Automatic Categorization of Essays Using Cloud Services

WAJID HUSSAIN
01-134141-141
AHMAD RAZA
01-134141-006

Bachelor of Science in Computer Science

Supervisor: Dr. Arif Ur Rahman

Department of Computer Science
Bahria University, Islamabad

December 2017

© Wajid Hussain & Ahmad Raza, 2017

Certificate

We accept the work contained in the report titled "Automatic Categorization of Essays Using Cloud Services", written by Mr. Wajid Hussain and Mr. Ahmad Raza as a confirmation to the required standard for the partial fulfillment of the degree of Bachelor of Science in Computer Science.

Approved by . . . :

Supervisor: Dr. Arif Ur Rahman (Assistant Professor)

_____

Internal Examiner:

_____

External Examiner:

_____

Project Coordinator: Dr. Sumaira Kauser (Assistant Professor)

_____

Head of the Department: Dr. Faisal Bashir (Associate Professor)

_____

December 8^{th}, 2017

Abstract

The number of newspapers publishing news stories online has grown in the past few years. Therefore, preserving news articles for the future is required because of various reasons including cultural heritage, evidence of activities as well as scientific and historical. However, the stories may be lost because of the constant change in technologies used to present and publish. Certain individuals or institutes may be interested to collect information related to a specific topic or event. The idea came from various research papers to overcome the need by capturing the stories which are relevant to a topic. Information is provided to the system through a list of keywords. The tool automatically compares the keywords after extracting metadata from the articles and automatically identifies the category of the article and then stores on cloud services.

The tool has the functionality of automatic categorizing the texture data and saving it on cloud. User can give input as a text file, web link or direct paste it in text box. After this user may choose to open and read details about an article of their choice.

Acknowledgments

First of all, We are grateful to Almighty Allah for his blessings upon us. Who gave us the ability and courage to complete our project on time. He is the only one who we always looked at in the event of happiness and trouble and He always helped us in the time of need. With his blessing upon us we have completed our Work.

We would specially thank to our supervisor Dr. Arif Ur Rahman. Who remained the source of guidance till the end of this project. He gave us a continuous advice on the content of the report and project. He gave us too much time to guide each and every step of this project. We could not achieve the desired results without the guidance of our supervisor.

We would like to thank our parents who supported us in all our endeavors and saw successful persons in us.

Last but not the least we would like to thank all our friends whose silent support led us to complete this task and enjoy the four years stay at the university.

WAJID HUSSAIN

AHMAD RAZA
Islamabad, Pakistan

Dec 2017

"Coming together is a beginning, Keeping together is a progress,
Working together is a Success."

Henry Ford

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

ACECS	Automatic Categorization of Essays Using Cloud Services
AWS	Amazon Web Services
API	Application Programming Interface
GUI	Graphical User Interface
JSON	JavaScript Object Notation
NLP	Natural Language Processing
RAKE	Rapid Automatic Keyword Extraction
S3	Simple Storage Services
SNS	Simple Network Services

Chapter 1

Introduction

This chapter describes the introduction of project in detail. It presents the problem description, project objectives and scope.

1.1 Introduction

Cloud computing has emerged as a new paradigm of computing. It is already in use in many technologies and companies are developing more and more cloud based solutions for users [1]. It lets developers focus on solving the core problem while the supporting services are offered by the cloud service providers. Thus, supporting programmers in building better solutions. It decreases the costs by helping users choose a payment model which better suits their individual needs.

The current proposal focuses on the development of an application that gets the data from one or more sources and stores data on the cloud in a specific format which includes metadata. The data that is used to represent other data is known as metadata. There are three different types of metadata: descriptive metadata, structural metadata, and administrative metadata. The descriptive metadata of an article describes it for purposes such as discovery and identification. Information such as title, summary, author, and keywords can be included in the descriptive metadata. The structural metadata of are about data containers and indicates how composite objects are joined, for example how the pages are ordered to form chapters. It describes the types, versions, relationships, and other characteristics of digital materials. The administrative metadata of provide information to help manage a resource, such as when and how it was created, the type of file and other technical information, and who can access it. Metadata should have two parts explicit and implicit. In explicit metadata is already available with an article such author name, date of publication, source from which a story is downloaded. Implicit metadata is contained within the text of an article and needs to be extracted to make it explicit.

1.2 Project Background

Earlier the metadata data are stored manually on the machines or computers, the two main parts of metadata explicit and implicit data are store manually that have some information like author name, title of the article and table of contents etc. and analyze the data and identify its category manually which is very hectic and time taking process to identify its category.

Cloud computing is a strategy for conveying information technology(IT) benefits in which assets are recovered from the Internet through online devices and applications, rather than an immediate association with a server. As opposed to keeping records on a restrictive hard drive or nearby stockpiling gadget, cloud-based capacity makes it conceivable to spare them to a remote database. For whatever length of time that an electronic gadget approaches the web, it approaches the information and the software programs to run it. There are many cloud providers including, Amazon AWS, Google, Microsoft, IBM and Oracle.

Amazon Web Services is a far reaching, developing distributed computing platform offered by Amazon.com. Web services are in some cases called cloud services or remote processing services. AWS is a safe cloud services platform, offering process control, database storage, content delivery and other functionality to enable organizations to scale and develop. Investigate millions of customers clients are right now utilizing AWS cloud items and answers for construct advanced applications with expanded adaptability, versatility and reliability quality.

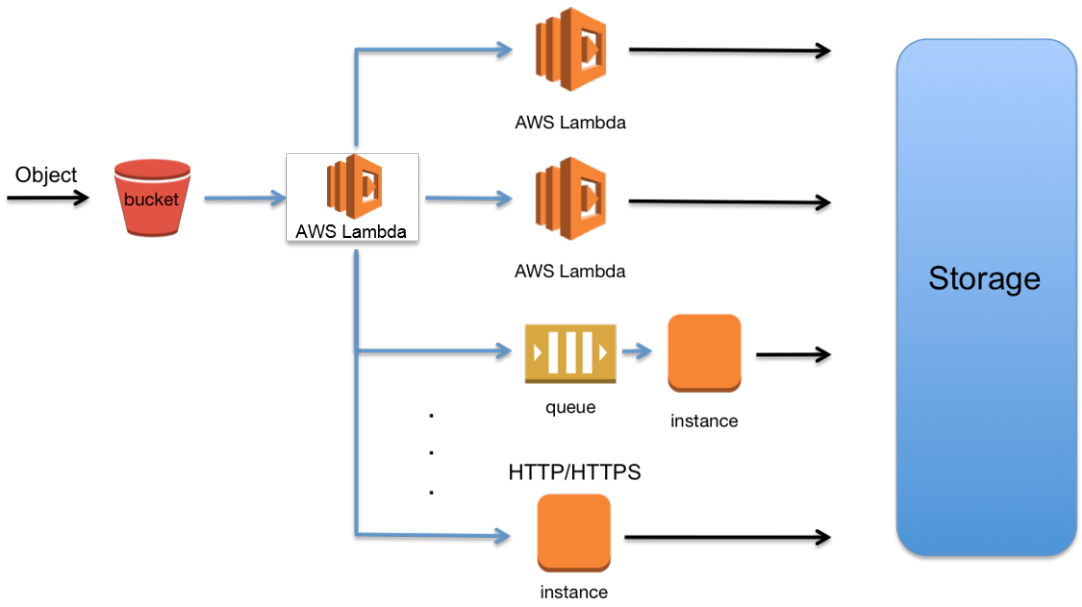


Figure 1.1: AWS-Infrastructure

The AWS Developer Tools help you safely store and form control your application's

source code and consequently assemble, test, and convey your application to AWS or your on-premises condition [2].

1.2.1 AWS CodeCommit

AWS Code Commit is a completely overseen source control benefit that makes it simple for organizations to have secure and very versatile private Git stores. CodeCommit eliminates the need to work your own particular source control framework or stress over scaling its infrastructure. You can utilize CodeCommit to safely store anything from source code to parallels, and it works flawlessly with your current Git devices.

1.2.2 AWS CodeBuild

AWS CodeBuild is a completely overseen manufacture benefit that gathers source code, runs tests, and delivers programming bundles that are prepared to convey. With CodeBuild, you don't have to arrangement, oversee, and scale your own form servers. Code Build scales ceaselessly and forms various forms simultaneously, so your manufactures are not left holding up in a line. You can begin rapidly by utilizing pre-packaged form situations, or you can make custom form conditions that utilization your own form apparatuses. With CodeBuild, you are charged incrementally for the register assets you utilize.

1.2.3 AWS CodePipeline

AWS CodePipeline is a persistent coordination and nonstop delivery service for quick and dependable application and foundation refreshes. CodePipeline assembles, tests, and sends your code each time there is a code change, in view of the discharge procedure models you characterize. This empowers you to quickly and dependably convey highlights and updates. You can without much of a stretch form out a conclusion to-end arrangement by utilizing our pre-constructed modules for well known outsider administrations like GitHub or incorporating your own custom modules into any phase of your discharge procedure. With AWS CodePipeline, you pay for what you utilize. There are no forthright expenses or long haul responsibilities.

1.2.4 AWS CodeDeploy

AWS CodeDeploy is an administration that mechanizes code organizations to any occurrence, including Amazon EC2 cases and cases running on-premises. AWS CodeDeploy makes it less demanding for you to quickly discharge new highlights, encourages you maintain a strategic distance from downtime amid application organization, and handles the many-sided quality of refreshing your applications. You can utilize AWS CodeDeploy to computerize programming organizations, disposing of the requirement for blunder

inclined manual operations, and the administration scales with your framework so you can undoubtedly send to one case or thousands.

1.3 Objective

This application design to help make the system that store the source and other data on the cloud in a specific format that include some metadata explicitly and implicitly that consist analysis component which identify its category. Automatic categorization is the main feature of this application which is done by tokenization.

1.4 Problem Description

Cloud storage is where information is remotely kept up, overseen, and went down. The administration enables the clients to store records on the web, with the goal that they can get to them from any area by means of the Internet. But when user want to store the data on cloud they don't know what is the actual category of the article which they are saving on cloud. Automatic Categorization of Essays Using Cloud Services is the software which collects the texture data from the user and allow them to store it on cloud by detecting automatically its correct category. The tool will use Amazon S3 service for storing the data on cloud and use tokenization technique to automatically categorize it.

1.5 Project Scope

Data can be in various formats including text, audio, video, images, and animations. The focus of the current scope is on Texture data. The texture data serve as some article or essays or any stories in texture form are only dealing with. Every one of the information is introduced as writings, expressions, or paragraphs. It includes listing vital attributes, emphasizing significant figures and identifying important features of data. Textual presentation information refers to information displayed in composed, paragraph form. He alternative refers to graphs or other types of visual graphs.

Chapter 2

Literature Review

This project is basically based on an implementation. The review of the literature focuses on the subject, i-e "Automatic categorization of Essays using cloud services". First, gathered research papers related to this topic, and then studied each research paper. In the research papers studied about the work done by the researchers. After reading some articles, they were only the closest articles to this subject. The main attention has been to find researchers who work on Cloud and his work. After studying this, analyze Cloud work and how the data is backed up on the cloud and what security risks are involved.

2.1 Digital Preservation

Digital preservation is a formal endeavor to ensure that digital information of continuing value remains accessible and usable. Digital content is a combination of files and meta data. Preservation metadata is a key component of digital preservation. Digital content includes images, text, sound, videos and maps etc. This requires some identifications and description captured as metadata. There are different practices for preserve digital contents. Sound preservation published by sound direction project which describes the audio preservation work-flow. For digital preservation of text and images there are different formats which are described in Table 2.1 but over focus is only on texture data available in a given document. Cloud computing can offer several benefits:

- Cloud flexibility allows emerging service providers to perform relatively fast and low-cost testing and trial. There are already some of these cloud services pilot activities and community sharing learning opportunities.
- The deployment of cloud storage services now has more flexibility and more options than previous years, and is therefore more relevant to archives (see Public, Community, Private, and Hybrid Clouds).

- Cost savings can be achieved through easier purchasing and economies of scale, especially for smaller repositories. In economic pressure, these are important.
- Cloud services can provide simple, automated replication for multiple locations that are necessary for enterprise recovery planning and professional management of digital storage access; in addition, experts can increase access to other proprietary tools, programs, work-flows, and service protocols for digital protection requirements To build.

Table 2.1: Formulas for Digital Preservation

Still-Images	Moving-Image	Sound	Texture	Web-Archive
svg	mpeg4	wave	nitf	arc
tiff	avi	mp3	xml	warc

2.2 Natural Language Processing

Natural Language Processing (NLP) is a domain of computer science which focuses on the processing of natural languages e.g. English, Urdu, Chinese. The typical tasks in NLP are tokenization, Lemmatization, named entity recognition, parts of speech identification, word and sentence boundary identification **"The Ultimate Introduction to NLP"** [3]. Tokenization is demonstration of separating an arrangement of strings into pieces, for instance, words, watchwords, expressions, pictures and different components called tokens. “NLP is a field that covers computer understanding and manipulation of human languages, and it’s ready with possibilities for news gathering. “Anthony Pesce said in NLP in the kitchen. You usually cheer about it in the context of analyzing large pools of legislation or other document sets, attempting to discover patterns or root out corruption."

There are libraries available for NLP. Some of the commonly used ones are:

- Stanford Core NLP[1] by using this tool which deal with a human natural language use to analysis it. Their parts of speech (Noun, verbs, adjective etc.). By using it we can identify the story contains the company name, person, location etc. how many stop words are using all these things are captured by this NLP tool. We eliminate those words which provide any identity of anything. Program automatically extract the noun phrase and eliminate those words which have like person name, address, location etc. will be eliminated.
- OpenNLP[2] supports the most widely recognized NLP tasks, for example, tokenization, sentence division, grammatical feature labeling, named element extraction,

[1] https://stanfordnlp.github.io/CoreNLP/
[2] https://opennlp.apache.org/

piecing, parsing, language detection and reference resolution. OpenNLP gives a charge line content, filling in as a special passage point to every single included instrument. The script is located in the bin directory of OpenNLP binary distribution. Included are versions for Windows: OpenNlp.bat and Linux or perfect frameworks: OpenNlp.

- NLTK[3] is a main stage for building Python projects to work with human language data. It gives easy to-utilize interfaces to more than 50 corpora and lexical assets, for example, WordNet, alongside a suite of content preparing libraries for arrangement, tokenization, stemming, labeling, parsing, and semantic thinking, wrappers for modern quality NLP libraries, and a dynamic dialog discussion. NLTK is appropriate for linguists, engineers, students, instructors, scientists, and industry clients alike. NLTK is accessible for Windows, Mac OS X, and Linux. The best part is that NLTK is a free, open source, group driven venture. Natural Language Processing with Python gives a functional prologue to programming for language processing. Composed by the makers of NLTK, it manages the reader through the essentials of composing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more.

- Lucene[4] is a full-text search library in Java which makes it simple to add seek search functionality to an application or site **"The apache lucene core"**[4]. It does as such by adding substance to a full-content file. It then allows you to perform queries on this index on this file, returning outcomes positioned by either the significance to the inquiry or arranged by a subjective field, for example, a record's last adjusted date. The substance you add to Lucene can be from different sources, similar to a SQL/NoSQL database, a file system, or even from sites. Lucene can accomplish quick inquiry reactions on the grounds that, rather than looking through the content specifically, it looks through a record. This would be what might as well be called recovering pages in a book related to a keyword by searching the index at the back of a book, rather than looking through the words in each page of the book.

2.3 Automatic Text Categorization

Automatic categorization of text data is the core part of this application. Twinword[5] API is used for this purpose in the development of this application. Twinword Propose related categories for each blog or article. It is Sentiment Analysis' free API returns sentiment analysis comes about with score for the given content. Since it allows to discover the

[3] http://www.nltk.org/
[4] https://lucene.apache.org/
[5] https://www.twinword.com/

tone of a client remark or post **"Automatic keyword extraction from documents based on multiple content-based measures"**[5]. Twinwords Word Associations API gets word relationship with semantic separation score. Since it expects to work with something other than just synonyms, clients can get related words of a similar family like "cats" and "dogs." This API permits to discover equivalent words and related words for single word or a phrase. It is text analysis API that can understand and relate words similarly as people do. It first applies tokenization technique on given texture data and extract keywords from it. After extraction of keywords Twinword compare them with other related articles. Then after analyzing the data it place the article in its correct category.

2.4 Amazon Web Services

Amazon Web Services (AWS) is a comprehensive, evolving cloud computing platform provided by Amazon Web services are sometimes called cloud services or remote computing services. The first AWS offerings were launched in 2006 to provide online services for websites and client-side applications **"Programming Amazon Web Services: S3, EC2, SQS, FPS, and SimpleDB "** [2].

- AWS offers a huge range of services to suits your application requirements.These database services are fully managed, just a few minutes to start in a few minutes. AWS database services include Amazon RDS, support for six popular database engines, Amazon Aurora, MySQL and PostgreSQL-compliant data, Amazon Dynamo Db, NoSQL database services fast and flexible, Amazon's Redshift data warehouse service and Amazon Elasticache , Memory cache service, support for Memcached and Redis. AWS also provides the AWS database migration service, which makes it easy and cheap to migrate your database to the AWS cloud **"Programming Amazon Web Services: S3, EC2, SQS, FPS, and SimpleDB "** [2].

- Amazon Web Services (AWS) is a secure cloud services platform that offers computing power, database storage, content delivery, and other features to help businesses scale and grow

2.4.1 Amazon S3

Amazon Simple Storage Service is storage for the Internet. It is designed to make web-scale computing easier for developers. Amazon S3 has a simple web services interface that you can use to store and retrieve any amount of data, at any time, from anywhere on the web. It gives any developer access to the same highly scalable, reliable, fast, inexpensive data storage infrastructure that Amazon uses to run its own global network of web sites. The service aims to maximize benefits of scale and to pass those benefits on to developers

"Programming Amazon Web Services: S3, EC2, SQS, FPS, and SimpleDB " [2]. Companies today need the ability to simply and securely collect, store, and analyze their data at a massive scale. Amazon S3 is object storage built to store and retrieve any amount of data from anywhere – web sites and mobile apps, corporate applications, and data from Internet of things (IoT) sensors or devices. It is designed to deliver one percent durability, and stores data for millions of applications used by market leaders in every industry. S3 provides comprehensive security and compliance capabilities that meet even the most stringent regulatory requirements. It gives customers flexibility in the way they manage data for cost optimization, access control, and compliance. S3 is the only cloud storage solution with query-in-place functionality, allowing you to run powerful analytic directly on your data at rest in S3.

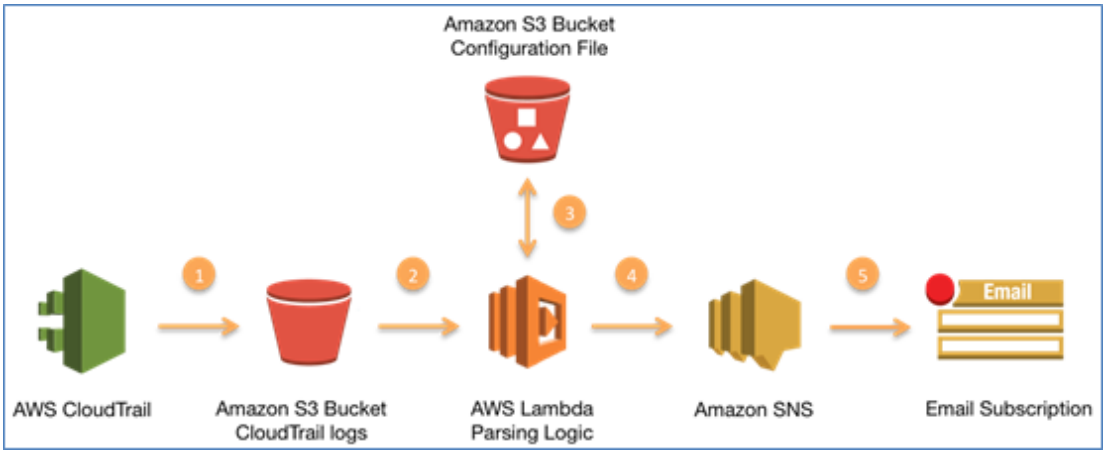


Figure 2.1: Amazon S3

2.4.2 Amazon S3 Bucket and logs

Amazon S3 integrates with Cloud Trail, which captures specific Amazon S3 API calls from its AWS account and passes the log file to your designated Amazon S3 bay. Cloud Trail captures API calls made by Amazon S3 or Amazon S3 APIs.

2.4.3 AWS Lambda

AWS Lambda starts the code only when it's needed and automatically scales from several requests per day to thousands per second. You pay only for the calculated time of consumption - there is no charge when the code is not running. With AWS Lambda, you can run code for virtually any type of application or backend - all at zero administration. AWS Lambda runs its code in a high-availability computing infrastructure and performs all administration of computing resources, including server and operating system maintenance,

capacity and automatic scaling, code monitoring, and logging. Just enter the code in one of the languages supported by AWS Lambda (currently Node.js, Java, C sharp and Python).

2.4.4 AWS SNS

AWS SNS is a Web service that is used to coordinate and manage the sending or sending of messages to or from a registered user. In Amazon SNS, there are two types of clients - publishers and subscribers - also known as producers and consumers. The editor communicates with the user asynchronously by generating and sending a message to the user, which is a logical access point and a communication channel. Subscribers (ie Web server, email address, Amazon SQS queue, AWS Lambda function) consume or receive messages or notifications on one of the supported protocols (c'ie Amazon SQS, HTTP / S, email, SMS, Lambda) To the subject.

2.5 Developer's Work

After reading these research papers, it was finally decided to do this semester project. Automatically classify the tests and divide the metadata into explicit and implicit. We developed this application in the JAVA programming language using Eclipse IDE.

Chapter 3

Requirement Specifications

This chapter presents a review of existing systems and establishes requirements specifications for the proposed tool.

3.1 Existing System

An open format is established by Amazon for storing data on cloud known as Amazon S3 service. Amazon S3 is cloud storage for the Internet. To upload your information (photographs, recordings, reports and so forth.), you initially make a bucket in one of the AWS Regions. You would then be able to transfer any number of items to the bucket. As far as execution, buckets and objects are assets, and Amazon S3 gives APIs to you to manage them. For instance, you can make a bucket and upload objects utilizing the Amazon S3 API. You can likewise utilize the Amazon S3 console to perform these operations. The console inside utilization the Amazon S3 APIs to send requests to Amazon S3. Amazon S3 bucket names are all around special, regardless of the AWS Region in which you make the bucket. You determine the name at the time you make the bucket. AWS Lambda is a compute service that gives you a chance to run code without provisioning or managing servers. AWS Lambda executes your code just when required and scales naturally, from a couple of requests for every day to thousands every second.

3.2 Proposed System

The main purpose of this project is to store the data on cloud by creating a bucket using S3 service of AWS and automatically categories the data stored in bucket after splitting the metadata in explicit and implicit parts. In this project we use these following technologies:

- JAVA

- Eclipse IDE
- AWS Dynamodb
- AWS S3
- AWS Lambda

with following key feature:

1. Accurate result
2. User friendly

To store and automatic categories the text file use JAVA application to create a form in which user select specific file from the personnel system or can cut paste the text in the text bar. This project is basically focus on automatically categories the essays and their storing on cloud. From users point of view, if they don't want to read the complete article or essay to define its category than they can use this application for automatic categorization.

3.3 Product Function

- User can browse by clicking on browse button and can select the file from personal system.
- User can cut paste the text in the text file wants to automatic categories.

3.3.1 Requirement Specification

- User specification of the system which can have a contact with system directly or indirectly are identified below.
- Developer who have complete control of all aspects of his application. Any time he makes changes in application like update or change visual interaction.
- Technical user who have full command on databases and cloud computing and uses of other programmed applications. This is naive user of application.
- Non-Technical user are end users who don't know anything about cloud computing and databases and only wants to automatic categories the articles and their storing on cloud.

3.3.2 System Requirement

1. **Functional Requirements**
   (a) Support large text files.
   (b) Select the file from personal system or cut paste text in text bar.
   (c) Only for text files (only focus on textual data).
   (d) Connected with cloud.
2. **Non-Functional Requirements**
   (a) **Usability: -** Automatic categorize the articles and store it on cloud efficiently.
   (b) **Performance: -** This tool automatic categorize the article and store it on cloud in its correct category in less than 5 second.
   (c) **Capacity: -** Application have a read, write capacity and can store almost 1000 articles.
   (d) **Operations: -** By using this software user can view different articles save on cloud as well as automatic categorize them.
   (e) **Security: -** Articles which user automatic categorize and store on cloud by using this tool cannot be view by any other person.
   (f) **Attractive layout: -** Application layout is self-explanatory user can understand it easily.

3.4 User Scenarios

Deliver a summary of the main purposes that the system will execute. Establish the purposes to be comprehensible to users. User scenario of this application:

- The user who want to use this application have internet connection.
- User install ACECS application.
- First user agree the licensee and key terms ACECS of application.
- After agreed user will able to install ACECS.
- After this user run the application and select specific file from directory or by name.
- After chose the file ACECS automatic categories the article or essay and store it on cloud.
- When a user wants to close the application the message will display "Do you really want to close the application?" Yes or No

3.5 Use Cases

3.5.1 Use Case 1

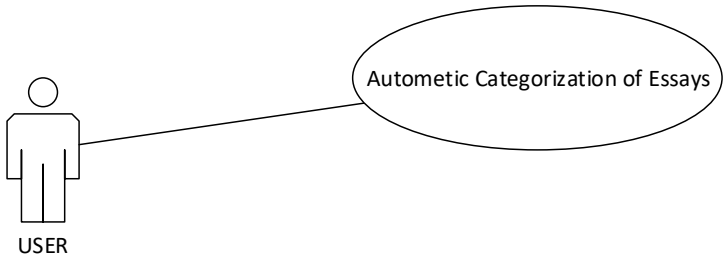


Figure 3.1: Use Case 1

User can easily automatic categorize the essays by applying one of following three methods in figure 3.1.

- By giving the link of website.
- By choosing the file containing texture data.
- Direct paste the texture data in the textbox. After taking the input system extract the keywords and match them with other related articles. After this application place the article in its correct category.

3.5.2 Use Case 2

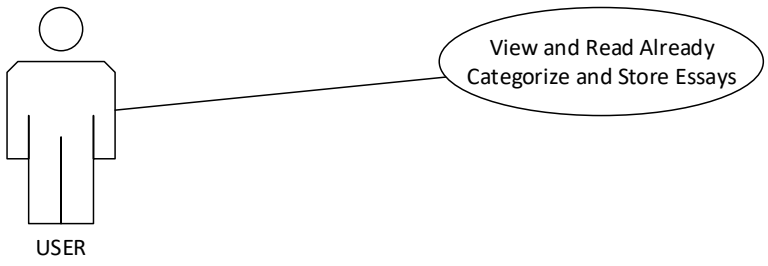


Figure 3.2: Use Case 2

Use Case 2 as presented in figure 3.2 shows that user can easily access the articles from the home page of application after storing articles on cloud. Users can read the articles which are stored in different categories. This function gives easy accessibility of data.

Chapter 4

Design

This chapter presents the system design and the functionality of the software developed. Moreover, the type of data which will be used as input, the various modules and graphical user interface (GUI) of the system are presented.

4.1 System Architecture

The Architecture of this system is interactive and simple. It provide user friendly interface which contains a different websites links user can select the different links by using web crawling. The high level diagram of this system as shown in the 4.3 and the Flow diagram of the given system as shown in 4.4. In which it display how the system performs its task step by step.

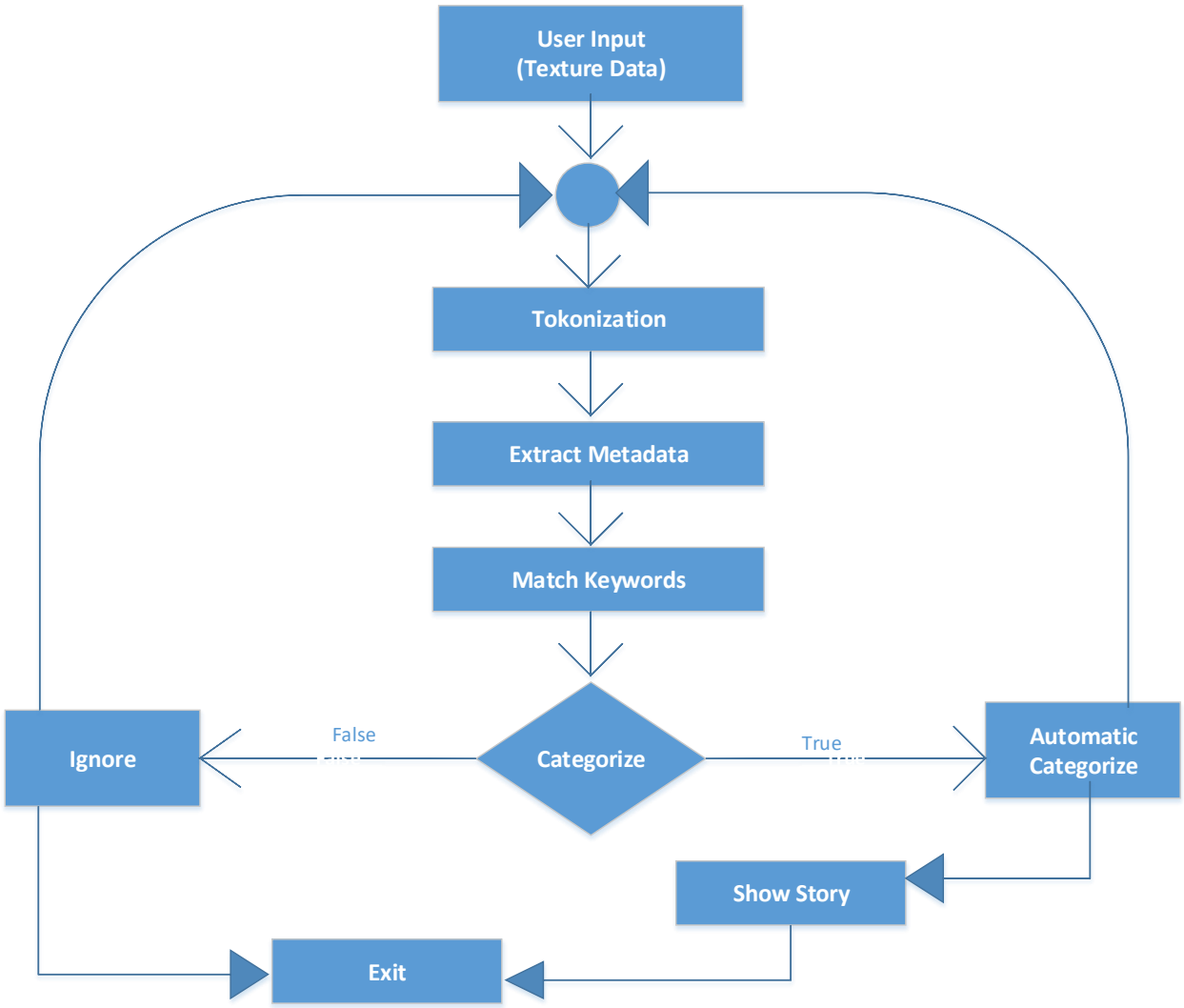
4.2 Design Constraints

1. Internet facility must be available.
2. Targeted users are educated enough to operate a computer and understand the system as well as the application to be perform.

4.3 Design Methodology

Using incremental model [6],The incremental process model is a process of software development where the product is designed, implemented and tested incrementally. At each increment a new and little more things or feedback is added until the product is to be develop. It includes both maintenance and development. When all the requirement is to be done then product is said to be finished. This model comprises the elements of both the

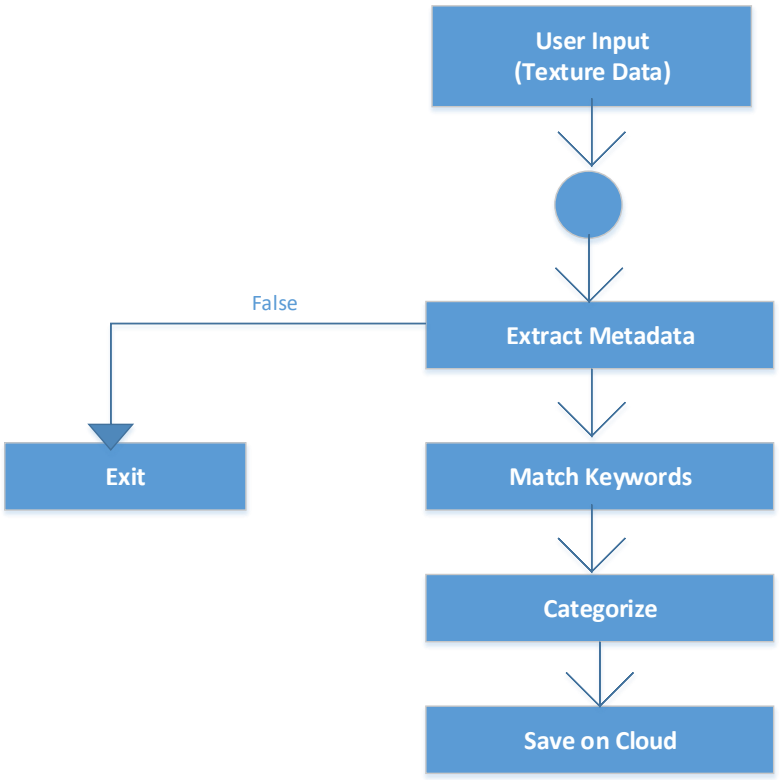
Figure 4.1: System Flow Diagram



waterfall model and Iterative idea of prototyping. Incremental model was used to develop the software. The goals of the project were achieved in the following four phases.

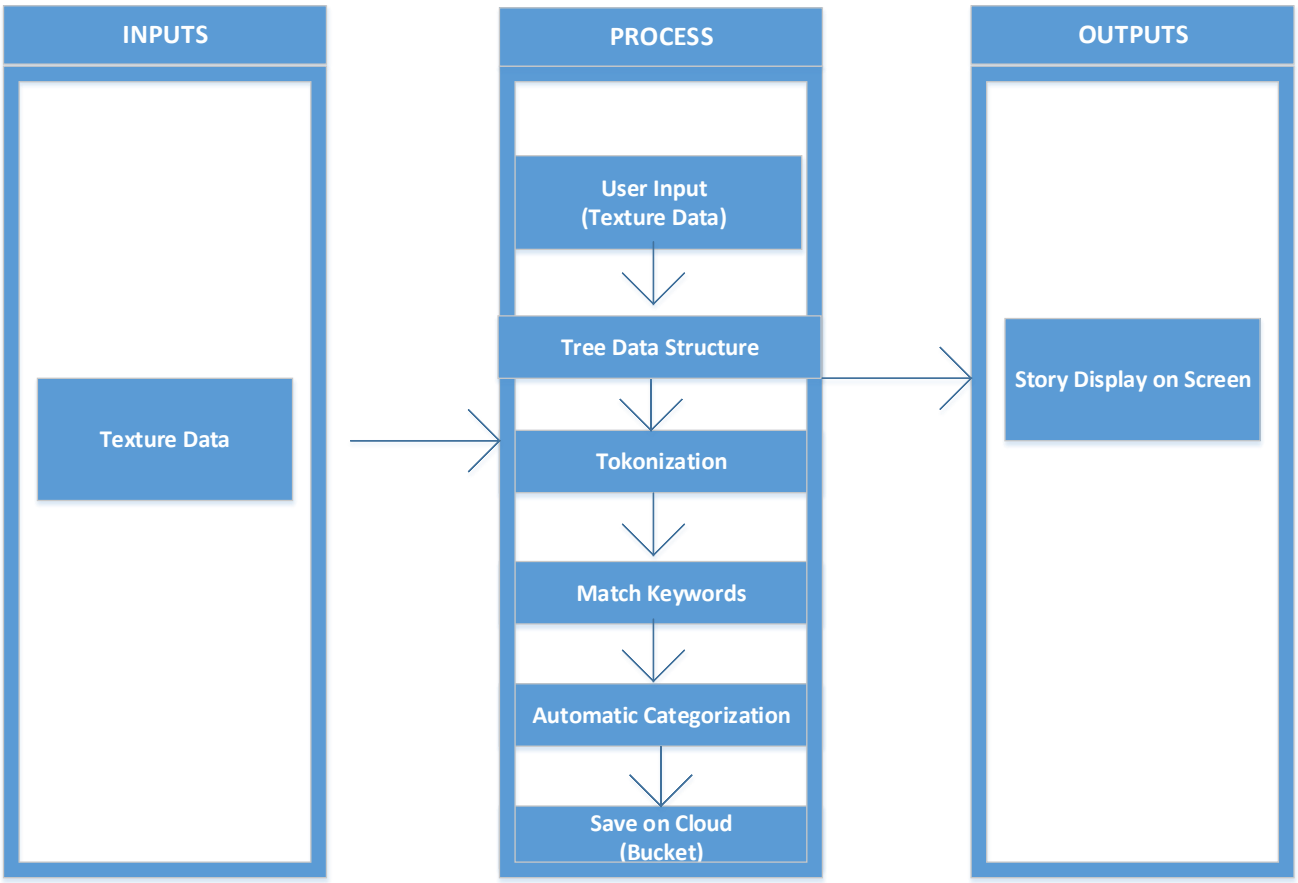
- Take texture data as an input from textbox or file.
- Extract Meta data from the given data for example publish date, publisher name, title of article etc.
- Apply tokenization technique and match keywords.
- Automatic categorize the texture data. The flow diagram of automatic categorization is in figure 4.2.

Figure 4.2: System Flow Diagram of Categorization



4.4 High Level Design

Figure 4.3: High level design



High level diagram which shown in 4.3. In which user will input the link where he/she want to extract the story. Then after selection it will move to the processing part where it process the link accordingly. Then after being process it will display the output score result of comparison.

4.5 Low Level Design

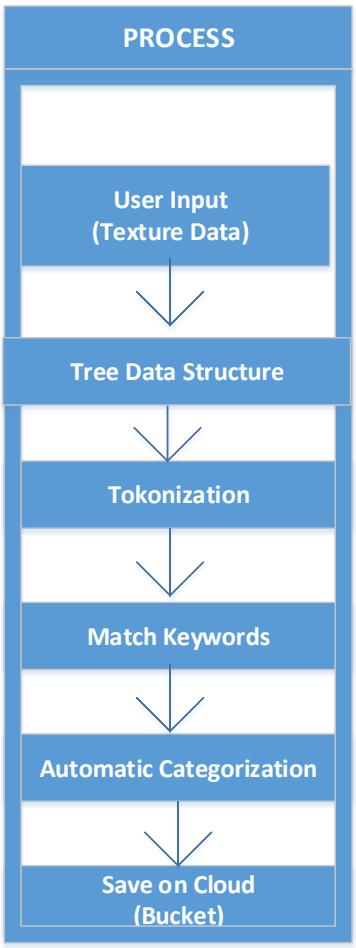
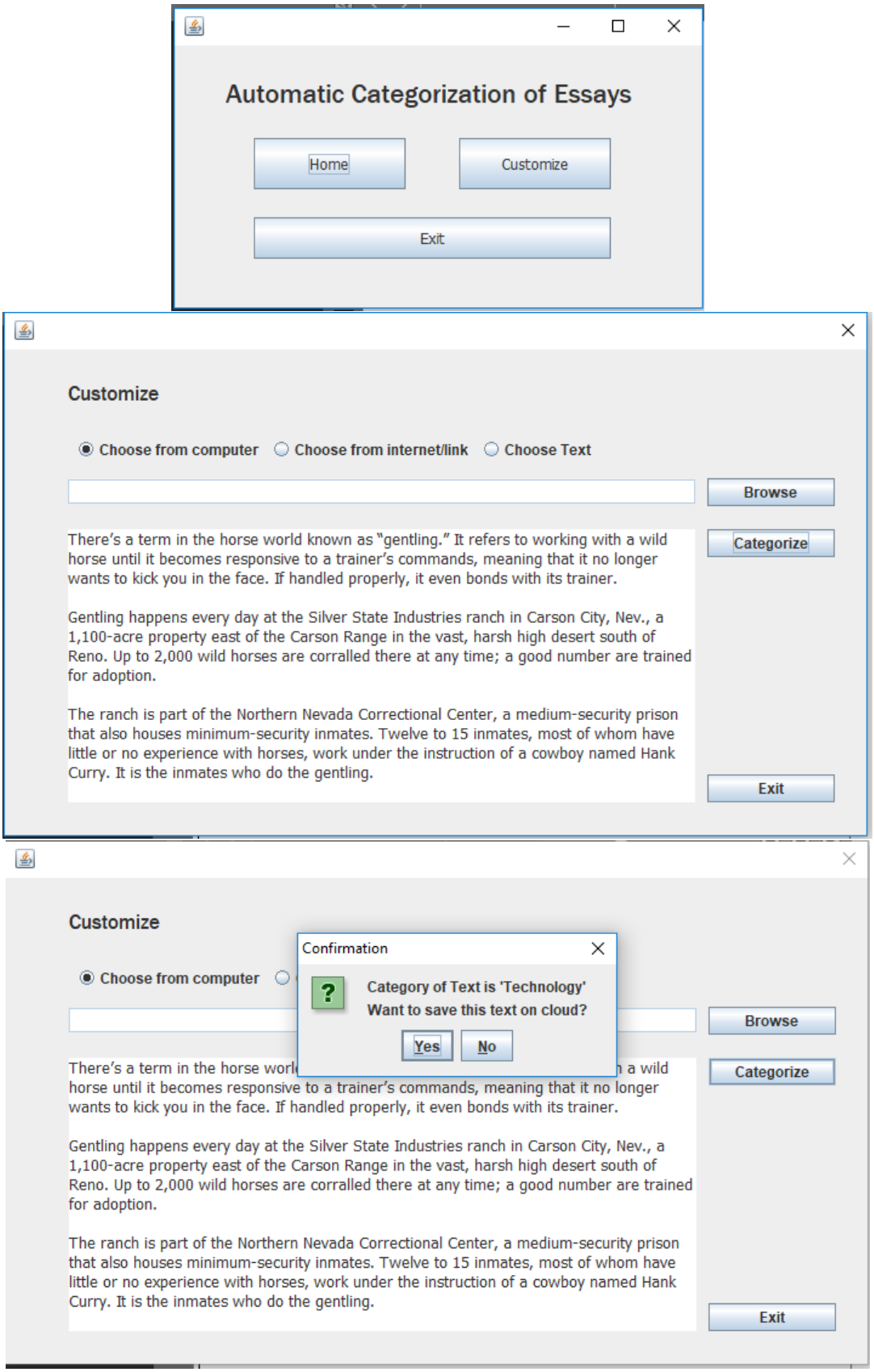


Figure 4.4: Low level design

This diagram shows the processing part of the application. From the first part where we obtain the texture data in the form of text file, web links by using Jsoup API which can copy all the text from the website and paste it into the text file Which shown in the 4.4 and 4.3. process will make a tree of words by using tokenization of words which was extracted from the site.After tokenization application automatically categorized the articles by using twinword api and save it on cloud. After this User can view the categorized articles on Home page of the application.4.2 explains the categorization process of the application.System take an input in the form texture data.Extract the metadata and apply tokenization technique. Match the extracted keyword with other related articles, categorize it and save it on cloud by using Amazon Dynamodb service.

4.6 GUI Design

Figure 4.5: GUI Design



The Graphical user interface of this application is simple and consistent because it is used for general user easy to understand and easy to analysis the steps taken by the application. User will easily identify which button will perform which action easily. GUI design which is meant to be look like is shown in 4.5. In this we have three windows:

- In first windows there is three buttons user will click Home button to view different articles store on cloud and can click Customize button for automatically categorizing some texture data. For closing the application user can click Exit button.
- In second window user can select a file from the computer by clicking on browse button can choose a link by using radio buttons as well as texture data. User can extract Meta data and publish the article on cloud by using this window.
- In third window there is a text box where user can paste the texture data and after this when user clicks on publish button data will automatically store on cloud in its correct category.

Chapter 5

System Implementation

Implementation is the process of putting a decision or plan into reality, effect or in execution. This chapter describe techniques and software components which are perform to develop this project into execution.

5.1 System Architecture

The system Architecture of this application is generic can be applicable to any text file which is user input. We just need to extract metadata and apply tokenization on given input and match the keywords. Then after analyzing the texture data will automatically store on cloud in its correct category. Application uses the components, tools and techniques of java language. In which all the built-in Libraries and API's are used. Tools and techniques which are used in this Project for example Extract metadata, Tokenization, Keywords Extraction, Automatic categorization, Stopwords and Cloud Storage, Stanford Core NLP tools, AWS Dynamodb, API (Scala,Twinword).

5.2 Tokenization

Tokenization is the demonstration of separating an arrangement of strings into pieces, for example, words, watchwords, expressions, images and different components called tokens. Tokens can be singular words, expresses or even entire sentences. During the time spent tokenization, a few characters like punctuation marks are disposed of. We use this technique in which sentence are chopped into chunk of words and now each word is deal with 1 token and throw away certain characters like punctuation etc. Token is a sequence of characters from a specific document join to make a semantic unit of processing.

5.3 Stop Words

Stopwords will be words which are filtered through earlier or after processing to natural language data. We eliminate Stop words mention in a table A.1 from our story by utilizing Core NLP tool in light of the fact that these words give additional data or may be some may be sometime they provide unnecessary information which may cause story out of there limit so by utilizing the tool NLP we wipe out these sorts of words and makes our story to the point.

Table 5.1: Common English Stopwords

a	into	don't	she'd	she'll	about	she's
down	wasn't	we	above	we'd	is	we'll
isn't	during	after	each	few	should	she
again	it	its	it's	were	in	was
against	for	shouldn't	we're	so	what's	weren't
all	we've	itself	further	i've	such	from
am	me	let's	than	what	that	some
an	more	has	when	that's	hadn't	had
and	hasn't	the	most	when's	only	herself
any	mustn't	have	where	their	hers	once
are	haven't	theirs	my	where's	hers	once
aren't	myself	having	which	them	hers	once
as	he	no	themselves	while	why's	they
at	nor	he'd	who	then	on	here's
be	he'll	not	there	whom	these	why
because	of	her	who's	there's	here	off
been	he's	or	they're	would	won't	they'll
before	other	him	wouldn't	they've	you	this
being	himself	ought	our	his	those	you'd
below	how	ours	you'll	through	to	your
between	i	out	you're	too	ourselves	how's
both	over	i'd	under	yours	you've	shan't
but	if	own	yourself	until	i'm	very
by	same	i'll	up	yourselves	do	does
cannot	can't	could	couldn't	doing	did	didn't

5.4 AWS Dynamodb

Amazon DynamoDB - [2] otherwise called Dynamo Database or DDB - is a completely overseen NoSQL database benefit gave by Amazon Web Services. DynamoDB is known for low latencies and versatility. Amazon DynamoDB is a completely managed NoSQL database benefit that gives quick and unsurprising execution with consistent adaptability. DynamoDB gives you a chance to offload the regulatory weights of working and scaling a

distributed database, with the goal that you don't need to stress over equipment provisioning, setup and design, replication, software fixing, or cluster scaling. With DynamoDB, we make database tables that can store and recover any measure of information, and serve any level of demand movement. We scale up or downsize our tables' throughput limit without downtime or performance degradation, and utilize the AWS Management Console to monitor resource utilization and performance metrics. By using DynamoDB we erase expired things from tables automatically to decrease storage usage and the cost of cost of storing data that is never again applicable. DynamoDB automatically spreads the information and traffic for our tables over a sufficient number of servers to deal with your throughput and storage requirements, while keeping up reliable and quick execution. All of our data is stored on solid state disks (SSDs) and naturally duplicated over numerous Availability Zones in an AWS area, giving implicit high accessibility and information strength.

5.5 API

The following APIs are used in the project.

5.5.1 Json

Json is a (JavaScript object notation) java API use in our project. This API is an open-standard format use human understandable text use to spread data consisting of values and attributes. This is basically how the program is interacting with the human.

5.5.2 Jsoup

Jsoup is a java library use to work with the HTML [7]. Jsoup is use in our project to extract all the data or story from the News story website and store it in a system or a file system. All the content like heading, hrefs, paragraph will be extracted and create a copy of that story in a system files.

5.5.3 Scala

Scala java API is use in our project. This API is use for the extraction of metadata from the given texture data as an input. It is use for web scraping and it's another extension of java scala is basically for data extraction from web links it parse the complete html page and fetch the data according to our requirements like an article extractor in our application or default web page extractor. It removes all the tags of html page and parse the specific story or article that we are interested for.

5.5.4 NLP

NLP (Natural Language Processing) is use in our project for tokenization. Tokenization is demonstration of separating an arrangement of strings into pieces, for instance, words, watchwords, expressions, pictures and different components called tokens.

5.5.5 Twinword

Twinword API is use in our project for automatic categorization of texture data. Twinword is text analysis API that can understand and relate words similarly as people do. It attracts the key words from the given data and after analyzing it suggest the correct category of the article.

5.6 Libraries

The following libraries are used in the project.

1. aws-java-sdk 1.11.218v
2. commons-io 2.6v
3. httpclient 4.5.3v
4. httpcore 4.4.8v
5. ion-java 1.0.3v
6. jackson-all 1.9.0v
7. joda-time 2.9.9v
8. jsoup 1.10.3v
9. org.eclipse.jface 3.13.1v
10. org.eclipse.jface.text 3.12.0v
11. org.eclipse.text 3.6.100v
12. org.eclipse.ui.forms 3.7.101v
13. org.eclipse.ui.workbench 3.110.1v
14. org.eclipse.osgi 3.12.50v
15. org.eclipse.swt.win32.win32.x86-64 3.6.100v

5.7 Keyword Extraction

We read different articles related with same topic for example cricket match there are quite similar words like player, bowler, keeper etc. When we provide article related with cricket to this application it will automatically place it in sports category on cloud by matching keywords use in the given article and use in related articles [5]. Twinword API is used for keyword extraction in our application and suggesting its correct category.

5.8 Automatic Categorization

We have created a publish option in the application by using this user can automatically categorize the texture data which is provided to this application as an input [5]. It is done by matching the keywords extracted from the input article and compare it with the quite similar articles. Then after the automatic categorization data is store on cloud in its correct category. For this purpose application use AWS DynamodB service. Amazon DynamoDB is a completely managed NoSQL database benefit that gives quick and unsurprising execution with consistent adaptability.

Chapter 6

System Testing and Evaluation

In this Chapter different testing systems are used for assessment and validation of this application. Testing plays an important role in the product software development process. It helps to validate the system will meet its requirements and the working of the application. Every project has few constraints and these constraints will be investigated during the test cases which are talk about in this chapter.

6.1 Usability Testing

Usability testing gives the information about how much time will it require performing a specific task of the system. Usability testing is assessed by the target audience group of the application. In our application, our audience is general Users. This application is use for automatic categorization of texture data and storing it on cloud. Usability testing tells that the system is performing the tasks that it is planned to do or, on the other hand not. The application performs the task the client want to perform from the application or not [8].

6.1.1 Easy to use

This project is easy to use and project is self-explanatory that the user can easily Interpret what the system is intend to do.

6.1.2 Easy to learn

Our system is very simple basic and consistent, visible and clear. The system is very easy to learn for the new users. It has very simple and easy GUI and self-explanatory new user can easily understand it.

6.2 Software Performance Testing

Software performance testing use to check how efficiently the system performs the task through this application. This will help us to determine the system capability, reliability and efficiency. Following steps were taken to increase the application performance.

- Comparison between keywords extracted from the texture data given by the user and other related articles is done on the run time before suggesting its category will increase the speed and consume less time.
- If there is no matching between the keywords extracted from article with any other article it will simply not suggest any category in suggested category box.
- System will not take or waste time to extract or checking the videos and images because they are not relevant for system.
- System will take some time 1 to 2 minutes for automatic categorization of given content and storing it on cloud.

6.3 Compatibility Testing

Compatibility Testing is a kind of non-functional testing. Compatibility implies on what conditions the system will perform well with no issues. This testing technique will help us to know the compatibility with which hardware and software resources need to use this application. Following are the compatibility feature should be having:

- Processor must be quick reason it performs to many processes to be taken so processing must be fast for example core i3 or more and RAM 4 GB.
- This application is developed on the Java. So, system must have the JDK platforms to run this type of application.
- This application is cloud based. So, user must have strong internet connection.

6.4 Exception Handling

In this system there are many exceptions are to be handle.

- HTTP exception case show if there is any issue regarding the internet connection exception will be shown.
- Exception will be displayed if the user input is not text related file for example some jpg file.
- Time out Socket exception is there as well.

6.5 Load Testing

Load testing is the way toward putting request on a software system or registering gadget and measuring its reaction. Load testing is performed to determine a system's behavior under both normal and anticipated peak load conditions. Load testing is that to test the framework under the strange circumstance applying stress to a software and decide the behavior of the system under this kind of circumstance. In this system stress may apply when Internet connection is disconnected over and over it will apply too much load on a system. In other case when a given article is too large it will too long to compare the keywords and automatic categorizing it. And the last thing is when the processing speed is slow then it applies all the load on the processor it will slow down the speed and consume lots of the time.

6.6 Stress Testing

Stress testing is a type of software testing that is utilized to decide the stability of a given framework. It put more prominent accentuation on heartiness, accessibility, and mistake dealing with under a substantial load, as opposed to on what might be viewed as right conduct under typical conditions. We apply stress test to our application by categorization on about 500 articles. The application successfully performs the automatic categorization on them it took about 20 seconds to complete its automatic categorization of each file.

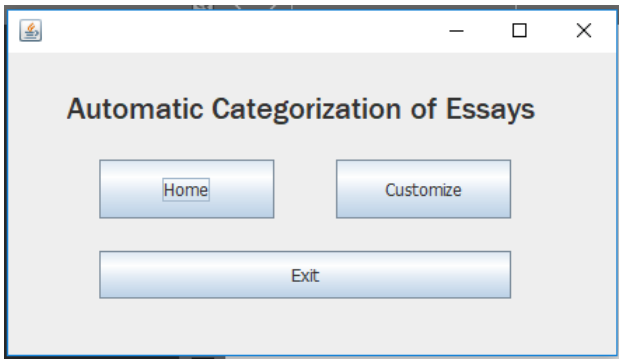
6.7 Security Testing

Security testing is a procedure expected to uncover imperfections in the security instruments of a data framework that ensure information. Ordinary security prerequisites may incorporate particular components of privacy, authentication, verification, accessibility, approval and non-repudiation. As over system automatic categorize the texture data and this data can be extracted from the authorized news websites where security is their priority. Our Application not categorize and display those types of data which will break the security terms and condition. No one can misuse this type of data.

6.8 Installation Testing

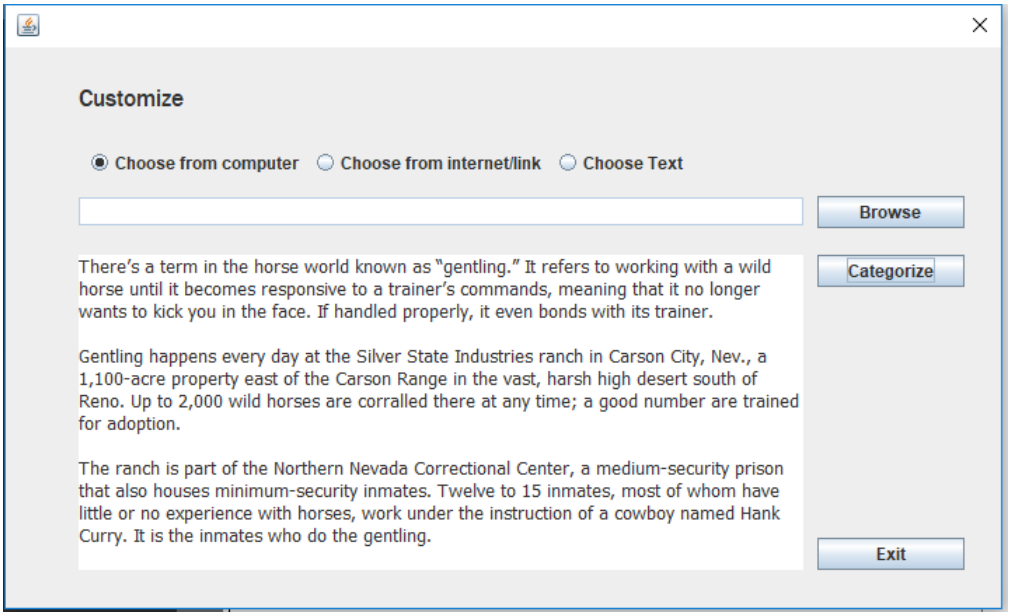
Installation testing refers to the testing of installation of the application. To run the project, there must be the Microsoft operating system installed in the system. This project is developing on Java. In this way, system must have JDK platform and java platforms. These things must need to run this sort of application.

Figure 6.1: GUI Main Window



6.9 Graphical User Interface Testing

Figure 6.2: GUI choose file or internet link



As shown in the Figur 6.2 6.1 where the Graphical user interface is displayed. GUI is developed in such a way that it is remain consistent in all the processes made by the application. The user interface of our system is self-explanatory. Our system contains tabs, radio buttons, text box etc. All things perform accurately. From one window user can select the file from personal computer and second one is use for pasting the texture data in the textbox which user want to automatically categorize. The test on textbox in second window is successful user can enter multiple words in it.

Chapter 7

Conclusions and Future Work

7.1 Conclusion

This project is all about to take texture data as an input and automatically categorize the text and store it on cloud. User can save their files on cloud after automatic categorizing which is very easily accessible from anywhere when strong internet connection is available. User can direct paste the content in the textbox and can also select the file from the hard drive. This application is cloud based that's why user must have internet connection to run this application. It is very hard manually to read long and large number of articles first and then decide their correct category. And after manually categorizing it's even hard to store the files on cloud. By using this application user can automatically categorize the texture data and store it on cloud in its correct category which can reduce the time and complexity as well. The main method is used in this algorithm to divide the problem in to small chunks and overcome those problems to address the main issue. So that's why, we divide the text into sentences or in the form of words and compare the extracted key words with the other related articles. After comparing we suggest the correct category of the article given by the user to this application. We made table of categories on cloud by using the Amazon Web Services account and store different articles in their correct category in the table. We almost give 500 articles to this application for automatic categorization and store them on cloud and this application successfully done its task.

7.2 Future Work

- The focus of this application is only on the texture data given to it. In future we will further expand this application, so it will also categorize other formats like videos, images etc.
- This application only analyzing the texture data and after extracting the keywords compare it with other related articles but in future after expanding this application

will also automatic categorize the videos after separating different frames of data given to it and will compare it with other related videos.

- In further we can create a graph for it. So user can see how much articles save on cloud in each category.
- We can also give a user summary of the article which he/her automatic categorize and store it on cloud by using this application. It will save the time of user of reading a long article.

Appendix A

User Manual

ACECS is user friendly software which is helpful for automatic categorizing the texture data and storing it on cloud for technical users and also for non-technical users.

- Run ACECS directly.
- Main window will pop up which have two main buttons Home and Customize.
- Select the customize button for categorizing the texture data.
- Select browse button for selecting the texture file directly from the computer hard disk.
- If user want to extract the data from the website he/she can give the link by clicking on link button.
- Option of Direct copy paste the texture data in textbox is also available for the user.
- When data is loaded click on categorize button for automatic categorizing the data and storing it on cloud.
- When data is loaded click on categorize button for automatic categorizing the data and storing it on cloud.
- View different categorized texture data by selecting the Home button from Main window.
- Click on specific category to view the articles store in it.
- Click on exit button to close the application.

References

[1] Latanya Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, October 2002. Cited on p. 1.

[2] James Murty. *Programming Amazon Web Services: S3, EC2, SQS, FPS, and SimpleDB*, pages 1–27. O'Reilly Media; 1 edition, 2008. Cited on pp. 3, 8, 9, and 24.

[3] Owen Fitzpatrick Richard Bandler, Roberti. *The Ultimate Introduction to NLP: How to build a successful life*, chapter chap 2, pages 5–10. HarperCollins; Reprint edition (January 3, 2013), March 19, 2013. Cited on p. 6.

[4] Apache Software Foundation. The apache lucene core. 18 October 2017. Cited on p. 7.

[5] Kun Yue, Wei-Yi Liu, and Li-Ping Zhou. Automatic keyword extraction from documents based on multiple content-based measures. *Computer Systems Science and Engineering*, 26(2):133, 2011. Cited on pp. 8 and 27.

[6] Dapeng Liu. *SERA '11 Proceedings of the 2011 Ninth International Conference on Software Engineering Research, Management and Applications*, chapter chap 4. 2011-08-10. Cited on p. 15.

[7] Pete Houston. *Instant Jsoup How-To*, chapter 5, page 26. Packt Publishing, 2013. Cited on p. 25.

[8] Boris Beizer. *Software Testing Techniques*. Van Nostrand Reinhold; 2 edition, (June 1990). Cited on p. 29.