Kaggle Task

Let’s Revisit Some Basic Concepts

GOAL	Put into practice the different IR-related techniques. Succinctly describe a solution to a problem.
TOTAL POINTS	10 points
HINT	You are allowed to take advantage of existing libraries and any programming language of choice. Use a sample of the given set of (development) instances for your own testing; that should help you explore what works and what doesn’t in your matching strategies. Start with a simple submission (e.g., exact matching) so that you have a baseline to use for further exploration.

As stated by Sharma et al. (2019), in community question answering sites “*the best answers are up-voted and these answers are a valuable learning resource for many topics. Duplicate questions on this site are not uncommon, particularly as the number of questions asked grows. This poses an issue because, if treated independently, duplicate questions may prevent a user from seeing a high-quality response that already exists and responders are unlikely to answer the same question twice. Identifying duplicate questions addresses these issues. It reduces the answering burden for responders and makes it possible to direct users to the best responses, improving the overall user experience*.”

So here is your challenge: can you identify duplicate questions?

Learning Objective

- Demonstrate your understanding of core IR-related techniques.
- Develop skills to present the solution to a problem.

Detailed Specifications

You will use a set of development instances from the Quora dataset. Each instance is a tuple of the form:

<pair id, question 1 id, question 2 id, question 1, question 2, is_duplicate[1]>

You will also have access to a test set of instances of the form

<pair id, question 1 id, question 2 id, question 1, question 2, ?>

Your task is to propose and implement a duplicate detection strategy to determine if any pair of questions in the test set are (or not) duplicates, i.e., if they both express the *same user information need*.

Duplicate Detection Strategy. Since the project goal is to apply text processing and content matching techniques you either already know or want to get familiar with, you can start for example, by considering exact-matching, stopword removal, and stemming/lemmatization (initial strategy). You will then identify a limitation of your initial strategy, think about how to address it and implement the corresponding solution. This will be the iteration of your original strategy.

[1] This is a binary feature, 1 indicates question 1 and question 2 are duplicates; 0 indicates otherwise.

Kaggle Submission. Anytime you want to test your solution, you can submit your prediction for the set of test instances to the Kaggle Competition site: Kaggle Site - IN425:Information Retrieval Q3. As per project requirements, you must make an *initial* submission and an *enhanced* one. Still, you are welcome to make intermediate submissions until the submission deadline for this project.

Your Kaggle submission file should include a tuple of the form <pair id, is_duplicate_prediction> for each of the instances in the test set.

Write Up. Using Association for Computing Machinery (ACM) - Generic Journal Manuscript Template - Overleaf, Online LaTeX Editor, you will write a short report describing your enhanced solution. To be sure you develop the right skills on the use of Latex and Templates for other class assignments (and your own thesis in the future), you will need to:

- Update **title** information
- Update **author** information
- Include 1 to 2 **CCS classifications** for your report
- Include 1 to 2 additional **keywords** for your report
- Existing **set up** lines should be updated to reflect the following information:
  - %% Rights management information. This information is sent to you
  - %% when you complete the rights form. These commands have SAMPLE
  - %% values in them; it is your responsibility as an author to replace
  - %% the commands and values with those provided to you when you
  - %% complete the rights form.
  - \setcopyright{rightsretained}
  - \copyrightyear{2023}
  - \acmYear{2023}
  - \acmDOI{}
  -
  - %% These commands are for a PROCEEDINGS abstract or paper.
  - \acmConference[IN4325-Q3-23]{ IN4325-Q3-23: Information Retrieval}{2023}{TU Delft}
  - \acmBooktitle{ IN4325-Q3-23: Information Retrieval,
  - TU Deflt}
  - \acmPrice{}
  - \acmISBN{}
- **Abstract** should include the following sentence: "In this manuscript, we discuss the strategy proposed to detect duplicate questions in community question-answering sites like Quora."
- **Introduction** should match the outline below (the content of which you can tweak). The discussion of your enhanced duplicate detection solution should be at most 2 paragraphs
  - [first paragraph] Question answering sites, like Quora or StackExchange, offer users access to information based on the wisdom of crowds. How users express their information needs can differ; this translates into duplicates, but non-exactly matching, questions. As stated by Sharma et al. (2019), "*if treated independently, duplicate questions may prevent a user from seeing a high-quality response that already exists and responders are unlikely to answer the same question twice.*"
  - [your paragraph] To address this issue, we have designed and developed a duplicate detection strategy…
  - [your second paragraph if needed]
  - [concluding paragraph]. Initial evaluations conducted using the Quora Dataset reveal [any insights on false positives and false negatives – not in numbers but lessons learned from them]

- **References** should include at least Sharma et al. (2019). If you use other papers for inspiration, be sure to cite them as well.
  - To get the bibtext for this reference, we encourage you to take advantage of Google Scholar or https://dblp.org/ or https://scholar.google.com

Grading Criteria

- Baseline Submission (3 Points)
  - Friday, February 24th, by the end of the day, you will need to make your initial submission on the Kaggle site
- Enhancement Submission (4 points)
  - Wednesday March 8th, by the end of the day, you will need to make your final, enhanced submission on the Kaggle site.
- Report (3 Points)
  - Wednesday March 8th, by the end of the day, you will need to submit your report via Brightspace. For full credit, i.e., 3 points, make sure to comply with all report requirements.

References

Sharma, L., Graesser, L., Nangia, N., & Evci, U. (2019). Natural Language Understanding with the Quora Question Pairs Dataset. *arXiv preprint arXiv:1907.01041*.