03/
[Cached](#) [Similar](#) [Inlinks](#) [IP:50.116.20.103](#)

[What is 'Http'? What Does 'Http' Do for Web Users?](#)
netforbeginners.about.com/od/h/f...tps.htm Words: **internet protocol web transfer protocols**
What Does 'Http' Do for Web Users. About.com About Tech Internet Basics. Glossary of Internet

Query results on Yioop with Question Answering System

Answer: a protocol for secur commun
Answer: a stateless protocol
Answer: a stateless protocol
Answer: the languag us on the web
Answer: an exampl of a stateless protocol layer
Answer: a stateless protocol
Answer: a stateless protocol
Answer: rpc

 what is http

1.98260 seconds. Showing 1 - 8 of 8

[HTTPS - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/HTTPS Words: **certificate server certificates user**
is a protocol for secure communication ... network which is widely used on the Internet. In its ... on the internet, HTTPS provides authentication of the website
[Cached](#) [Similar](#) [Inlinks](#) [IP:198.35.26.96](#) Score:9.84

[Hypertext Transfer Protocol - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Hypertext...Protocol Words: **methods web protocol rfc server**
page. RFC 7230 , HTTP/1.1: Message Syntax and Routing RFC ... 7231 , HTTP/1.1: Semantics and Content RFC 7232 , HTTP/1.1: Conditional Requests RFC 7233 , HTTP/1.1: Range Requests RFC 7234 ,
[Cached](#) [Similar](#) [Inlinks](#) [IP:198.35.26.96](#) Score:9.84

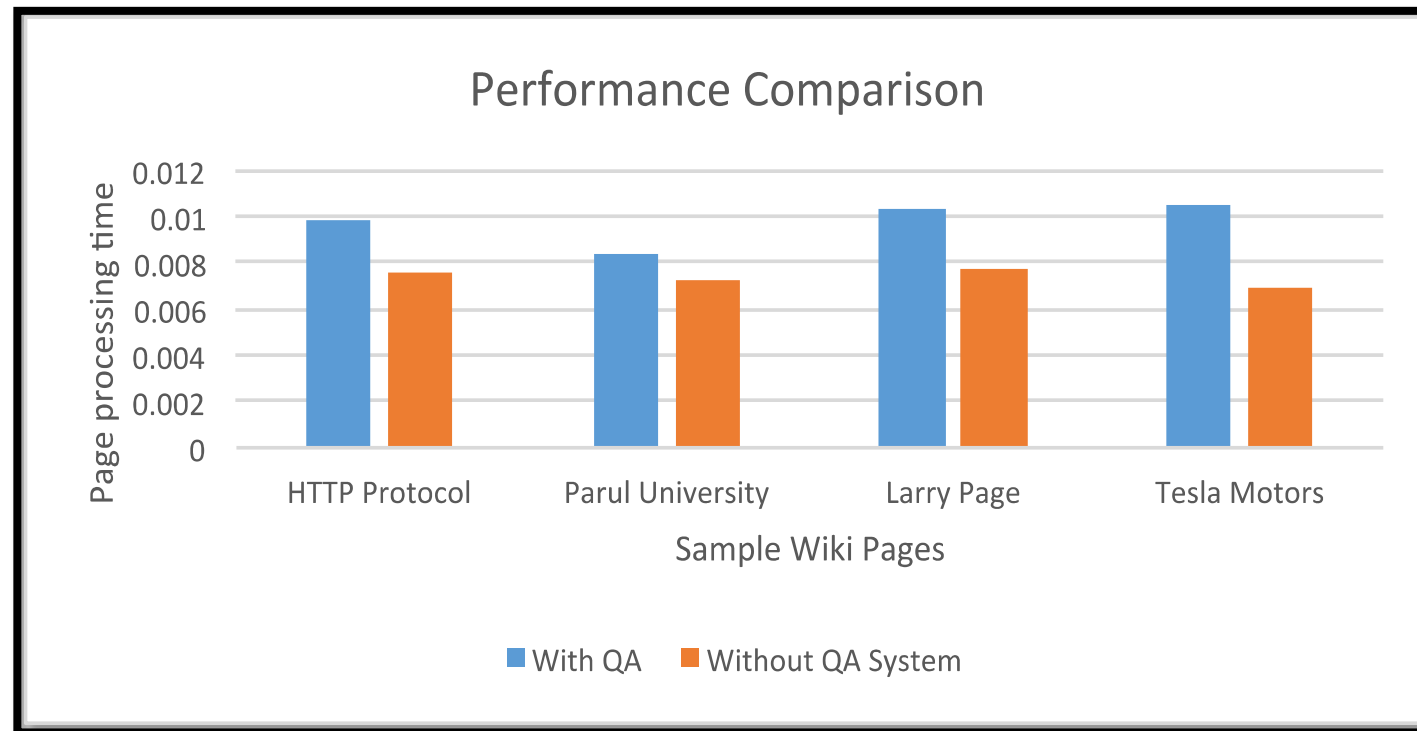
[Hypertext Transfer Protocol - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/HyperText...Protocol Words: **methods web protocol rfc server**
Status codes 7 Persistent connections 8 HTTP session state 9 ... page. RFC 7230 , HTTP/1.1: Message Syntax and Routing RFC ... 7231 , HTTP/1.1: Semantics and Content RFC 7232 , HTTP/1.1: Conditional Requests RFC 7233 , HTTP/1.1: Range Requests RFC 7234 ,
[Cached](#) [Similar](#) [Inlinks](#) [IP:198.35.26.96](#) Score:9.75

[Internet - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Internet Words: **internet web networks early protocol**
Shopping Television Voice over IP World Wide Web search ... The Internet is the global system of interconnected ... IP networks, is a proper noun. The designers of early
[Cached](#) [Similar](#) [Inlinks](#) [IP:198.35.26.96](#) Score:9.68

[Stateless protocol - Wikipedia, the free encyclopedia](#)

Observation on performance

- Trade off between space and time
 - [Subject – Predicate – Object] -> “Offset information in Paragraph”
 - [Subject – Predicate – Object] -> “Answer”



Areas of Improvement

- Overall accuracy of Question Answering System depends on the accuracy of individual components
- Junk triplet removal strategy
- User feedback
- Limitation on the Source of Information for Question Answering System

Conclusion

- Question Answering System will help Yioop in answering the questions asked by the user
- Performance lag only at crawl time
- Integration with Web Interface helps the user/developer to see the triplets generated for individual web pages

Questions?