Scientific Knowledge Base Creation via Intention Mining

Hamza Ali

Bachelor in Computer Science

Air University

Islamabad, Pakistan

[181058@students.au.edu.pk](mailto:181058@students.au.edu.pk)

Muhammad Huzaifa

Bachelor in Computer Science

Air University

Islamabad, Pakistan

[181122@students.au.edu.pk](mailto:181122@students.au.edu.pk)

**Solving a user’s query has been tough since the use of the internet, as the internet is the source of information and spams. For getting a solution, one relies on google but sometimes it gives false information via spamming tags. For that, one needs to have a search engine that would yield solutions from authentic sites and scholarly articles. For that, we would be using web crawling, NLP and numerous data processing python tools for giving authentic solutions for students, teachers and any other tech seekers. Our web-app would simply input search query along with certain categories like Linux, Linux → server (sub category). After the input field, user’s queries would be entertained. In some cases, if the tools fail to gather relevant data, queries can either be posted to our forums, which would be held till a certain time, and after then (unsolved) it will be posted to stack overflow. This process will give out authentic solutions to user’s queries.**

***Keywords— knowledge base, knowledge graphs, NLP(****key words****)***

# Introduction

Getting authentic information has been quite difficult now-a-days due to a large number of scams/ fake meta tags. The huge internet is useful if one knows how to use it for themselves. For a developer/ tech seeker, one would always consider relevant and up to date solutions for their queries, therefore they depend on getting that without wasting time in browsing through each link. Similarly, for newbies, they would get many links for their queries, each with relevant or fake meta tags, but they can’t distinguish between the both. In both cases, they want a platform that can provide authentic and up to date solutions to their problems. They just need to provide the queries and relevant categories to our server, and let the server handle the rest, which would crawl the web and academic datasets for the solutions. Provided the case that if the solution could not be found, then the queries (with unique tokens) would be posted to our forums where our experts will answer it for them.

# Background

Techseeker (seniors) use only reputed sites for their solution, be it the new technology or some help with the research, but for our juniors, they need certain guidance to distinguish between the reputed sites and unauthentic sites. For that, we are developing a knowledge base for scientific documents, which would include different tech terminologies, from AI, BigData, and Linux to certain small queries. A knowledge base is a self-serve customer service library that includes information about a product, service, or some specific topic. A knowledge base allows you to create self-service customer support content around recurring topics, issues, and themes. A knowledge base doesn’t simply have to be a reservoir of help center articles. It can also include functions like a frequently-asked questions (FAQ) section, a user forum, articles, white papers, how-to articles, video tutorials, case studies, and dictionaries or glossaries — essentially, anything that helps customers understand and use your products or services.

# Key Points

## Abbreviation

* KB → Knowledge Base
* NLP → Natural Language Processing
* UI → User Interface
* FAQ → Frequently Asked Questions
* FYP → Final Year Project

## Glossary

* Knowledge Base → A self-serve online library of information about a particular product, service, department, or topic.
* Knowledge Graph → A graph representing subject, action and object
* Stop words → common words that are unnecessary in building graphs like as, before, did, doesn’t, etc. For more info, see [stopwords](https://countwordsfree.com/stopword).

# goals and objective

By taking input and categories from the user, our goals will be:

* Extracting the required information for specific topic from available datasets,
* Extraction of entity pairs from grammatical patterns from text using NLP library SpaCy,
* Developing relationships among text,
* Knowledge graph formation

1. Working

Our main focus is to gather relevant data from the datasets that would be used to plot the knowledge graphs and then apply queries as per user’s request. Implementation of the knowledge graph is completed but the queries.

* 1. Data cleaning

First, we need to clean the data. For that, we would

1. Create token and replace unnecessary punctuations with null
2. Eradicate stopwords tokens from the rest.
   1. Finding entities

After data cleaning, we would get the different entities from cleaned data. Here, we would getting the subject and the object

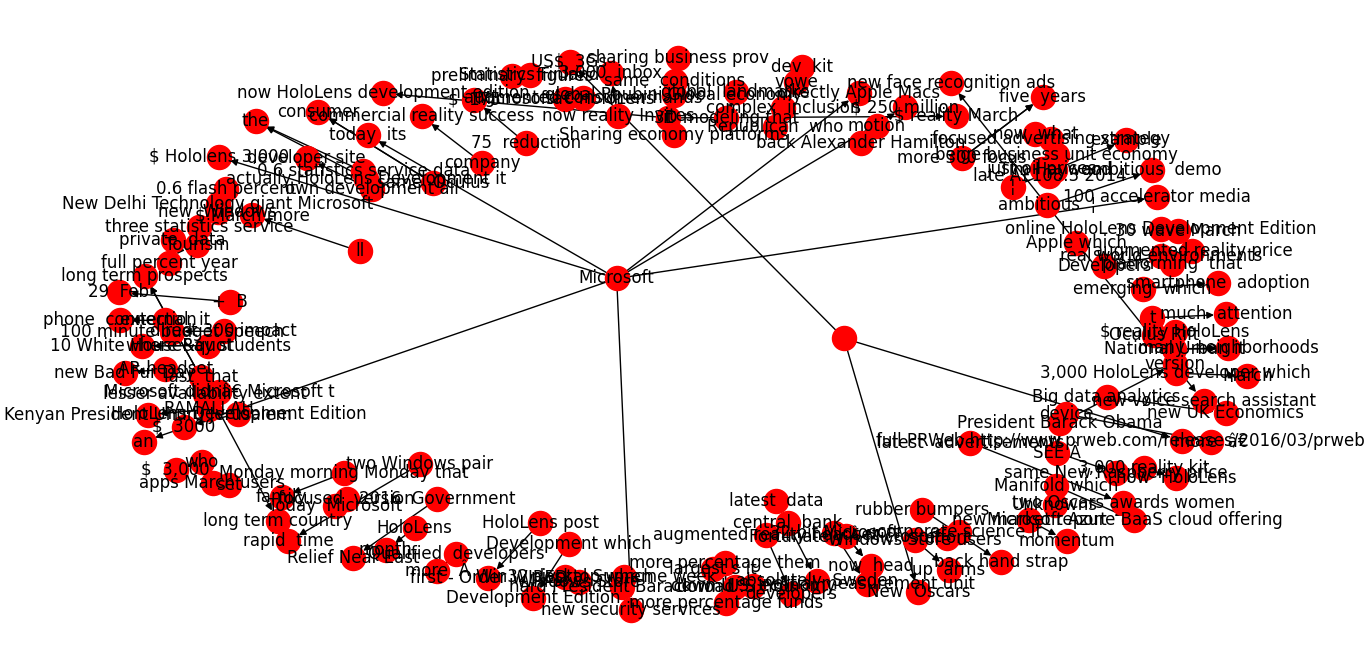
* 1. Finding relations

After finding the entities successfully, we would need to find the relation between those entities. For that. we would use matcher class, that would match the vocabulary and find the possible matches for those entities.

* 1. Create the Knowledge Graph

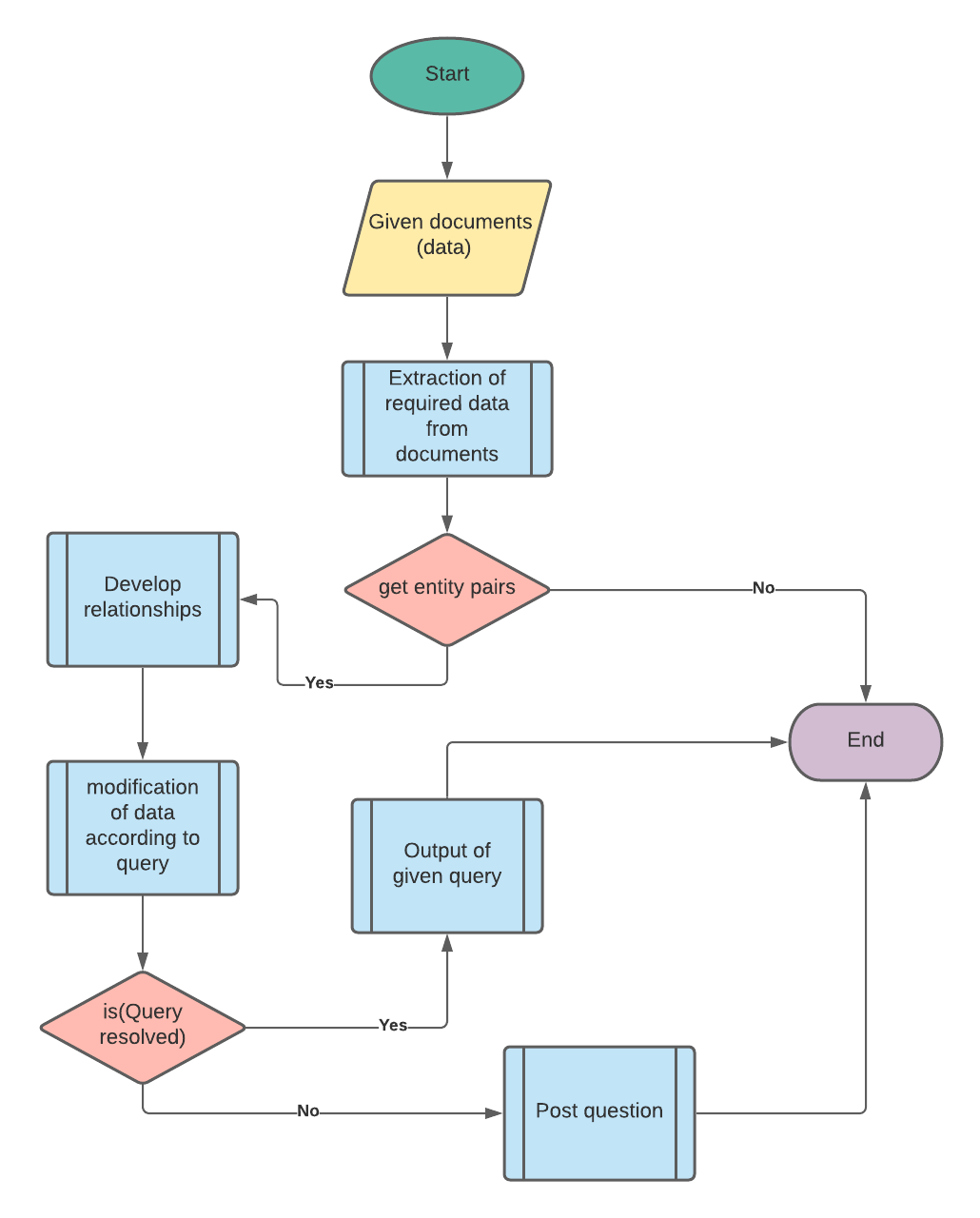
Next, we would be plotting the graph. We would need to create a list of source and target from entities, and edge from relation. For plotting, we would create a dataframe (using pandas library) of source and target.

* 1. Output



1. Flow Diagram

Here’s the visualization of the project. After the necessary knowledge graph, we would use query [4] entered by the user. If that query can be solved by the datasets, then it would be shown, else the question would be posted to a forum where users have to wait for the solution and would be notified if the solution is found.



1. Proposed Approach
   1. Data Set Exploration

in our dataset, we have found the following limitations

* Tackling of large
* Web Crawling (request approval)
* Faster execution with higher accuracy

1. Conclusion

Creation of knowledge base will not only benefit every student, or teacher but every tech geek around the internet. Searching out for any solution from installation and new technology is troublesome when one doesn’t know what the internet has to offer on the back of each website. There is a dire need for a platform that would provide solutions to all academic problems from its datasets and crawling features in a short span. For this, we will be creating a simple web-app with an easy UI that would only need user’s input and categories for optimizing search for web and datasets. Our datasets would only be the combination of recent tech-related articles, where users may search for different developing languages. For crawling, we will be utilizing the authentic and trust-worthy sites rated by many developers. Both of these features will be used to enhance the solution quality to each queries. There is a possibility where the internet and datasets fail to give out quality solutions, in that case, that query would be forwarded to the experts. In short, the user will need to input the query and it’ll be entertained with faster and accurate execution.

1. References

[1] Rabah A. Al-Zaidy C. Lee Giles Automatic Knowledge Base Construction From Scholarly Documents

[2] Deepanshi Dhingra <https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/>

[3] Aravind CR <https://neptune.ai/blog/web-scraping-and-knowledge-graphs-machine-learning>

[4] G. Xiao, L. Ding, B. Cogrel, & D. Calvanese. Virtual knowledge graphs: An overview of systems and use cases. Data Intelligence 1(2019), 201-223. doi: 10.1162/dint\_a\_000