PROJECT REPORT

TITLE BASED SCIENTIFIC JOURNAL CLUSTERING

YOSEFINA OKTAVIANI SANTOSO
15.K1.0014

Faculty of Computer Science
Soegijapranata Catholic University
2019

APPROVAL AND RATIFICATION PAGE

TITLE BASED SCIENTIFIC JOURNAL CLUSTERING

by

YOSEFINA OKTAVIANI SANTOSO– 15.K1.0014

This project report has been approved and ratified

by the Faculty of Computer Science on January, 3, 2019

With approval,

Supervisor,

R.Setiawan Aji Nugroho, ST.,.McomIT.,Ph.D
NPP : 058.1.2004.264

Examiners,

1.)

Hironimus Leong , S.Kom, M.Kom
NPP : 058.1.2007.273

2.)

Herawati

Rosita Herawati, ST., MIT
NPP : 058.1.2004.263

Dean of Faculty of Computer Science,

Erdhi Widyarto Nugroho, ST., MT
NPP: 058.1.2002.254

STATEMENT OF ORIGINALITY

I, the undersigned:

Name : YOSEFINA OKTAVIANI SANTOSO

ID : 15.K1.0014

Certify that this project was made by myself and not copy or plagiarize from other people, except that in writing expressed to the other article. If it is proven that this project was plagiarizes or copy the other, I am ready to accept a sanction.

Semarang, January, 3, 2019

YOSEFINA OKTAVIANI SANTOSO
15.K1.0014

ABSTRACT

Scientific journals develop very rapidly along with the development of science. Reporting from labs.semanticscholar.org/corpus, the number of scientific journals has reached over 39 million. A large number of scientific journals makes it challenging to grouping scientific journals. Grouping becomes more difficult because each scientific journal can have more than one topic. Therefore, special methods are needed to group the scientific journals. One of the well-known topic modeling methods is Latent Dirichlet Allocation (LDA). This research is an implementation of the LDA algorithm to do topic modeling in scientific journals. The topic modeling in this study uses the title as a corpus. Various titles are processed into a bag of words in the pre-processing process so that they can be used to distribute. The results of the distribution stage are used for sampling with the Gibbs Sampling method. Through the sampling process, testing can also be done to determine the optimal parameters. The testing in this study used perplexity to find the most optimal number of iterations and topics. The result from this research is that the LDA Algorithm successfully performs topic modeling in scientific journals by generating a list of keywords for each topic and grouping documents on each topic. The optimal parameters based on the results of perplexity comparison are 3 topics and 500 iterations.

Keyword: Topic Modeling, LDA, perplexity, scientific journal

OUTLINE

This project that titled “Title Based Scientific Journal Clustering” is consists of 6 chapters.

Chapter I discuss the background, project scopes, and objectives from this research about scientific journal clustering.

Chapter II is a literature study that discusses journals from around the world regarding topic modeling, LDA, and perplexity.

Chapter III is about research methodology. There are 7 methods used in this research. The method The stages start from the search for research materials to the output results and search the number of topics and iterations that are most appropriate to produce the most optimal value.

Chapter IV is an analysis of corpus data used and steps that must be taken. In addition, there is also a work process design that is described in the form of a flowchart.

Chapter V explains the implementation of a program from the process of processing corpus data into a bag of words to how to obtain output in the form of a list of keywords and a list of documents on each topic.

Chapter VI is the conclusions on whether LDA is capable or not to do modeling topics in scientific journals and the number of iterations and topics to produce optimal values.

TABLE OF CONTENTS

ILLUSTRATION INDEX

INDEX OF TABLES