الجمهورية الجزائرية الديموقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليـــــم العـــــــــالي والبحــــث العلمـــــي
Ministry of Higher Education and Scientific Research
جــــــــامعة بــــرج بوعريريـــج
University of Bordj Bou Arréridj

Faculty of Mathematics & Computer Science

Computer Science department

Graduation Thesis

For obtaining the master's degree in information and communication technologies

Theme: ***The role of social media on the improvement of business processes (in the digital strategy)***

Realized by:

- MESSAGIER Abderraouf

Presented publicly in: 11/07/2021

in front of the jury:

President	Dr. AKHROUF Samir	MCA in the U. El Bachir El Ibrahimi-BBA.
Examiner	Dr. ATTIA Abdelouhab	MCA in the U. El Bachir El Ibrahimi-BBA.
Supervise by:	Dr. BARKAT hadj	MCA in the U. El Bachir El Ibrahimi-BBA.

Class of: 2020 / 2021

Dedications

I dedicate this modest work

To my dear parents who supported me during my studies, with their attentions and their encouragement, especially my mother may Allah bless her soul.

To my dear three sisters for their supportiveness, kindness, trustworthiness.

To my dear brother for being a role model.

To my friends and colleagues in Master 2 and to all those who have helped me.

Messagier Abederraouf

Thankfulness

With the help of Almighty Allah, I was able to accomplish this modest work.

I warmly thank my thesis supervisor Mr. Hadj Barakat to have accepted to supervise this work and to have especially keeping up to date through every step of this journey.

I sincerely thank all the teaching staff of the specialty (information and communication technologies)at the University of Bordj Bou Arréridj.

I would like to thank all the cleaning staff of MI department for their hard work to keep our college at its best.

I want to thank all the people who have contributed from near or far to the accomplishment of this work.

Last but not least, I want to thank me for believing in me, I want to thank me for doing all this hard work.

ملخص

الموضوع الذي تتناوله الأطروحة هو إستخراج بيانات الويب عن طريق عملية تجريف الويب، و العمل مقترح من خلال هذه الأطروحة هو الإستجابة لمشكلة محددة و في هذه الحالة هي تصفية البيانات و إنتقاء البيانات المحتاجة من خلال إقتراح نظام تجريف الويب الذي هو تطبيق حاسوبي يسمى hashtag search application.

في كثير من الأحيان يعجز التجار المقبلين على بداية مشروع في أي مجال على تجميع البيانات المحتاجة لنجاح مشروعهم الخاص، خاصة في مجال بيع وصناعة الألبسة فصاحب المشروع يحتاج ان يكون على علم بما هو رائج في الأسواق فهذا يساعده على منافسة الشركات الناجحة فلهذا يحتاج لوسيلة تساعده على الحصول على المعلومات المهمة لضمان البيع والبقاء في السوق.

من بين التكنولوجيا الأكثر شيعا وإستعمالا في وقتنا هذا مواقع التواصل الاجتماعي فهي تساعد الفرد و الجماعة على التواصل و الإتصال، من مميزات مواقع التواصل الاجتماعي الإحتواء على كم هائل من المعلومات و البيانات لكن كيف يقوم صاحب المشروع بتصفية كل هذه المعلومات للحصول على البيانات التي تهمه فقط.

hashtag search application هو تطبيق يساعد على إستخراج البيانات من موقع التواصل الاجتماعي Instagram بإستعمال خوارزمية عملية تجويف الويب، يعمل هذا التطبيق كمحرك بحث متصل بالInstagram يقوم التطبيق بإستخراج البيانات من الموقع ويستعمل الHashtag لتحديد المحتوى المراد بحثه.

يحتوي التطبيق على خاصيتين فالتطبيق يجلب للمستعمل المنشورات المشهورة و المنشورات التي نشرت مؤخرا في Instagram ومنه يصبح المستعمل على الدراية بمتطلبات الأشخاص و بالمنتوجات المشهورة التي في السوق والتي تعد أكثر مبيعا و بذلك من خلال تحليل هذه البيانات يمكن لصاحب المشروع إنجاح مشروعه.

الكلمات المفتاحية : إستخراج البيانات، المنشورات المشهورة، المنشورات التي نشرت مؤخرا، البحث عن المعلومات، خوارزمية عملية تجويف الويب.

Abstract

The theme addressed by this thesis is the extraction of web data by the web scraping method, the goal that need to be achieved through this thesis is to respond to a very specific problem which is to extract existing digital data by proposing a web scraping system which is an application named hashtag search application.

In many cases, business owners who are about to start a project in any field are unable to collect the data they need for the success of their own project, especially in the field of selling and manufacturing clothing cause the project owner needs to be aware of what is trendy in the markets, as this helps him to compete with successful companies, so he needs a tool to help him obtain important information to ensure sales and stay in the market.

Among the most common and widely used technology in our time social media sites help the individuals and the companies to communicate and stay in touch. One of the advantages of social media sites is that they contain a huge amount of information and data, but how does the business owner filter all this information to obtain the data that interests him only.

Hashtag search application is an application that helps to extract data from the social media site Instagram using the algorithm of the web scraping, this application works as a search engine connected to Instagram, the application extracts data from the site and uses the hashtag to determine the content to be searched.

The application contains two features. The application brings to the user the trendy posts and the latest posts in Instagram, and with the information brought with the hashtag search application the user will become familiar with the demands of people and the trendy products that are in the market which are best-selling, so by analyzing this data, the project owner can make his project a success.

Key words: Data extraction, web scraping, business owners, trendy posts, the latest posts, algorithm of the web scraping, search engine.

Résumé

Le thème abordé par cette thèse est l'extraction de données web par la méthode de web scraping, le but qui doit être atteint à travers cette thèse est de répondre à une problématique bien spécifique qui est d'extraire des données numériques existantes en proposant un système de web scraping qui est une application nommée hashtag search application.

Dans de nombreux cas, les propriétaires d'entreprise qui sont sur le point de démarrer un projet dans n'importe quel domaine sont incapables de collecter les données dont ils ont besoin pour le succès de leur propre projet, en particulier dans le domaine de la vente et de la fabrication de vêtements, car le propriétaire du projet doit être conscient de ce qui est à la mode sur les marchés, car cela l'aide à rivaliser avec les entreprises prospères, il a donc besoin d'un outil pour l'aider à obtenir des informations importantes pour assurer les ventes et rester sur le marché.

Parmi les technologies les plus courantes et les plus utilisées à notre époque, les sites de médias sociaux aident les individus et les entreprises à communiquer et à rester en contact. L'un des avantages des sites de médias sociaux est qu'ils contiennent une énorme quantité d'informations et de données, mais comment le propriétaire d'entreprise filtre-t-il toutes ces informations pour obtenir les données qui ne l'intéressent que lui.

Hashtag search application est une application qui permet d'extraire des données du site de réseau social Instagram en utilisant l'algorithme du web scraping, cette application fonctionne comme un moteur de recherche connecté à Instagram, l'application extrait les données du site et utilise l'hashtag pour déterminer le contenu à être cherché.

L'application contient deux fonctionnalités. L'application apporte à l'utilisateur les publications à la mode et les dernières publications sur Instagram, et avec les informations apportées avec l'application de recherche d'hashtag, l'utilisateur se familiarisera avec les demandes des personnes et les produits à la mode les plus vendus sur le marché, ainsi en analysant ces données, le maître d'ouvrage peut réussir son projet.

Mots clés : Extraction de données, web scraping, chefs d'entreprise, posts à la mode, les derniers posts, algorithme du web scraping, moteur de recherche.

Summary

List of Figures

I. Chapter one: General introduction and Problematic

II. Chapter two: State of the Art

III. Chapter three: Architecture and Modeling

IV. Chapter four: Implementation and Results

GENERAL INTRODUCTION

Over the last decade technology developments are increasing rapidly. The biggest business (Amazon, Ali BABA, Google, Facebook …) were born two decades ago, they are too dependent on technology and specially on social media.

Some of the social media that many of us already know, such as Facebook, Instagram and Twitter are one of best information spreading and searching facilities. Because it is not only providing the information itself, but also provide the information by users' social network (transparency).

Business is a big domain, it has so many layers such as business environment, management, organization, marketing, finance, accounting, it also covers so many fields that's have a big value of improving the economy like (Transport Industry, Computer Industry, Petroleum industry, Clothing industry …), but we are going to focus on marketing in the clothing industry.

Clothing industry is one of the biggest industries ever just talking by numbers, the global apparel market is value about 1.5 trillion U.S. dollars in 2020, so to have a place among the giants' companies like (Gucci, Dior, Nike…) you must have a solid marketing strategy.

Although social media give you access to Important information, but this information comes in a large amount and also can be irrelevance that's why in my end of studies' project I'm seeking to create a system that can search the relevant information (in Instagram), filter this kind of information, and present it to user.

As part of my end of study project at the Faculty of Mathematics and computer science, at the University of Bordj Bou Arreridj, and for the "Master's" degree in computer science, it is the work presented in this document. And it's about the conception and the realization of Social Media Web Scraping (Instagram).

To carry out this work I have used the method «UML» for modelling the information system, it is an approach on the level of conception and development of a system which presents a flexible approach and a formalism rich enough to simplify the study.

I have also chosen the programming language Python version 3.9 and used the development environment integrated (PyCharm Community Edition) for programming my application.

The Application that I am seeking to create is a tool for extracting data from Instagram (popular posts and latest posts) for future treatment and analysis.

My thesis contains four chapters each chapter touches a certain area, starting by chapter one where there will be a brief introduction to the theme of the thesis with the problematic, next in chapter two there will be a detailed description of the proposed model jumping to the third chapter also known as the conceptional side of the thesis, and finally in the last chapter there will be the fully presentation of the application.

Chapter 1:

General introduction

And

Problematic

1. Context:

1.2. Introduction:

Social media is a vast subject, particularly when it comes to fashion. You must keep track of multiple platforms, plan various types of content, and collaborate with various social media influencers.

In the world of social media, fashion brands prosper and expand. Instagram, for example, is highly visual platforms that help you target potential customers and raise brand awareness.

Fashion marketing is the method of controlling the flow of merchandise from the initial collection of designs to be manufactured to the introduction of goods to retail customers, the goal of fashion marketing is to maximize a company's sales and profits.

1.3. Definitions:

➢ *Social media:*

Social media is any interactive medium that allows users to easily build and share content with the public. Facebook, and Instagram are only a few examples of social media websites and applications.

“Instagram (commonly abbreviated to IG or Insta) is an American photo and video sharing social networking service created by Kevin Systrom and Mike Krieger in April 2012, Instagram became the 4th most downloaded mobile app of the 2010s.” [1]

➢ *Marketing:*

Marketing is an operation (activity), set of institutions, and processes for developing, interacting, distributing, and sharing value-added services for consumers, companies, partners, and society as a whole. [2]

➢ *Clothing industry:*

Clothing industry or garment industry also called apparel and allied industries, centre on both trading and production of outerwear, underwear, footwear…ext. [3]

In general, the textile industry is a group of related industries that manufacture fabric from a variety of natural (cotton, wool, etc.). It is a major contributor to the economies of many countries, encompassing small and large-scale operations all over the world. [4]

1.4. Social Media Process Flow:

An overview of how companies can use social networking and social media to build communities, participate in discussions, and monetize their efforts.

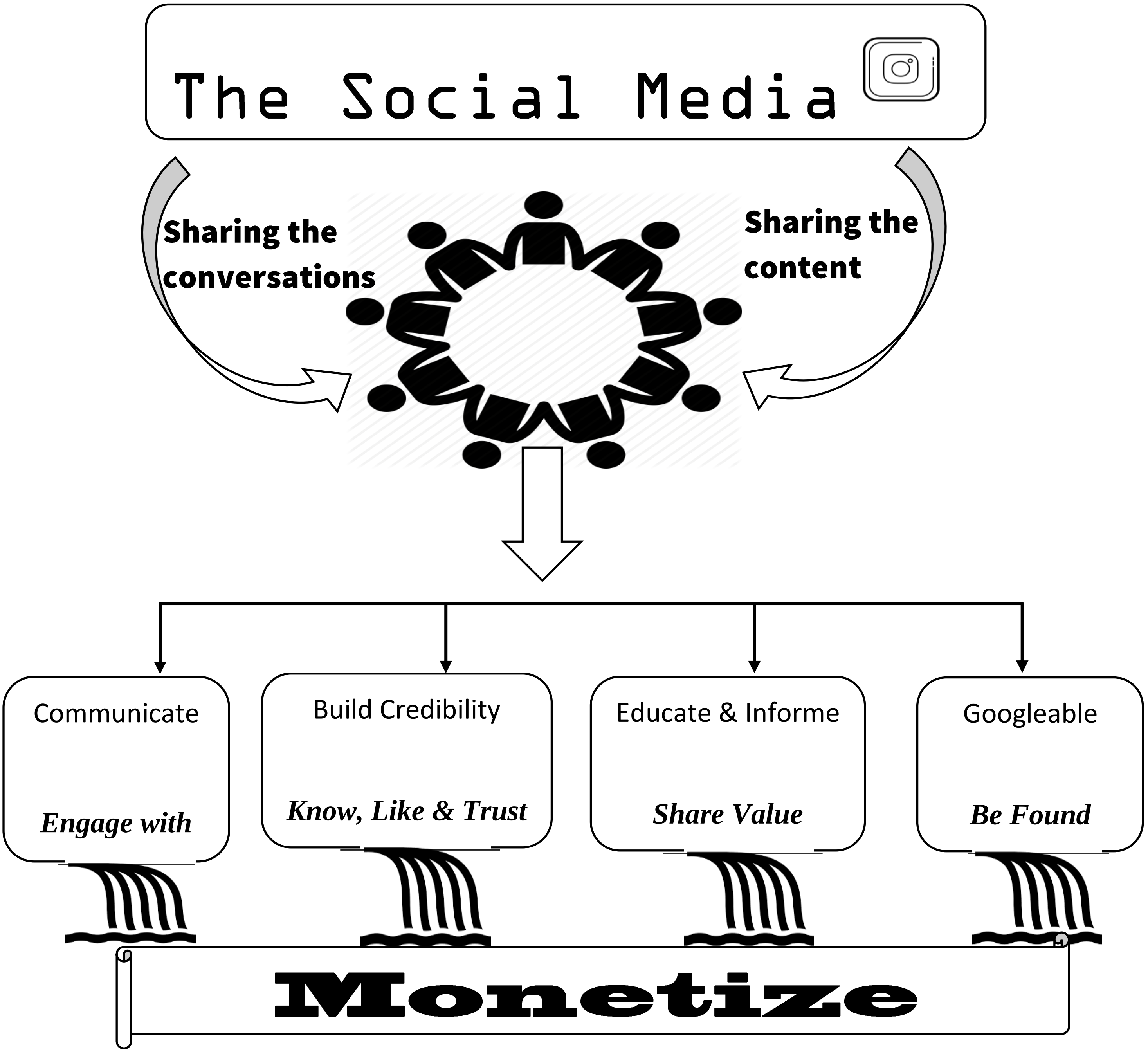


Figure 1: Social Media Process Flow

2. Problematic:

We can all agree that in today's world, social media is the refugee for us to find information in any type of field, weather finding something to buy or how to fix something or even looking for a personal advice.

Over the last decade, social media has evolved into a powerful marketing tool that has not only added a new dimension to marketing but also opened up plenty of new opportunities to the marketers to create brand awareness among consumers. It has become now widely regarded as the most accessible, interactive, and transparent form of public relations.

Although social media is considered the Mecca of information and public audience, it is hard to collect certain information in certain field or target a special type of audience weather by age or personal interests. So, the problem that we are facing is how to filter this information to have a higher efficiency to marketing your product.

3. Objective:

To improve our marketing strategy and elevate our business (clothing business), there is some goals need to be achieved such as:

- Filtering information by finding posts that have relation with clothing hashtags
- Targeting the audience that have interests in clothing
- Finding the most popular posts to keep up with latest fashion trends
- Finding the posts that have strong engagement for potential influencers

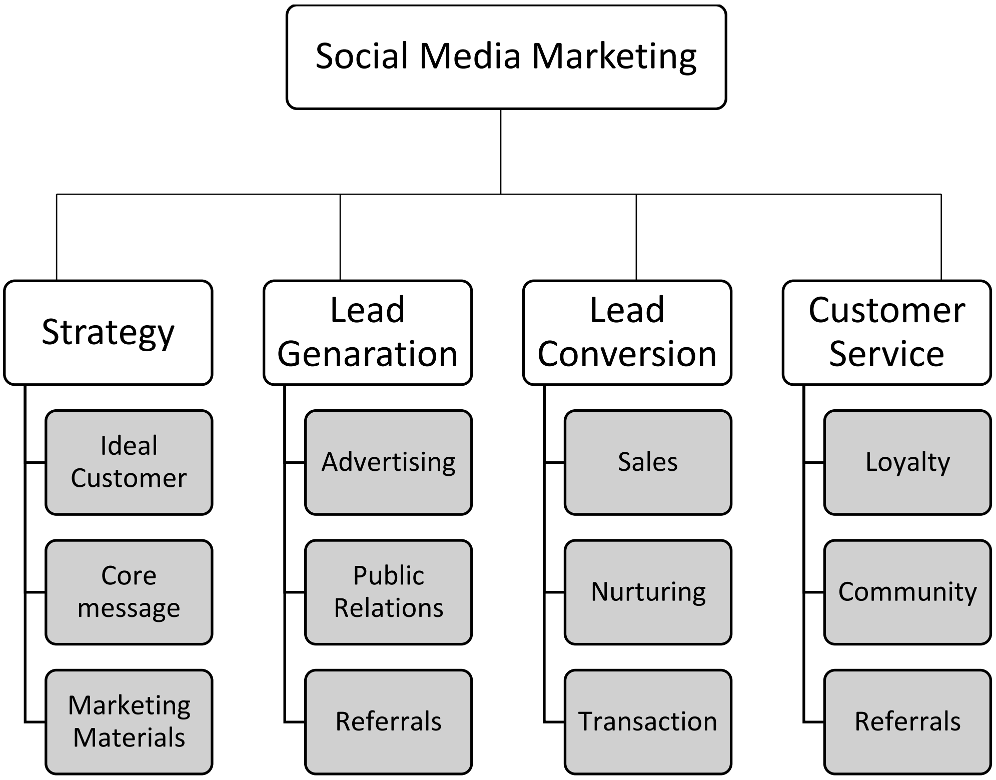


Figure 2: The organization chart of Social Media Marketing [5]

4. Contribution:

We can all agree that in order for any business to succeed in any industry what so ever you have to build you strategies and insights from data.

Data is the new differentiator. Market analysis and business strategies are built around it.

Whether you're starting a new project or developing a new plan for an existing company, you'll almost always need to access and analyze a large amount of data.

Web scraping has become a popular strategy for e-commerce companies, particularly when it comes to delivering rich data-based insights. It is a powerful tool, it let you to knows your audience and help you deliver what the audience actually likes.

Scrapping not only gives numbers, but also sentiment and behavioural analysis, so the business can know what audience types and choice of product they want. Giving your clients something, they do relate to, give you the advantage of having a loyal costumer.

5. Plan of thesis:

In my thesis there is more to it then this chapter, after the general introduction and the first chapter there will be three other chapter and general conclusion to finalize the thesis.

In the second chapter I will explain and define the types of Information research systems and the web scraping technique and how it works, and the objectives of this proposed model.

In the third chapter I will touch the conceptional side of the application, and the modulization aspect using UML.

In the fourth chapter I will talk about the tools and the programming language used for the creation of the application, supported with some interfaces for demonstration.

Finally, a general conclusion will conclude my work for an overall idea of the thesis.

Chapter 2: State of the Art

1. Existing models:

1.1. Introduction:

The study of existing models is the most accurate and remarkable in the project of data extraction, it presents the nature of the business, facilitate its study and consists of the examination of each document, its circulation, and its characteristics in order to establish an information system that meets the needs and the requirements of the user.

1.2. Information research system:

- ### *Definition:*

Briefly, it is a system which allows us to find an information relevant to a query in a large collection of documents, which means filtering the data according to the user needs and domain.

- ### *Research strategies:*

Keyword research:

Keyword research is a practice search engine optimization (SEO), professional use to find and search terms that users enter into search engines when looking for products, services or general information. Keywords are related to queries, which are asked by users in search engines.

A Keyword search looks for words anywhere in the record. Keyword searches are a good substitute for a subject search when you do not know the standard subject heading. [6]

Search by navigation:

Among the existing search methods, we find navigation search systems. This is exploratory research that allows a user who often does not have prior knowledge of a certain information, to discover them according to its needs, and this by proposing exploration criteria.

Some of the navigation systems or as we call theme search engines we can mention (google, Bing, yahoo…)

web scraping method:

It is the most developed method, and it generally means extracting data from websites using automated processes implemented using a bot or web crawler. It is a form of copying in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis. [7]

1.3. Architecture of an information research system:

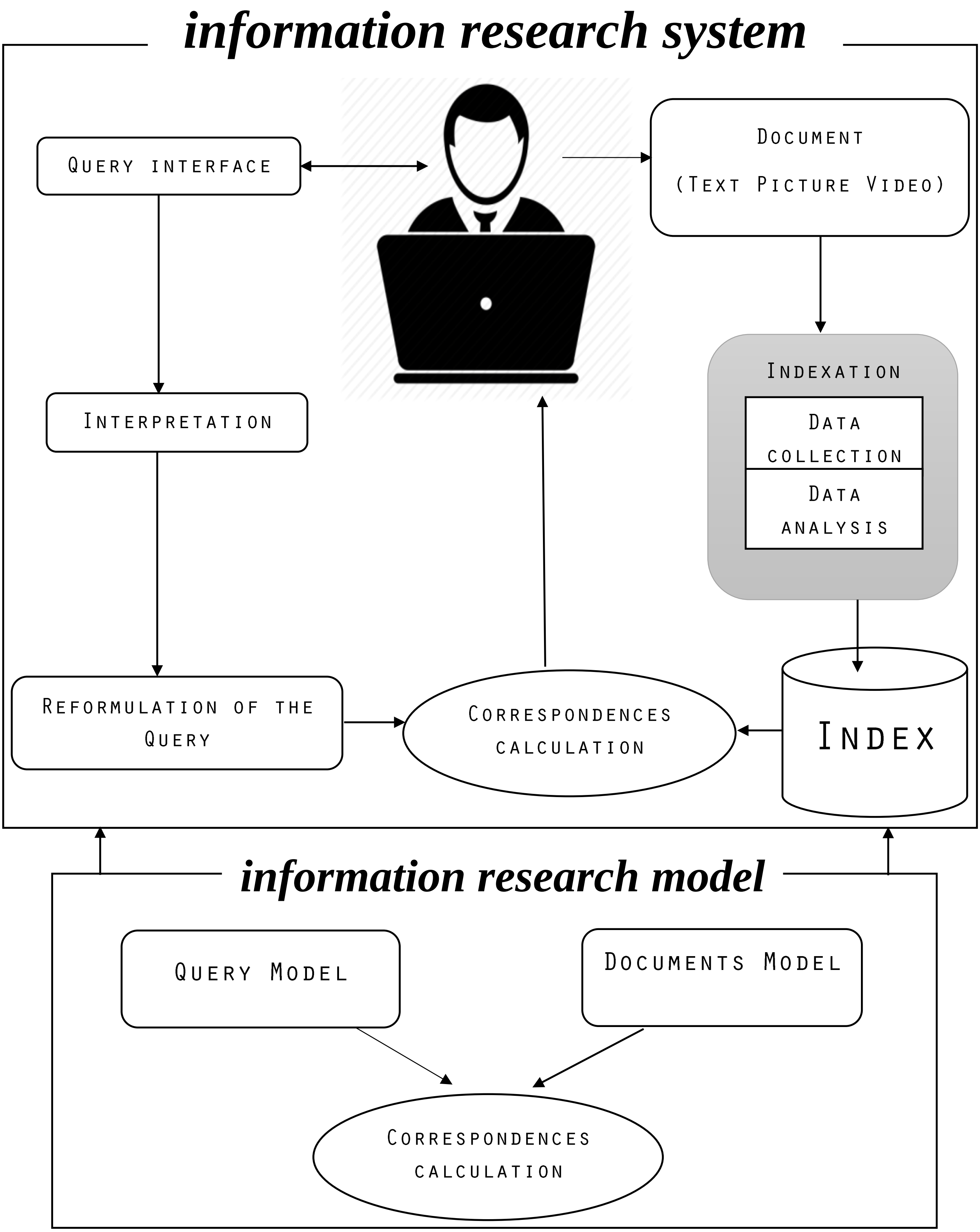

information research system
Query interface
Document
(Text Picture Video)
Interpretation
Indexation
Data collection
Data analysis
Reformulation of the Query
Correspondences calculation
Index
information research model
Query Model
Documents Model
Correspondences calculation

Figure 1: General architecture of information research system [8]

2. Brief description of the proposed model:

2.1. Web Scraping:

➢ Definition:

The term “scraping” refers to the extraction of data, and web scraping can be defined as a process or technology used to collect or extract a large amount of data from one or more websites in a short amount of time and it is widely recognized as an efficient and powerful technique for collecting heterogeneous and big data through a program, another website or script, this process can be done manual by a user, or by using a bot or a crawler (indexing robot). [9]

An API interface acts as an intermediary between the execution script of the scraping program and the site targeted for data extraction, in order to manually process the data collected from websites.

In conclusion the scraping program works in three steps, first it connects to the targeted website through a protocol and then on the second step the program recovers the filtered data that has been analysed and lastly the data will then be exported in different formats such as CSV and JSON as user needed.

➢ How web scraping works:

First step (Request - Response):

The first simple step in any web scraping program (also known as a "scraper") is to ask the website you are scraping for content from a specific URL. In return, the scraper obtains the requested information in HTML format. HTML is the type of file used to display all textual information on a web page.

Second step (Analyse and extract):

HTML is a markup language with a simple structure. As for analysis, it generally applies to any computer language. It is the process of taking the code as text and producing a structure in memory that the computer can understand and use.

So, HTML analysis basically involves embedding HTML code and extracting relevant information such as page title, page paragraphs, page headers, links, bold text, etc.

Third step (Download data):

The last part is where you download and save the data to a CSV, JSON, or database so that it can be retrieved and used manually or used in any other program.

And by applying all this steps, you can now extract specific data from the web and store it in a database or spreadsheet for later retrieval or analysis. [10]

2.2. Web scraping architecture:

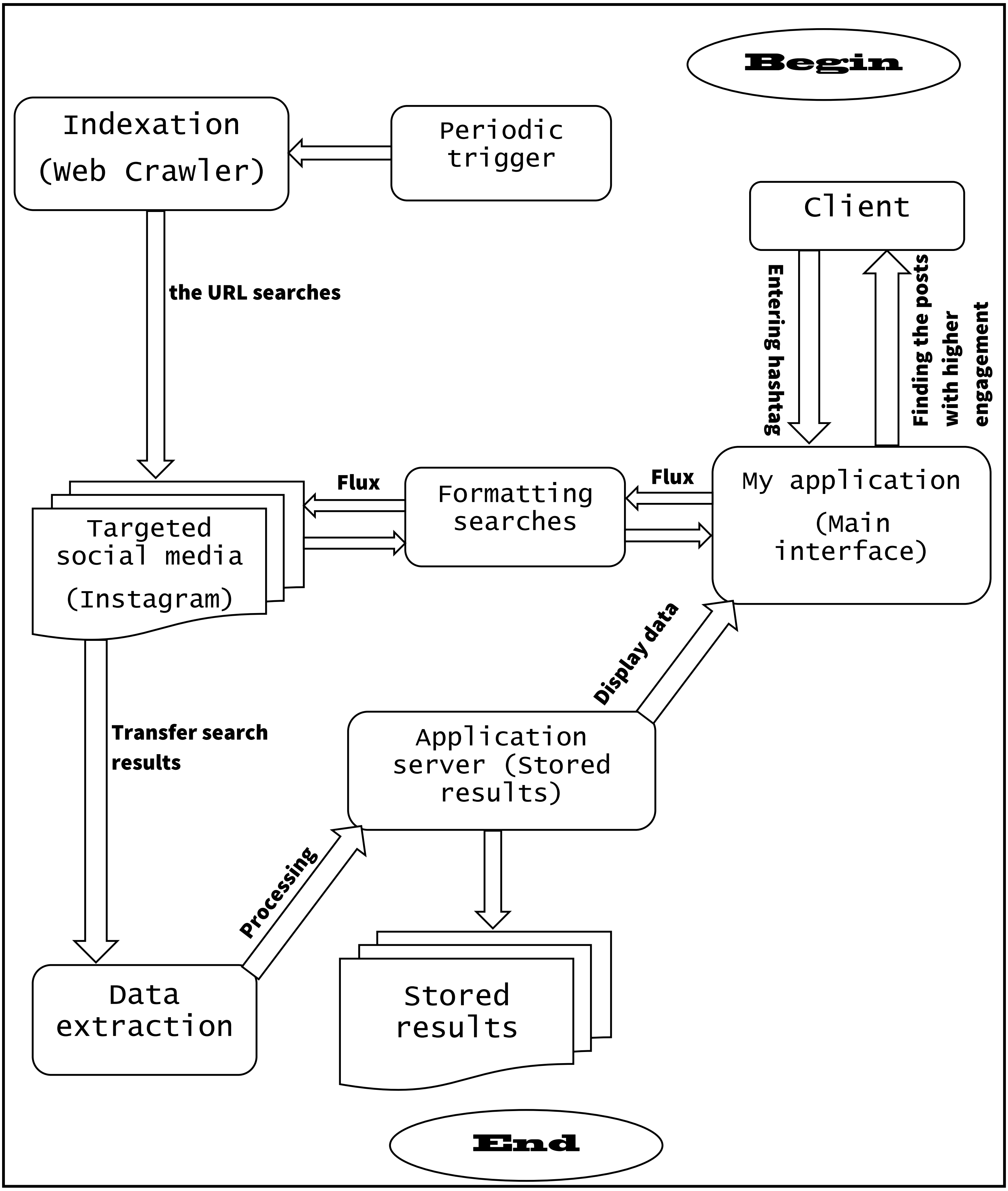


Figure 2: Detailed architecture of a web scrapping social media site (Instagram)

3. Objectives of the proposed model:

On and all, web scraping has many uses in all the fields but in garments industry it can be so beneficial, improving your marketing strategy is a key for your business success and web scrapping can make that happened, and that by helping you retrieve the data you need to have, to get a better knowledge over your surrounding (customers, fashion trends, fashion influencers).

I have chosen Instagram as a social media platform cause it's the most go to place for fashion inspiration or clothing information in general.

So basically, the user will give the programme certain hashtags that have relation to clothing domain such us (#Mensfashion, #OOTD, #Sneakerhead), then the programme will gather the data from the website and the query now will provide us with most popular post and latest posts, and by analysing this data you have now a chance to find a potential influencer to advertise your product and keep up with fashion world and be aware of the latest fashion trends.

4. Conclusion:

In this chapter, I started by defining the general system of my thesis and that's the Information research system and among this system's methods.

I have chosen the model that I will be using in my graduation thesis and that model is called web scraping and, I defined its process with its architecture.

And finally, I have stated the objective of this model which means how this model going to work.

Chapter 3: Architecture and Modeling

1. Introduction:

Conception is about determining in a detailed and precise manner what the system is capable of doing. In this chapter, I present the phases of the overall approach proposed for application development using UML language.

2. Definition:

2.1 UML:

The Unified Modeling Language (UML) shortly is a general-purpose, developmental, software engineering modeling language that provides a standard way to view the system design. [11]

UML is defined as a graphical and textual modeling language intended to understand and describe needs, specify and document systems, sketch software architectures, design solutions and communicate points of view. It is a formal and standardized language which, thanks to its graphic representation, allows solutions to be conceived and facilitates their understanding. [12]

Its versatile nature and its flexibility make it a universal modeling language and an essential standard for the analysis and conception activities, and in particular makes it possible to:

- Understand and describe the needs.
- Specify a system.
- Establish the software architecture.

Given the variety of formalisms used by object analyses and conception methods, UML represents a real factor of progress through the standardization effort.

UML's objective is to provide a standard notation, which can be used by all object-oriented methods, and to select and integrate precursor notation elements. [13]

A wide range of applications is covered by UML. Therefore, it provides constructs for a wide range of systems and activities (e.g., distributed systems, analysis, system design and deployment).

UML is a notation that resulted from the unification of OMT from: [14]

- Object Modeling Technique OMT [James Rumbaugh 1991] - was best for analysis and data-intensive information systems.

- Booch [Grady Booch 1994] - was excellent for design and implementation. Grady Booch had worked extensively with the Ada language, and had been a major player in the development of Object-Oriented techniques for the language. Although the Booch method was strong, the notation was less well received (lots of cloud shapes dominated his models - not very tidy)

- OOSE (Object-Oriented Software Engineering [Ivar Jacobson 1992]) - featured a model known as Use Cases. Use Cases are a powerful technique for understanding the behaviour of an entire system (an area where OO has traditionally been weak).

UML contains new notions, such as extension mechanisms and a constraint language, that were not present at the time in another major methods.

3. UML diagrams:

UML 2 has many types of diagrams, divided into two categories (Structure diagrams, Behavioral diagrams), Some types represent structural information such as (Class diagram) and the rest represent the behavioral side (Use case diagram).

UML in its version 2, offers thirteen complementary diagrams which allow the modeling of a project throughout its life cycle.

The thirteen types of UML diagrams are divided into two categories as mentioned:

3.1. Structure diagrams:

Structure diagrams show the pieces in your system's static structure. It shows the things in the system – classes, objects, packages or modules, physical nodes, components, and interfaces. [15]

The six UML structural diagrams are structured approximately around the key groups of items that you'll encounter when designing a system.

The six structural diagrams presented like this: [12]

- Class diagram: it represents the conceptual architecture of a system through a simulation cantered on the concepts of classes and associations.
- Object diagram: shows instances of structural elements and their links at runtime and helps to illuminate the class diagram.
- Component Diagram: Indicates complex structures with their supplied and required interfaces.
- Deployment diagram: defines the physical deployment of objects on hardware resources.
- Package diagram: specifies the logical organization of the model and the relationships between packages.
- Composite structure diagram: designates the internal organization of a complex static element and describes the collaboration of instances.

3.2. Behavioral diagrams:

The five behavioral diagrams of the UML are used to depict, specify, develop, and describe a system's dynamic features. It demonstrates how the system interacts with itself and other elements (users, other systems). *from[https://www.coursehero.com/file/p53caoa/]*

The seven behavioral diagrams presented like this: [12]

- Use case diagrams: represents the structure of the major functionalities required by the users of the system.
- Activity diagram: shows the flow of actions and decisions within an activity and graphically represents the behavior of a use case.
- State machine diagram: describes the internal behavior of an object using a finite state machine.
- Sequence diagram: shows the vertical sequence of messages passed between objects within an interaction.
- Communication diagram: refers to the communication between objects in the plan within an interaction.
- Interaction Overview Diagram: combines activity and sequence diagrams to combine interaction fragments with decisions and flows.
- Timing diagram: represents the states and interactions of objects in a context where time has a strong influence on the behavior of the system.

3.3. Structure Vs Behavioral:

In conclusion the structure of the objects, classes, or components that exist in the problem domain is specified using static modeling. These are represented by the terms class, object, and component.

Dynamic modeling, on the other hand, relates to the representation of object interactions in real time. Sequence, activity, collaboration, and status are all used to express it.

3.4. General view on UML diagrams:

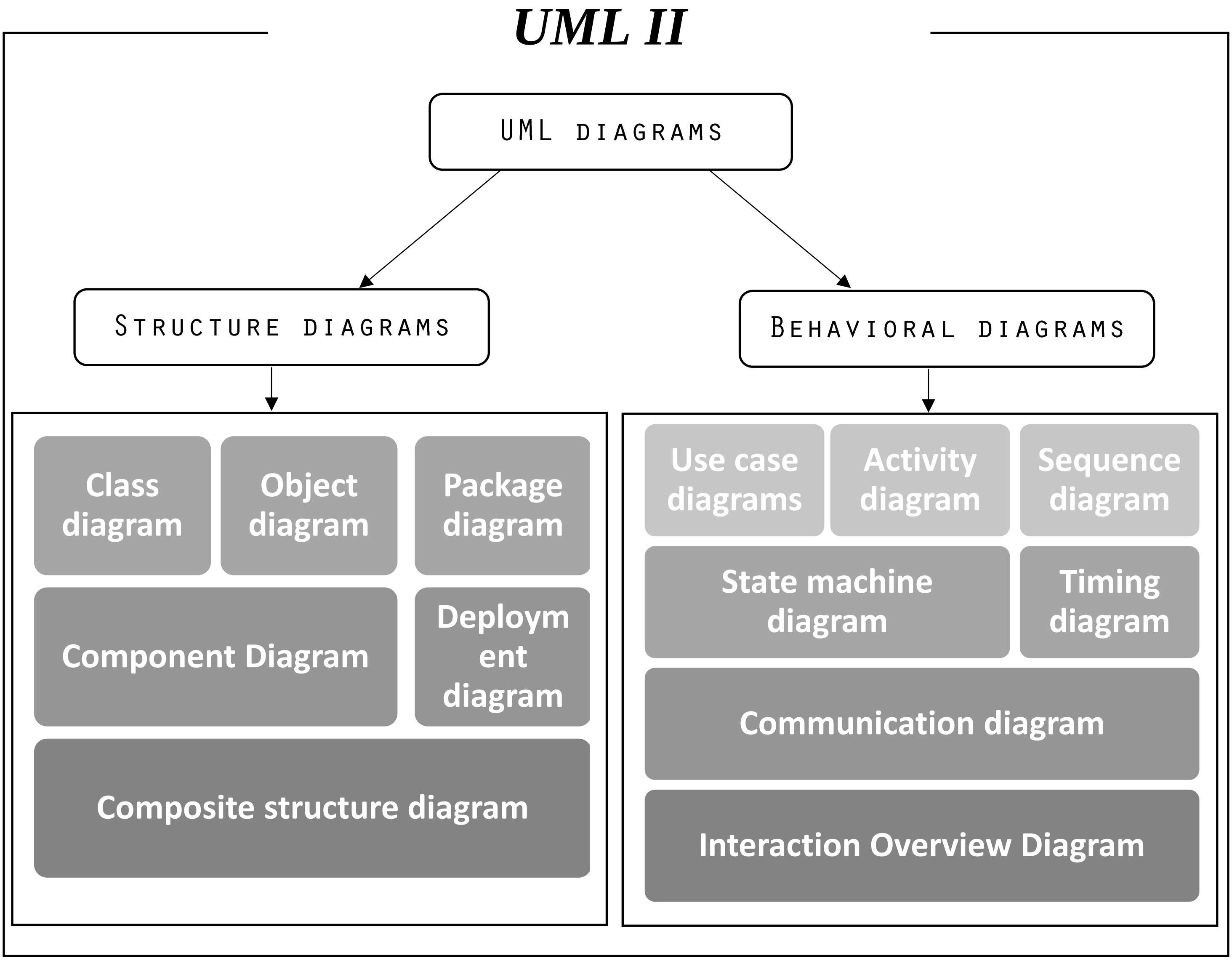


Figure 1: UML 2 Diagrams

4. Advantages of UML:

[16]

- You know exactly what you are getting
- You will have lower development costs
- Your software will behave as you expect it to. Fewer surprises
- The right decisions are made before you are given poorly
- written code. Less overall costs
- We can develop more memory and processor efficient
- systems
- System maintenance costs will be lower. Less relearning
- takes place
- Working with a new developer will be easier.
- Communication with programmers and outside contractors
- will be more efficient

5. Use case diagram:

5.1. Definition:

a use case model specifies a system's functional requirements. It's a representation of the system's planned functionality (use cases) as well as its environment (actors). You can use this model (use case) to connect what you require from a system to how it meets those needs.

Use cases are a way to collect and describe the needs of the actors in the system. They also make it possible to express the needs of the users of a system.

So, in conclusion the roles of use case diagrams are:

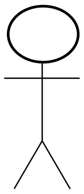
- Collect, analyze and organize the requirements.
- Identify the main functionalities of a system.

5.2. Elements of a Use Case:

The elements used for representing use case diagram:

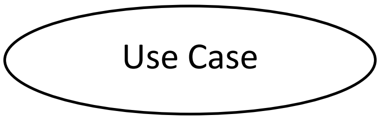
Actors: The users that interact with a system. An actor can be a person, an organization, or an outside system that interacts with your application or system. They must be external objects that produce or consume data. [17]

the representation of actor:



Use cases: characterize the interactions that occur between actors and IT systems while the execution business processes

the representation of use case:



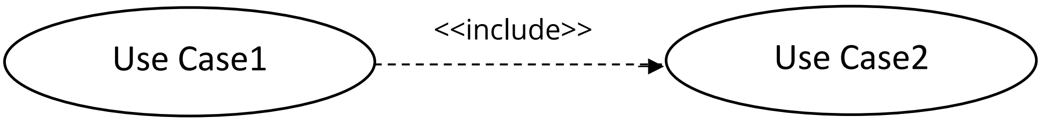
System boundary: A system boundary defines the scope of what a system will be. It represented by box that sets a system scope to use cases.

5.3. Relationships in Use Cases:

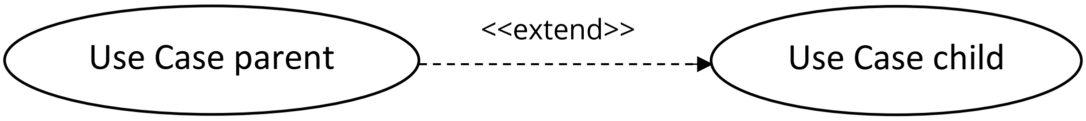
There are different types of relationships that exist between use cases. A relationship between two use cases is essentially a dependency between the two.

Relationships between use cases can be one of the following: [18]

Include: When a use case is depicted as using the functionality of another use case in a diagram, this relationship between the use cases is named as an include relationship.



Extend: In an extend relationship between two use cases, the child use case adds to the existing functionality and characteristics of the parent use case.



5.4. System interaction Use Case:

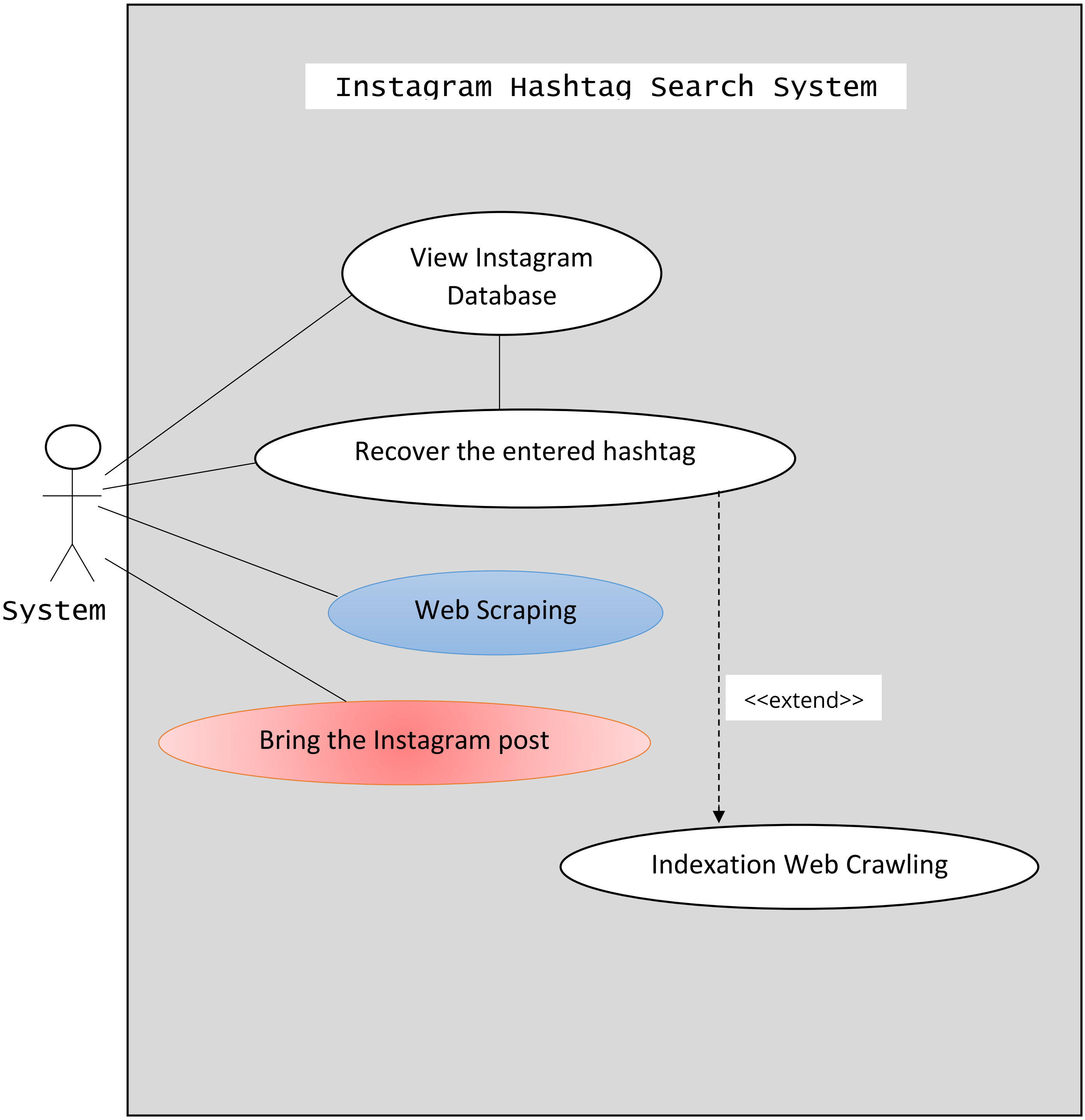


Figure 2: System interaction Use Case Diagram

5.5. User interaction Use Case:

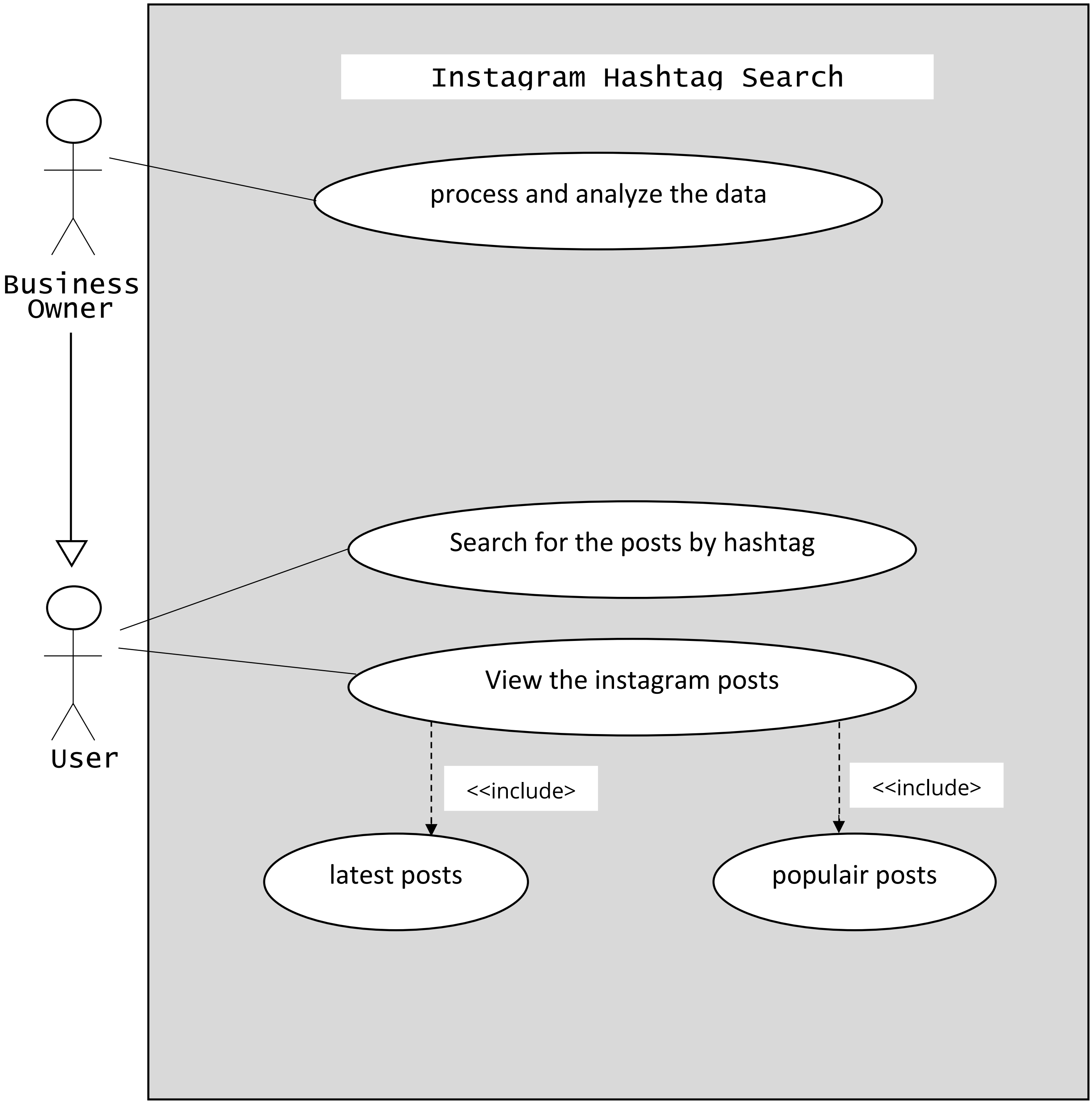


Figure 3: User interaction Use Case Diagram

6. Sequence diagram:

6.1. Definition:

[19]

The Sequence Diagram illustrates object collaboration on time sequence perspective. It demonstrates how objects interact with one another in a certain use case scenario.

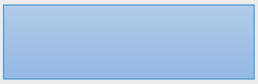
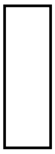
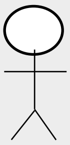
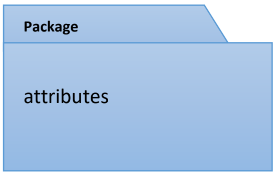
A sequence diagram is an interaction diagram that details how operations are performed: what messages are sent and when they are sent. The emphasis is on communication.

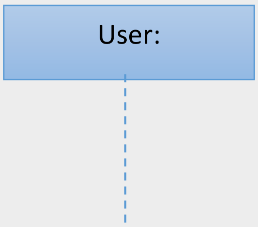
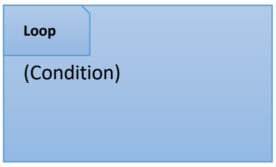
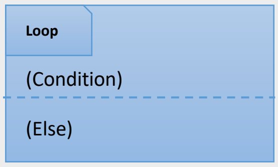
So, sequence diagram has Two dimensions:

- Time
- Objects

6.2. Sequence diagram symbols and components:

A sequence diagram shows, as parallel vertical lines (lifelines), different processes or objects that live simultaneously, and, as horizontal arrows, the messages exchanged between them, in the order in which they occur. This allows the specification of simple runtime scenarios in a graphical manner**.** [20]

Symbol	Name	*Description*
	Object symbol	In UML, this represents a class or object. The object symbol displays how an object will interact with the rest of the system. This shape should not include any class attributes.
	Activation box	Represents The amount of time it takes an object to execute a task. The activation box will grow longer as the work progresses.
	Actor symbol	Indicates the user that interact with the system or the entities that are external to the system.
	Package symbol	this element is used to include interactive diagram elements. This rectangular form, often known as a frame, features a small inner rectangle for labeling the diagram. (Used in UML 2.0)

	Lifeline symbol	It represents the passage of time and it extends downward as the time goes onward. The successive events that occur to an object during the charted process are shown by this dashed vertical line. Lifelines may begin with an object symbol or an actor symbol.
	Option loop symbol	It used for modeling if/then situations, which means a situation that will only happen if specific circumstances are met.
	Alternative symbol	Signifies a decision between two or more message sequences, which mean you have to choose one of the conditions. It designated by rectangle shape with a dashed line within to represent options.

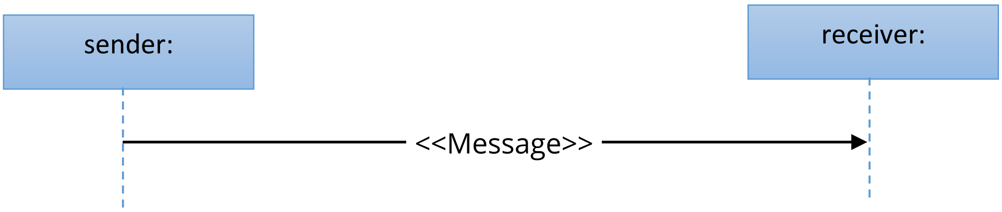
6.3. Types of Messages in Sequence Diagrams:

Message Arrows: In a sequence diagram, an arrow from the Message Caller to the Message Receiver specifies a message.

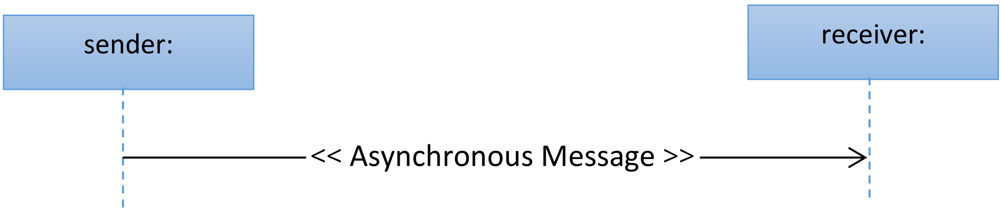
A message can be sent in any way, such as left to right, right to left, and back to the Message Caller. While the arrow can be used to describe the message being delivered from one object to another, different arrowheads can be used to signify the sort of message being delivered or received.

Here is the type of arrows and their functionality in the sequence diagram:

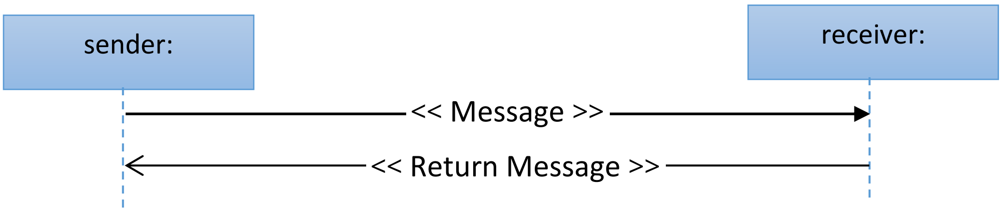
Synchronous message: A solid line with a solid arrowhead is used to represent this arrow. You use this type of arrow when you send a message and you must wait for a reply before proceeding. Both the call and the reply should be shown in the diagram.



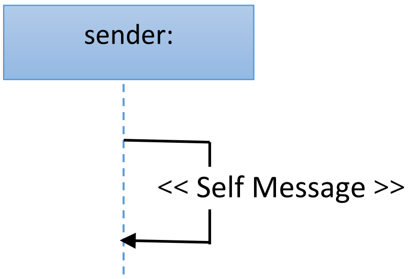
Asynchronous message: A solid line with a lined arrowhead is used to represent this arrow. The sender of asynchronous message does not need to wait for a response before proceeding. So, in this case only the call should be included (one way message).



Return message: A dashed line with a lined arrowhead represents this arrow. A return message indicates that the message receiver has completed processing the message and is handing control back to the message caller.



Self-Message: A solid line with a lined arrowhead is used to represent this arrow and it goes from the sender to itself. It is referred by a message arrow that begins and ends on the same lifeline.



6.4. System web scraping sequence:

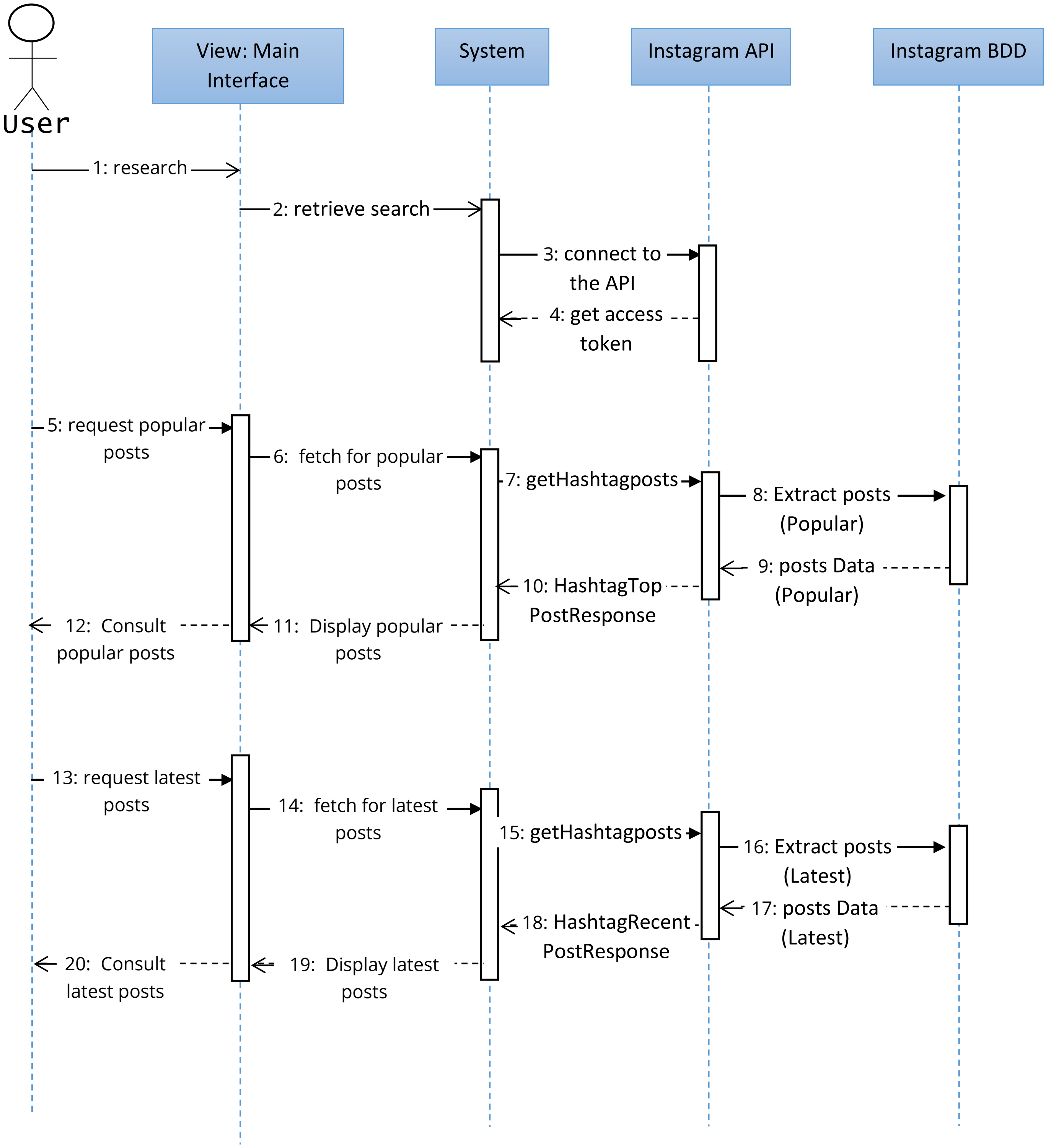


Figure 4: System web scraping sequence diagram (search by hashtag the latest and most popular posts in Instagram)

7. Class diagram:

7.1. Definition:

Almost all object-oriented methods use the class diagram as a key modeling methodology. The types of items in the system and the many sorts of static relationships that exist between them are shown in this diagram.

So, in general class diagram explain how to design a system from three separate perspectives: conceptual, specification, and implementation.

Any organization can benefit from class diagrams in a variety of ways, such as:

- Establish data models for information systems, regardless of how simple or complicated they are.
- Obtain a better understanding of the overview of an application's schematics.

7.2. Class diagram components:

The class is represented as a rectangle with its name, attributes, and operations separated into different containers.

The following are essential features of a UML class diagram:

Class Name: The class name is primarily required for the class's graphical representation. It can be found in the uppermost container.

These rules must be taken by consideration in the creation of class name:

- A capital letter should always be used to begin a class name.
- The first and uppermost compartment should always include a class name in the centre alignment.
- The name of the class should always be printed in bold.

Attributes: A class's attribute is a named parameter that indicates that the object is being modeled. This component is located directly below the class name container in the class diagram (Middle section).

The attributes characteristics are:

- Attributes must have a descriptive name that explains how they are used in a class.
- In most cases, the attributes are written alongside the visibility factor.
- The accessibility of a class attribute is expressed by visibility.
- The four visibilities are: public, private, protected, and package. characterized by +, -, #, ~ respectively.

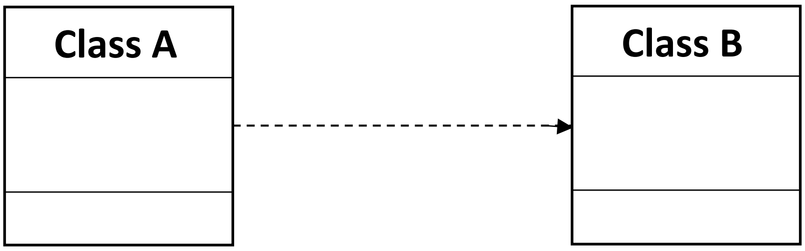
Operations: A class's operation Also known as methods, each operation takes its own line, shown in list format. The operations focus on the interaction between a class and data. This component is located in the lowermost container.

7.3. Relationships in Class Diagram:

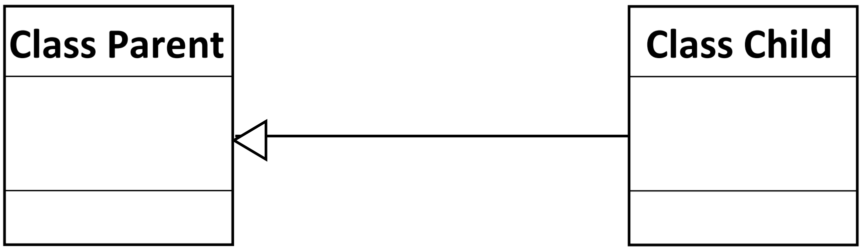
[21]

Dependency: dependency is the relationship between two or more than two classes in which a change in one of these can force changes in the other.

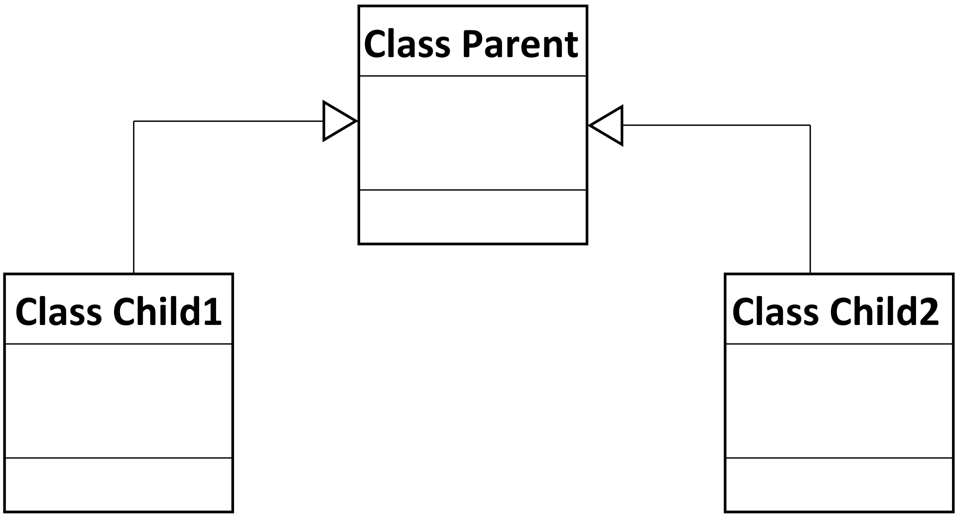
However, a weaker relationship will always be created. Dependency demonstrates that one class depends on a different class.



Generalization: A generalization helps a subclass (child) to connect to the superclass(parent). Which means that the subclass is inherited from the superclass. Generalization cannot be used to model implementation of the interface.



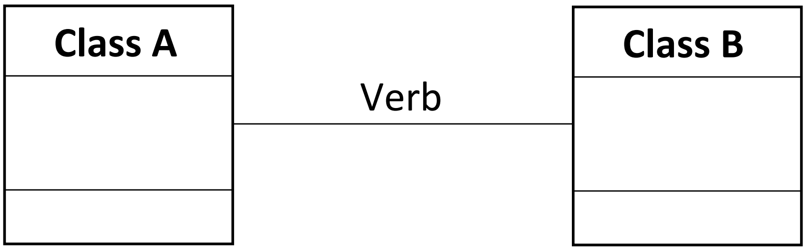
Class diagram enables multiple superclass heritage.



Association: This kind of relationship is used to defines relationships that are static between classes A and B.

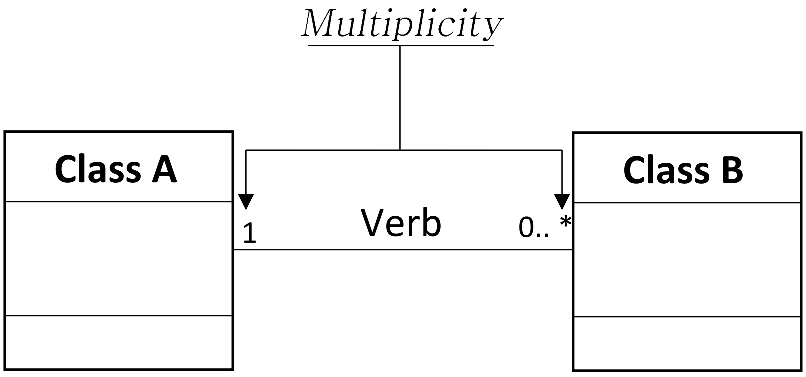
Here are some Association rules that should be followed in class diagram:

- The name of an association should be a verb or phrase that specifies the nature of the relationship between two classifiers.
- It should be given a name that reflects the function of the class at the receiving end of the association path.



Multiplicity: A multiplicity is a factor that is linked to an attribute. When a class is created, it indicates how many instances of attributes are created. If no multiplicity is given, one is assumed to be the default multiplicity.

PS: the symbol (*) represent that to many instances of attributes are created.



7.4. Class Diagram:

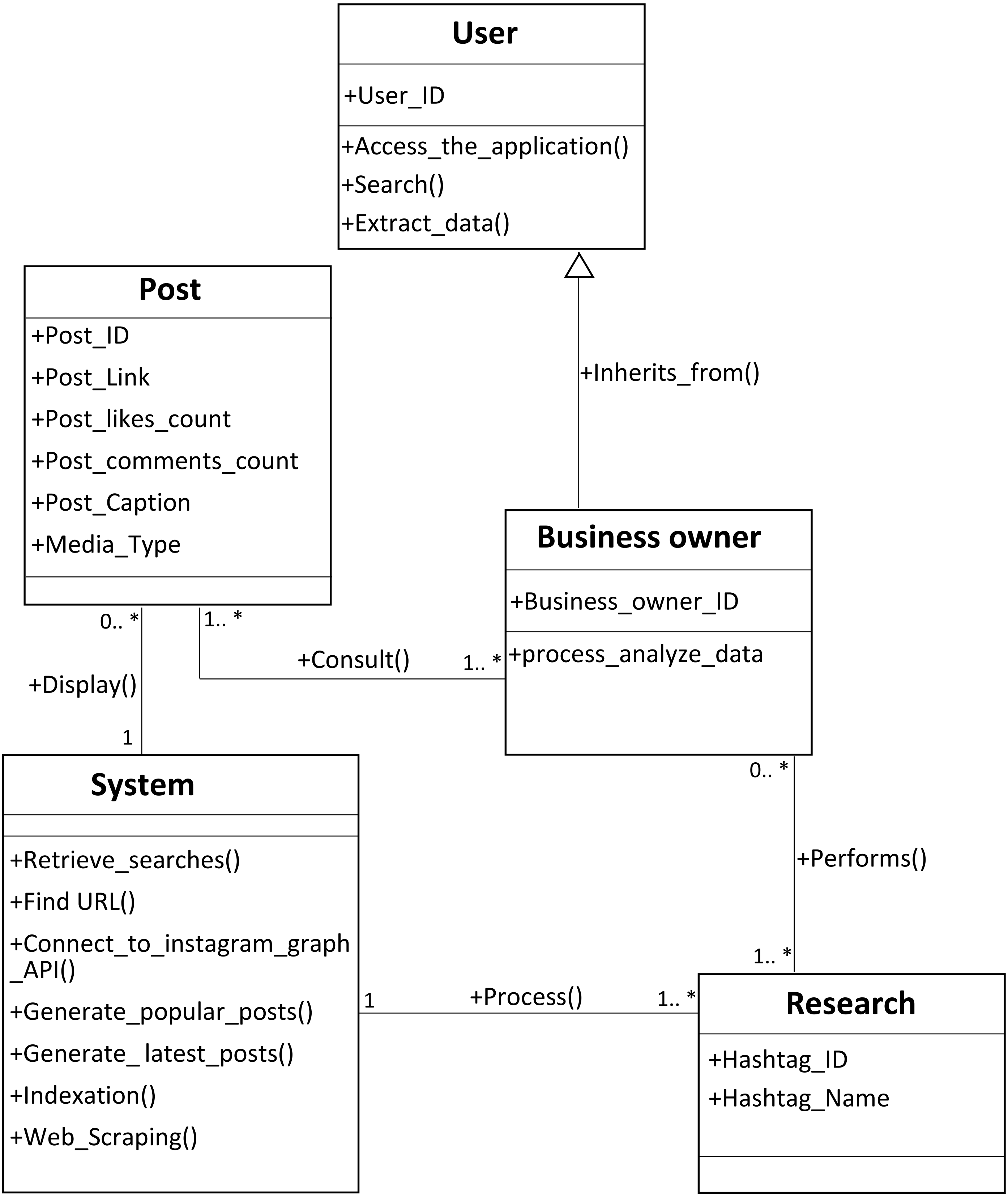


Figure 5: System web scraping class diagram (search by hashtag the latest and most popular posts in Instagram)

8. Conclusion:

To conclude, in this chapter there was the representation of the theoretical aspect of the application (Hashtag Search), or in another way the conceptional side.

in order to establish a good conception of the information collect and develop an efficiency and precision of location of the data I used the UML method for the detailed study.

Also, I have explained some method of UML based on the work of these methods and their definitions.

Finally, I chose three diagrams from the UML method according to my application aspect (Use Case, Sequence and Class diagrams), to show in details how does my application works.

Chapter 4: Implementation and Results

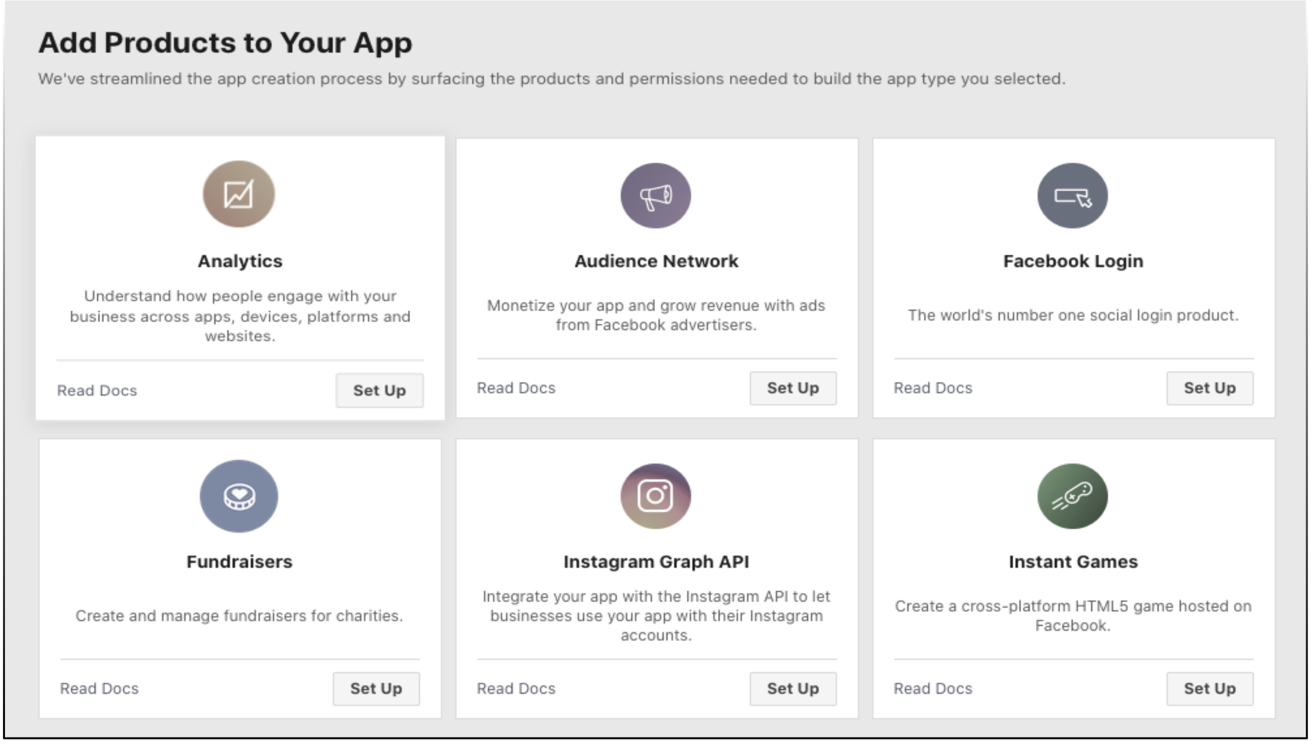
1. Introduction:

As we already know the web is incredibly large and ever-changing source of diverse types of data - some beneficial and some not -.

This data can be full with information that we can have a lot use of it, but in order to gain this source of power we need to collect and keep this data for future processing and analysis. Doing it manually would probably take a lot of time and efforts, to facilitate this operation some websites and companies provide public avenues (APIs) for interested parties to log in and request data such as (Instagram API).

2. Instagram graph API:

The Instagram Graph API allows Instagram Professionals — Businesses and Creators — to use your app to manage their presence on Instagram. The API can be used to get and publish their media, manage and reply to comments on their media, identify media where they have been @mentioned by other Instagram users, find hashtagged media, and get basic metadata and metrics about other Instagram Businesses and Creators. [22]



Add Products to Your App

We've streamlined the app creation process by surfacing the products and permissions needed to build the app type you selected.

Analytics

Understand how people engage with your business across apps, devices, platforms and websites.

Audience Network

Monetize your app and grow revenue with ads from Facebook advertisers.

Facebook Login

The world's number one social login product.

Fundraisers

Create and manage fundraisers for charities.

Instagram Graph API

Integrate your app with the Instagram API to let businesses use your app with their Instagram accounts.

Instant Games

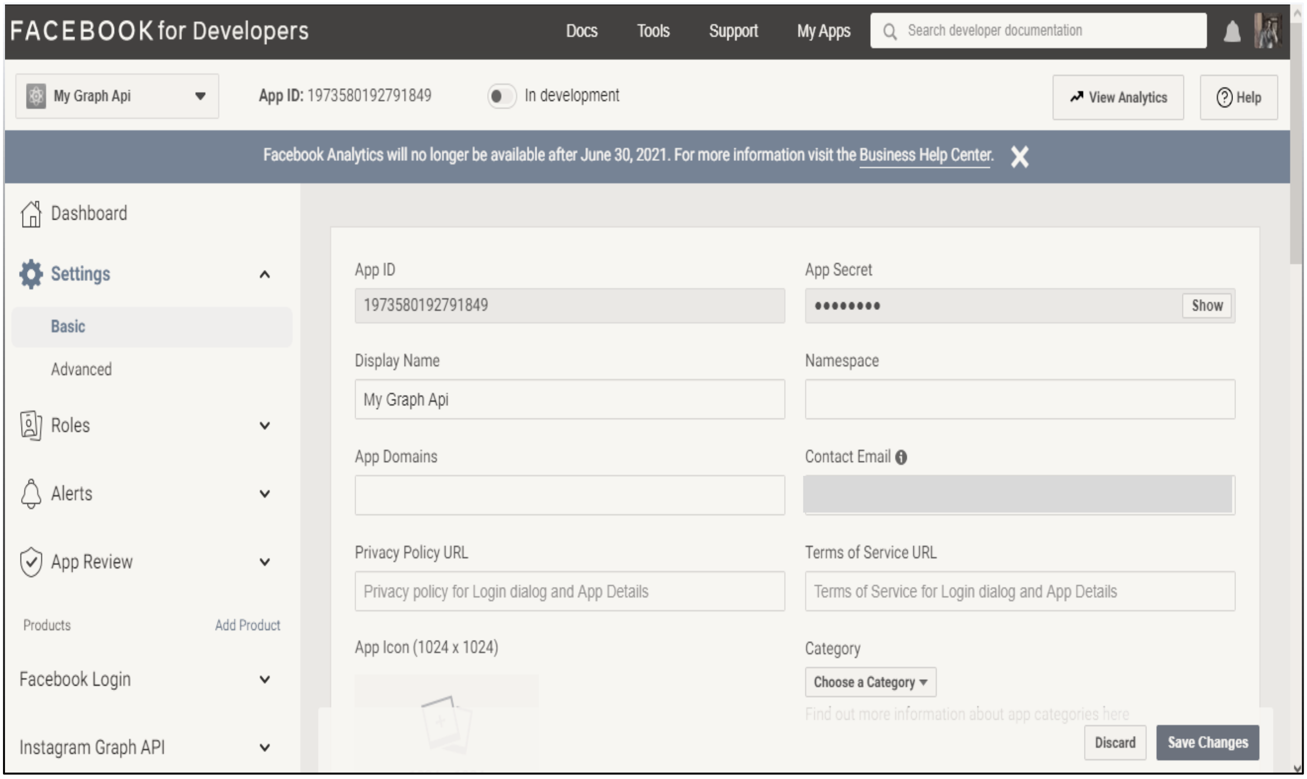
Create a cross-platform HTML5 game hosted on Facebook.

Figure 1: Facebook App Dashboard

2.1. Creation of Facebook Developers Application:

In order to get access to the Instagram API the only portal you have is through Facebook Developers by creating application that allows you to explore the graph API.

To access your application, you will have a unique App ID and App secret.



App ID: 1973580192791849 In development

Facebook Analytics will no longer be available after June 30, 2021. For more information visit the Business Help Center.

App ID: 1973580192791849

App Secret: ••••••••

Display Name: My Graph Api

Namespace:

App Domains:

Contact Email:

Privacy Policy URL: Privacy policy for Login dialog and App Details

Terms of Service URL: Terms of Service for Login dialog and App Details

App Icon (1024 x 1024)

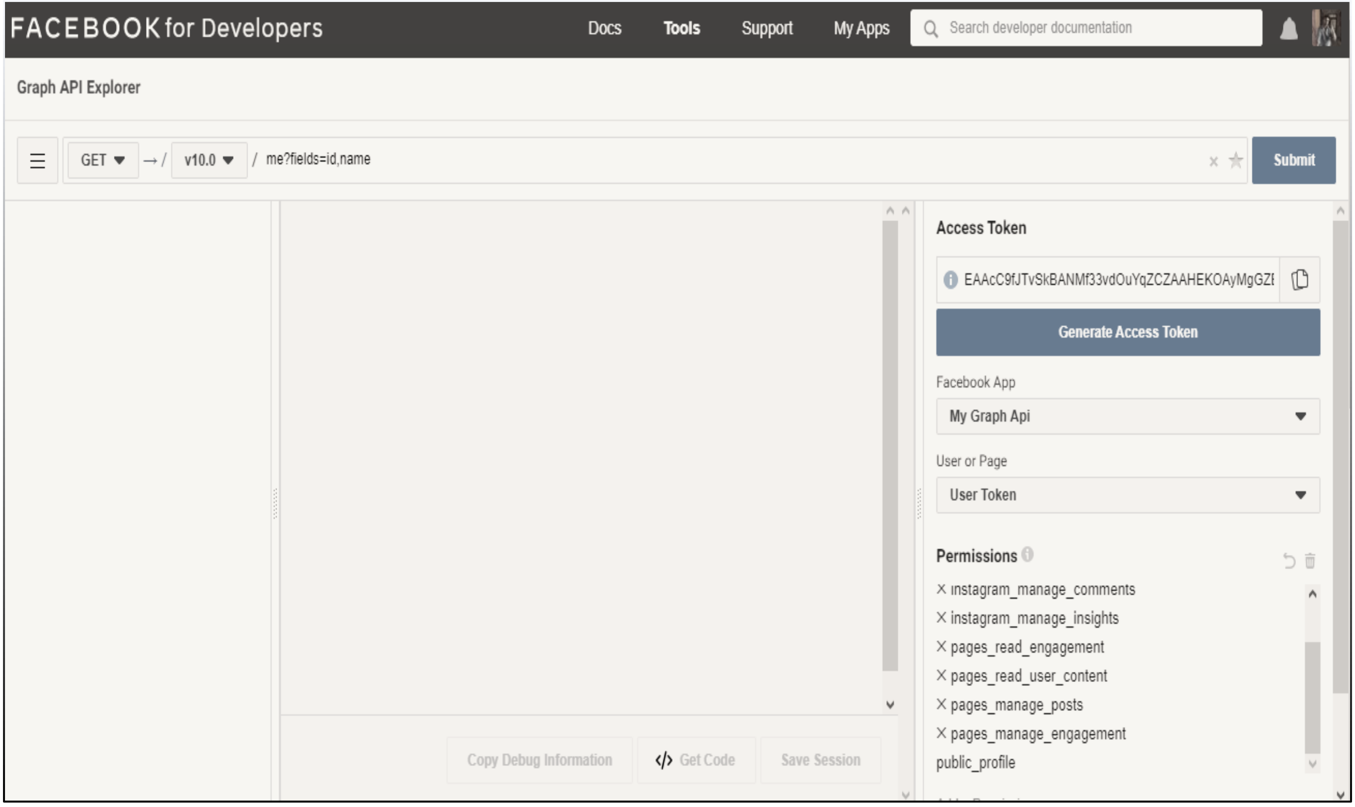
Category

Find out more information about app categories here

Figure 2: Facebook Developers Application (My Graph API)

After the creation of the application now you can explore the graph API with the given tool by Facebook Developers (Graph API Explorer) and add Permissions that will provide a way for your app to access data from Instagram.

The application will also provide you with Access Token, that will expire in one day, this Access Token is the key to scrapping Instagram because it contains all the information of the graph API.



GET → / v10.0 / me?fields=id,name

Access Token: EAAcC9fJTvSkBANMf33vdOuYqZCZAAHEKOAyMgGZI

Facebook App: My Graph Api

User or Page: User Token

Permissions

× instagram_manage_comments
× instagram_manage_insights
× pages_read_engagement
× pages_read_user_content
× pages_manage_posts
× pages_manage_engagement
public_profile

Figure 3: Graph API Explorer

Permissions: Obtaining permissions for your app entails selecting the permissions that your app requires in order to function properly.

the permissions my app needs to function:

- pages_show_list
- instagram_basic
- instagram_manage_comments
- instagram_manage_insights
- pages_read_engagement
- pages_read_user_content
- pages_manage_posts
- pages_manage_engagement
- public_profile

3. Technical presentation of the "Hashtag Search" application:

The Hashtag Search application is an Instagram posts extraction application using the web scraping technique that I have developed to meet the needs of the customer and specially to respond to our initial problem, I have detailed the theoretical and conceptual side of the application in the previous chapter and now in this chapter I will approach the technical aspect of our data extraction system.

3.1. Functional path:

The Hashtag Search is an application that allows the user to search the posts of Instagram through hashtag names and get the latest and the most popular posts in the platform.

The application gives the user a narrow view on the fashion industry and the advantage to keep being updated to the latest trends and what are the people's opinion in the fashion industry.

1. The user enters the application and types the name of the hashtag.
2. Then the user has the choice of hitting the latest posts or popular posts button.
3. Then the application will generate the posts according to the user choice.
4. The user can choose the link that has been given by the application and access the Instagram post.

4. Development environment:

4.1. Programming languages:

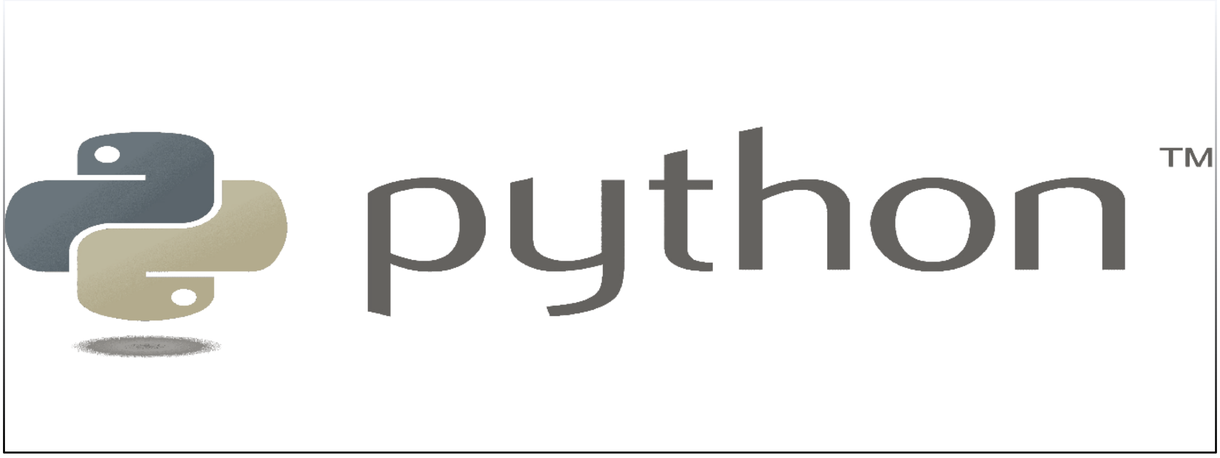
Python: Python is an interpreted, interactive, object-oriented programming language. It incorporates modules, exceptions, dynamic typing, very high-level dynamic data types, and classes. It supports multiple programming paradigms beyond object-oriented programming, such as procedural and functional programming. Python combines remarkable power with very clear syntax. It has interfaces to many systems calls and libraries, as well as to various window systems, and is extensible in C or C++. It is also usable as an extension language for applications that need a programmable interface. Finally, Python is portable: it runs on many Unix variants including Linux and macOS, and on Windows. [23]

Quotes about Python:

Python is used successfully in thousands of real-world business applications around the world, including many large and mission critical systems. Here are some quotes from happy Python users:

"Python is fast enough for our site and allows us to produce maintainable features in record times, with a minimum of developers," *said Cuong Do, Software Architect, YouTube.com.*

"Python has been an important part of Google since the beginning, and remains so as the system grows and evolves. Today dozens of Google engineers use Python, and we're looking for more people with skills in this language." *said Peter Norvig, director of search quality at Google, Inc.* [24]



4.2. Technologies libraries and Packages:

JSON:

json — JSON encoder and decoder: **Source code:** Lib/json/__init__.py

JSON (JavaScript Object Notation), specified by RFC 7159 (which obsoletes RFC 4627) and by ECMA-404, is a lightweight data interchange format inspired by JavaScript object literal syntax (although it is not a strict subset of JavaScript 1).

json exposes an API familiar to users of the standard library marshal and pickle modules. [25]

Requests:

Requests is a simple, yet elegant HTTP library. Requests allows you to send HTTP/1.1 requests extremely easily. There's no need to manually add query strings to your URLs, or to form-encode your PUT & POST data — but nowadays, just use the json method! [26]

PyQt:

PyQt is a set of Python bindings for The Qt Company's Qt application framework and runs on all platforms supported by Qt including Windows, macOS, Linux, iOS and Android. PyQt6 supports Qt v6, PyQt5 supports Qt v5 and PyQt4 supports Qt v4. The bindings are implemented as a set of Python modules and contain over 1,000 classes. [27]

PyQt5 comprises PyQt5 itself and a number of add-ons that correspond to Qt's additional libraries. Each is provided as a source distribution (sdist) and binary wheels for Windows, Linux and macOS. [28]

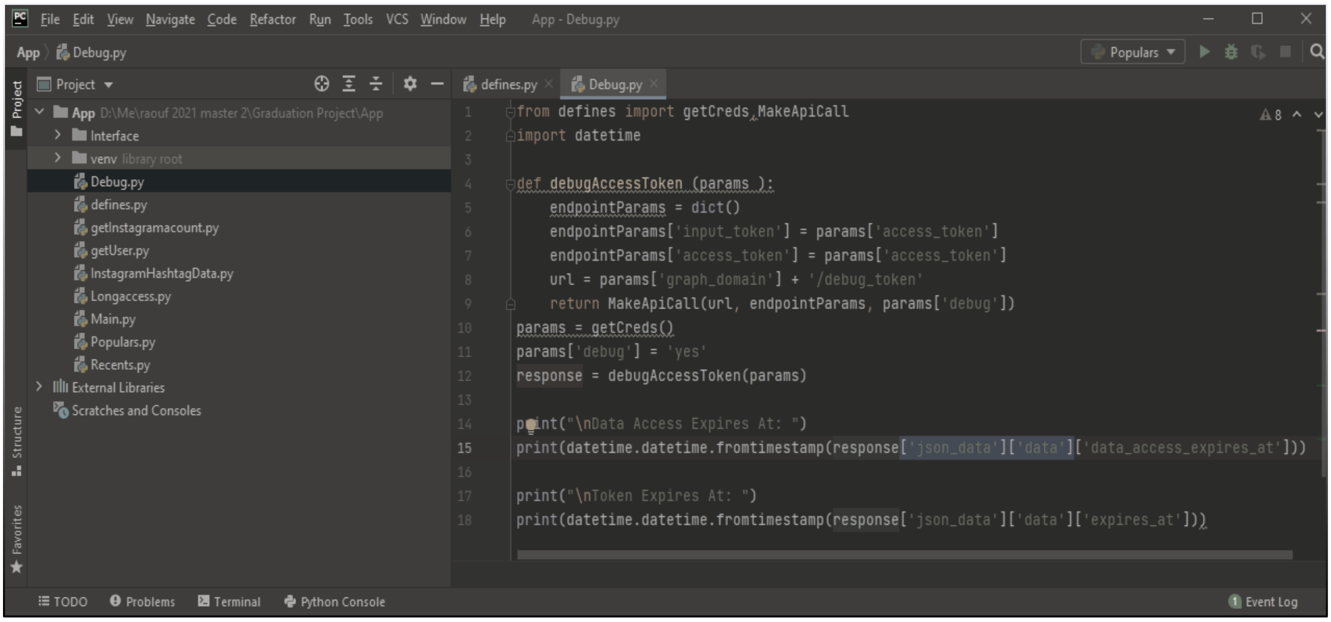
4.3. Development tool:

PyCharm:

PyCharm is an integrated development environment (IDE) used in computer programming, specifically for the Python language. It is developed by the Czech company JetBrains (formerly known as IntelliJ).

It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django as well as data science with Anaconda.

PyCharm is cross-platform, with Windows, macOS and Linux versions. The Community Edition is released under the Apache License, and there is also Professional Edition with extra features – released under a proprietary license. [29]



```python
from defines import getCreds,MakeApiCall
import datetime

def debugAccessToken (params ):
    endpointParams = dict()
    endpointParams['input_token'] = params['access_token']
    endpointParams['access_token'] = params['access_token']
    url = params['graph_domain'] + '/debug_token'
    return MakeApiCall(url, endpointParams, params['debug'])
params = getCreds()
params['debug'] = 'yes'
response = debugAccessToken(params)

print("\nData Access Expires At: ")
print(datetime.datetime.fromtimestamp(response['json_data']['data']['data_access_expires_at']))

print("\nToken Expires At: ")
print(datetime.datetime.fromtimestamp(response['json_data']['data']['expires_at']))
```

Figure 4: PyCharm Community Edition 2020.3

5. Algorithm of the operation:

```
Begin
    Retrieve the user search
    (Exemple: [hashtag_name] = ‘fashion’ )
    For each target Instagram URL
        Attach hashtag search to URL
        Execute the search
        Retrieve the request response (MakeApiCall)
        Extract data to text (getCreds)
    End For
    For each URL search result
        Represent structured data based (JSON).
        For each post
            If User chose the popular posts then
                [post_type] = ‘top_media’
                print(″------ Post Info ------″)
                print(″Link to post: ″ + post[’permalink’])
                print(″Likes: ″ + post[’like_count’])
                print(″Comments: ″ + post[’comments_count’]
                print(″Post Caption: ″ + post[’caption’])
                print(″Media Type: ″ + post[’media_type’])
            End If
            If User chose the recent posts then
                [post_type] = ’recent_media’
                print(″------ Post Info ------″)
                print(″Link to post: ″ + post[’permalink’])
                print(″Likes: ″ + post[’like_count’])
                print(″Comments: ″ + post[’comments_count’]
                print(″Post Caption: ″ + post[’caption’])
                print(″Media Type: ″ + post[’media_type’])
            End If
        End For
    End For
    Show all posts found with respective information
End
```

6. Demonstration and result of the application:

6.1. The main interface:

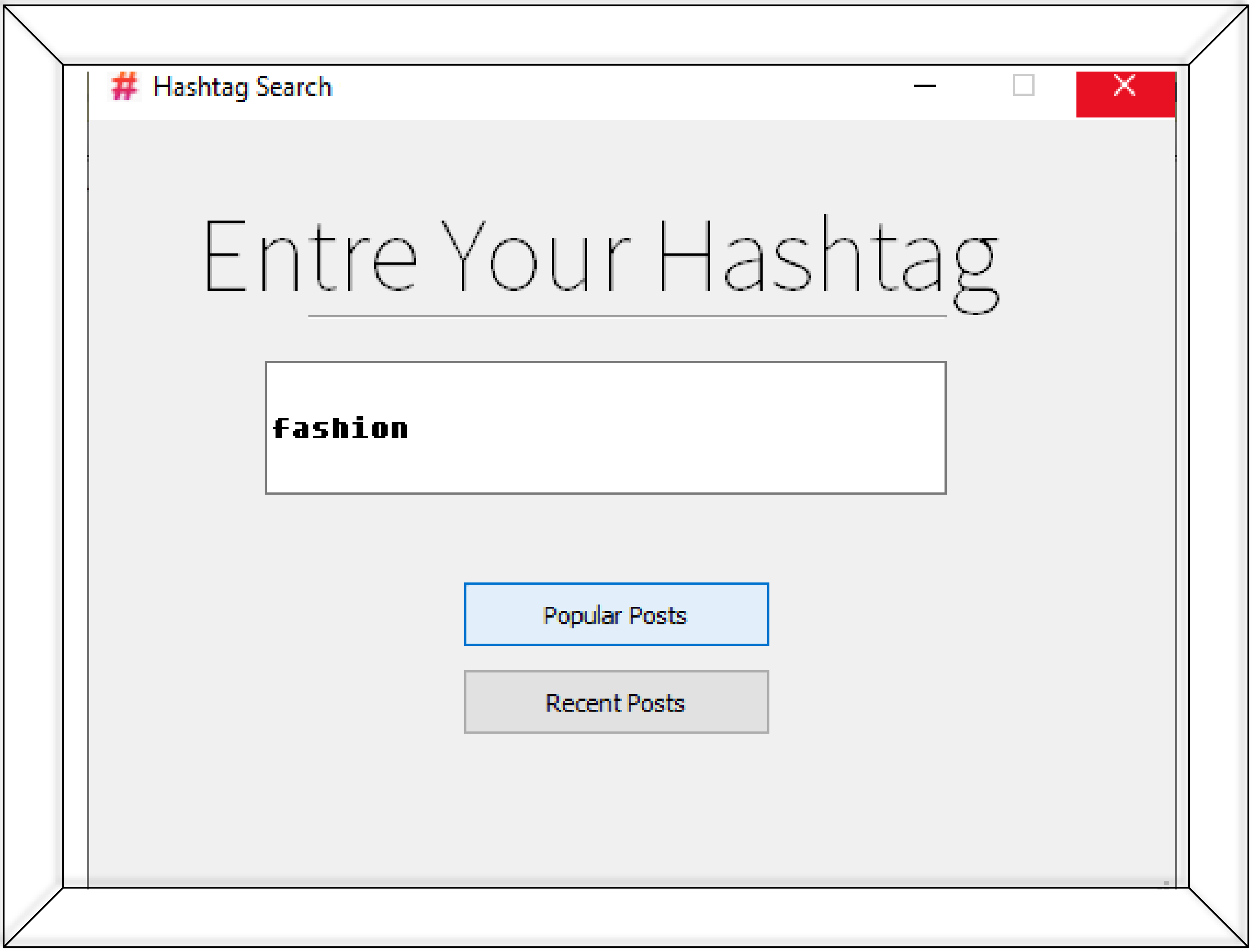
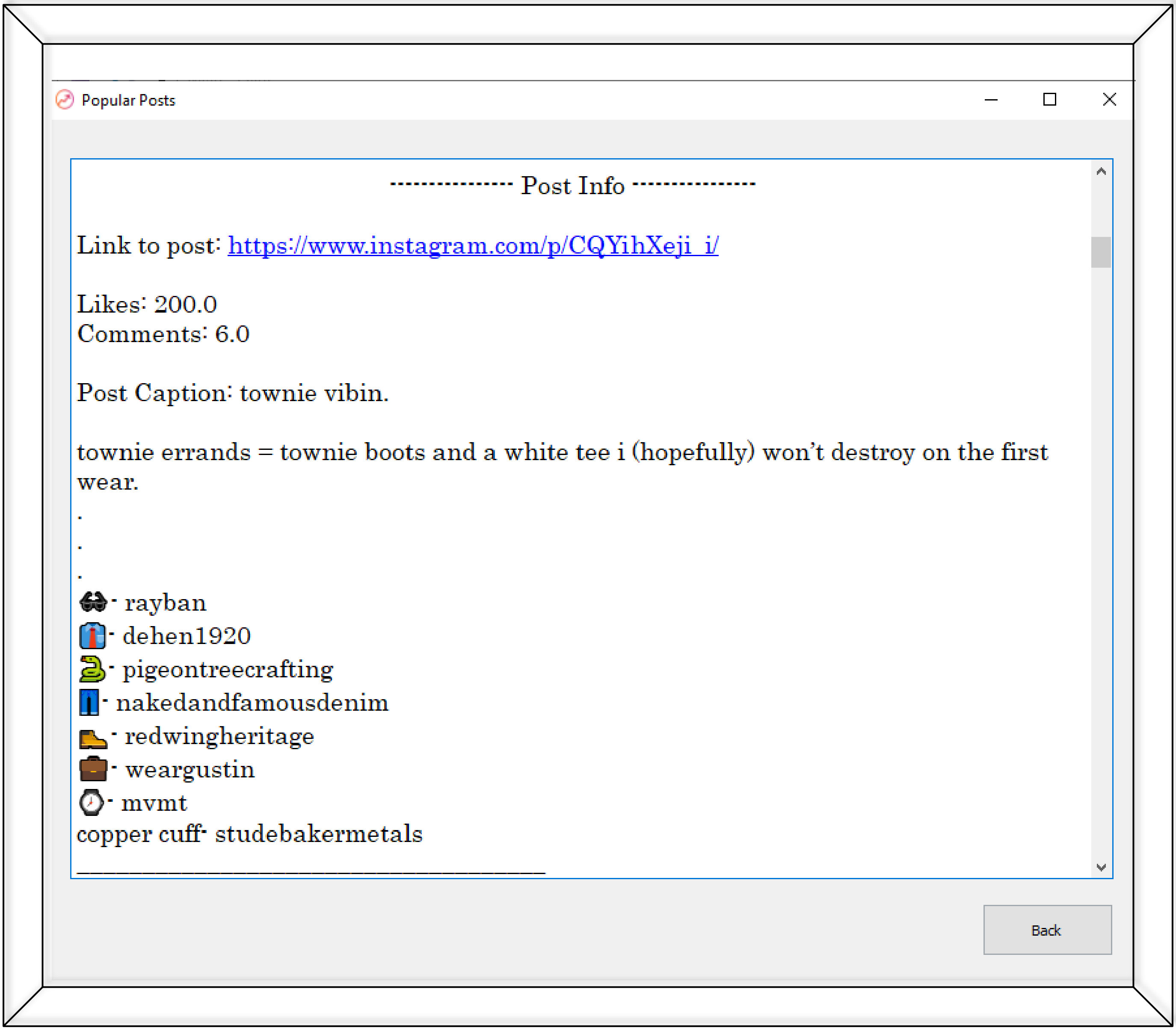


Figure 5: The main interface of the application

6.2. The popular posts interface:



--------------- Post Info ---------------

Link to post: https://www.instagram.com/p/CQYihXeji_i/

Likes: 200.0
Comments: 6.0

Post Caption: townie vibin.

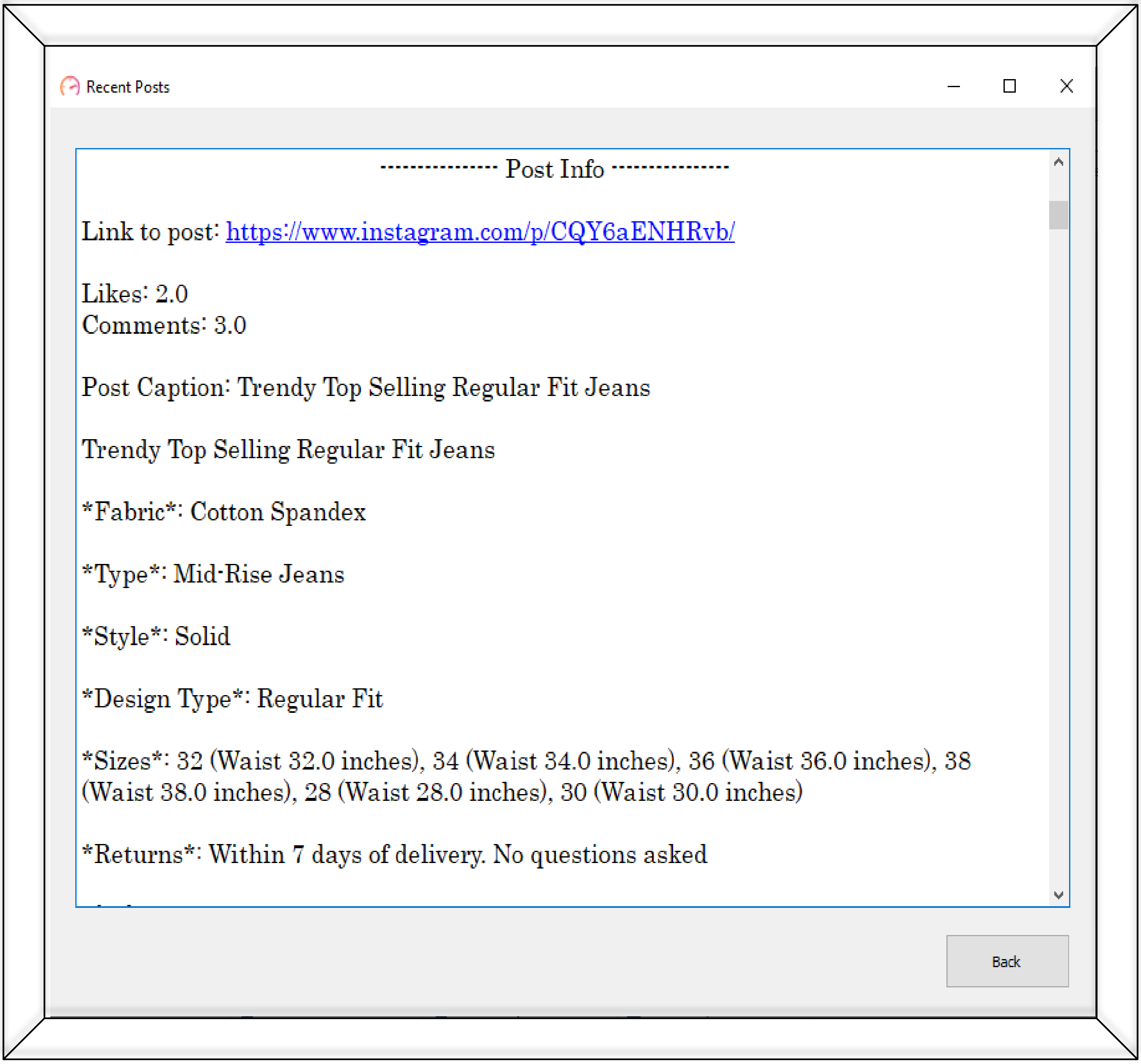
townie errands = townie boots and a white tee i (hopefully) won't destroy on the first wear.

.
.
.
🕶- rayban
👔- dehen1920
🐍- pigeontreecrafting
👖- nakedandfamousdenim
🥾- redwingheritage
💼- weargustin
⌚- mvmt
copper cuff- studebakermetals

____

Figure 6: The results of a hashtag search (popular posts)

6.3. The recent posts interface:



--------------- Post Info ---------------

Link to post: https://www.instagram.com/p/CQY6aENHRvb/

Likes: 2.0
Comments: 3.0

Post Caption: Trendy Top Selling Regular Fit Jeans

Trendy Top Selling Regular Fit Jeans

Fabric: Cotton Spandex

Type: Mid-Rise Jeans

Style: Solid

Design Type: Regular Fit

Sizes: 32 (Waist 32.0 inches), 34 (Waist 34.0 inches), 36 (Waist 36.0 inches), 38 (Waist 38.0 inches), 28 (Waist 28.0 inches), 30 (Waist 30.0 inches)

Returns: Within 7 days of delivery. No questions asked

Figure 7: The results of a hashtag search (recent posts)

7. Conclusion:

In this chapter, I have presented my web scraping application as well as the development and implementation environment, focusing on the techniques and programming tools used to create a system that aims to extract Instagram posts in two different types the popular ones and the latest ones to offer them to the user.

Finally, I ended this chapter by demonstrating my work supported by some interfaces that show how the application function.

GENERAL CONCLUSION

We are living in an era where the web is growing at an insane speed and of course, all of the technologies that surround it are advancing at the same rate, as we have seen in the devolvement of Information research systems.

As I have mentioned in my thesis on the role of data in all the domains specially in the clothing industry, data is considered as a source of power in today's world. Companies have now understood that the internet can be a great means of expansion, it has also been my main occupation of my project on to think about how to extract and process specific data using several techniques.

In order to dominate in the market specifically in garments industry, you have to be aware of what people want and what's trendy cause people tend to follow trends so they don't experience the fear of missing out, that's why you need to gather data from the web to be exact from the social media due to the "fashion people" located there (Instagram).

Web scraping, crawling is only at the start of certain growth. It is enough to observe the number of projects, articles and conferences on the subject to understand that a phenomenon emerges.

The Web is certainly not being used to its fullest today. However, finding the data that only concern you can help you gain both time and storage. The future tools developed will have to deal with an incalculable amount of data. We can already think that the provision of information through APIs provided will be an efficient way to manage this mass of data.

Finally, scraping is certainly to be done on the regulatory manner to clarify the situation and make it possible to defend against illegal attacks.

In the first chapter, I defined the terms (social media, marketing, clothing industry) and explained how social media process flow, I also enumerated both the problematic and the objective of my thesis then I approached how social media contribute in the marketing of a business, finally I established how extracting and collecting data (web scraping) can have a huge impact in garments industry.

In the second chapter, I briefly defined Information research system with its different research strategies then I summed up its architecture with a schema, next I dealt with the proposed model of my thesis which is web scraping from its definition to how it works then to sums it up with its architecture, in addition I have cited the different objectives of the web scrapping in my thesis theme.

In the third chapter, I first dealt with conceptual method I have chosen which is UML touching several aspects of this method, after the In-depth explanation of the UML I showed the importance and advantages of this method in the conception of my application that extract posts from Instagram according to a hashtag that have relation with the clothing industry to facilitate the understanding of the market and offer the best choices to the business owner.

I presented in the second part of this chapter my post extraction system with a conceptual study which is the starting point of my post extraction application called “The Hashtag Search” application.

Finally in the fourth and last chapter, I presented my scraping application as well as the development and implementation environment, in focusing on the programming techniques and tools used to create a system that aims to extract posts from Instagram in both types, posts that have many likes (popular posts) and posts that are up to date (recent posts).

Finally, I ended this chapter by demonstrating my work supported by some interfaces that show how the application function.

References

[1] *Wikipedia contributors. "Instagram." en.wikipedia.org/wiki/Instagram.*

[2] *pressbooks.senecacollege.ca › introbusinessbam107*

[3] *www.innovationintextiles.com › rss*

[4] *graduateway.com › analysis-of-textile-industry-essay*

[5] *https://roundpeg.biz/2010/12/social-media-organization-chart-strategy/*

[6] *Wikipedia contributors. "Keyword research." en.wikipedia.org/wiki/Keyword_research*

[7] *Onlinecourses24x7.com/udemy-100-free-web-scraping-for-beginners-withpython-scrapy-bs4/*

[8] *Mawloud Mosbah Distance Measurements in the Context of Content Image Search (CBIR). information research [cs.IR]. Université 20 août 1955 Skikda (Algérie), 2017. French. fftel-02948637f*

[9] *Wikipedia contributors. "Web scraping." en.wikipedia.org/wiki/Web_scraping*

[10] *https://celadonsoft.com/ai-ml/complete-beginners-guide-to-web-scraping*

[11] *https://www.xpcourse.com/uml-modeling-tutorial*

[12] *[Grady Booch, James Rumbaugh, Ivar Jacobson, UML User Guide, 2000 (ISBN 2-212-09103-6)]*

[13] *https://www.visual-paradigm.com/guide/uml-unified-modeling-language/what-is-use-case-diagram/]*

[14] *https://www.visual-paradigm.com/guide/uml-unified-modeling-language/what-is-uml/*

[15] *https://www.coursehero.com/file/p53caoa/*

[16] *https://www.slideshare.net/MarwaAliEissa/introduction-to-the-unified-modeling-language*

[17] *https://www.lucidchart.com/pages/uml-use-case-diagram*

[18] *https://www.developer.com/design/creating-use-case-diagrams/*

[19] *https://www.scribd.com/presentation/365368680/Software-Developmentlecture-6/]*

[20] *Wikipedia contributors. "Sequence diagram." en.wikipedia.org/wiki/Sequence diagram.*

[21] *https://www.coursehero.com/file/92725657/object-orienteddocx/*

[22] *https://developers.facebook.com/docs/instagram-api/*

References

[23] *https://docs.python.org/3/faq/general.html#what-is-python*

[24] *https://www.python.org/about/quotes/*

[25] *https://docs.python.org/3/library/json.html/*

[26] *https://pypi.org/project/requests/*

[27] *https://riverbankcomputing.com/software/pyqt/intro /*

[28] *https://www.riverbankcomputing.com/static/Docs/PyQt5/introduction.html#pyqt5-components/*

[29] *Wikipedia contributors. "PyCharm." en.wikipedia.org/wiki/PyCharm.*