November 2023

Keyword Extraction and Analysis from Free-form Text

Prerna Kakkar

Shrey Nagpal

Shubham Sharma

Souvik Paul

Keyword Extraction and Analysis from Free-form Text

ABSTRACT

Log files and reports describing outages and bugs include a number of free-form text fields that include crucial information about the outages or bugs. However, such information may often be lost because it does not have a defined structure. For such information to become useful, developers often manually label the data to find patterns. This is a labor intensive process. This disclosure describes natural language processing (NLP) techniques to automatically extract and store relevant structured information from free-form text fields. Such information, once in structured format, can be used to analyze the data and identify trends. Free-form information is recast in structured format, and insights obtained therefrom can be used to analyze the data and to identify trends such as common types of bugs, affected users, critical components or binaries, etc.

KEYWORDS

- Natural language processing (NLP)
- Large language model (LLM)
- Lemmatization
- Sentiment analysis
- Root cause analysis
- System outage
- Server outage
- Free-form text
- Log-file analysis
- Keyword clustering

BACKGROUND

Log files and reports describing outages and bugs include a number of free-form text fields that include crucial information about the outages or bugs. However, such information may often be lost because it does not have a defined structure. For such information to become useful, developers often manually label the data to find patterns. This is a labor intensive process.

DESCRIPTION

This disclosure describes natural language processing (NLP) techniques to automatically extract and store relevant structured information from free-form text fields. Such information, once in structured format, can be used to analyze the data and identify trends. For example, the techniques can be used to identify common types of bugs, affected users, critical components or binaries, etc.

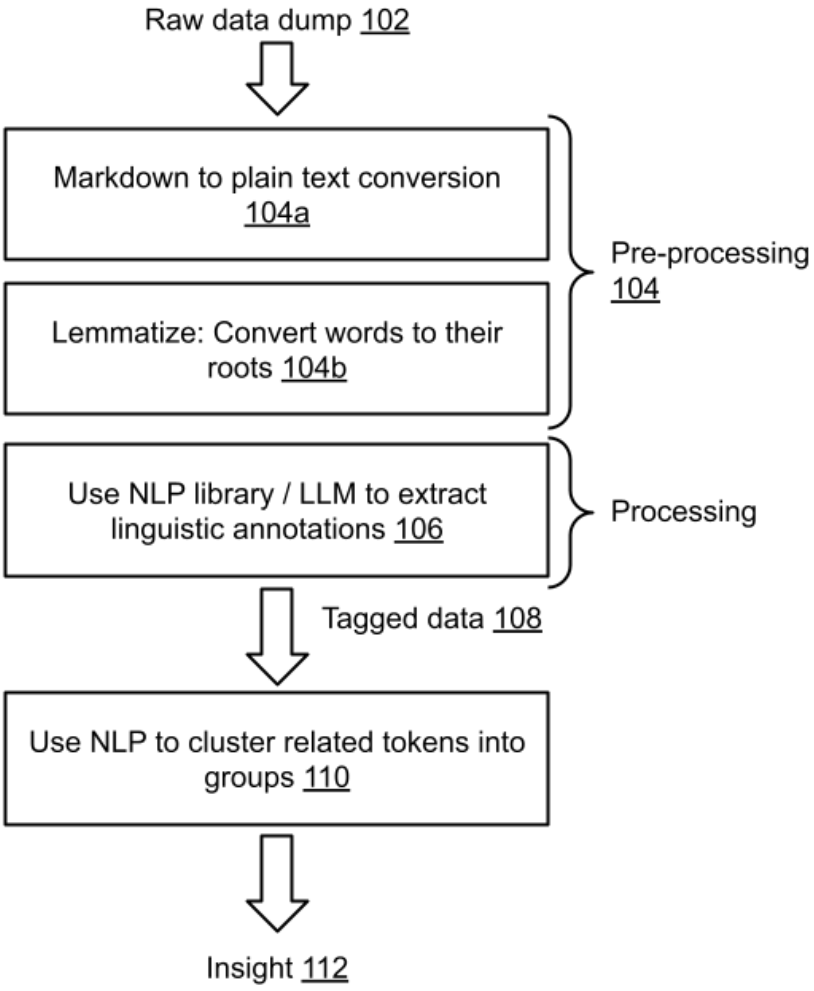

Fig. 1: Keyword extraction and analysis from free-form text descriptive of outages/ bugs

The techniques, illustrated in Fig. 1, include:

- *Dataset creation* (102): Data is ingested from tools along the developer workflow pipeline, e.g., a reliability dashboard, to obtain free-form text data available in outages and bugs at various stages.
- *Pre-processing data* (104):
    - Markdown to plain text conversion (104a), and
    - *Lemmatization* (104b), which entails grouping together the inflected forms of a word such that they can be analyzed as a single item. A word group is identified

by the lemma (or dictionary form) of the word. For example, the words ‘woman’ and ‘women’ are treated as one after lemmatization.

- *Extraction of keywords* (106):
    - Natural language processing (NLP) procedures find, e.g., using large language models (LLMs), linguistic annotations such as nouns, adjectives, adverbs, etc. in the free-form text of the bug report or outage log file. Linguistic annotations such as nouns, noun phrases, etc. are extracted from the free-form text.
    - Data is tagged with tokens found using NLP (108). The tokens are used for analyzing data.
- *Grouping of keywords* (110): A custom NLP model trained on technical documentation and on outages can be used to determine groupings for tokens. These groupings are usable as fields to filter data.
- *Presenting insights* (112): Insights obtained from keyword extraction and analysis are presented in the form of dashboards for ease of use. Insights can be used for root cause analysis, to identify critical focus areas, to prioritize tasks that can reduce the number of outages, etc.

The described techniques can be valuable for software developers, team leaders, and reliability engineers that currently use manual labeling of data to find patterns. Valuable information can be extracted from free-form fields and can be used to improve analysis. Teams can prioritize their efforts towards components that need more attention. The described techniques can be utilized for any free-form fields in a dataset and have the quality of a data analytics solution that first targets incident data.

CONCLUSION

This disclosure describes natural language processing (NLP) techniques to automatically extract and store relevant structured information from free-form text fields. Such information, once in structured format, can be used to analyze the data and identify trends. For example, the techniques can be used to identify common types of bugs, affected users, critical components or binaries, etc.

REFERENCES

1. "Inflection" available online at https://en.wikipedia.org/wiki/Inflection accessed on Oct. 21, 2023.
2. Lemma (morphology) available online at https://en.wikipedia.org/wiki/Lemma_(morphology) accessed Oct. 21, 2023.