

Assignment 04

Question 1.

Explain in natural language what character sequences are matched by the following regular expressions.

(a) /Ha(ha)+!~/gi

(b) /(+49 |0)1615[0-9]{4}[5-9]{3}/g

Solutions

- (a) This regular expression matches a string that contains the pattern 'Ha', followed by one or more occurrences of 'ha' and includes an exclamation mark '!' at the end. The 'gi' flags make the search case-insensitive (i) and perform a global search (g) to find all matches in the input string.
- (b) This regular expression matches a sequence of phone numbers that start with either "+49" or "0", followed by the digits "1615", four numerical digits in the range 0 to 9, and then three numerical digits in the range 5 to 9. Flag "g" is used to allow iterative searches (namely global search).

Question 2.

Try the same search in several different search engines (Google, Bing, Baidu, DuckDuckGo, Ecosia, or others). What similarities or differences do you observe in the results and their ranking, as well as in the user experience, design, and advertisements? Give examples of the queries you searched for, and limit your answer to 2 - 3 paragraphs.

Solutions

I have tried to search a video game called "Chained Echoes" in Google, Bing, and Ecosia respectively.

They all have around 10 results on the first pages and covers all sorts of information such as texts, video, images, and the other frequently related search terms. And most results on the first page are similar. However, differences exist:

1. Ads: first noticeable difference is that Ecosia has ads and they are positioned at top while other two search engines do not have ads.
2. Design: Bing shows more information for each result while the other two keep them unfolded. For example, they all have the result of "Chained Echoes on Steam" but only Bing will provide a "Download" quick access, screenshots in-game, system requirements and more, the other two shows the textual introduction to the game instead. Bing looks the fanciest with more images and videos highlighted at front. Ecosia is the minimalist among then and is the only one who has a colour theme (green), which aligns with its environmental friendly impression.
3. User experience: Google among them has the most practical design which keeps a perfect balance between the aesthetics and the functions. It is the only which highlights the fundamental facts of this search term - at top of the result page we have the game title in bold format and large front size and its icon at the left corner, with a key point "video game" below the title, telling us what fundamentally the search item is. And on the right of this top row, we have "overview, reviews, videos, gameplay, guides" bottom which cluster the related results, and it covers most information the users might need, so very clean and clear.

Question 3

Search the web for a Computer-Assisted Language Learning (CALL) application that uses Human Language Technology. Both commercial applications and applications from research projects are acceptable. Next, describe the application in two or three paragraphs, being sure to include the following information:

- the name of the application you chose, and whether the system is a commercial application or a research project,
- the intended target users (i.e., the kind of people expected to use the application),
- the type of language training offered by the app (e.g., listening comprehension, grammar practice, etc.), and what if any feedback is offered, and
- whether or not the app has the task of processing well-formed (i.e., native) language, language produced by learners, or both.

Finally, speculate on which of the language technologies we have covered in class might be used in the CALL application's language processing pipeline. Examples of technology we have covered include dialogue systems, spell checking, language models, sentence segmentation, tokenization, POS tagging, parsing, and automatic speech recognition.

Solutions

- Name: **Duolingo**
- Type: **Commercial**
- **Target users:** young people between 16 to 22 years old who wish to make learning a new language (Ch Daniel, 2024).
- **Type of language training:** vocabulary spelling, listening comprehension, grammar practice, pronunciation.
- **Both.** Duolingo primarily focuses on processing language produced by learners (namely learner generated content). The app is designed to evaluate and provide feedback on the language input provided by learners during various exercises and activities. Native language is used as a gold standard to check the input.
- **Dialogue system** is applied by the interactive exercises embedded in an interface of simulate dialogue.
- **Spell checking** is also included as part of the exercises.
- **Language models** are used to understand and evaluate whether users are using the correct vocabulary, grammar and such.
- **Sentence segmentation** is uses as Duolingo breaks down sentences into smaller parts for users to complete. This breaking down process naturally involves tokenization.

[1] Ch Daniel. (2023, December 29). Duolingo Users and Growth Statistics (2024). SignHouse. <https://www.usesignhouse.com/blog/duolingo-stats>