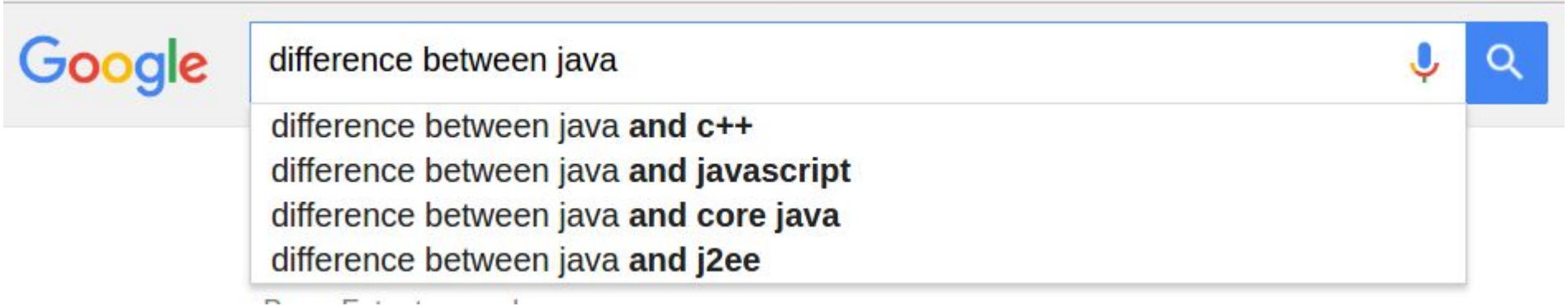
Difflet : Difference Aggregator

Guide: Prof. Radhika Mamidi
Semester Project: Monsoon ‘16 (CSE703)
IIIT-Hyderabad

Utsav Chokshi	201505581
Harshendra Avabratha	201505520
Md Tareque Khan	201505521

The problem

Very often we do searches on the Internet like



Existing Solution

The top result from google - A blog result

Question: What are the differences between C++ and Java.

Answer:

Java doesnot support pointers. Pointers are tricky to use and troublesome.

Java does not support multiple inheritances because it causes more problems than it solves. Instead Java supports multiple interface inheritance, which allows an object to inherit many method signatures from different interfaces with the condition that the inheriting object must implement those inherited methods. The multiple interface inheritance also allows an object to behave polymorphically on those methods.

Java does not include structures or unions.

Java does not support destructors but adds a finalize() method. Finalize methods are invoked by the garbage collector prior to reclaiming the memory occupied by the object, which has the finalize() method. This means you do not know when the objects are going to be finalized. Avoid using finalize() method to release non-memory resources like file handles, sockets, database connections etc because Java has only a finite number of these resources and you do not know when the garbage collection is going to kick in to release these resources through the finalize() method.

Some of the other sources can be quora, stackoverflow.

Issues with the present solution

- There is no single place
- No uniform presentation
- Takes some effort
- Content is static

Difflet

An alternative search engine to find difference between two entities.

Features

- Show point-by-point difference between two related entities.
- Dynamic content generation.
- Results in seconds.
- Shows images and video results as well.

The Wikipedia :

free encyclopedia

- Semi-structured Data
- Plethora of information in just the infoboxes
- Categorised pages

Finding similarity

Why to find Similarity ?

- To check whether two entities belong to same category or not.
- Ex : Idli vs. India (No meaning of finding difference between such entities)

Approaches Explored

- Extracting category from Wikipedia Page
  - Not always meaningful !
  - Ex : India => BRICS Nations , South Asian Countries
- Building hypernym hierarchy and finding common ancestor
  - Used NLTK library
    - Leacock-Chodorow (LC) Similarity (based on shortest path)
    - Wu-Palmer Similarity (lowest common ancestor approach)
  - Did not work for Locations like countries

Finding similarity (contd.)

Conclusion and Final Approach

- An efficient solution is to use Word2Vec to find the similarity between two words.
- Implementing/Integrating Word2Vec is big challenge in itself.
- Final approach : Have category tagged for major/popular entities.

Plethora of information in Infoboxes

Independence from United Kingdom	
• Dominion	15 August 1947
• Republic	26 January 1950
Area	
• Total	3,287,263[14] km²[b] (7th) 1,269,346 sq mi
• Water (%)	9.6
Population	
• 2016 estimate	1,293,057,000[15] (2nd)
• 2011 census	1,210,854,977[16][17] (2nd)
• Density	389.7/km² (31st) 1,009.2/sq mi
GDP (PPP)	2016 estimate
• Total	$8.727 trillion[18] (3rd)
• Per capita	$6,664[18] (122nd)
GDP (nominal)	2016 estimate
• Total	$2.384 trillion[18] (7th)
• Per capita	$1,820[18] (141st)
Gini (2009)	33.9[19] medium · 79th
HDI (2014)	▲ 0.609[20] medium · 130th
Currency	Indian rupee (₹) (INR)
Time zone	IST (UTC+05:30) *DST is not observed*
Date format	dd-mm-yyyy
Drives on the	left
Calling code	+91
ISO 3166 code	IN
Internet TLD	.in
other TLDs	[show]

INFOBOX at WikiPage about **INDIA**

INFOBOX at WikiPage about **PAKISTAN**

Independence from the United Kingdom	
• Conception[12]	29 December 1930
• Declaration	28 January 1933
• Resolution	23 March 1940
• Dominion	14 August 1947
• Islamic Republic	23 March 1956
Area	
• Total	881,913 km²[a] (36th) 340,509 sq mi
• Water (%)	3.1
Population	
• 2016 estimate	201,995,540[11] (6th)
• Density	260.8/km² (55th) 675.6/sq mi
GDP (PPP)	2016 estimate
• Total	$984.205 billion[14] (26th)
• Per capita	$5,084[14] (136th)
GDP (nominal)	2015 estimate
• Total	$270.961 billion[14] (42nd)
• Per capita	$1,427.08[14] (153rd)
Gini (2008)	30.0[15] medium
HDI (2014)	— 0.538[16] low · 147th
Currency	Pakistani rupee (Rs) (PKR)
Time zone	PST (UTC+5[b])
Drives on the	left[17]
Calling code	+92
ISO 3166 code	PK
Internet TLD	.pk
Website	

Data Extraction : Wptools

1. Python module that retrieves a given page from wikipedia
2. Extracts metadata.
3. Extracts infoboxes, images

Cons:

1. Data extracted is in MediaWiki Format. So it needs to be parsed.
2. No proper mediawiki parser available online.
3. Regex solution becomes unmaintainable

Data Extraction : DBpedia

- DBPedia provides Structured Data extracted from Wikipedia in form of RDF Triples
- RDF : Resource Description Framework
- Example of RDF Triples for “Mike Smith Knows John Doe”

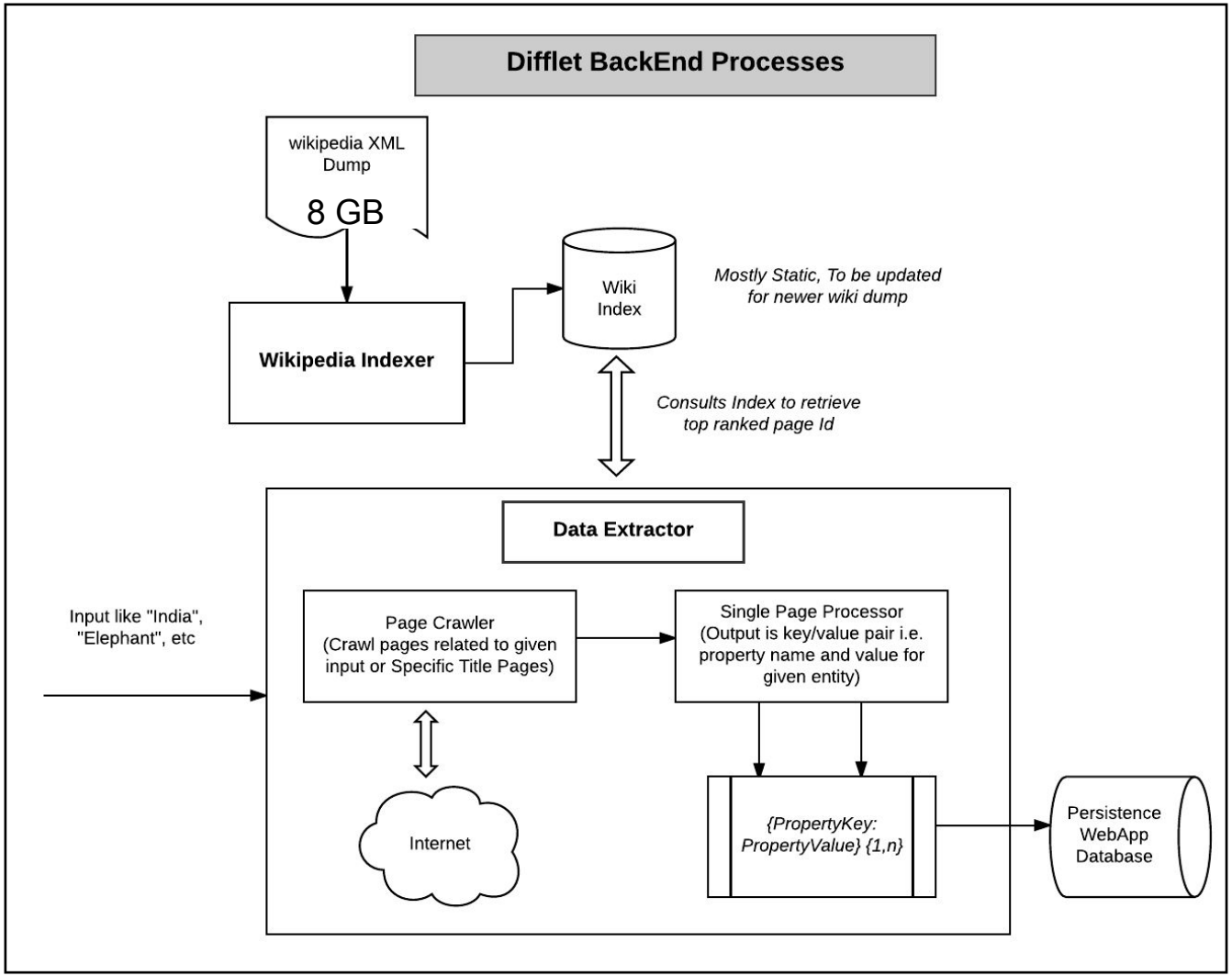
```
uri://people#MikeSmith12 http://xmlns.com/foaf/0.1/knows uri://people#JohnDoe45
```

- Another example : “India has population 1293057000”
- Parsed RDF Triples on the fly and generated Property-Value pairs to store it in database.
- Above processing is done only, when first time query is fired.

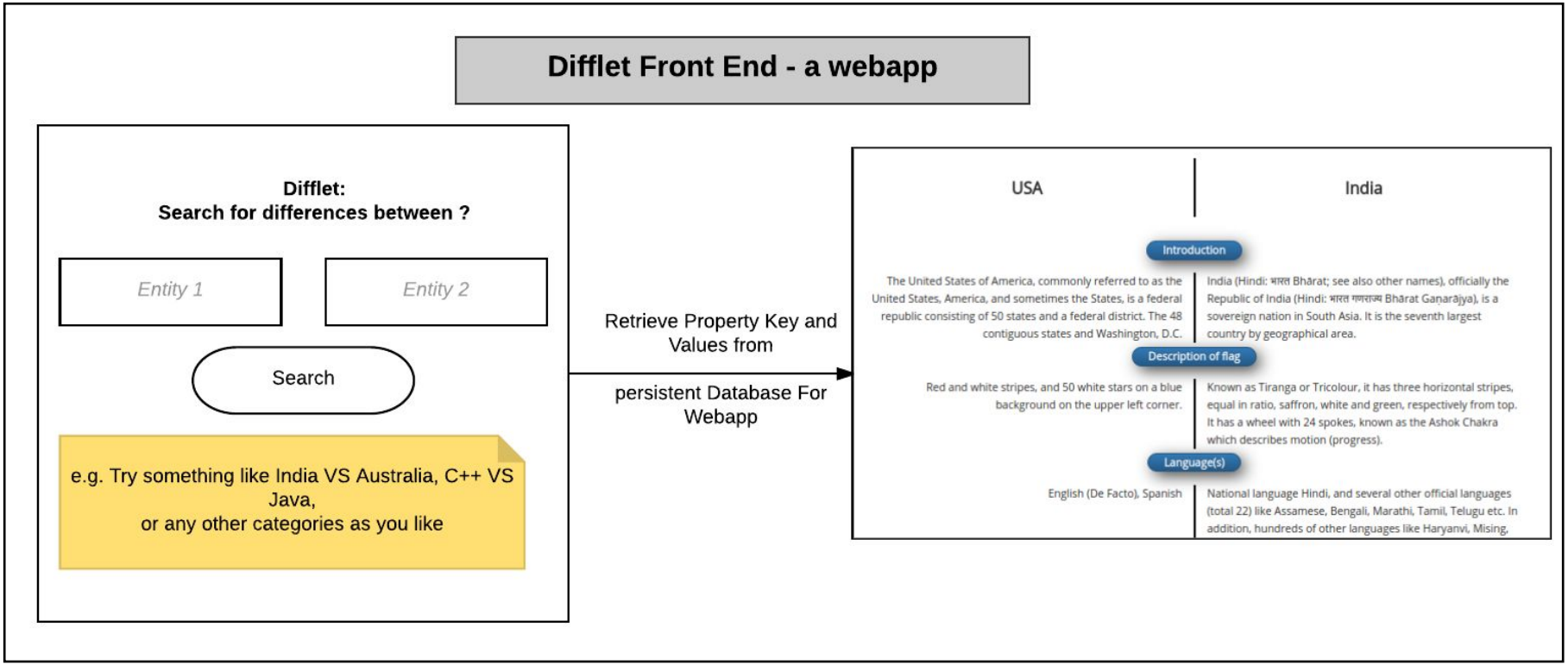
Final Solution !!!

Mixture of data from
DBpedia
+
WPtools Parsing

Architecture: Back-end



Architecture: Front-end



Front End

Web2py:

Model : mysql db schema

View : display the html

Controller : contact back-end, get data, give database uptodate, get image and video

Other Technologies & interactions

- **html5/css3/js for UI beautification**
- **Social sharing**
- **Recent and Popular difflets**

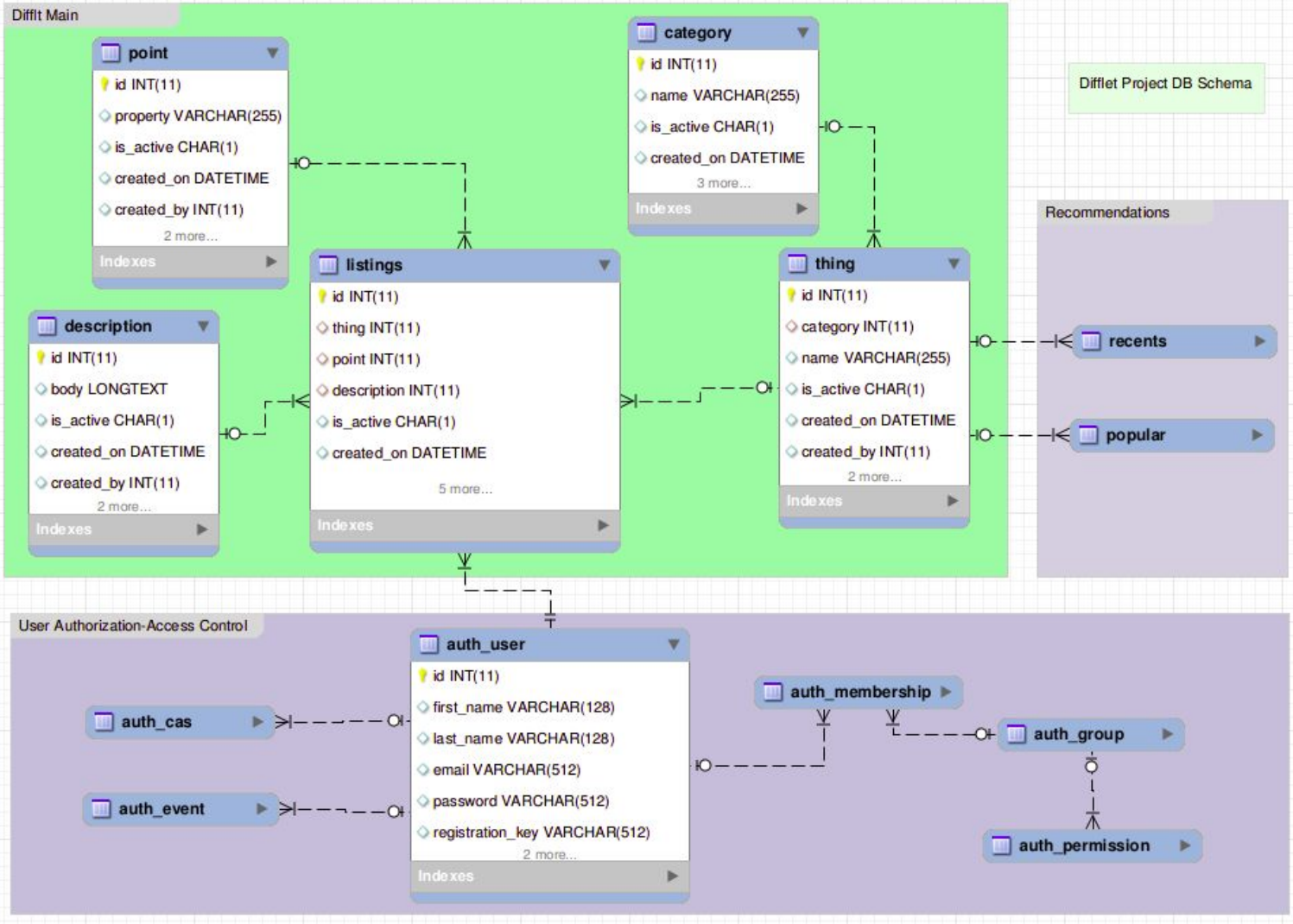
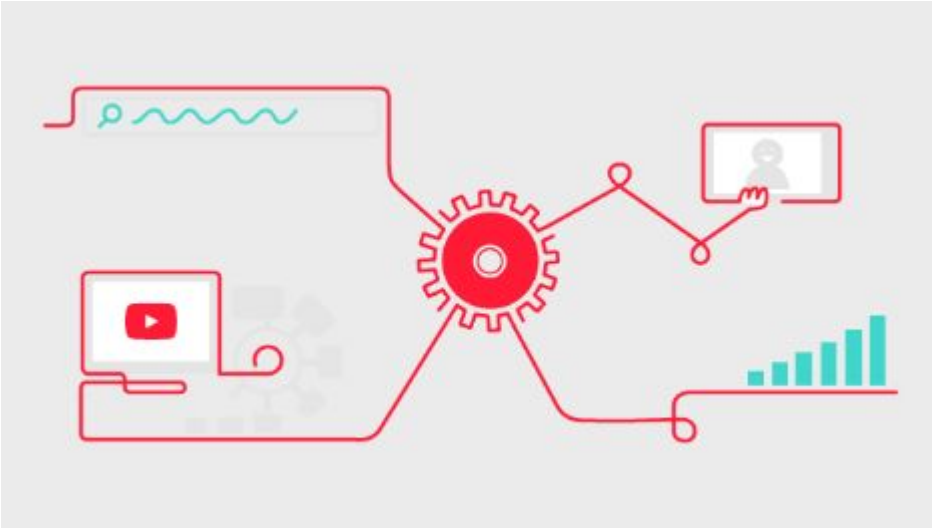

Diffit Main
point
id INT(11)
property VARCHAR(255)
is_active CHAR(1)
created_on DATETIME
created_by INT(11)
2 more...
Indexes
category
id INT(11)
name VARCHAR(255)
is_active CHAR(1)
created_on DATETIME
3 more...
Indexes
Diffiet Project DB Schema
Recommendations
listings
id INT(11)
thing INT(11)
point INT(11)
description INT(11)
is_active CHAR(1)
created_on DATETIME
5 more...
Indexes
thing
id INT(11)
category INT(11)
name VARCHAR(255)
is_active CHAR(1)
created_on DATETIME
created_by INT(11)
2 more...
Indexes
description
id INT(11)
body LONGTEXT
is_active CHAR(1)
created_on DATETIME
created_by INT(11)
2 more...
Indexes
recents
popular
User Authorization-Access Control
auth_user
id INT(11)
first_name VARCHAR(128)
last_name VARCHAR(128)
email VARCHAR(512)
password VARCHAR(512)
registration_key VARCHAR(512)
2 more...
Indexes
auth_cas
auth_event
auth_membership
auth_group
auth_permission

Image & Video

Image taken from Wikipedia via wptools

Videos taken from Youtube Data API



	Ganges	Nile
Image		
City	Rishikesh,Haridwar,Farrukhabad,Kanpur,Jajmau,Allahabad,Mirzapur,Varanasi,Ghazipur,Buxar,Ballia,Patna,Hajipur,Munger,Bhagalpur	Jinja, Uganda,Juba,Khartoum,Cairo
Mouth Country	Bangladesh	Egypt
Mouth Location	Bay of Bengal	Mediterranean Sea
Source Elevation	3892	2700
Name	Ganges	Nile

Sample Output

Further Scope

1. Get similarity and identify whether the two things are differentiable or not at the time user input the query

2. Make difflets socially editable and customizable.

Deployment

https://mdtareque.pythonanywhere.com/difflet/

Partially deployed as youtube sdk not installed on pythonanywhere and disk-space of 100 MB available.

Source Code: https://github.com/mdtareque/difflet

Demo

Q & A

Thanks