การตรวจจับการคัดลอกผลงาน (Plagiarism detection)

การคัดลอกผลงานเช่นรายงาน บทความ งานวิจัย หรือวรรณกรรม เป็นปัญหาทั้งในวงการศึกษาและวงวรรณกรรม การตรวจจับการคัดลอกผลงานนั้นจำเป็นต้องอาศัยคนอ่านงานทั้งสองชิ้นและนำมาเปรียบเทียบกัน แต่จำนวนของชิ้นงานซึ่งมีอยู่มากทำให้การตรวจจับด้วยคนทำได้ลำบาก การใช้คอมพิวเตอร์อย่างแพร่หลายก็มีส่วนทำให้การคัดลอกงานทำได้ง่ายขึ้น และเพิ่มภาระในการตรวจจับให้มากขึ้นไปอีก

อย่างไรก็ดี เราสามารถใช้คอมพิวเตอร์ช่วยในการตรวจจับผลงานที่เกิดจากการคัดลอกได้ วิธีในการตรวจจับด้วยคอมพิวเตอร์นั้นมีหลายวิธี แต่หลักการพื้นฐานในการตรวจจับทั้งหมดตั้งอยู่บนสมมติฐานที่ว่า ผลงานที่เกิดจากการคัดลอกควรจะมีความคล้ายคลึงกับชิ้นงานต้นฉบับมากกว่าผลงานที่เขียนขึ้นใหม่ หากผลงานสองชิ้นใด ๆ มีความคล้ายคลึงกันสูง ย่อมเป็นไปได้สูงที่จะเกิดการคัดลอกขึ้นระหว่างผลงานทั้งสองชิ้นนั้น หากให้ผลงานหนึ่งชิ้นแทนด้วยเวกเตอร์ของความถี่ของคำที่ปรากฏในชิ้นงาน วิธีหนึ่งในการวัดความคล้ายคลึงของผลงานสองชิ้นคือการใช้ค่าความคล้ายเชิงโคไซน์ (cosine similarity) ของเวกเตอร์ที่ใช้แทนชิ้นงานทั้งสอง

สมมติให้มีตัวอย่างชิ้นงานสองชิ้น คือ

1. Plagiarism is not a crime per se but is disapproved more on the grounds of moral offence, and cases of plagiarism can involve liability for copyright infringement.

2. Plagiarism is not actually a crime by itself but it is disapproved as it is immoral, and may violate the copyright law.

กำหนดให้คำศัพท์ที่ปรากฏในผลงานชิ้นใดชิ้นหนึ่งหรือทั้งสองชิ้นอยู่ในเซต $V = \{\text{plagiarism}, \text{crime}, \dots\}$ ซึ่งมีจำนวนทั้งหมด $|V| = 32$ คำ (ไม่แยกตัวพิมพ์ใหญ่จากตัวพิมพ์เล็ก) ได้แก่

- a, actually, and, because, but, by, can, cases, copyright, crime, disapproved, for, grounds, immoral, infringement, involve, is, it, itself, law, liability, may, moral, more, not, of, offence, on, per se, plagiarism, the, violate

words	...	but	by	can	cases	copyright	crime	...
v_1	...	1	0	1	1	1	1	...
v_2	...	1	1	0	0	1	1	...

Figure 1: เวกเตอร์ของชิ้นงาน

หากเรียงคำศัพท์ทั้งหมดตามลำดับตัวอักษร แล้วให้ v_1 แทนเวกเตอร์ของความถี่ของแต่ละคำศัพท์ใน V ที่ปรากฏในชิ้นงานแรก และ v_2 แทนเวกเตอร์ของความถี่ของคำศัพท์ใน V ที่ปรากฏในชิ้นงานที่สองดังรูปที่ 1

ค่าความคล้ายเชิงโคไซน์ของ v_1 และ v_2 จะคำนวณได้จากสมการ (1)

$$
\begin{aligned}
cosine(v_1, v_2) &= \frac{v_1 \cdot v_2}{|v_1||v_2|} \\
&= \frac{\sum_{i=1}^{|V|} (v_1[i] * v_2[i])}{\left(\sqrt{\sum_{i=1}^{|V|} v_1[i]^2}\right)\left(\sqrt{\sum_{i=1}^{|V|} v_2[i]^2}\right)}
\end{aligned} \tag{1}
$$

ซึ่งจะได้ค่าความคล้ายเชิงโคไซน์ของ v_1 และ v_2 เท่ากับ 0.52