Research Protocol for Data Mining in Testing big data: Ten years of Mining Repositories

1. Research Objective

The objective of this research is to conduct systematic data mining from digital content platforms such as StackExchange, LinkedIn, Medium, Dev.to, and other relevant sources, to collect information that answers the following research questions:

- What tools are widely used for testing in Big Data systems?
- What testing methods are applied in Big Data systems?

This data will be organized into a structured dataset, aiming to provide a foundation for further analysis, replicability, and the expansion of knowledge in the area of testing for Big Data systems.

2. Research Questions

- Tools Used in Big Data System Testing:
  Identify popular tools used for E2E testing, integration, and unit testing in Big Data systems, including frameworks, libraries, and automation tools such as Selenium, JUnit, TestNG, and Apache JMeter.

- Testing Practices and Processes in Big Data Systems:
  Catalog methodologies and strategies for testing, such as automation, data validation, and load testing.

3. Data Sources and Mining Tools

Data collection will be conducted from multiple digital content sources. The primary sources and methods include:

- StackExchange API: Extraction of discussions, questions, and answers relevant to testing in Big Data systems.
- LinkedIn: Mining of articles, posts, and publications by experts in data engineering and testing.
- Medium: Automated search for technical articles using the Google Custom Search API.
- Dev.to: Analysis of posts by developers about frameworks, tools, and testing practices.

4. Search Strings

The search strings are carefully designed to capture relevant information based on the following criteria:

- Focus on technical keywords related to Big Data system testing.
- Temporal filters to ensure the currency of the information.
- Inclusion of multiple platforms, such as Medium, LinkedIn, and Dev.to.

Example Search Strings:

Medium:

```
site:medium.com ("big data" OR "data quality") AND test*
daterange:20140101-20240101
```

LinkedIn:

```
site:linkedin.com ("big data" OR "data quality") AND test*
daterange:20140101-20240101
```

Dev.to:

```
site:dev.to ("big data" OR "data quality") AND test* daterange:20140101-20240101
```

Stack Exchange Search String:

```
terms = [ "big data",  "data quality",  "test",  "testing",  "tools"]
from_date = int(datetime(2014, 1, 1).timestamp())
to_date = int(datetime(2024, 12, 31).timestamp())
```

5. Inclusion and Exclusion Criteria

To ensure the relevance and quality of the collected data, the following criteria are defined:

Inclusion Criteria:

- Publications with technical details on specific testing tools for Big Data systems.
- Articles and discussions describing testing practices and frameworks applied to Big Data systems.
- Content updated from 2014 to 2024, related to automation, data validation, and load testing.

Exclusion Criteria:

- Generic publications without focus on Big Data systems or specific testing practices.
- Irrelevant or redundant data, such as promotional content about tools without technical specifications.

6. Protocol Phases

Definition of Keywords: Identification of technical terms relevant to the research scope. Examples:

- Tools: "Big Data test tools", "Integration testing tools", "Test frameworks", "Automation tools", "Load testing tools".
- Practices: "Best practices for Big Data testing", "Test strategies for data pipelines", "Big Data testing methods".

Data Mining:

- StackExchange API: Automated searches for technical topics related to Big Data testing tools and practices, with a particular focus on StackOverflow.
- Google Custom Search API: Automated searches for technical articles in Medium, LinkedIn, and Dev.to, with temporal filters to capture data from 2014 to 2024.
- Scraping and APIs: Tools such as BeautifulSoup and Selenium will be used for HTML content analysis when necessary.

Data Processing and Filtering:

- Normalization: Conversion of unstructured data into a tabular format.
- Deduplication: Removal of duplicate records to ensure data integrity.
- Analysis: Focus on detailed and relevant publications, prioritizing content that explicitly mentions testing tools and practices.

Creation of Final Dataset: Each record will include:

- Source: Medium, LinkedIn, Dev.to, StackExchange.
- Category: Tool or practice.
- Description: Extracted technical details.
- Relevance: Evaluation based on number of mentions, interactions, and content quality.
- Reference Link: URL for original consultation.

Data Validation and Analysis:

- Peer review to ensure the quality of the collected data.
- Visualization of trends (graphs and tables) to highlight the most mentioned tools and practices.

7. Ethical Considerations

- Compliance with Usage Policies: The research will respect the Terms of Service of the platforms and APIs used, ensuring that all data sources are accessed ethically and responsibly.
- Data Privacy: No personal data will be collected. The information extracted will be exclusively of a public and technical nature, focusing on the technical content of the publications.