# Глава 2. Выбор и обоснование алгоритма

## 2.1. Формальная постановка задачи

*Содержательная постановка задачи*

Имеется фиксированное множество документов. Требуется разбить заданное множество документов на группы в соответствии с их семантической схожестью.

Предусмотреть 2 режима работы:

1. автоматическое выявление групп документов
2. выбор одной группы документа (из заранее определенных групп) на основании его содержания.

В качестве критериев схожести документов при кластеризации используется мера TF-IDF.

**Обозначения и формальная постановка задачи**

Обозначения:

*D* – исходное множество документов коллекции,

*di* ∈ *D* – *i*-й документ множества *D*,

 – *j*-е слово документа *di*,

 – число вхождений *tj*-го слова в документ *di*,

 – общее число слов в документе *di*, 

 – частота *tj*-го слова (term frequency) в документе *di*,

 – обратная частота документа (inverse document frequency) – инверсия частоты, с которой слово  встречается в документах коллекции

– число документов *D*, в которых встречается слово  (когда )

 – булева переменная, равная 1 – если *di*-й документ относится к таксону *pk* и 0 – в противном случае,

 – оценка важности слова в контексте документа (мера TF-IDF)

- евклидово расстояние между *di*-м и *dl*-м документами на основании меры TF-IDF



**Цель** – суммарное расстояния между документами, попавшими в один таксон, должно быть минимальным.

## 2.2. Алгоритм

### 2.2.1. Кластеризация

*Шаг 1*. Считать содержимое всех файлов и объединить слова в единый список, предварительно определив основу каждого слова;

*Шаг 2*. Строим матрицу схожести документов, используя выбранную меру расстояния и TF-IDF;

*Шаг 3*. Используя алгоритм Прима, строим минимальное остовное дерево и затем удаляем ребра больше заданной длины, тем самым разбивая множество документов на группы;

### 2.2.2. Классификая

*Шаг 1*. Считать содержимое всех файлов и объединить слова в единый список, предварительно определив основу каждого слова;

*Шаг 2*. Считываем файл, содержащий описание классов и добавляем их ключевые слова в список слов;

*Шаг 3*. Считаем частоту слов в каждом документе;

*Шаг 4*. Нормализуем матрицу;

*Шаг 5.* Сравнимаваем расстояния i-го документа с j-ой группой и определяем самою подходящую.