งานวิจัยนี้มีวัตถุประสงค์เพื่อหาสูตรสำหรับการคำนวณจำนวนของเอ็นแกรมแบบปิดจากขนาดของข้อมูลข้อความภาษาไทย โดยวิธีการวิจัยเริ่มจากการหาจำนวนเอ็นแกรมจากแฟ้มข้อความจำนวน 400 แฟ้ม แล้วใช้ 300 แฟ้มแรกในการหาสูตรและส่วนที่เหลือใช้ในการทดสอบสูตรที่ได้ การหาสูตรทำโดยใช้กฎของฮีฟและการวิเคราะห์การถดถอย สำหรับการวิเคราะห์การถดถอยได้ทดสอบตัวแบบบสามตัวแบบ ได้แก่ ตัวแบบเชิงเส้นอย่างง่าย ตัวแบบเอ็กโปเนนเชียลและตัวแบบพาวเวอร์ ด้วยตัววัด R^2 จะได้ว่าตัวแบบพาวเวอร์เป็นตัวแบบที่ดีที่สุด จากนั้นงานวิจัยนี้ได้นำตัวแบบที่ได้ไปพยากรณ์จำนวนเอ็นแกรมแบบปิดโดยวัดความถูกต้องของการพยากรณ์ด้วยตัววัด MAPE ผลปรากฏว่าสูตรที่ได้จากตัวแบบพาวเวอร์ดีกว่าสูตรที่ได้จากการใช้กฎของฮีฟ

187711

The purpose of this research is to find closed-form formulas for determining numbers of closed n-gram from size of text files. The research starts with determining closed n-gram with cutoff 2 to 21 from 400 text files. It then uses the first 300 to derive formulas and the rest to test fitness of formulas obtained. Formulas for each cutoff are derived using the well known Heap law and regression analysis. Based on measure of goodness of fit called the coefficient of determination or R^2, the power regression model is the best model described by given data among simple linear, exponential and power regression models. The research uses formulas obtained to predict numbers of closed n-gram and measures prediction errors in terms of mean absolute percentage error or MAPE. The results show that formulas of power regression model are better than of the Heap law.