**Assignment 2**

**Scenario:** Assume you are a software developer, and currently you do not have enough knowledge and background in NLP. However, your manager considers you as a smart employee and believe that you have the potential to learn new skills by yourself within short period of time. Therefore, your manager assigned you a project which requires you to perform basic NLP tasks such as text summarisation, sentiment analysis and identifying the text similarity between words or sentences.

In this situation, before you start the project, you want to learn some NLP concepts in advance. You want to explore what are the common challenges faced by the software developers and what is the possible solution to those problems. Now your first step is to find out the common NLP challenges faced by the developers. You know that developers often share their challenges in the discussion forums such as Stack Overflow (SO). Therefore, you want to design a knowledge base system which will document a comprehensive list of challenges by software developers posted in Stack Overflow and their corresponding solutions (answers in SO). Now as there are almost 20974 posts on NLP in Stack Overflow so far, you want your system to categorise these challenges and solutions based on some criteria. So you would need to similar questions to look for patterns to categorise these challenges.

When you start working on your project assigned by your manger, you can read the documentations prepared by you. You will have a knowledge base system of basic NLP tasks performed by the software developers, the common challenges they face and the corresponding solutions to those challenges. This would significantly save your time to look for the challenges and solutions from the online sources. You trust your system over the answers provided by the ChatGPT because you know that the solutions (answers) you collected from Stack Overflow are provided by the experienced software developers and there are less chances that these answers are wrong as opposed to the answers by ChatGPT.

In this assignment, you will learn how to design such system. Here are some basic steps you should follow:

1. **Data collection:**

- Use *Stack Exchange API* to download the posts from Stack Overflow (SO) with the following tags: *[nlp]*

- There will be other tags associated with [nlp]. Consider collecting them too.

- Collect other information such as date/time of the posts, no. of views etc. Generate graphs using popular python libraries to visualise the data.

Your dataset must include the following information in four different columns:

* title of the posts (1st column)
* description of the posts (2nd column)
* tags of the posts (3rd column)
* At least one Accepted answer (4th column)
* More accepted answers (5th column) (optional)

[Accepted answers are indicated with a green tick symbol in the Stack Overflow site. There is a filter option on Stack Overflow which helps you to see the posts that do not have any accepted answer.]

1. **Pre-processing:** Perform some pre-processing on each column of the dataset. For example, remove the punctuation marks, special symbols, convert to lower case, remove screenshots from the posts and answers, tokenization etc.
2. **Graphical representation of the dataset:**

* Provide a visual representation of most frequent terms (often known as Word cloud) used in the titles of the posts. Hint: use WordCloud in python. Do some preprocessing such as removing stop words before generating the word cloud. Note that if you do not carefully pre-process your dataset, the word cloud might not truly represent the important terms in the titles of the Stack Overflow posts.

1. **Categorisation of the posts:**

* At first read some posts (the title and the accepted answers) from the dataset to plan how you want to categorise the posts (i.e., titles/questions).
* Identify posts that discussed about the issues and categorise them based on certain keywords and concepts.

Suggestion: There are different ways to categorise the posts. Categorisation can be performed by identifying similar words between different posts. Some example strategies on how you can think of categories:

- What are the posts that has the term “how to” or “how”? Identifying posts with the “how to” or “how” term might help you to categorise the questions that are related to the **implementation issues** (i.e., your category name) faced by the software developers in NLP tasks. Therefore, the category name can be “Implementation issues”. For example, in the following post [How can I use BERT for long text classification?](https://stackoverflow.com/questions/58636587/how-can-i-use-bert-for-long-text-classification), the developer is trying to know how to implement BERT for text classification. This is an implementation issue. Again, the following post [How to config nltk data directory from code?](https://stackoverflow.com/questions/3522372/how-to-config-nltk-data-directory-from-code) can be categorised as implementation issues as well.

-Task based categories are also possible. For example, the following post [How to compute the similarity between two text documents?](https://stackoverflow.com/questions/8897593/how-to-compute-the-similarity-between-two-text-documents) , the developer wants help to perform the task: **calculating text similarity**. In the following post [How to get rid of punctuation using NLTK tokenizer?](https://stackoverflow.com/questions/15547409/how-to-get-rid-of-punctuation-using-nltk-tokenizer), the developer needs help to perform the task – **tokenization.** Following are some more examples of different tasks:

[How do I do word Stemming or Lemmatization?](https://stackoverflow.com/questions/771918/how-do-i-do-word-stemming-or-lemmatization) (task: Stemming/Lemmatization)

[How to determine the language of a piece of text?](https://stackoverflow.com/questions/39142778/how-to-determine-the-language-of-a-piece-of-text) (task: language identification)

-See what are the posts that has the term “what”- this will help to categorise the questions to basic **understanding issues.** For example: [What do spaCy's part-of-speech and dependency tags mean?](https://stackoverflow.com/questions/40288323/what-do-spacys-part-of-speech-and-dependency-tags-mean)

-Categorisation can be possible based technical terms. For example, categorising posts related to nlp libraries such as spaCy, NLTK, hugging face transformers, Gensim, Word2Vec, FastText, LDA topic modelling.

suggestion: You can define rule in your program to look for some predefined terms in the title of the post using text similarity.

**Rubrics**

|  |  |
| --- | --- |
| **Criteria** | **Pts** |
| **Dataset Collection** | |
| The dataset contains of at least 20,000 Stack Overflow posts on the following tag: [nlp] | 10 |
| The dataset has at least four columns: titles of the posts, description of the posts, tags, and accepted answers | 20 |
| **Data Preprocessing** | |
| At least 4 preprocessing tasks have been performed on the collected data (e.g., stop word removal, tokenisation, converting to lower case) | 10 |
| **Data Visualisation** | |
| WordCloud consists of important and distinguishable NLP terms rather than noise words | 5 |
| **Data Categorization** | |
| Clear description on how the posts were categorised | 10 |
| Each category consists of at least 10 posts or more | 10 |
| At least 100 posts have been categorised out of 20000 posts (in case of individual project). For group projects, each member should categorise at least 100 posts. | 10 |
| GitHub repository link to the dataset, codes, and details of categories in the report are included. | 10 |
| **Implementation** | |
| You python code is easy to understand, well commented and executable | 15 |

Total = 100