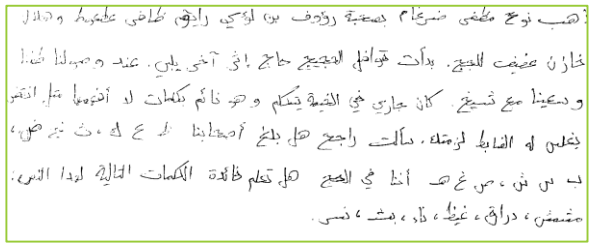
Алгоритм автоматического определения автора рукописного арабографического документа по почерку

Павлов В. А.
СПбГУ
2015 г.

Научный руководитель: к. ф.-м. н. Шалымов Д. С.

Задача

Цель: создание компьютерной системы для автоматического распознавания автора рукописного арабографического документа



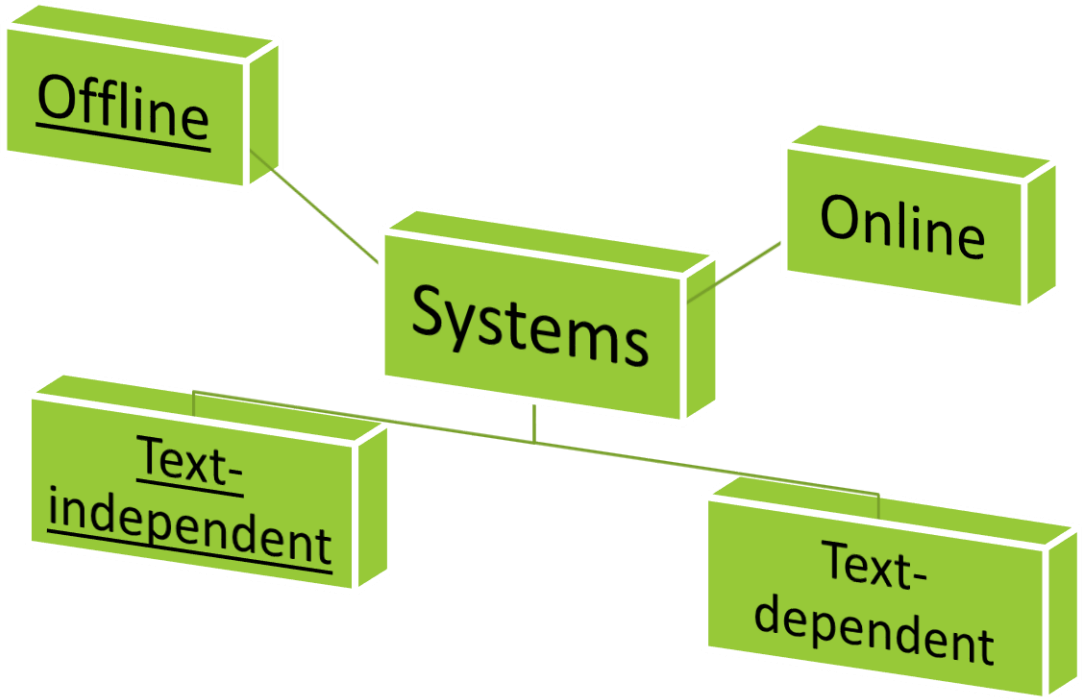
Входной документ

Система

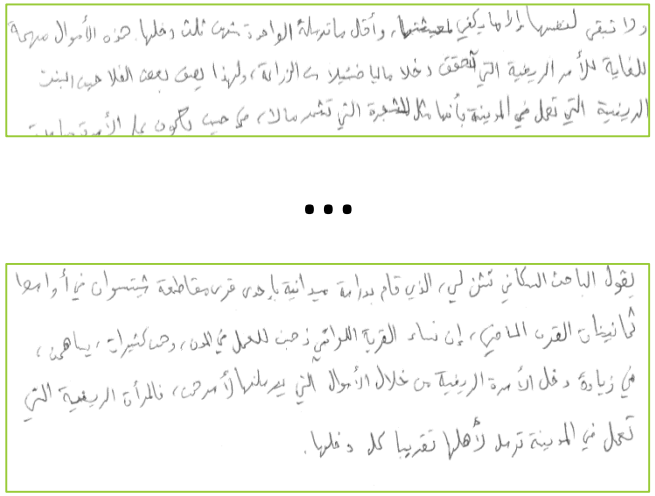
Существующие системы

Арабский язык:
Системы со средней точностью до 96%

Персидский язык:
Система B.Helli и M.E. Moghaddam с точностью около 100% при наличии достаточного количества тренировочных данных

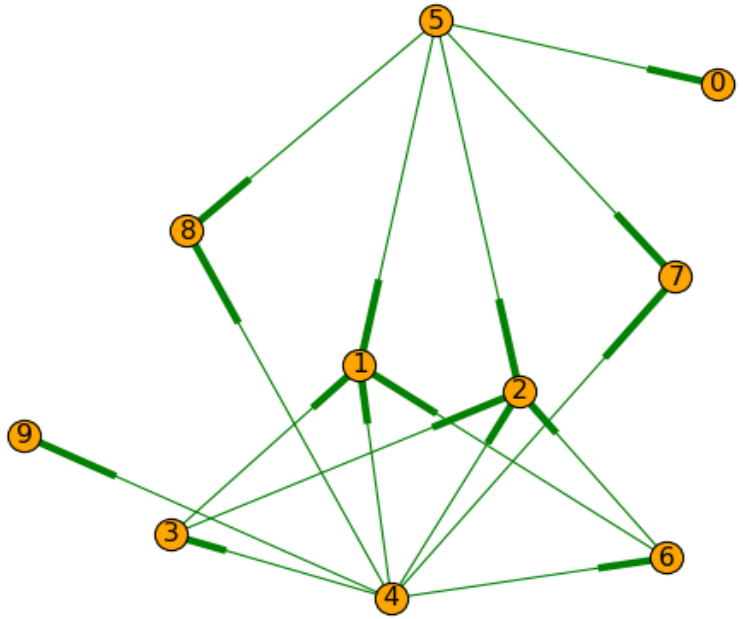


Основные концепции системы



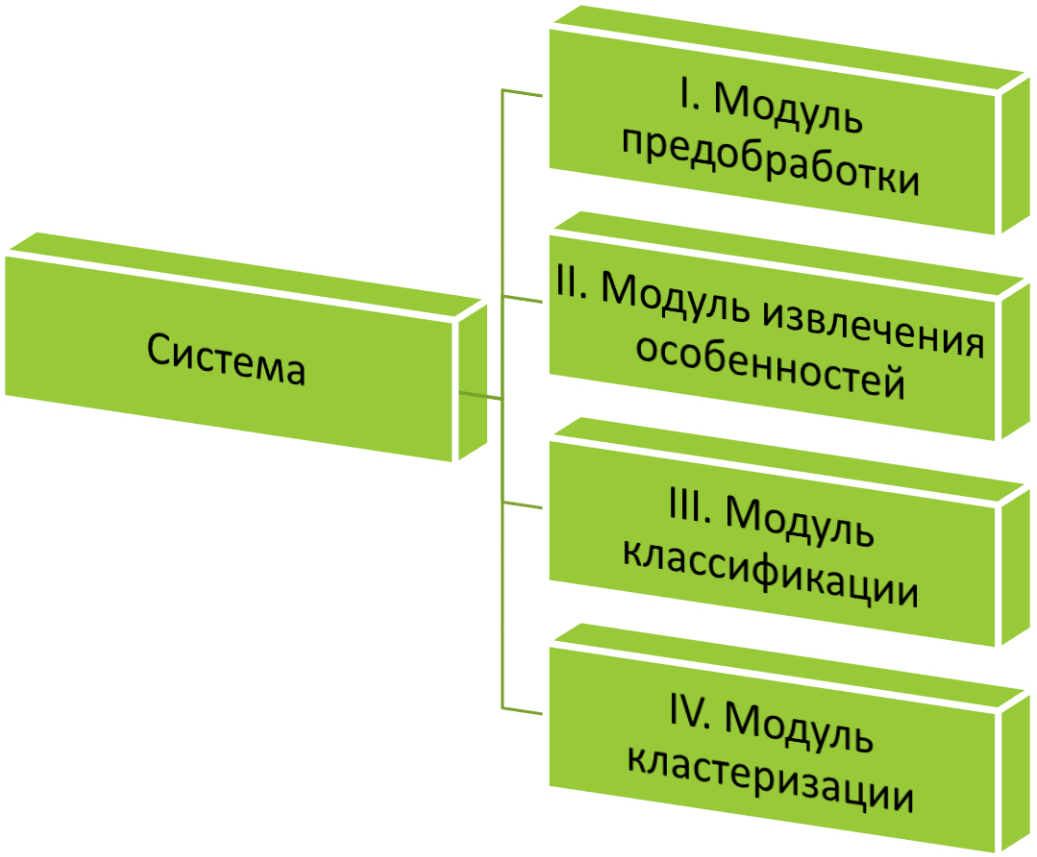
...

Рукописи автора



Граф отношения особенностей (Feature Relation Graph - **FRG**)

Устройство системы

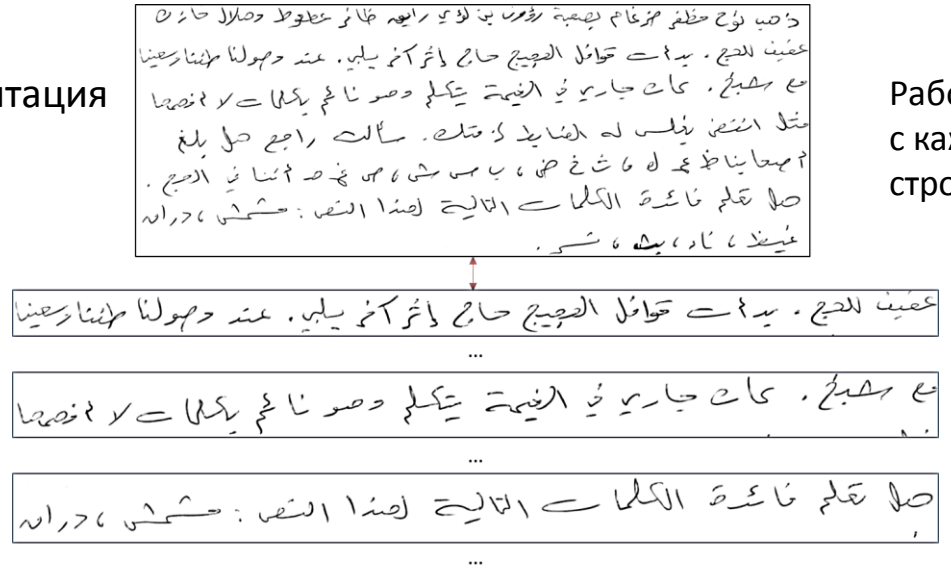


I. Модуль предобработки

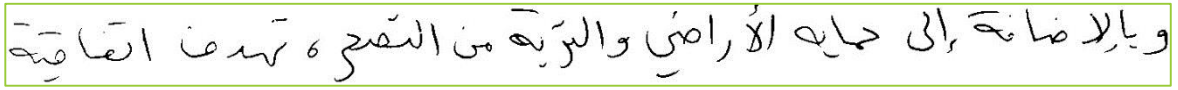
Требования на входные данные: сегментированные на строки тексты(ручная + готовая)

Описание работы: удаление лишних пикселов и наращивание линии текста для удаления нежелательных разрывов и/или сгущений

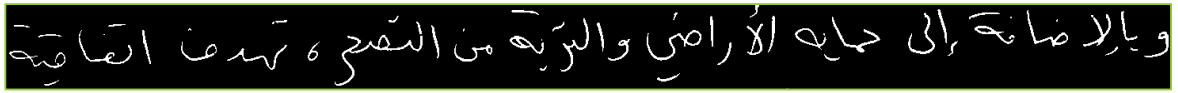
Сегментация



Работа с каждой строкой



Предобработка(Библиотека OpenCV)



II. Модуль извлечения особенностей

Процесс работы: использование Gabor и XGabor фильтров

Gabor filter – детекция определенно ориентированных локальных паттернов (параметры: $\lambda, \theta, \psi, \sigma, \gamma$)

XGabor filter – детекция криволинейных паттернов определенного устройства (параметры: λ, r_x, r_y)

Проблема: подбор параметров фильтров для входных текстов

Gabor filtering($\lambda = 8, \theta = 0, \psi = 0, \sigma = f(\lambda), \gamma = 1$)

XGabor filtering($\lambda = 8, r_x = 3, r_y = 1$)

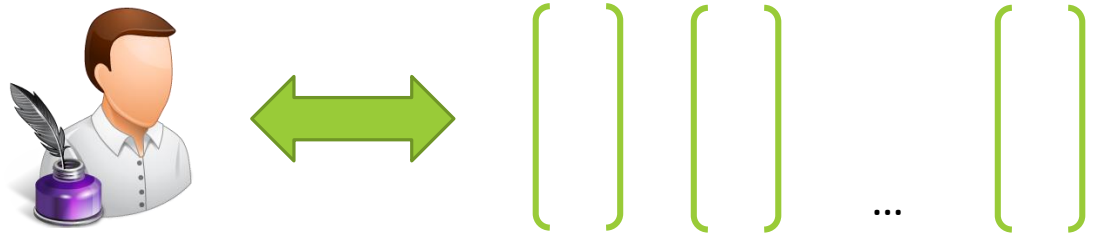
II. Векторы особенностей

- Серия применений различно ориентированных фильтров Gabor к входной строке
- Серия применений настроенных на различные кривые фильтров XGabor к входной строке

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Вектор особенностей

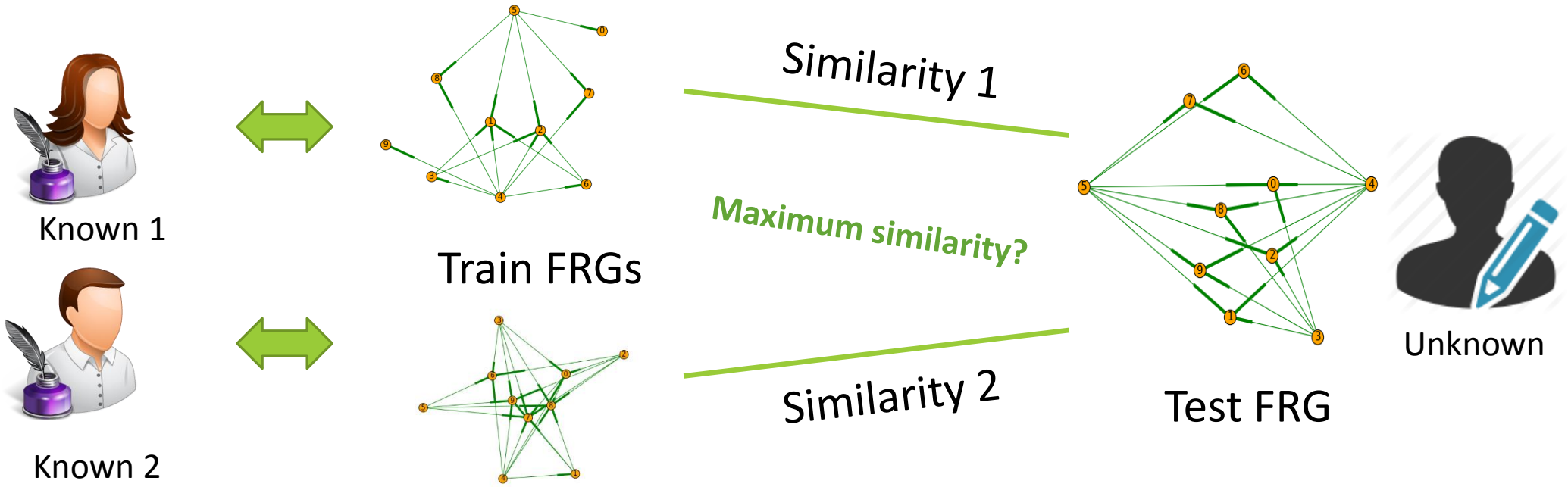
- Итог: автору сопоставляется набор $n-$мерных векторов



III. Модуль классификации

Назначение: генерирование FRG графов и работа с ними

Описание работы: создание FRG графов эмпирическими методами на основе полученных векторов особенностей и вычисление специальных мер сходства между двумя графами

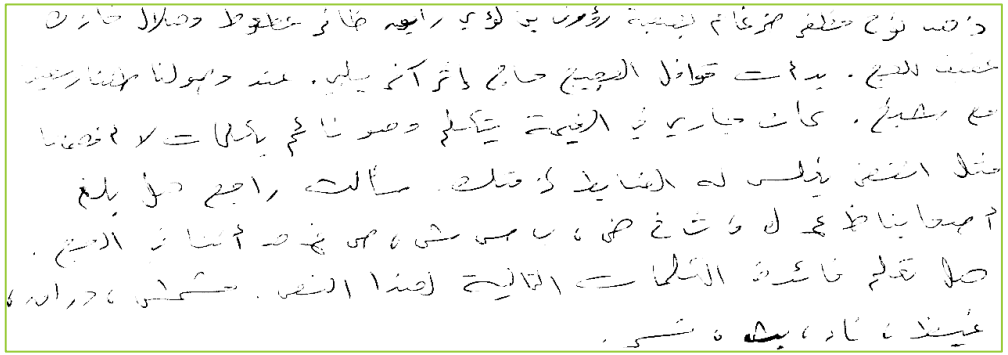


III. Feature Relation Graph

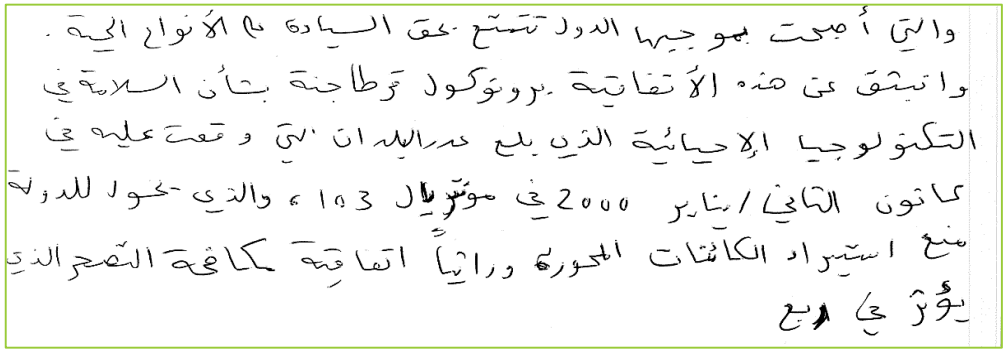
По какому принципу строится FRG?

- Вершина x – особенность x
- Ребро (x, y) – тенденция особенности y численно превосходить особенность x

A



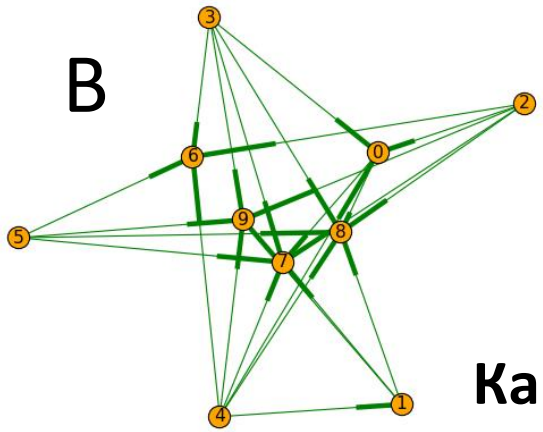
B



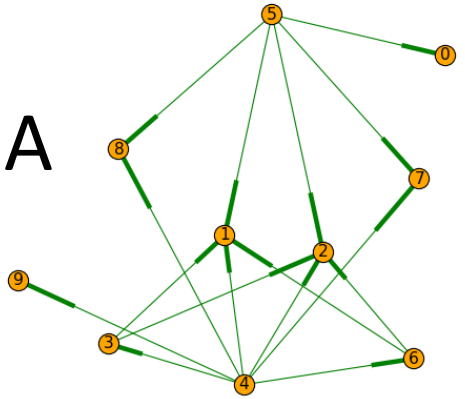
Замечание:

- FRG - ориентированный ацикличные графы

B

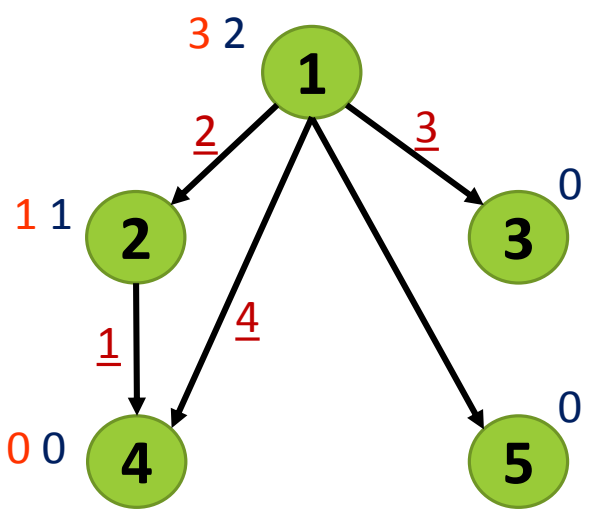


A

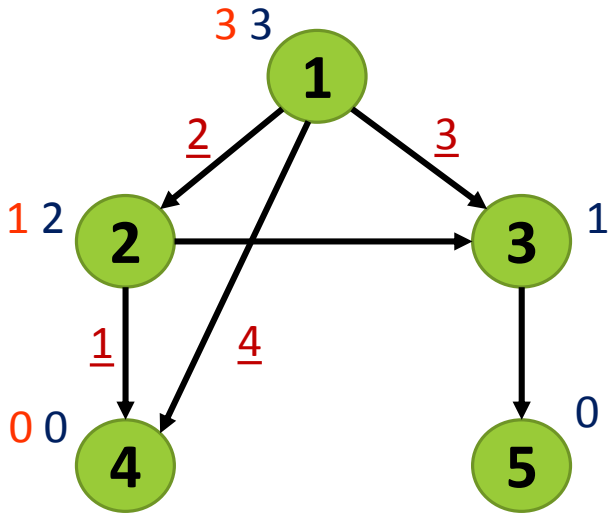


Как сравнить?

III. Сравнение двух почерков



G_1



G_2

Similarity(G_1, G_2) = **4**

Алгоритм:

- (*)Вычисление высоты для каждой вершины
- (*)Сортировка общих ребер графов по возрастанию

$(a, b) > (c, d) \Leftrightarrow \text{height}(a) > \text{height}(c)$ in some graph

- (*)Вычисление величины T для каждой вершины отсортированных общих ребер графов, где

$$T(v) = \begin{cases} 0, & \text{если } v - \text{лист} \\ \sum_{u \text{ in } nieghbors(v)} 1 + T(u), & \text{иначе} \end{cases}$$

- (*)$Similarity(G_1, G_2)$ вычисляется как сумма величин T по вершинам общих ребер графов

III. Эксперименты с алгоритмом

Материалы: база данных **KHATT(KFUPM Handwritten Arabic TexT Database)**

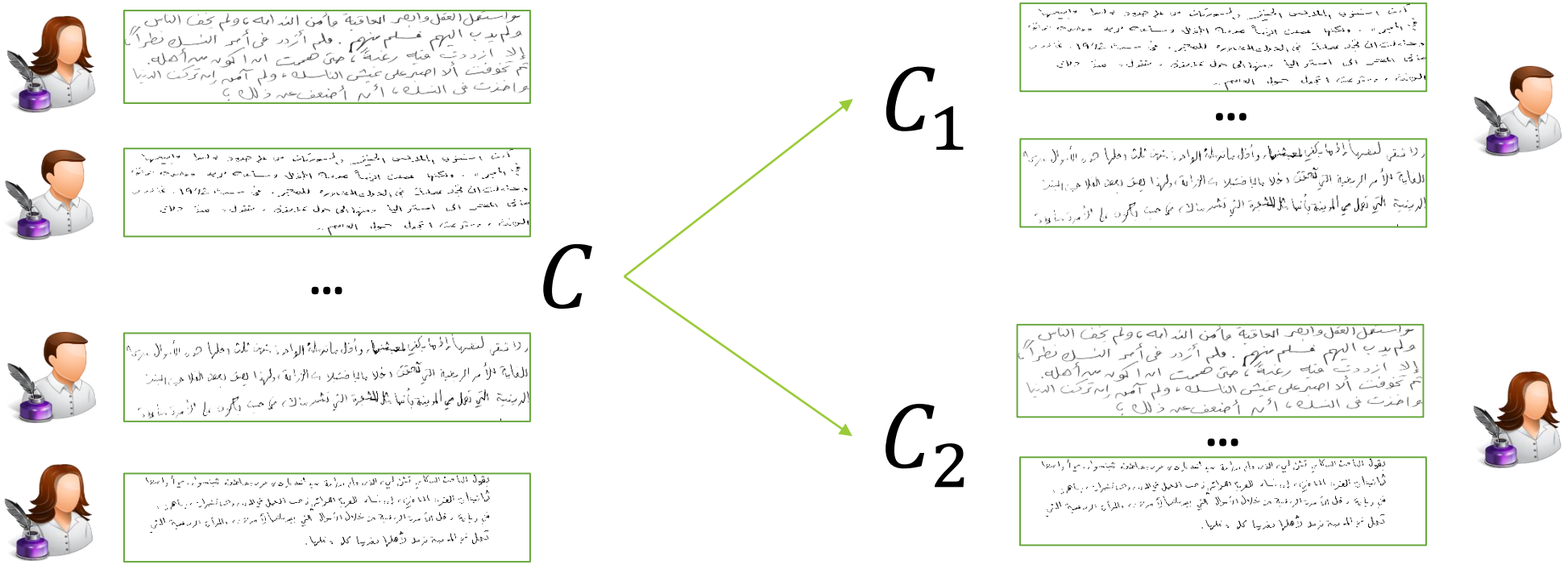
- Заранее сегментированные тексты

Authors	Train/Test	Features	Precision(%)
3	4/8	8	66
3	8/4	8	100
3	4/8	24	100
3	8/4	24	100
10	4/8	8	50
10	8/4	8	70
10	4/8	24	80
10	8/4	24	80
15	4/8	24	40
15	8/4	24	45
15	4/8	36	40
15	8/4	36	50
20	8/4	36	45
30	8/8	36	40

IV. Модуль кластеризации

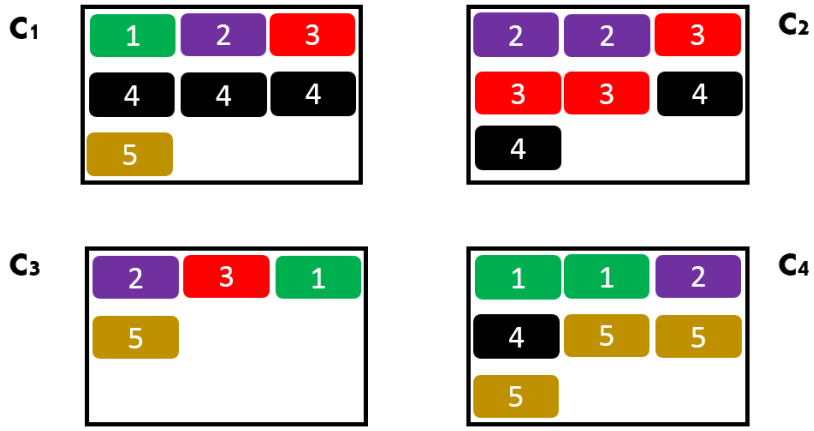
Назначение: кластеризация набора документов

Описание работы: разбиение множества документов на группы по потенциальному авторству



IV. Cost function

- Результаты кластеризации не всегда идеальны – **как оценить**?
- Минимизация функции: $Cost(C_1, C_2, \dots, C_k) \to \min$
- Алгоритм вычисления $Cost(C_1, C_2, \dots, C_k)$:
  - Для каждого автора A:
    - Выбираются «**материнские кластеры**» – кластеры **наименьшего размера** среди тех, что содержат **наибольшее количество рукописей автора.**
    - Ищется кластер X **без хозяина**.
      - X найден:
        - A помечается как хозяин X.
        - Rate A = $\frac{\text{Число документов A в X}}{\text{Общее число документов A}}$ (*)
      - X **не** найден:
        - Среди материнских кластеров ищется Y , чей обладатель B имеет наименьшее Rate.
        - U — обладатель большего числа документов в Y среди A и B. V – другой автор.
        - Rate U пересчитывается в соответствии с (*). Rate V: $= 0$.
  - $Cost(C_1, C_2, \dots, C_k) = \frac{\sum_{A\ in\ authors} Rate(A)}{Author\ amount}$



$Cost(C_1, C_2, C_3, C_4) = 0.36$

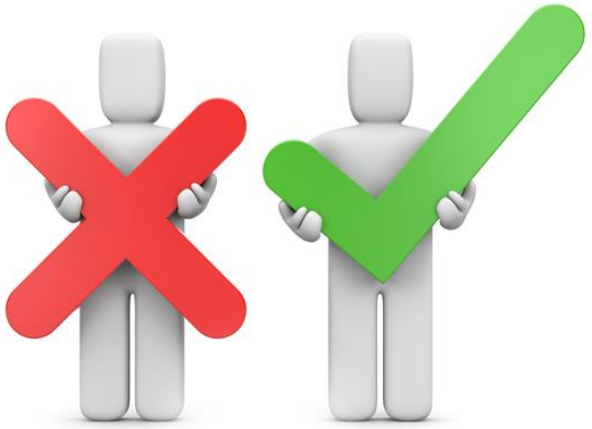
IV. Эксперименты с кластеризацией

Authors	Features	K-Means	Online K-Means	PAM	DBSCAN
3	15	33	33	60	33
3	32	66	33	60	33
3	64	66	50	66	66
10	15	10	10	55	33
10	32	20	20	65	40
10	64	20	25	65	55
20	15	5	5	45	12
20	32	8	5	53	20
20	64	10	8	55	25

Выводы

- Алгоритм идентификации показал **достойные результаты** на **арабографических рукописных текстах** при наличии **достаточного количества тренировочных данных и небольшого числа авторов**

- Алгоритм **может быть пригоден** для использования при кластеризации документов лишь **небольшого** числа авторов с использованием алгоритмов кластеризации **PAM**



Результаты

- **Реализована система** для автоматической идентификации автора арабографического рукописного документа по почерку его создателя
- Была **проведена настройка системы** для работы с арабографическими рукописными документами
- **Проведен ряд тестов** для анализа работы алгоритма для решения **задачи классификации** арабографических рукописных документов
- В систему была **добавлена функциональность кластеризации** входного множества документов с использованием алгоритмов K-Means, Online K-Means, PAM, DBSCAN
- Был **произведен ряд экспериментов** для установления точности используемого алгоритма при решении **задачи кластеризации**
- Подача заявки и **прохождение на конференцию** ICDAR-2015(International Conference on Document Analysis and Recognition)