# Определение архетипа

Будем обозначать группу неточных повторов (near duplicates) некоторого документа как . Количество элементов в обозначим как Группа состоит из повторов: , Везде будем обозначать — индекс группы неточных повторов в наборе всех групп неточных повторов, — индекс неточного повтора в группе.

Каждому можно сопоставить интервал , где - это координата первого символа в (расстояние в символах от начала текста), - координата последнего символа . Будем обозначать Соответственно, длина вычисляется как . Обозначим через среднюю длину повтора в группе:

.

Определим функцию , действующую из декартового произведения множества текстовых строк в множество нтервалов следующим образом. возвращает в случае вхождения строки в строку все интервалы, которые соответствуют этим вхождениям. В противном случае она возвращает . Например:

"123 456 789 456", "456"

"123 456 789", "abc" .

Для группы нечетких повторов будем считать архетипом упорядоченное множество строк таких, что:

1. – интервалы, соответствующие строкам архетипа, непересекаются.
2. – все строки, из которых состоит архетип группы, входят во все нечёткие повторы, образующие данную группу.
3. – для любого элемента длина архетипа должна быть не меньше половины его собственной длины.

# Примеры

1. Пусть группа содержит описания годов рождения писателей:

* Писатель Гоголь родился в 1809 г.
* Писатель Достоевский родился в 1821 г.
* Писатель Лев Николаевич Толстой родился в 1828 г.

Рассмотрим упорядоченный набор строк "Писатель ", " родился в ", " г." . . Средняя длина элементов равна 40, . Выполнимость всех условий архетипа очевидна, значит .

2. Модифицированный предыдущий пример – мы включили в элементы группы (обозначим ее имена-отчества писателей:

1. Писатель Николай Васильевич Гоголь родился в 1809 г.
2. Писатель Фёдор Михайлович Достоевский родился в 1821 г.
3. Писатель Лев Николаевич Толстой родился в 1828 г.

Претендент на архетип остаётся тот же, но увеличилась средняя длина элементов группы – она составила 52 символа. Условие 3 не выполнилось, следовательно S не является архетипом.