# Track 4 Design Document

Version: 5/18/2017
Student Name: Xin Xie

The Theme: Image Search

This document is designed for cs 503 BitTiger, Starting from Spring 2017/4/25

## 1. Introduction

Purpose of this document is a design document for  Capstone CS503，it was realized prototype implement of and Machine learning for searching image by image.

## 2. Architecture Design

User choose image from web (url) or upload it from local, sytem will present images matching their search.

2.1 User customize the photo interface

2.2  Image recognize

2.3 Image search

2.4 Getting informaition

## 3. System Design

The system design chart(image 1)

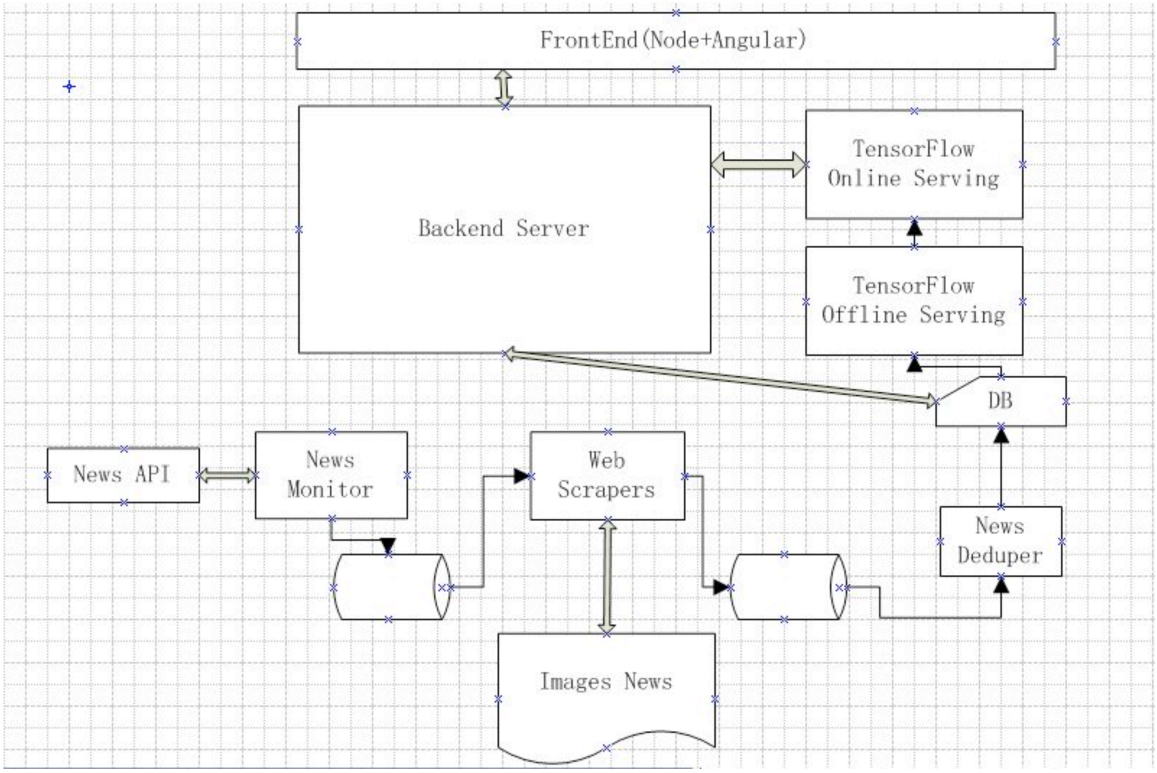


*image 1*

The design specifies how the date of the software is going to be stored.
Front-end: AngularJS, Nodejs, Boostrap
Back-end: Python
Database: Mongodb, Redis
Others: CloudAMQP, Xpath,

## 4. Scraping and Parse

4.1 Scraper basic flow:

4.1.1 Request Web server to retrieve HTML content → request
4.1.2 Parse the HTML into structured data →lxml
4.1.3 Use XPath and Regex to extract useful information →lxml,re
4.1.4 Store the information →pymongo

4.2 Parse:

The function of the parser is to extract the text and images found in the document. The text from the following locations in the document is extracted and added to the database table KEYWORDS. It also stores the location and image to which the word is tied. All the commonly used words such as *if, on, and the* which are called as *stopwords* are eliminated.

4.2.1 Image file name: The file name of the image often gives a clue to what the image contains. All the path names are extracted and stored. For example if the image is in images/faculty/name.jpg, then images, faculty and name are separately stored as keywords for file name.

4.2.2 Image ALT text: The image ALT tag in HTML usually is explicitly designed to be the textual alternative to the image.

4.2.3 Page Title: Since images are used for enhancing the Web page's content and the title is usually designed to provide a high level description of the page, they are semantically related.

4.2.4 Image caption: words that are found surrounding an image in the HTML document is extracted as caption for the image. It is usually in the same paragraph that the image is embedded in. The number of words that we want to associate with an image that surround the text is configurable.

4.2.5 Anchor Text: Terms found in the anchor tags for referenced images are relevant.

The images themselves are resized to a thumbnail version using ImageMagick and stored in the database. This saves disk space. In order to make sure that small gif images are not resized to a much bigger size, thus distorting the images, a check of file attributes such as size in bytes, height and width in pixels is done before making a thumbnail version. Since we store the location of the page from which an image has come, it can be linked to the original page and image in the results screen. Images that appear multiple times within a page are probably functional images like clipart and may not be relevant. A count of the number of times the same image is found in a page is kept so that they can be eliminated. Finally images that are smaller than a minimum size that is configurable, for example, 1024 bytes are eliminated. This is because in all probability they are small functional images such as navigation buttons.

## 5. UI

User Interface:

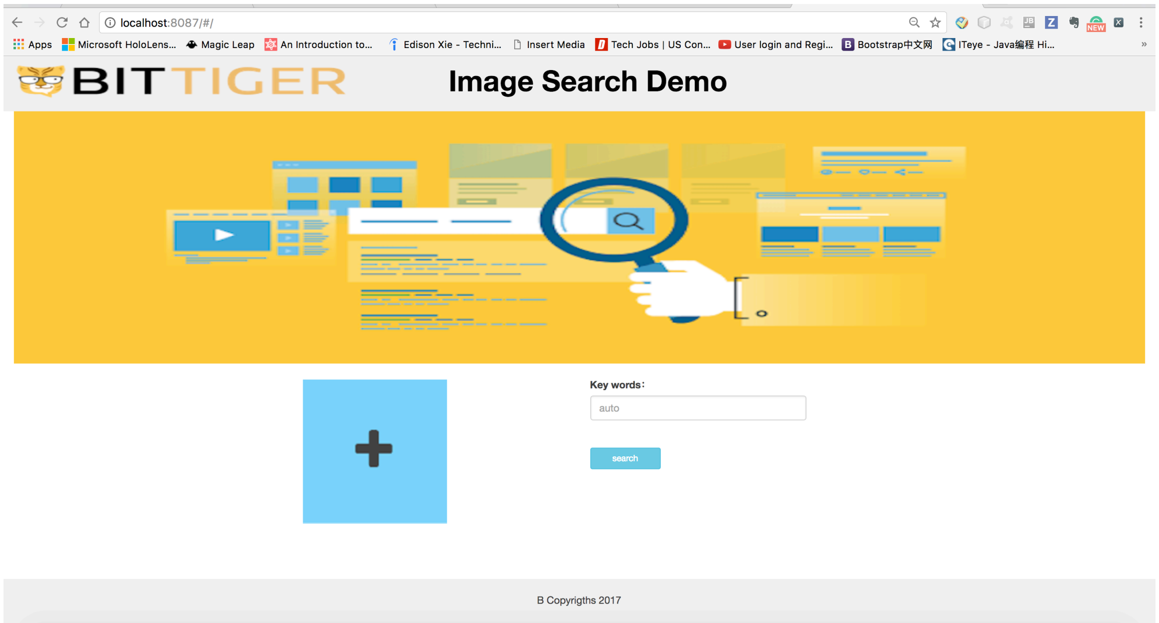


*image 2*

Responsive design:

Eg: iPhone6(image 3)

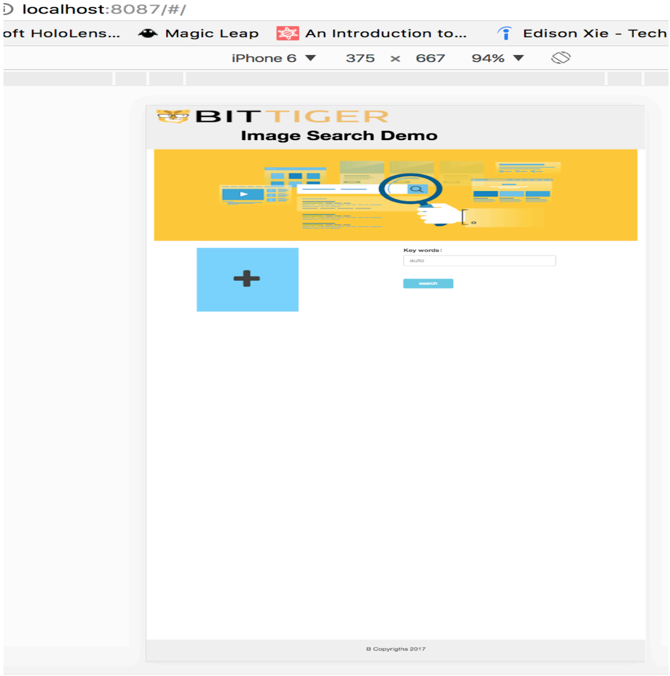
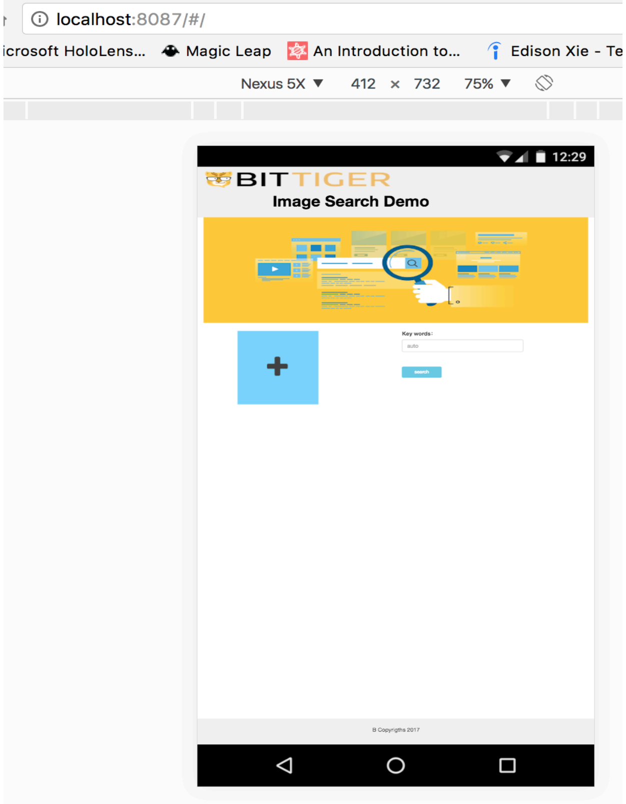


*image 3*

## Nexus 5X(image 4)



localhost:8087/#/

Nexus 5X ▼ 412 × 732 75% ▼

BITTIGER

Image Search Demo

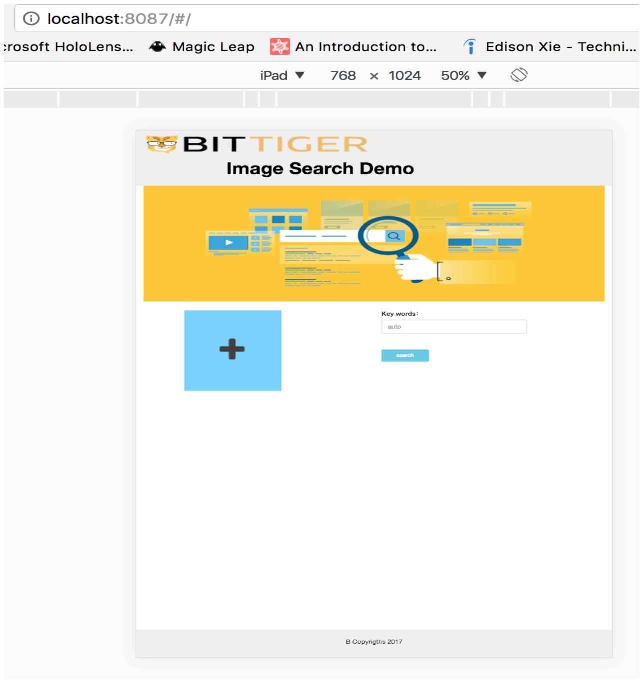
Key words:

auto

© Copyrigths 2017

*image 4*

## iPad(image 5)



localhost:8087/#/

iPad ▼ 768 × 1024 50% ▼

BITTIGER

Image Search Demo

Key words:

auto

© Copyrigths 2017

*image 5*

## 6. References

List of books, papers, URLs, tools

https://www.filestack.com/docs/google-integration

http://tineye.com/

http://www.gazopa.com/

http://similar-images.googlelabs.com/

http://www.picitup.com/

http://www.tiltomo.com/

http://labs.ideeinc.com/upload/

http://www.incogna.com

BitTiger learning materials