# CSC 791 Homework 1 - Reading Assignment Summary

Neela Krishna Teja, Tadikonda ( 200109991 )

August 26, 2016

i Nguyen, A. T., Nguyen, T. T. T. N., Lo, D., and Sun, C. (2012). Duplicate bug report detection with a combination of information retrieval and topic modeling. Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering - ASE 2012, 70. Retrieved from `http://dl.acm.org/citation.cfm?doid=2351676.2351687`

ii
1. *Techincal issue:* The actual root cause of a bug. Various bugs can share a single Technical issue. In other words two bug reports can share similar topics.
2. *Technical topic:* Perspective based reports with different description of the root cause, suggestions, tips for the fixes e.t.c with characteristics of latent, semantic features.
3. *Latent:* Details that are not obvious but inherent in the context and understood by comprehension.
4. *Ensemble Averaging:* It is Linear weighted averaging of candidates. In this paper DBTM and BM25F a combined using Ensemble Averaging.

iii
1. *Motivational Statements* Two bug reports addressing same issue, but describe it from two different perspectives. This adds common technical topics between them among other technical topics reported.
2. *Related Work* Provided previous work on this problem - some ML techniques like Information Retrieval Model, Vector Space Model using simple Natural language Processing, binary classification ( applying linear regression ), Support Vector machine, extended BM25F known as REP.
3. *Informative Visualizations:* Figures showing the variables derived in Topic Model with dependencies between them. Some figures showing the accuracy comparisons across several models, over various sources of bug reports.
4. *Baseline Results:* Results comparing the accuracy of the DBTM against previous models like REP, T-Model, BM25F, and RTM+BM25F over various bug report sources like OpenOffice, Mozilla and Eclipse. Where DBTM displays higher accuracy than other models at all places.
5. *Sampling Method:* Gibbs sampling method. It is used for training the DBTM with identified historical bug report data.
6. *Patterns:* The DBTM sensitivity is analyzed by varying number of topics from 20 to 400 in steps of 10 and measured the top 10 accuracy.
7. *Negative Results:* Using less number of topic gives less accuracy as the comparison results in more bugs going into one duplicate group.

iv
1. To further improve the time efficiency, bug report groups can be strategically chosen for comparison with bug report groups. For example, we can use the topic proportions for a new bug report, against the bug report groups topics to strategically pick bug report groups tp do the comparison, instead of going bruteforce with each new bug on the group.
2. The paper can be improved by providing any further enhancements that could have been made by them to improve the efficiency, atleast a line or 2 on the open ended questions and future work.
3. A quick overview on the preproessing of the text document into words

v two A. T. Nguyen, T. T. Nguyen, J. Al-Kofahi, H. V. Nguyen, and T. N. Nguyen. A Topic-based Approach for Narrowing the Search Space of Buggy Files from a Bug Report. In ASE11, pp. 263-272. IEEE CS, 2011.

- This paper presents an LDA approach to finding the bug reports corresponding to a source file.

three C. Sun, D. Lo, S.-C. Khoo, and J. Jiang. Towards more accurate retrieval of duplicate bug reports. In ASE11, pages 253262. IEEE CS, 2011.

- This paper develops an extended BM25F technique to find the duplicate bugs and presents its results.

four C. Sun, D. Lo, X. Wang, J. Jiang, and S.-C. Khoo. A discriminative model approach for accurate duplicate bug report retrieval. In ICSE10. ACM, 2010.

- This paper also talks about the duplicate bug identification using Support Vector machine for making a discriminative model.