# Time Series Based Text Summarization

1181203334, Mak Yen Wei, FYP1

# The Norm

# Finding the right answers on open forums is hard

- Around half of the searches don’t get the right answers (Microsoft CEO)
- Searching won’t be reliable without proper keywords.
- Particularly true in the **”Software Engineering“** domain.

-

# Hypothesis

# Faq Generation

- Leveraging summarization, topic modelling, sentiment analysis models.
- Put together a framework that uses these models to generate FAQs.

# Objective

# To Propose

A scalable framework
for FAQ Generation

## To Gather

Dataset from online forums in the software engineering domain

## To Evaluate

Multiple proposed architecture for FAQ Generation

# Scope

# Platforms

Stackoverflow (FYP1), Github, Reddit, Twitter (FYP2)

# Technologies

Python, FastAPI, ReactJs, Tensorflow, Keras

# Language

English

# Literature Review

# FAQ Processing

**Problem: Traditional methods are based on the use of extensive manual classification, it takes tremendous amount of time and effort.** (Raazaghi, 2015) (Hu, Yu, & Jiau, 2010) (Henß, Monperrus, & Mezini, 2012) (Razzaghi, Minaee, & Ghorbani, 2016).

**Open Source Forums** (Hu et al., 2010) (Makino, Noro, & Iwakura, 2016)

- A great source of information for questions findings
- Question abstraction or question finding can be done in these forums

**Neural Networks** (Duan, Tang, Chen, & Zhou, 2017)

- CNN and RNN have been researched to handle questions generation and questions retrieval tasks.

**Context Matching**

- Different formats of words might have the same meaning. (Makino et al., 2016) (Kothari, Negi, Faruquie, Chakaravarthy, & Subramaniam, 2009)
- For EG: (What should i do when my credit card is stolen vs How do i report a stolen credit card)
- Two of the sentences might seems drastically different but in reality in the terms of FAQ, the answer will be the same.
- Levenshtein distance is a traditional and naive way of scoring similarity, as semantic knowledge is not taken into account. (S. Zhang, Hu, & Bian, 2017)
- Knowledge bases are used to overcome the semantic problem (Zhou et al., 2013).
- WEKOS is another method proposed, it uses the CBOW (continuous bag-of-words) model to learn continuous word representation. Quite similar to what chatGPT has. (Othman et al., 2019).

# FAQ Processing (cont.)

**Rankings**

- DMN is a deep learning model proposed to generate matching scores based of 2 matrices input and to rank them. (Gupta & Carvalho, 2019)
- In the DMN model, words are represented as a dot product of embeddings of every words in the sentences.
- Multi-hop attention network are also effective on reasonings tasks such as questions answering. (Gupta & Carvalho, 2019).
- An encoder and decoder network that attends to different parts of the input is used, again very similar to chatgpt architecture.
- Document classifiers are also used to query and rank correspondings sentences. (Makino et al., 2016)

**Semantic Networks** (WANJAWA & MUCHEMI, 2020)

- Semantic network are used to learn and understand relationships and connections between different concepts using machine learning algorithms.
- It's essentially a knowledge graph.

**Topic Modeling**

- Topic modeling also called topic mining can be useful to find questions answers pairings.
- Topic modeling discovers an "abstract" topic in a collection of texts.
- LDA is mostly used to achieve topic modelling (Henß et al., 2012),
- A pipeline proposed by (Henß et al., 2012) used topic modelling to perform prediction on question and answer pairings.

# Stop Words

Stop words are words that are filtered out before or after processing of natural language data (text).

Stop words Characteristics:

1. Used for connecting important words.
2. Not important to the meaning of the sentence.
3. Occurs frequently in a document.

**Problem: There is no single universal list of stop words used by all domains. The stopwords used in technical languages differs from the general stopwords list used by general application such as the NLTK library** (Gerlach et al., 2019) (Sarica & Luo, 2020)

**TFIDF model**

- (Gerlach et al., 2019) proposed a pipeline in generating stopwords list for a specific domain.
- Phrases can be detected using the Mikolov algorithm (Mikolov,Sutskever, Chen, Corrado, & Dean, 2013)
- Term frequency-inverse-document-frequency (TFIDF) is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

# Sentiment Analysis

**Sentiment analysis is the process of determining whether a piece of writing is positive, negative, or neutral.**

**Usage**

- While most sentiment analysis work are based of current social medias such as Twitter, Facebook, and Instagram, sentiment analysis can be applied to any text.
- (Alqaryouti, Siyam, Monem, & Shaalan, 2019) uses sentiment analysis to further investigate and understand the customers needs and wants for government entities.

**Grananularity** (Alqaryouti et al., 2019)

- The level of granularity refers to the level of detail at which the sentiment is expressed.
- Three levels: sentence, document and aspect level

**Aspect based** (Liu, Chatterjee, Zhou, Lu, & Abusorrah, 2020) (Xue & Li, 2018) (M.Abdelgwad, A Soliman, I.Taloba, & Farghaly, 2022) (Alqaryouti et al., 2019)

- To analyze a piece of text to determine the sentiment expressed towards a particular aspect, rather than the overall sentiment of the text.
- To identify the sentiment expressed towards a particular aspect of an entity.

# Sentiment Analysis (cont.)

**Traditional Methods** (Syam Mohan E, 2021)

- Lexicon and corpus based approaches, unsupervised learnings.
- Lexicon approach uses a predefined list of words coupled with the annotated polarity values of the word's sentiment.
- Corpus based approach however focuses more in predicting the word to be a prefixer or suffixer of the sentiment word.
- Both methods are not very accurate.

**Machine learning and deep learning**

- Attention and non attention based LSTM models are proven to be effective in ABSA (Syam Mohan E, 2021) (Liu et al., 2020)
- The SVM model coupled with PCA (Feature selection) model are very popular in the polarity classification work. (Syam Mohan E, 2021) (Zainuddin, Selamat, & Ibrahim, 2018) (Al-Smadi, Qawasmeh, Al-Ayyoub, Jararweh, & Gupta, 2017)
- CNN were also used with Word2Vec to predict polarity. (Syam Mohan E, 2021) (Kumar, Pannu, & Malhi, 2020) (Rezaeinia, Rahmani, Ghodsi, & Veisi, 2019) (B. Zhang et al., 2019)

# Summarization

**Text summarization is the process of automatically generating a shorter version of a text that preserves the most important information.**

**Two main types of summarization: Extractive and Abstractive.** (Sharma & Sharma, 2022) (Nallapati, Zhou, Dos Santos, Gulcehre, & Xiang, 2016)

**Evaluation of Summarization** (Sharma & Sharma, 2022)

$$CompressionRate = \frac{Lengthof Summary}{Lengthof OriginalText}$$

**Extractive Summarization**

- Selecting and concatenating important sentences from the original text. (Sharma & Sharma, 2022)
- Sentence ranking model using the **TFIDF** model.
- CN-Summ is a neural network model with a graph-like architecture to connect sentences that shared common significant nouns. Thus generating a summarization extractively. (Antiqueira, Oliveira, da F. Costa, & Nunes, 2009)
- LDA model is also used in the field of extractive summarization. (Mashechkin, Petrovskiy, Popov, & Tsarev, 2011)

# Summarization (cont.)

**Abstractive Summarization**

- generating new sentences that summarize the meaning of the original text.
- Common approaches involves using a sequence to sequence neural network. (RNN, LSTM)
- The recurrent connections in RNN's model allows the model to basically store information from a previous step in a memory cell, keeping track of contextual information as the model goes on (Sharma & Sharma, 2022).
- AMR was one of the best RNN models for text summarization. AMR is a neural net model that produces a single graph to represent time series information to form a summary (Banarescu et al., 2013).
- Hybrid models that uses both extractive and abstractive models are also researched (Sharma & Sharma, 2022)

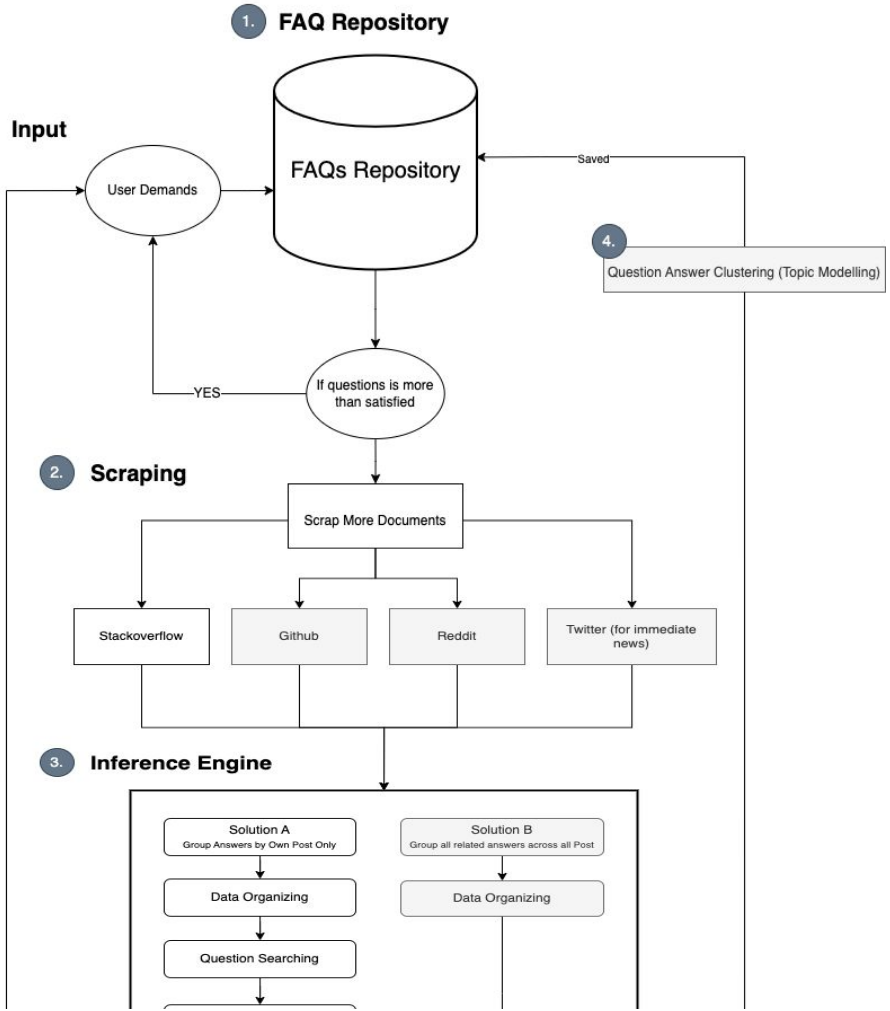
# Evaluation Metrics

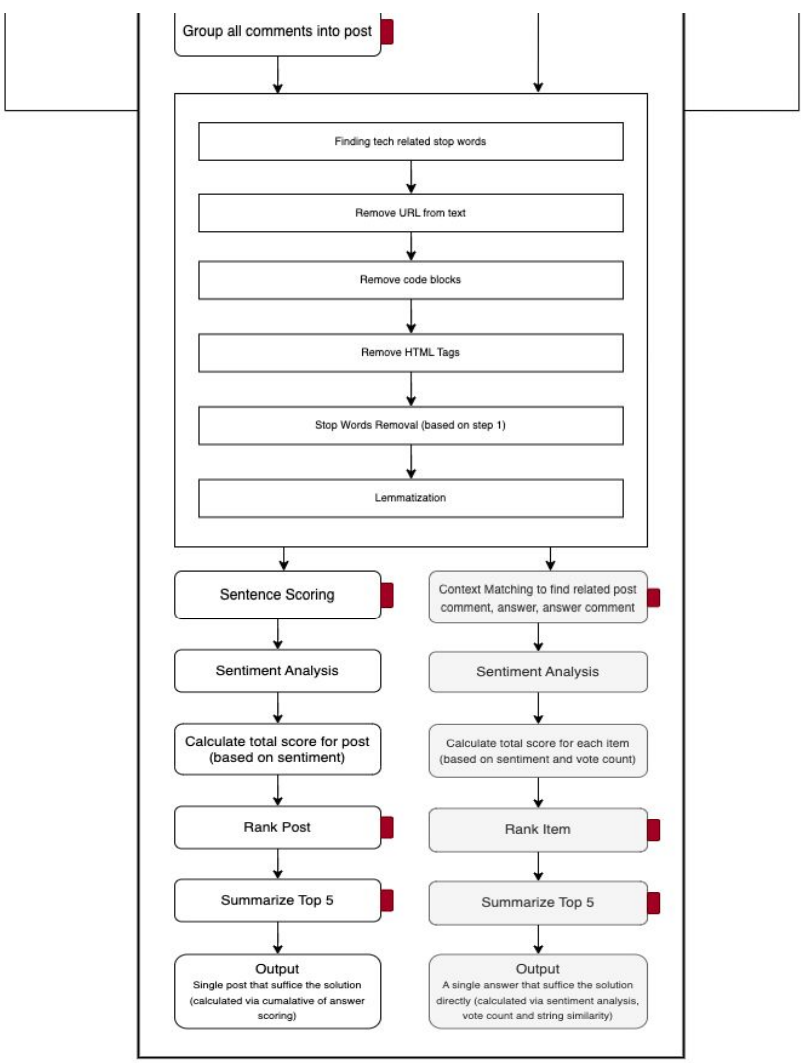
**Evaluating the success rate of extracting FAQ can be subjective, not all datasets coincide with the same domain.**

**Therefore, a manual evaluation approach are used to evaluate the success rate of a question finding process** (Jijkoun & de Rijke, 2005).

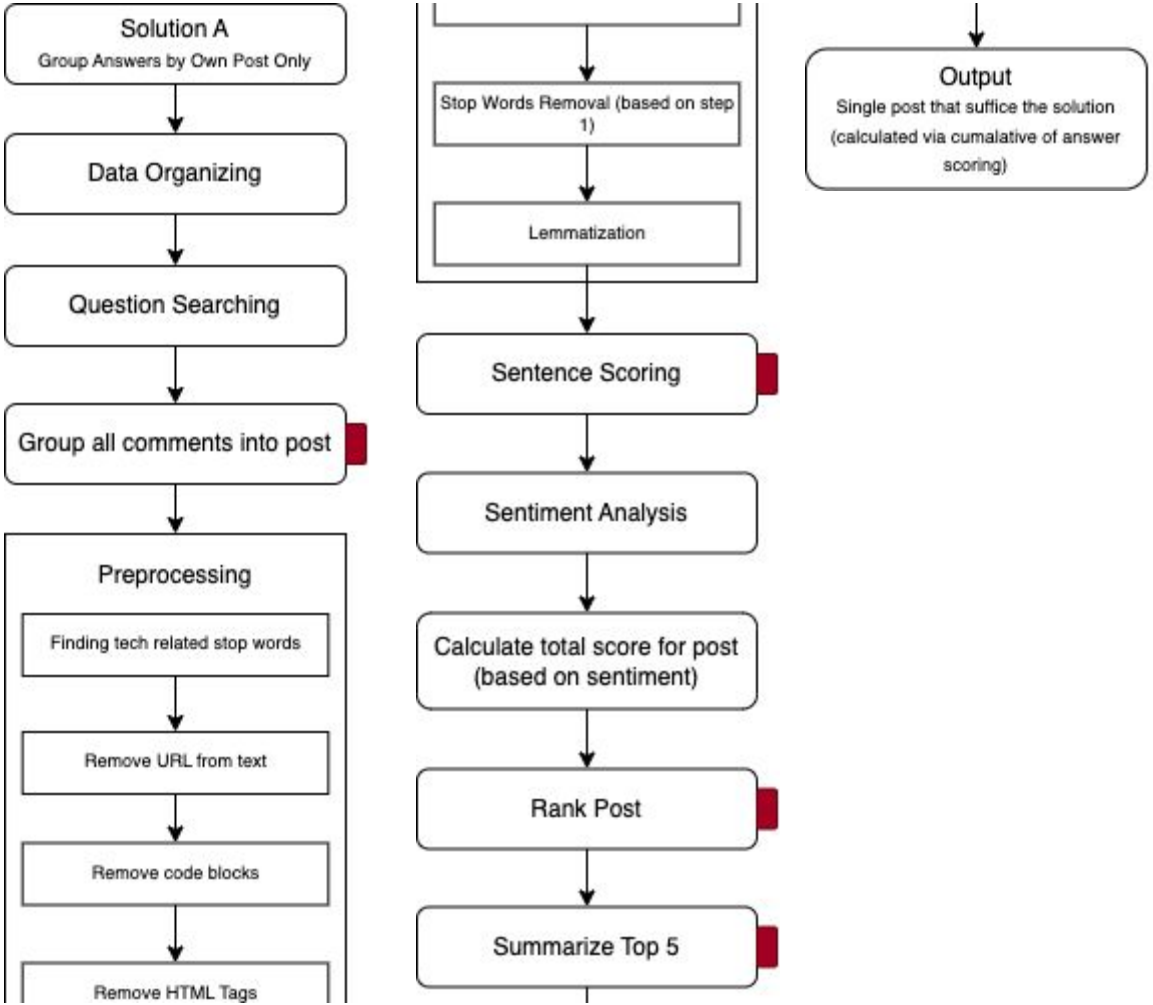
More reading will be done on this topic on FYP2

# Proposing Framework

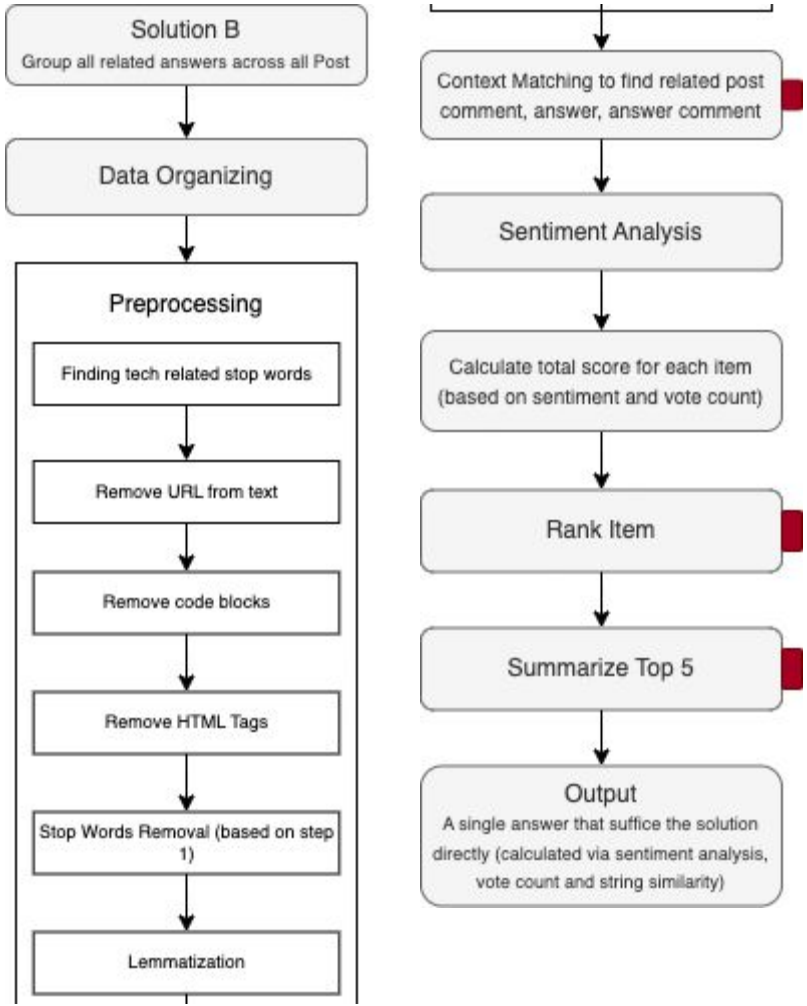

1. FAQ Repository
Input
User Demands
FAQs Repository
Saved
4.
Question Answer Clustering (Topic Modelling)
YES
If questions is more than satisfied
2. Scraping
Scrap More Documents
Stackoverflow
Github
Reddit
Twitter (for immediate news)
3. Inference Engine
Solution A
Group Answers by Own Post Only
Data Organizing
Question Searching
Solution B
Group all related answers across all Post
Data Organizing


Group all comments into post
Finding tech related stop words
Remove URL from text
Remove code blocks
Remove HTML Tags
Stop Words Removal (based on step 1)
Lemmatization
Sentence Scoring
Sentiment Analysis
Calculate total score for post (based on sentiment)
Rank Post
Summarize Top 5
Output
Single post that suffice the solution (calculated via cumalative of answer scoring)
Context Matching to find related post comment, answer, answer comment
Sentiment Analysis
Calculate total score for each item (based on sentiment and vote count)
Rank Item
Summarize Top 5
Output
A single answer that suffice the solution directly (calculated via sentiment analysis, vote count and string similarity)

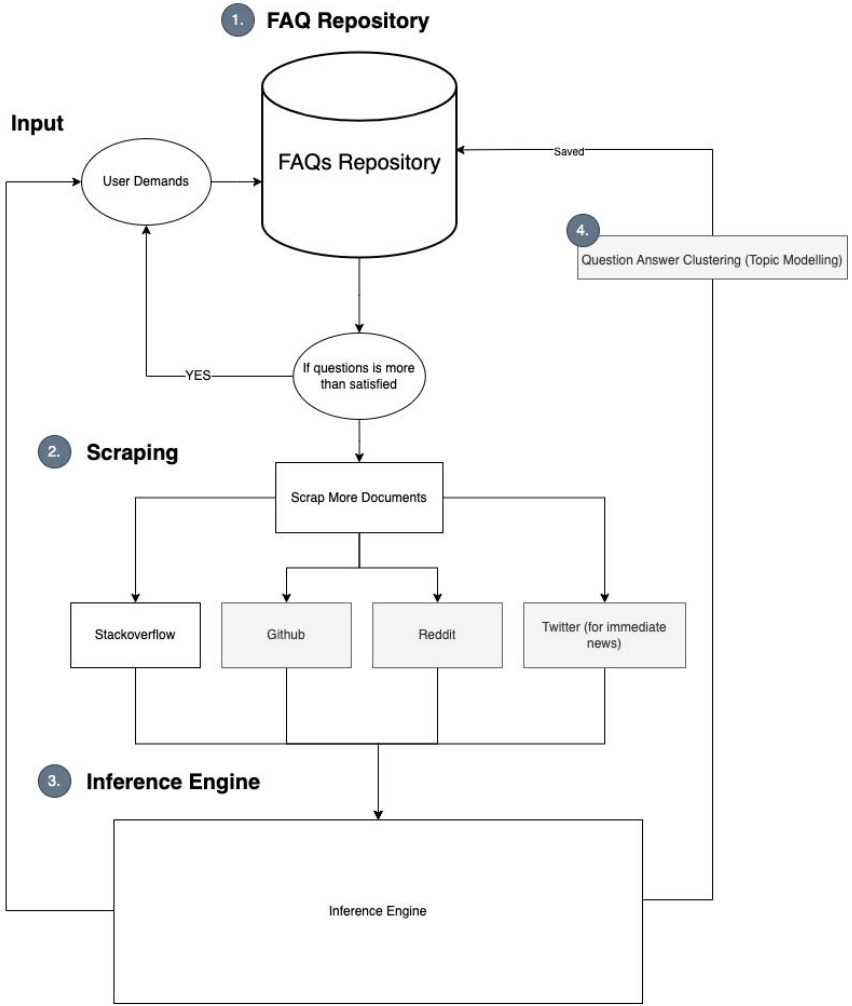
# Solution A


Solution A
Group Answers by Own Post Only
Data Organizing
Question Searching
Group all comments into post
Preprocessing
Finding tech related stop words
Remove URL from text
Remove code blocks
Remove HTML Tags
Stop Words Removal (based on step 1)
Lemmatization
Sentence Scoring
Sentiment Analysis
Calculate total score for post (based on sentiment)
Rank Post
Summarize Top 5
Output
Single post that suffice the solution (calculated via cumalative of answer scoring)

# Solution B

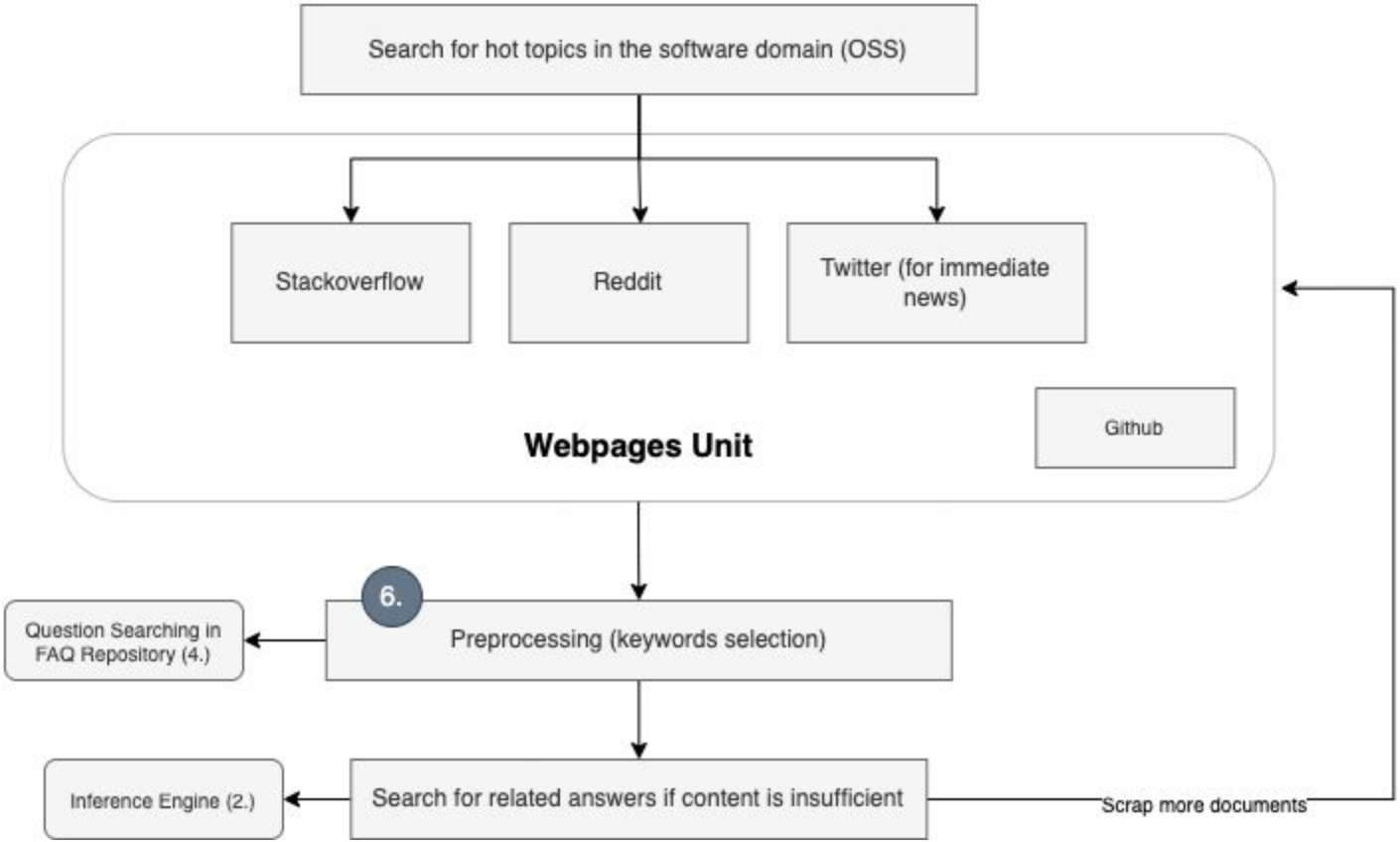

Solution B
Group all related answers across all Post
Data Organizing
Preprocessing
Finding tech related stop words
Remove URL from text
Remove code blocks
Remove HTML Tags
Stop Words Removal (based on step 1)
Lemmatization
Context Matching to find related post comment, answer, answer comment
Sentiment Analysis
Calculate total score for each item (based on sentiment and vote count)
Rank Item
Summarize Top 5
Output
A single answer that suffice the solution directly (calculated via sentiment analysis, vote count and string similarity)

# FAQs Repository


1. FAQ Repository
Input
User Demands
FAQs Repository
Saved
4.
Question Answer Clustering (Topic Modelling)
YES
If questions is more than satisfied
2. Scraping
Scrap More Documents
Stackoverflow
Github
Reddit
Twitter (for immediate news)
3. Inference Engine
Inference Engine

# Cronjobs

# 5. CronJobs


Search for hot topics in the software domain (OSS)
Stackoverflow
Reddit
Twitter (for immediate news)
Github
Webpages Unit
6.
Preprocessing (keywords selection)
Question Searching in FAQ Repository (4.)
Search for related answers if content is insufficient
Inference Engine (2.)
Scrap more documents

# FAQ Database Design

```
ISent {
  p__score: number;
  n__score: number;
  nu__score: number;
  comparative: number;
  tokens: string[];
  words: string[];
  positive: string[];
  negative: string[];
  neutral: string[];
}

IAns {
  answer: string;
  tags: string[];
  upvotes: number;
  downvotes: number;
  sentiment: ISent;
}

IFaq {
  question: string;
  answer: IAns[];
  tags: string[];
}
```

# Preliminary Results

# Note

The preliminary results done are based on a set of dataset scraped from stackoverflow with the query of ”React UseEffect“

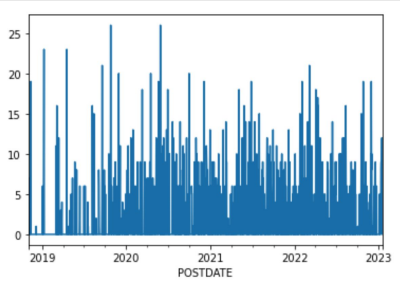
Only Solution A is tested in FYP1, solution B will be further explored in FYP2

# Data Exploration

## Unique Posts

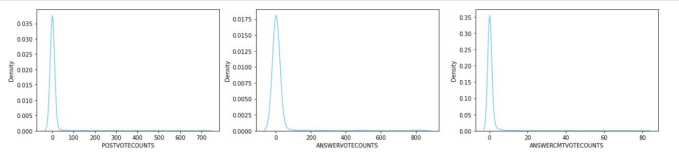
There are **600** Unique posts scraped

## Post time




The number of posts are span across years, determining that the topic subjected is long-lived

## Vote count




The majority of posts has very low vote counts, meaning that most of the posts are not active

# Post Matching

The query used is **”How to solve useEffect hook rerenders infinitely?“**

## Fuzzy Wuzzy - Partial Ratio and Token sort Ratio combined

| | postId | title | partial_ratio | token_sort_ratio | average |
|---|---|---|---|---|---|
| 0 | 73534338 | React UseEffect render infinite loop | 77 | 79 | 78.0 |
| 0 | 64651759 | react useEffect hook causes infinite loop | 71 | 71 | 71.0 |
| 0 | 58557877 | React useEffect hook infinity loop | 62 | 75 | 68.5 |
| 0 | 65558836 | React useEffect hook is causing infinite loop | 69 | 68 | 68.5 |
| 0 | 71835956 | React UseEffect hook Issue Rendering an Interval | 66 | 70 | 68.0 |
| ... | ... | ... | ... | ... | ... |
| 0 | 73334298 | React useEffect every 5 seconds with setInterv... | 51 | 50 | 50.5 |
| 0 | 60524845 | React useEffect and clearInterval | 45 | 56 | 50.5 |
| 0 | 59725069 | React useEffect hook register callback with ac... | 54 | 47 | 50.5 |
| 0 | 70258999 | React useEffect() infinite re-render for getti... | 46 | 55 | 50.5 |
| 0 | 61127662 | React useEffect restarting timer | 47 | 54 | 50.5 |

112 rows × 5 columns

# Post Matching (cont.)

### Spacy Similarity

| | postId | title | score |
|---|---|---|---|
| 0 | 65404350 | How to correctly add event listener to React u... | 0.902468 |
| 0 | 70710122 | How to use setInterval with react useEffect ho... | 0.899317 |
| 0 | 72632981 | How to fulfill React useEffect missing depende... | 0.871420 |
| 0 | 74609238 | How to stop the infinite loop inside this Reac... | 0.861767 |
| 0 | 67343507 | How to give a boolean indicator to React useEf... | 0.833936 |
| ... | ... | ... | ... |
| 0 | 70790681 | React useEffect dependency not triggering from... | 0.507214 |
| 0 | 72784960 | I dont understand why this infinite loop wont ... | 0.505654 |
| 0 | 70110116 | React useEffect does infinite loop | 0.504187 |
| 0 | 71157226 | Why is React useeffect not updated when props ... | 0.502435 |
| 0 | 73369425 | React useEffect invalid hook call | 0.500079 |

195 rows × 3 columns

```
195 number of titles found
```

# Preprocessing

## Evaluate Examples

Imagine if that was a Web Socket and we were scheduling a new heartbeat tick every time we remounted; the server would be very angry at our apps heart palpitations.

**Example 1**

'\n You have to convert the response to json <a href="https://google.com">Please Look at this link</a> with await <a href="https://google.com">await</a> response.json();\nand then use setState. \n\n<pre class="lang-js s-code-block"><code class="hljs language-javascript">useEffect(() => { \n console.log("useEffect TopTen has been called!"); \n const fetchdata = async () => {\n const response = await api.topTen(); // this calls axios(url)\n const responseData = await response.json();\n </pre>\n '

**Example 2**

# Stopwords Removal

| Original | Results |
| --- | --- |
| Imagine if that was a Web Socket and we were scheduling a new heartbeat tick every time we remounted; the server would be very angry at our apps heart palpitations. | imagine web socket scheduling new heartbeat tick every time remounted; server would angry apps heart palpitations. |

# Stemming

| Original | Results |
| --- | --- |
| Imagine if that was a Web Socket and we were scheduling a new heartbeat tick every time we remounted; the server would be very angry at our apps heart palpitations. | imagin if that wa a web socket and we were schedul a new heartbeat tick everi time we remount ; the server would be veri angri at our app heart palpit . |

# Lemmatization

| Original | Results |
| --- | --- |
| Imagine if that was a Web Socket and we were scheduling a new heartbeat tick every time we remounted; the server would be very angry at our apps heart palpitations. | Imagine if that wa a Web Socket and we were scheduling a new heartbeat tick every time we remounted ; the server would be very angry at our apps heart palpitation . |

# Extracting URL

Using example 2

[('https://google.com',
'Please Look at this link'), ('https://google.com', 'await')]

# Code blocks removal

Using example 2

’\n You have to convert the response to json <a
href=”https://
google.com”>Please Look at this link</a> with await <a href=”
https://google.com”> await</a>
response.json();\n and then use
setState. \n\n<pre class=”lang-js
s-code-block”></pre>\n ’

# HTML Tags Removal

Using example 2

’\nYou have to convert
the response to json Please Look at this
link with await await response.json();\nand then use
setState.\n\nuseEffect(() => { \n console.log(”useEffect
TopTen has been called!”); \n const fetchdata = async () =>
{\n const response = await api.topTen(); // this calls axios(url)\n
const responseData = await
response.json();\n setLoading(false);\n
setTopten(responseData.
data); \n setError(responseData.
error); \n };\n\n fetchdata (); \n}, []);\n\n ’

# Combining all preprocessing techniques

| Original | Results |
| --- | --- |
| '\n You have to convert the response to json <a href="https://google.com">Please Look at this link</a> with await <a href="https://google.com">await</a> response.json();\nand then use setState. \n\n<pre class="lang-js s-code-block"><code class="hljs language-javascript">useEffect(() => { \n console.log("useEffect TopTen has been called!"); \n const fetchdata = async () => {\n const response = await api.topTen(); // this calls axios(url)\n const responseData = await response.json();\n </pre>\n ' | You have to convert the response to json with await response.json();and then use setState. |

# Sentence Scoring

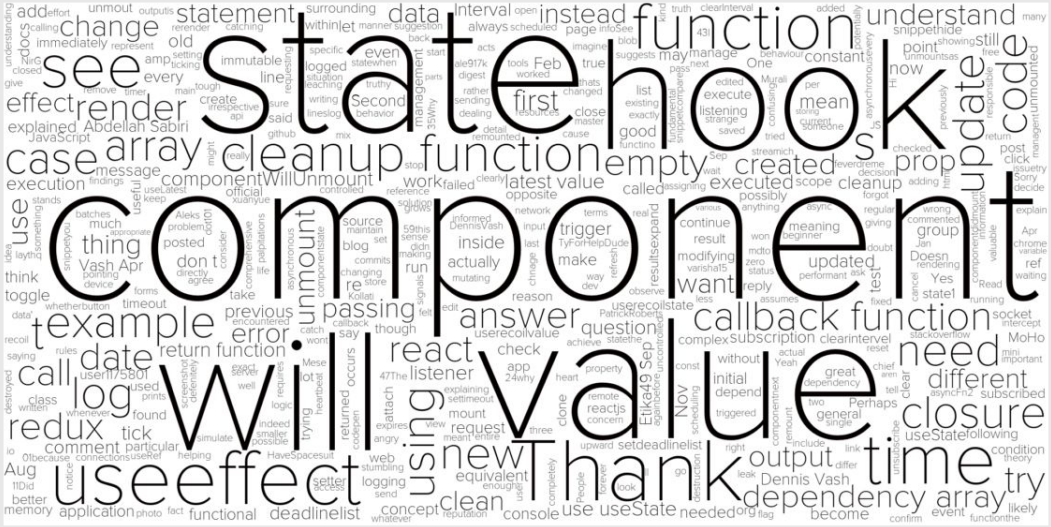
| | title | sentence_scores | sentence_list | avg_score |
|---|---|---|---|---|
| 13 | Confusing React useEffect hook behaviour | (48.428571428571495) | [ State update is asynchronous State is const... | 48.428571 |
| 4 | react useEffect hook cleanup | (1.9999999999999998, 30.428571428571406) | [I doubt, someone may explain that in more com... | 16.214286 |
| 15 | React useEffect hook doesn't clear interval | (16.0) | [ this may works for youclearintervel when tim... | 16.000000 |
| 21 | How can I make this React useEffect hook work ... | (2.25, 28.25) | [I tried your first sandbox and printed all th... | 15.250000 |
| 10 | React useEffect hook is causing infinite loop | (14.399999999999999) | [You can pass a dependency array as the last a... | 14.400000 |
| 2 | React useeffect hook | (4.7142857142857135, 4.142857142857142, 27.714... | [I deleted my answer, I think there were some ... | 12.190476 |
| 1 | react useEffect hook causes infinite loop | (10.333333333333336) | [Not the cause of issue, but what is ?, hook s... | 10.333333 |
| 17 | componentWillUnmount with React useEffect hook | (0.5714285714285714, 14.571428571428573, 12.42... | [What is it that you want to do when the compo... | 9.571429 |
| 6 | React useEffect hook dependency array | (13.0, 9.666666666666666, 2.0) | [even if you provide dependency array it will ... | 8.222222 |
| 3 | Setting hook state inside React useEffect hook | (12.333333333333337, 2.333333333333333) | [I dont think you need , also try making a san... | 7.333333 |
| 18 | React useEffect stop infinite loop | (6.000000000000001, 3.8000000000000007, 12.799... | [ is a new array on every render, yet the last... | 6.100000 |
| 11 | React useEffect invalid hook call | (4.5, 3.25, 9.0) | [you cannot use useEffect in a function, see t... | 5.583333 |
| 19 | React useeffect hook behaving not like i expected | (7.25, 3.5) | [I deleted my answer, I think there were some ... | 5.375000 |
| 0 | React useEffect hook infinity loop | (3.5714285714285703, 9.571428571428571, 0.1428... | [I dont think the second variation can cause y... | 4.892857 |
| 20 | React useEffect hook loop, dependency problem | (10.666666666666668, 0.16666666666666666, 3.16... | [ whenever the component is rendered, the depe... | 4.750000 |
| 5 | React useEffect hook load onsnapshot condition... | (12.875, 0.25, 2.25, 0.75, 1.125, 24.375, 0.75... | [commenting because I dont have time to flesh ... | 4.700000 |
| 16 | How can I escape React useEffect infinite loop? | (0.5, 7.0, 1.5) | [why is in the dependency array for ?, It is i... | 3.000000 |
| 9 | React useEffect fetch hook makes endless calls... | (3.5, 1.3333333333333333, 0.6666666666666666, ... | [Objects are compared by reference, not by the... | 2.750000 |
| 12 | React useEffect hook toggle issue | (3.5, 1.0, 3.5) | [Don't toggle the class manually Rerender the ... | 2.666667 |
| 7 | React useEffect hook running infinite loop des... | (0.6666666666666666, 5.0, 1.0) | [If the only changing piece is isAuth then why... | 2.222222 |
| 14 | componentWillUnmount lifecycle with react usee... | (2.0, 1.0) | [yes there is, everything is explained in the ... | 1.500000 |
| 8 | Using react useEffect hook | (0.3333333333333333, 0.3333333333333333, 3.0) | [What do you mean by your loading is not worki... | 1.222222 |
| 22 | React UseEffect hook firing every time I click... | 0 | [] | 0.000000 |

# Rank Post: Part 1

# Sentiment Analysis

| | title | avg_positive_sentiment | avg_negative_sentiment | avg_neutral_sentiment | sentence_list |
|---|---|---|---|---|---|
| 0 | React useEffect hook does not call after recoi... | 0.388967 | 0.159183 | 0.451850 | [One thing that stands out to me is that you a... |
| 1 | React useEffect hook does not fire when prop d... | 0.356100 | 0.113650 | 0.530250 | [Did you try to pass different photo prop valu... |
| 2 | react useEffect hook triggers only once althou... | 0.349400 | 0.164017 | 0.486583 | [because there is no value change for error &a... |
| 3 | React useEffect hook doesn't clear interval | 0.281500 | 0.159580 | 0.558960 | [this may works for you:clearintervel when tim... |
| 4 | How do I fire React useEffect hook only once a... | 0.280633 | 0.047675 | 0.671700 | [Why do you want to update from ?, Does the in... |
| 5 | Unexpected behaviour using React useEffect hook | 0.242813 | 0.160796 | 0.596396 | [The line where you have used useState hook, y... |
| 6 | Confusing React useEffect hook behaviour | 0.228324 | 0.098843 | 0.672843 | [1., State update is asynchronous., 2., State ... |
| 7 | react useEffect hook cleanup | 0.214458 | 0.230279 | 0.555279 | [I doubt, someone may explain that in more com... |
| 8 | Loading spinner with react useEffect hook and ... | 0.209583 | 0.219074 | 0.571343 | [why this statement : "I am using react hook u... |
| 9 | componentWillUnmount lifecycle with react usee... | 0.207567 | 0.058167 | 0.734267 | [yes there is, everything is explained in the ... |

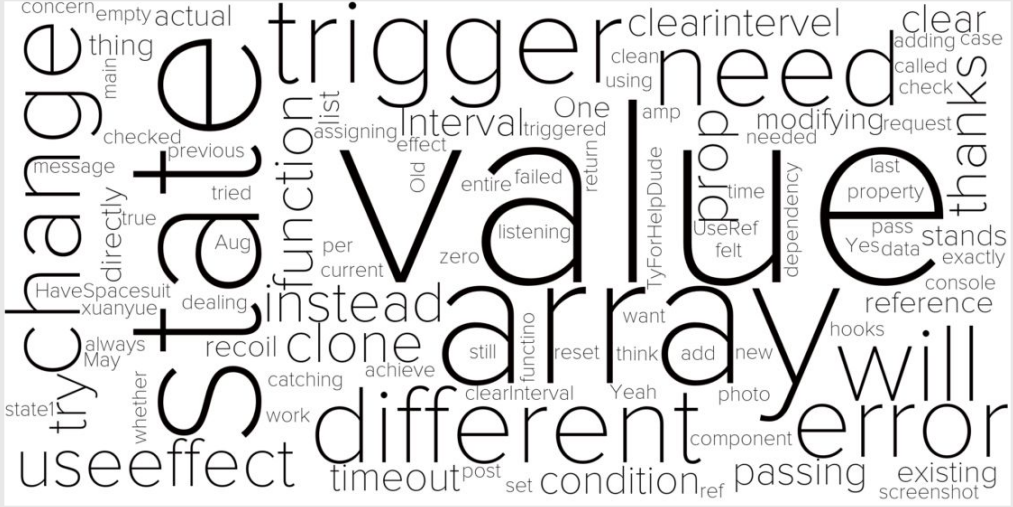
# Rank Post: Part 2

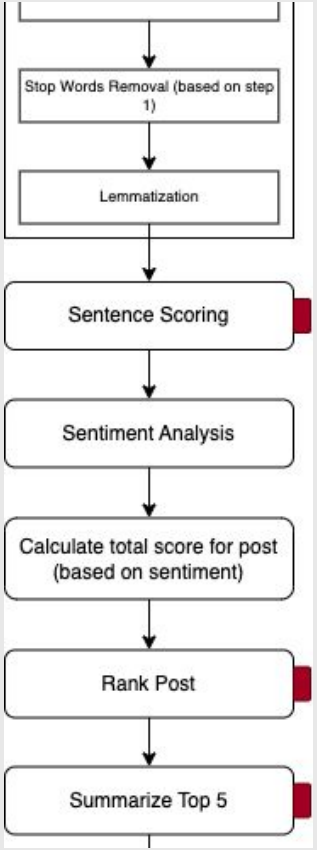


# Summarization

| | title | avg_positive_sentiment | avg_negative_sentiment | avg_neutral_sentiment | sentence_list | summary |
|---|---|---|---|---|---|---|
| 0 | React useEffect hook does not call after recoi... | 0.388967 | 0.159183 | 0.45185 | [One thing that stands out to me is that you a... | One thing that stands out to me is that you ar... |
| 1 | React useEffect hook does not fire when prop d... | 0.3561 | 0.11365 | 0.53025 | [Did you try to pass different photo prop valu... | Did you try to pass different photo prop value... |
| 2 | react useEffect hook triggers only once althou... | 0.3494 | 0.164017 | 0.486583 | [because there is no value change for error &a... | because there is no value change for error &am... |
| 3 | React useEffect hook doesn't clear interval | 0.2815 | 0.15958 | 0.55896 | [this may works for you:clearintervel when tim... | clearintervel when timeout is zero by adding c... |
| 4 | How do I fire React useEffect hook only once a... | 0.280633 | 0.047675 | 0.6717 | [Why do you want to update from ?, Does the in... | UseRef is exactly what I needed. What you need... |

# Rank Post: Part 3



# Survey on experts

Note 3 experts on the field of ”react“ is chosen

# How to fetch data using the useEffect Hook?

The result is relevant to the question, but the summarized output is not useful.

# useEffect hook doesn't work!

It is able to find the post that are related to my prompt.

# How do i use the Use effect hook with use state.

Maybe my query is abit out of context because useState is involved, but surprisingly the framework is able to find the post.

# DEMO TIME!

# Conclusions

- Solution A was built and tested, and the results can be improved further
- Solution B will be expected to perform even better
- Deep learnings model should be considered in FYP2
- More data sources needed to be added
- A working frontend and backend is in development, and will be the main focus on FYP2
- Summarization results are not as satisfactory, others algorithm should be considered
- Framework's speed is slow, improvements have to take place in FYP2
- Matching posts with deep learning should be tested
- Evaluation Methods should be further discovered with more literature reviews.
- Lastly the FYP Topic Name will be changed from "Time Series Based Text Summarization" to **"Forum Text Processing and Summarization"**

# The END