RAW PROBLEM STATEMENT

Problem Statement ID:

1782

Problem Statement Title:

An online system to automatically verify new title submissions by checking for similarities with existing titles

Description:

Background: Press Registrar General of India (PRGI) maintains a database containing approximately 160,000 titles. When a user submits a new title for verification, we need to check its similarity against the existing titles in our database. The goal is to ensure that the new title does not duplicate or closely resemble any existing title to avoid confusion and maintain uniqueness. Additionally, the system must enforce specific guidelines to ensure that certain words are not used, combinations of existing titles are not allowed, and titles with similar meanings or periodicity modifications are rejected.

Problem Description: Develop a system to automatically verify new title submissions by checking for similarities with existing titles in the database and ensuring compliance with specific guidelines. The system should reject titles that are too similar to existing ones, contain disallowed words, or violate other outlined rules. Additionally, the system should provide a probability score indicating the likelihood of a title being verified.

Requirements:

1. Similarity Check:
a. Implement a mechanism to check for similar-sounding names using phonetic similarity algorithms (e.g., Soundex, Metaphone).
b. Identify titles that have common prefixes or suffixes (e.g., The, India, Samachar, News).
c. Ensure that variations in spelling or slight modifications do not bypass the similarity check (e.g., Namaskar vs. Namascar).
d. Calculate a similarity percentage for each title comparison.

2. Prefix/Suffix Handling:
a. Maintain a list of disallowed prefixes and suffixes.
b. Reject any new titles that include these disallowed prefixes or suffixes if they cause the new title to resemble an existing title closely.

3. Guideline Enforcement:
a. Maintain a list of disallowed words (e.g., Police, Crime, Corruption, CBI, CID, Army).

b. Ensure that titles containing these disallowed words are rejected.
c. Prevent the creation of new titles by combining existing ones (e.g., if "Hindu" and "Indian Express" exist, "Hindu Indian Express is not allowed").
d. Check for titles with similar meanings in other languages and reject them (e.g., "Daily Evening" and "Pratidin Sandhya").
e. Disallow adding periodicity (e.g., daily, weekly, monthly) to existing titles to form new ones.

4. Verification Probability:
a. Provide a probability score indicating the likelihood of a title being verified.

5. Database Interaction:
a. Efficiently search and compare new titles against the database of 160,000 titles.
b. Track current applications and use them for future reference to reject similar titles submitted later.
c. Use indexing and optimised search techniques to handle the large dataset and ensure quick responses.

6. User Feedback:
a. Provide clear feedback to the user if their submitted title is too similar to an existing title, contains disallowed prefixes/suffixes, violates guidelines, or is created by combining existing titles.
b. Display the verification probability to the user.
c. Allow the user to modify their title and resubmit it for verification.

7. Scalability:
a. Design the system to handle an increasing number of titles and user submissions.
b. Ensure that the system remains performant as the database grows.

Expected Solution:

1. The system will provide the probability of a title being verified. For instance, if a title has a similarity score of 80%, the verification probability shall not be more than 100%-80%=20%

2. The system will reject any new title that is too similar to existing ones, contains disallowed words or prefixes/suffixes, combines existing titles, or has similar meanings in other languages.
3. The system will track current applications and use them for future reference, rejecting similar titles submitted later by other users.

Acceptance Criteria:

1. Accuracy:
a. The system correctly identifies similar-sounding titles and provides consistent results.
b. The system accurately rejects titles with disallowed prefixes, suffixes, and words.
c. The system prevents the creation of titles by combining existing titles and identifies titles with similar meanings in other languages.
d. The system disallows adding periodicity to existing titles.
e. The system provides an accurate verification probability score.

2. Performance:
a. Title verification is completed within a reasonable time frame (e.g., under 2 seconds per title).
b. The system can handle multiple title verification requests simultaneously without significant performance degradation.

3. User Experience:
a. Users receive clear and actionable feedback on why their title was rejected.
b. Users see a probability score indicating the likelihood of their title being verified.
c. The interface for title submission and feedback is user-friendly and intuitive.

4. Robustness:
a. The system handles edge cases and variations in spelling effectively.
b. The system is resilient to errors and provides meaningful error messages when issues occur.

Organization:

Ministry of Information and Broadcasting

Department:

Ministry of Information and Broadcasting

Category:

Software

Theme:

Miscellaneous

Dataset Link:

https://drive.google.com/drive/folders/13p8Z_XU9oJjzHiiJI8HSGBgFQFCOC76E?usp=sharing

---

END