WVTool (Word Vector Tool). Công cụ này giúp việc xử lý các thao tác xử lý ngôn ngữ tự nhiên nhanh và dễ dàng hơn.

3.1.2 Tính trọng lượng đặc điểm lớp trong báo cáo lỗi

Trong quy trình xử lý báo cáo lỗi, việc tính đặc điểm trọng lượng lớp vô cùng quan trọng, nó ảnh hưởng trực tiếp đến kết quả xác định sự giống nhau giữa hai báo cáo lỗi. Mỗi từ trong các báo cáo lỗi sẽ được xác định và chuyển sang mô hình không gian vector tương ứng với một trọng lượng. Phương pháp này được thừa kế và cải tiến từ Class-Feature-Centroid (CFC)

Bảng 1: Các công thức tính trọng lượng bên trong lớp inner

Tên công thức	Chức năng
EXP-DF (CFC)	$I^i_{inner} = b^{\frac{DF^j_{t_i}}{\|C_j\|}}$
TF	$I^i_{inner} = tf_{ijk}$
EXP-TF	$I^i_{inner} = b^{tf_{ijk}}$
EXP-TF-DF	$I^i_{inner} = b^{tf_{ijk} \times \frac{DF^j_{t_i}}{\|C_j\|}}$

(Guan *et al.*, 2009; Eui-Hong Han và George Karypic, 2000), và trọng lượng đặc điểm lớp (Zhang *et al.*, 2012). Trong CFC, trọng lượng của từ w_{ij} được tính như sau:

$$w_{ij} = b^{\frac{DF^j_{t_i}}{|C_j|}} \times \log(\frac{|C|}{CF_{t_i}}) \quad (3.1)$$

Trong đó t_i là từ (term) trong báo cáo lỗi, $DF^j_{t_i}$ là số báo cáo lỗi chứa t_i của lớp C_j,$|C_j|$ là số báo cáo lỗi trong lớp C_j, $|C|$ là tổng số lớp, CF_{t_i} là số lớp chứa t_i, và b là tham số lớn hơn một, dùng để điều chỉnh cho trọng lượng w_{ij}. Trong CFC, $b^{\frac{DF^j_{t_i}}{|C_j|}}$ xem xét đến số báo cáo lỗi chứa mức độ xuất hiện thường xuyên của một từ bên trong lớp. Công thức log xem xét mức độ giống như IDF (inverse document frequency) truyền thống. Phương pháp của chúng tôi được cải tiến từ CFC và trên cơ sở dựa vào (Guan *et al.*, 2009), khi đó mức độ thường xuyên của một từ tf_{ijk} của t_i trong báo cáo lỗi d_k, thuộc lớp C_j được tính như sau:

$$tf_{ijk} = \frac{fre(t_i)}{fre(t_i) + d + h \times \frac{dl}{dl_{avg}}} \quad (3.2)$$

Trong đó $fre(t_i)$ là số lần xuất hiện của t_i trong báo cáo lỗi d_k hoặc của lớp C_j, d là tham số điều chỉnh tránh cho mẫu số bằng 0, h là tham số ảnh hưởng đến chiều dài của báo cáo lỗi, dl là chiều dài của báo cáo lỗi d_k hoặc tổng chiều dài trong lớp C_j, dl_{avg} là trung bình của chiều dài các báo cáo lỗi. Nếu $t_i \in d_k$, khi đó dl_{avg} được tính như sau:

$$dl_{avg} = \frac{\sum_{d_m \in C} dl(d_m)}{\sum_{C_n \in C} |C_n|} \quad 3.3)$$

Trong đó $|C_n|$là số báo cáo lỗi trong Cn Nếu $t_i \in C_j$ nhưng $t_i \notin d_k$, khi đó:

$$dl_{avg} = \frac{\sum_{d_m \in C} dl(d_m)}{|C|} \quad (3.4)$$

Trong đó |C| là tổng số lớp, d và h là hai tham số, và nó có thể nằm trong một khoảng giá trị tùy theo tập dữ liệu. Tuy nhiên, trong nghiên cứu này chỉ xác định $0.3 \le d \le 0.8$ và $1.5 \le h \le 20.0$. Lý do d và h cho kết quả tốt nhất trong tf_{ijk} khi chúng tôi tiến hành thực nghiệm để tìm ra các giá trị tốt nhất của hai tham số này.

a. Chỉ số tác động bên trong lớp Inner

Với việc mở rộng thông tin dựa vào lớp, khi đó bốn công thức để tính chỉ số tác động bên trong lớp Inner được giới thiệu, và được tiến hành thực nghiệm để tìm ra một công thức tốt nhất. Bảng 1 cho thấy bốn công thức dùng để tính trọng lượng bên trong lớp Inner.

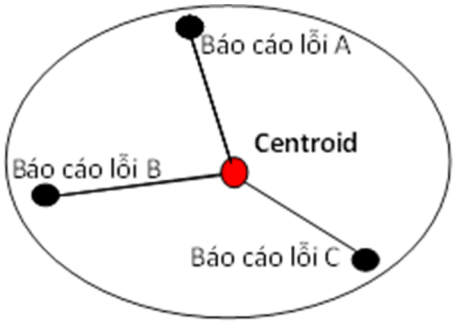
b. Chỉ số tác động bên trong lớp Inter

Để tăng cường độ chính xác trong việc phân loại báo cáo lỗi đối với chỉ số bên trong lớp I_{inter}, trong trường hợp này sử dụng theo phương pháp CFC:

$$I^i_{inte} = \log\left(\frac{|C|}{CF_{t_i}}\right) \quad (3.5)$$

Nếu từ t_i xuất hiện trong tất cả các lớp, khi đó $I^i_{inter} = 0$, do $|C| = CF_{t_i}$, Nếu từ t_i xuất hiện chỉ trong một lớp, khi đó $I^i_{inter} = log\,|C|$. Trong trường hợp này, t_i có sự phân biệt tốt nhất trong các lớp báo cáo lỗi trùng nhau.

3.1.3 Centroids và centroids mở rộng



Hình 5: Mô hình centroid