หัวข้อโครงการ การจำแนกภาษาบาลีและภาษาสันสกฤตในภาษาไทยด้วยการเรียนรู้ของเครื่อง

โดย นายประสบการ นำแสงวาณิช
นายอนุชัย แซ่ซ้ง
นายชาญชัย จีระ

สาขาวิชาการ / คณะ เทคโนโลยีสารสนเทศและธุรกิจดิจิทัล / บริหารธุรกิจ

อาจารย์ที่ปรึกษา ดร.ณัฐรฐนนท์ กานต์รวีกุลธนา

ปีการศึกษา 2566

บทคัดย่อ

โครงการวิจัยเรื่อง การจำแนกภาษาบาลี และภาษาสันสกฤตในภาษาไทยด้วยการเรียนรู้ของเครื่อง มีวัตถุประสงค์ 1) เพื่อทดสอบประสิทธิภาพของแบบจำลองการทำนายผลการจำแนกภาษาบาลี และภาษาสันสกฤต 2) เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการทำนายผลการจำแนกภาษาบาลี และภาษาสันสกฤต 3) เพื่อนำเสนอแบบจำลองที่มีประสิทธิภาพดีที่สุดในการทำนายผลการจำแนกภาษาบาลี และภาษาสันสกฤตโดยใช้การเรียนรู้ของเครื่อง (machine learning) คือ การจำแนกภาษาบาลี และภาษาสันสกฤต โดยมีแบบจำลอง (Model) ที่ใช้ดังนี้ ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ต้นไม้การตัดสินใจ (Decision Tree) และนาอีฟ เบย์ (Naive Bayes) โดยการจำแนกภาษาบาลี และภาษาสันสกฤต โดยมีจำนวนข้อมูล (data) ทั้งหมด 100,000 คำ แบ่งเป็นภาษาบาลี 50,000 คำ และ ภาษาสันสกฤต 50,000 คำ

ผลการวิจัยพบว่า ค่าความแม่นยำ (Accuracy) ของแบบจำลอง (Model) ทั้ง 3 แบบ โดยเรียงจากมากไปน้อยได้แก่ ต้นไม้ตัดสินใจ อยู่ที่ร้อยละ 0.69 ลองลงมาคือแบบจำลอง นาอีฟ เบย์ (Naive Bayes) อยู่ที่ร้อยละ 0.65 และลองลงมาคือแบบจำลอง ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) อยู่ที่ร้อยละ 0.47

คำสำคัญ : แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน แบบจำลองต้นไม้การตัดสินใจ แบบจำลองนาอีฟเบย์ ประสิทธิภาพแบบจำลอง การเรียนรู้ด้วยเครื่อง

Project Machine Learning the classification of Bali and Sanskrit in Thai language using machine learning

โดย Mr. Prasopkarn Namsaenwanich
Mr. Anuchai Saesong
Mr. Chanchai Jeera

Major / Faculty Information Technology and Digital Business / Business Administration

Adviser Mr. Natratanon Kanraweekultana (Ph.D.)

Academic Year 2023

ABSTRACT

This project explores the classification of Pali and Sanskrit languages in Thai texts through machine learning with three objectives: 1) evaluating the performance of different models in classifying Pali and Sanskrit languages, 2) comparing the effectiveness of these models, and 3) identifying the most effective model for this classification task. The models evaluated include the Support Vector Machine (SVM), Decision Tree, and Naive Bayes. The dataset consists of 100,000 words, evenly divided between the Pali and Sanskrit languages, with 50,000 words from each.

The findings reveal the accuracy of the three models, ranked from highest to lowest, as follows: the Decision Tree model demonstrated the highest accuracy at 69%, followed by the Naive Bayes model at 65%, and the Support Vector Machine model at 47%.

Keywords: Support Vector Machine, Decision Tree, Naive Bayes, Model Performance, Machine Learning