Predicting Future Developer Behavior in the IDE Using Topic Models

Extended Abstract for Journal-First Paper[1]

Kostadin Damevski
Virginia Commonwealth University
Richmond, VA, USA
kdamevski@vcu.edu

Hui Chen
Brooklyn College
Brooklyn, NY, USA
huichen@ieee.org

David C. Shepherd
ABB Corporate Research
Raleigh, NC, USA
david.shepherd@us.abb.com

Nicholas A. Kraft
ABB Corporate Research
Raleigh, NC, USA
nicholas.a.kraft@us.abb.com

Lori Pollock
University of Delaware
Newark, DE, USA
pollock@vcu.edu

1 BACKGROUND AND MOTIVATION

Interaction data, gathered from developers' daily clicks and key presses in the IDE, has found use in both empirical studies and in recommendation systems for software engineering. We observe that this data has several characteristics, common across IDEs:

- exponentially distributed - some events or commands dominate the trace (e.g., cursor movement commands), while most other commands occur relatively infrequently.
- noisy - the traces include spurious commands (or clicks), or unrelated events, that may not be important to the behavior of interest.
- comprise of overlapping events and commands - specific commands can be invoked by separate mechanisms, and similar events can be triggered by different sources.

These characteristics of this data are analogous to the characteristics of synonymy and polysemy in natural language corpora. Therefore, this paper (and presentation) presents a new modeling approach for this type of data, leveraging topic models typically applied to streams of natural language text. Specifically, we describe how to adapt and apply a popular probabilistic topic model, Latent Dirichlet Allocation (LDA) to IDE interaction data, and extend these definitions to a variant of LDA that takes time into account, called Temporal LDA. We describe a technique to train this model using historical IDE interaction data.

ICSE '18, May 27-June 3, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5638-1/18/05.
https://doi.org/10.1145/3180155.3182541

2 APPLYING TEMPORAL LDA TO INTERACTION DATA

To apply Temporal LDA to IDE interaction data, we need notions of a document and a word. The interaction log messages, as a whole, can directly be used as the words. We describe how the notion of a document in IDE interaction data analysis can be formulated as any window of interaction events and commands executed over a contiguous time interval. IDE interactions, as they are occurring during a developer's daily work, comprise of a stream (or sequence) of sessions (i.e., documents). We discuss how Temporal LDA, a model that has previously been proposed for predicting the topics distribution of a new tweet given a succession of historical tweets, can be applied.

3 APPLICATIONS

(1) The Temporal LDA model extracted from interaction data is interpretable, allowing for analysis by researchers to determine the relationships between extracted high-level IDE user behaviors. We present a set of extracted Temporal LDA topics related to debugging behavior in the IDE and discuss the transition tendencies between them, and the high-level developer behaviors they denote.

(2) The Temporal LDA model can be used for timely recommendation of IDE commands to developers. For this purpose we present evaluations on both MS Visual Studio and ABB Robot Studio, showing the precision of command recommendation on a held out data set. We also perform k-tail evaluation, which examines the capability of the model to predict commands previously unobserved for a specific user, mirroring the model's practical use for IDE command recommendations.

[1] https://doi.org/10.1109/TSE.2017.2748134