

20BCE069

Practical 1

NLP

Libraries

☐ **NLTK**

It is used for preprocessing the unstructured data which contains human-readable text.

Features:

- Used for tokenization: giving tokens to words which identifies the word that appear frequently.
- Filtering stop words
- Stemming : Text processing task in which you reduce words to their root, which is the core part of a word.
- Tagging parts of speech
- Lemmatizing
- Chunking: identifies the phrases

Cons:

- NLTK is a complicated solution with a harsh learning curve and a maze of internal limitations.
- For sentence tokenization, NLTK doesn't apply semantic analysis. Unlike *Gensim*, NLTK lacks neural network models or word embeddings.
- NLTK is slow, whereas spaCy is said to be the fastest alternative. However, it's possible to speed up execution using Python's multiprocessing module.

☐ **Spacy**

Features:

- Parts of speech tagging
- Making word predictions

- Morphology : morphology is the process by which a root form of a word is modified by adding prefixes or suffixes that specify its grammatical function but do not change its part-of-speech
- Lemmatization
- Tokenization
- Merging and splitting

Cons:

- less flexibility compared to NLTK

☐ **Gensim**

Features:

- Parallelized implementations of fastText, word2vec and doc2vec algorithm
- Latent semantic analysis (LSA, LSI, SVD)
- Non-negative matrix factorization (NMF)
- Latent Dirichlet allocation (LDA)
- tf-idf

Cons:

- designed primarily for unsupervised text modeling
- don't implement full NLP pipeline, should be used with other library like Spacy or NLTK

REGULAR EXPRESSION FOR EMAIL

The screenshot displays the regex101.com website interface. The browser's address bar shows the URL `regex101.com`. The website's header includes the logo "regular expressions 101" and navigation links for "@regex101", "donate", "sponsor", "contact", "bug reports & feedback", "wiki", and "what's new?".

The main content area is divided into several sections:

- SAVE & SHARE:** Includes a "Save Regex" button with a keyboard shortcut `ctrl+s`.
- FLAVOR:** A list of programming languages and their respective regex flavors, including PCRE2 (PHP >=7.3), PCRE (PHP <7.3), ECMAScript (JavaScript), Python (checked), Golang, Java 8, and .NET (C#).
- FUNCTION:** A list of functions including "Match" (checked), "Substitution", "List", and "Unit Tests".
- TOOLS:** A section for additional tools.
- SPONSORS:** A section for sponsors, featuring "DOPPLER" with the tagline "All your environment variables, in one place".

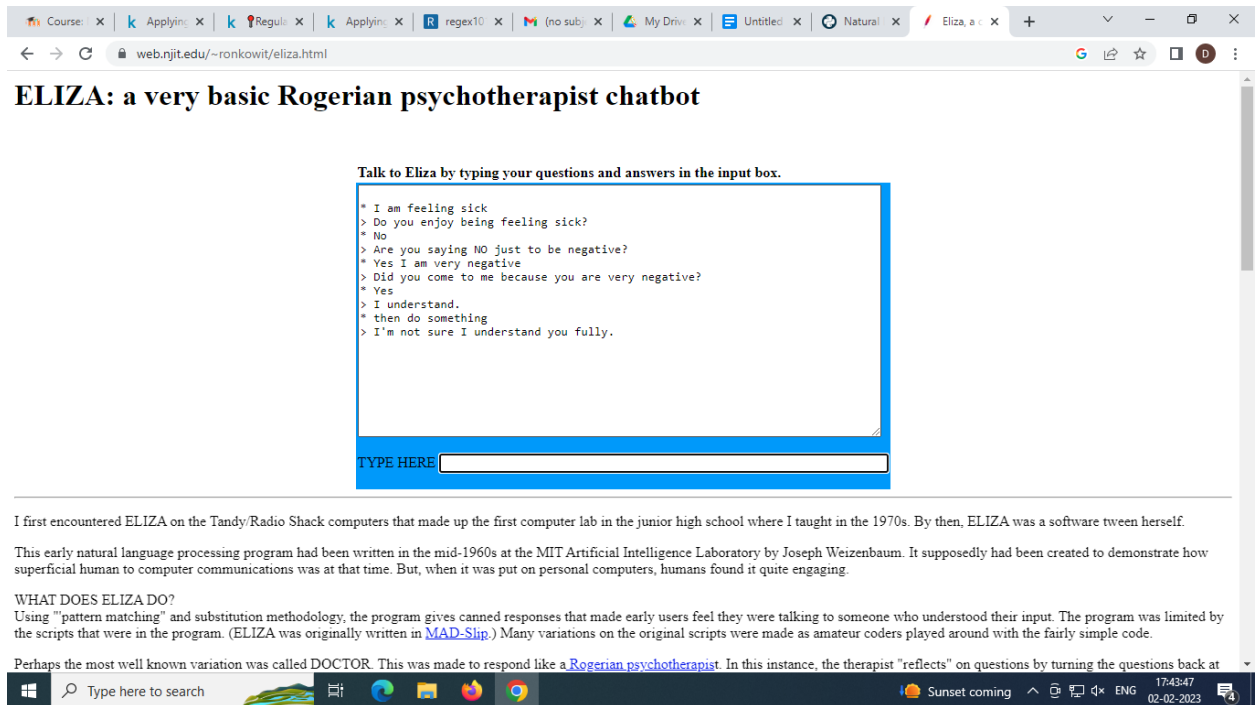
The central part of the interface is the "REGULAR EXPRESSION" editor, which contains the regex `^r"[0-9]{2}[bce]{3}[0-9]{3}@nirmauni.ac.in$`. Below this is the "TEST STRING" section, which contains the email address `20bce069@nirmauni.ac.in` and the text `xyz@gmail.com`. A green banner at the top of the editor indicates "1 match (21 steps, 0.0ms)".

The right sidebar provides detailed information about the match:

- EXPLANATION:** A list of tokens and their explanations. It includes a note that `^` asserts position at start of a line, and a note that `[0-9]` matches the previous token exactly 2 times.
- MATCH INFORMATION:** A table showing the match results. The first match is "Match 1" with a range of "0-23" and the text "20bce069@nirmauni.ac.in".
- QUICK REFERENCE:** A section for quick reference, including a search bar and a list of tokens: "All Tokens", "Common Tokens", "General Tokens", and "Anchors".

The bottom of the screenshot shows the Windows taskbar with the search bar, task view button, and several open applications (Edge, File Explorer, Firefox, Chrome). The system tray shows the date and time as "17:38:43 02-02-2023" and the weather as "27°C Mostly sunny".

TALK WITH ELIZA



Features:

- Eliza is a very simple chatbot to implement as it relies on a limited set of rules.
- It can simulate a conversation by asking questions, rephrasing, and reflecting the user's input back to them.
- Eliza can appear to be empathetic, supportive, and understanding by following the Rogerian therapy approach.
- It can handle a limited range of topics, such as emotions, relationships, and personal experiences.

Limitations:

- Eliza's responses are limited to a pre-programmed set of rules and patterns, which can make it repetitive and predictable.
- It cannot understand the meaning behind the user's input, which can lead to irrelevant or nonsensical responses.
- Eliza does not learn from user interactions or improve over time, making it less effective at understanding and adapting to individual users.
- It lacks the ability to provide helpful advice, guidance, or solutions to users' problems as it only reflects what the user says.
- It is not capable of handling complex conversations, such as those involving sarcasm, humor, or multiple topics.