


Алгоритм автоматического определения автора рукописного арабографического документа по почерку

Павлов В. А.
СПбГУ
2015 г.

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

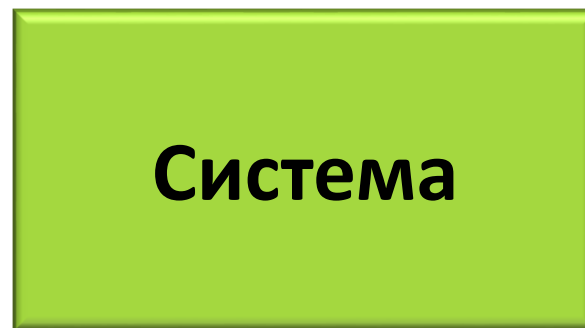


Задача

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

هـب نوح مكنى شرفاً مـصعبه روف بن لوكي رادق طلافى عطيه و دلال
خازن مـصعبه للصحف. بدأت قوائمه العجيبه حاجه اى اى يابن. عند وصلنا اننا
ودعينا مع شيخه كان جاري في الضيقه يتكلم و هو قائم كلمات له اضمها مثل القوم
بغضه له القابله لورمه. مـصعبه راجع هل بلغه اصحابنا لـ ع كـ مـ نـ بـ
بـ مـ نـ مـ نـ غـ هـ انا في الصحف هل تعلم طاقه الكلمات التاليه لودا التـ:
مـصعبه، دراق، غيطه، مـ، مـصعبه، مـصعبه.

Входной документ



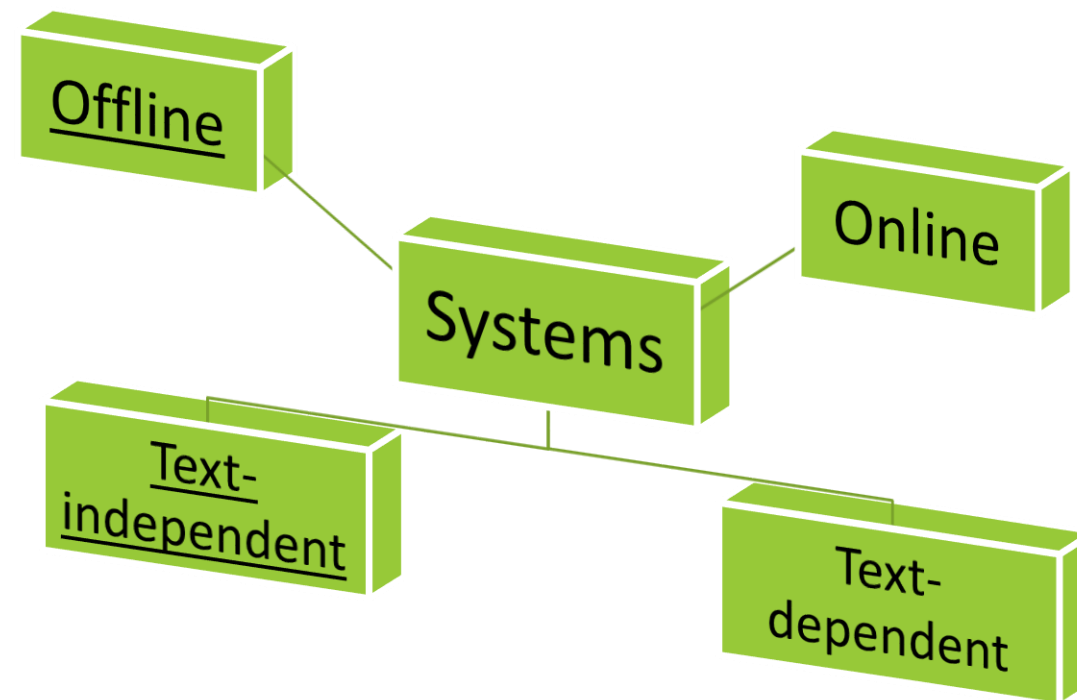
Существующие системы

Арабский язык:

Системы со средней точностью до 96%

Персидский язык:

Система V.Helli и M.E. Moghaddam с точностью около 100% при наличии достаточного количества тренировочных данных



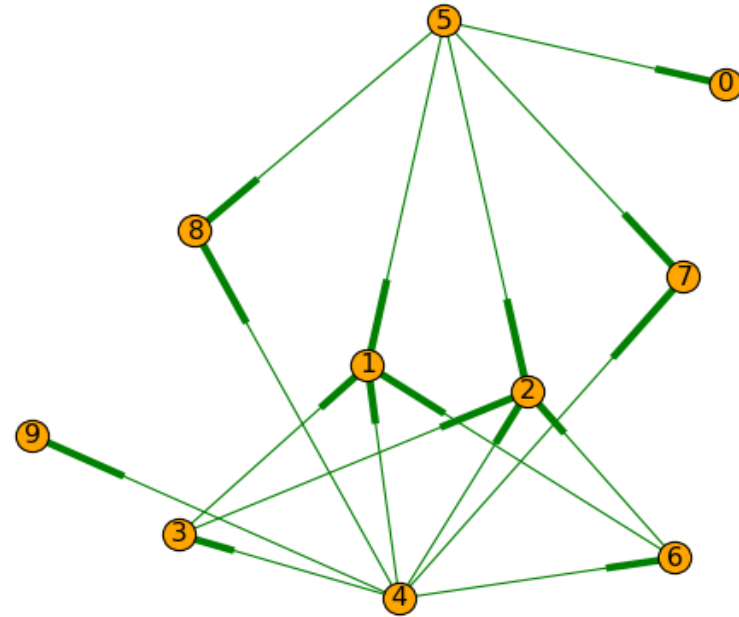
Основные концепции системы



ولا تنقري لنفسها إلا ما يكفي لمعيشتها، وأقل ما تنسأه الواحدة شتمت ثلاث دغليها. هذه الأسمال صريحة للغاية للأمر الريفيّة التي تصقّت دخلاً بالياً ضئيلاً من الزراعة، ولهذا يصعب دفع الفلاحيين البنيت الريفيّة التي تعمل في المدينة بأنماثل للشجرة التي تكسر سالماً، مما يجب أن يكون على الأسمال...

...

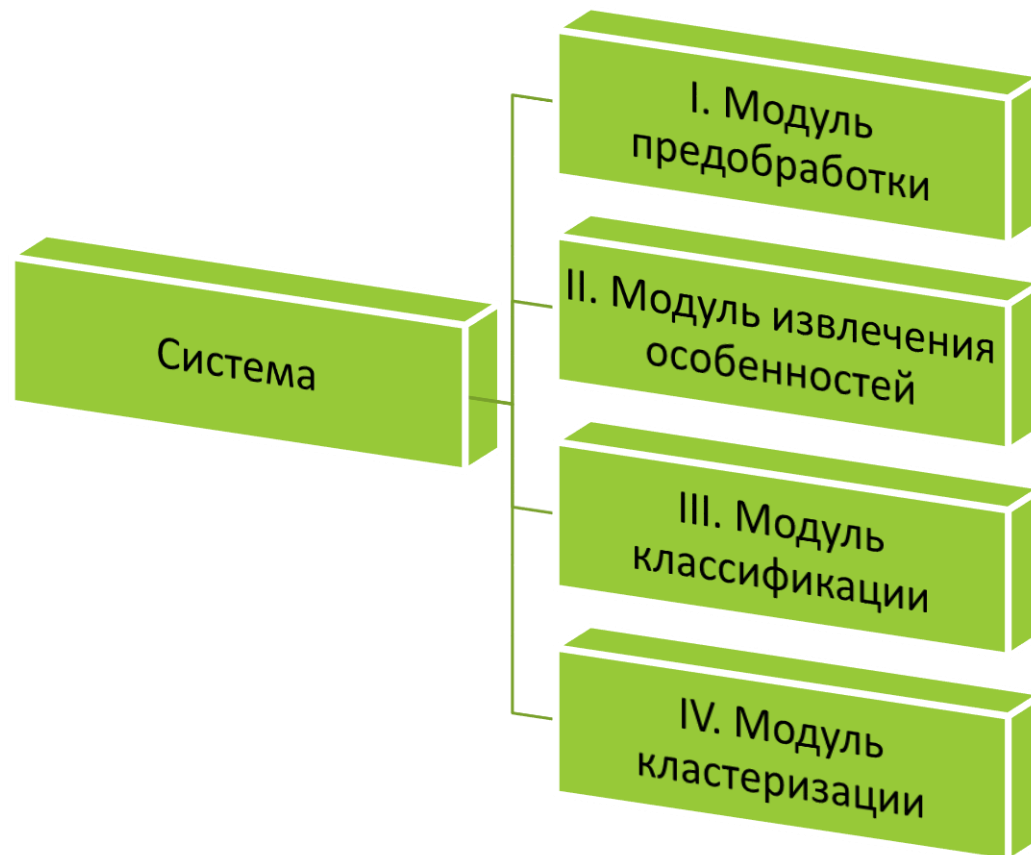
يقول الباحث السكّان تشن لي، الذي قام بدراسة ميدانية بإحدى قرى مقاطعة شينجوان نينغشيا وأنها ثماً أيضاً في القرن الماضي، إن نساء القرية اللواتي ذهبن للعمل في المدن، وحصاً كثيراً، يباهين، في زيادة دخل الأسرة الريفيّة من خلال الأسمال التي يربطنها لأمرهن، فالمرأة الريفيّة التي تعمل في المدينة ترمد لأهلها تقريباً كل دغليها.



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 пикселей и наращивание линии текста для удаления нежелательных разрывов и/или сгущений



II. Модуль извлечения особенностей

Процесс работы: использование Gabor и XGabor фильтров

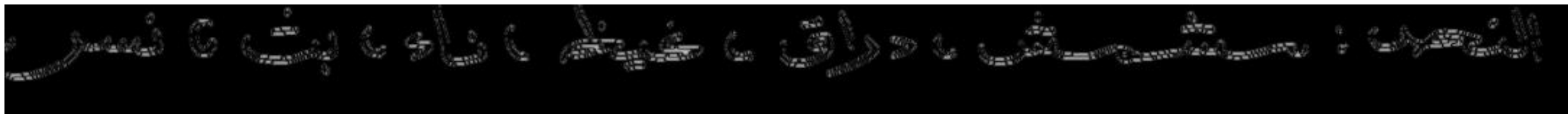
Gabor filter – детекция определенно ориентированных локальных паттернов (параметры: $\lambda, \theta, \psi, \sigma, \gamma$)

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



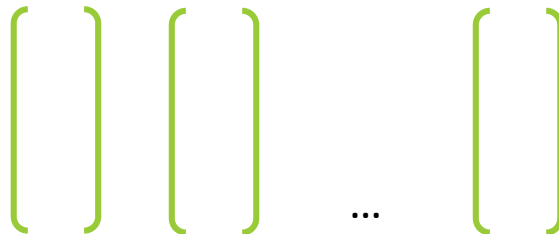
- Серия применений различно ориентированных фильтров Gabor к входной строке
- Серия применений настроенных на различные кривые фильтров XGabor к входной строке



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

Вектор особенностей

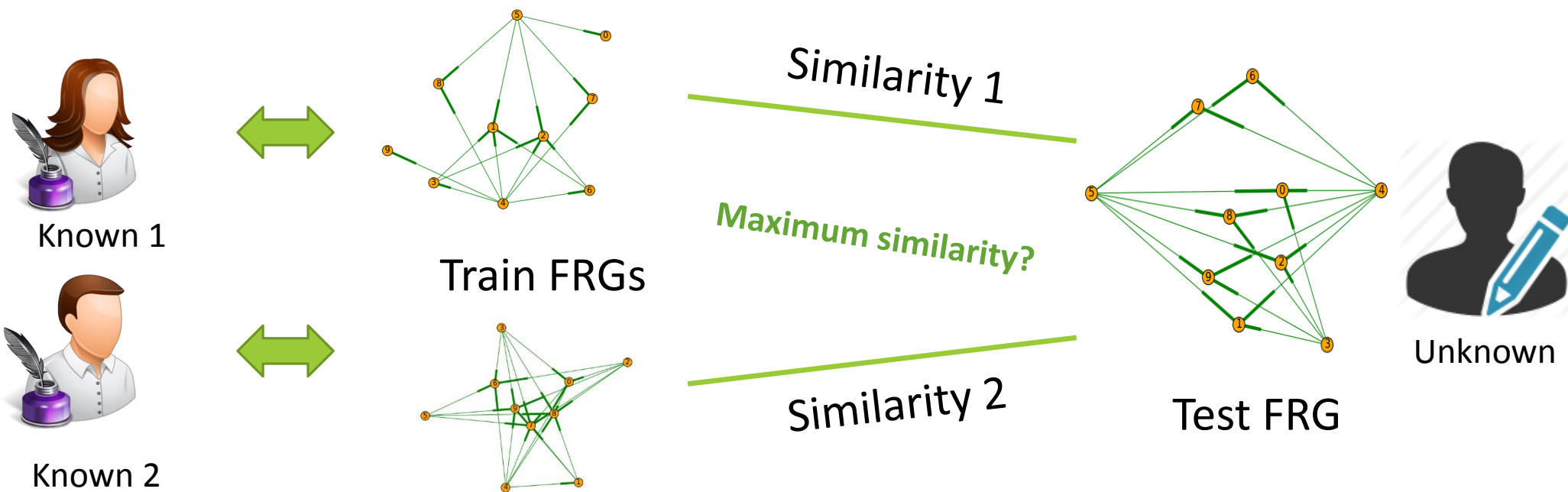
- Итог: автору сопоставляется набор n – мерных векторов



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



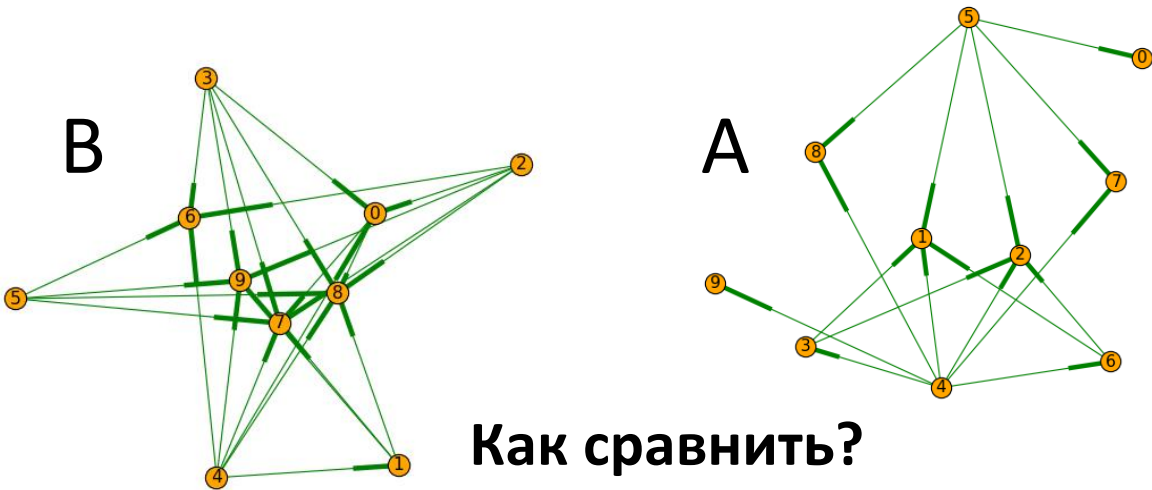
III. Feature Relation Graph

По какому принципу строится FRG?

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

Замечание:

- FRG - ориентированный ациклические графы



Как сравнить?

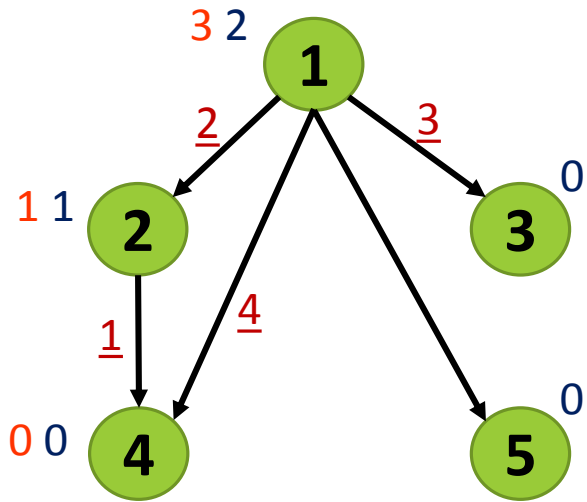
A

دائمًا نوضح مفهوم الرسوم البيانية باستخدام أمثلة بسيطة مثل العلاقات بين الأشخاص. عند دمجها مع الرسوم البيانية، يمكن استخدامها لتحليل الاتجاهات في البيانات. على سبيل المثال، يمكن استخدامها لتحليل العلاقات بين الأشخاص في الشبكات الاجتماعية. في هذا السياق، يمكن استخدامها لتحليل العلاقات بين الأشخاص في الشبكات الاجتماعية. في هذا السياق، يمكن استخدامها لتحليل العلاقات بين الأشخاص في الشبكات الاجتماعية.

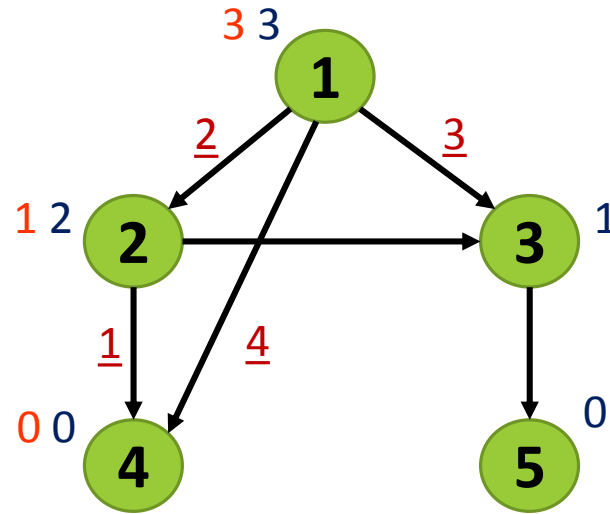
B

والتي أصبحت بموجبها الدول تتسع بحق القيادة في الأنواع الحديثة. وانتدقت عن هذه الاتفاقيات بروتوكول ترطاجنة بشأن السلامة في التكنولوجيا الإحصائية الذين بلغ عدد البلدان التي وقعت عليه في كانون الثاني/يناير 2000 في مؤتمر بال 103، والذي يحول للدولة منع استيراد الكائنات المحورة وراثياً اتفاقية مكافحة التصحر الذي يؤخر في وضع

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
 $(a, b) > (c, d) \Leftrightarrow height(a) > height(c)$ in some graph
- (*) Вычисление величины T для каждой вершины отсортированных общих ребер графов, где

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

- (*) $Similarity(G_1, G_2)$ 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



واسمعت العقل وادبر الحاقبة فأمن الله به ، ولم تحف الناس
 ولم يدب اليهم فلم ينهزم . فلم أررد في أمير السراج نظراً
 إلى أن أردت فيه رفعة ، حين صحبت ابنه أ كونه سراً له .
 ثم خوفت أن اصير على عيش النسله ، ولم آمنهم ولم تركت الدنيا
 فاجتهدت في النسله ، إنهم أضعف عن ذلك ؟



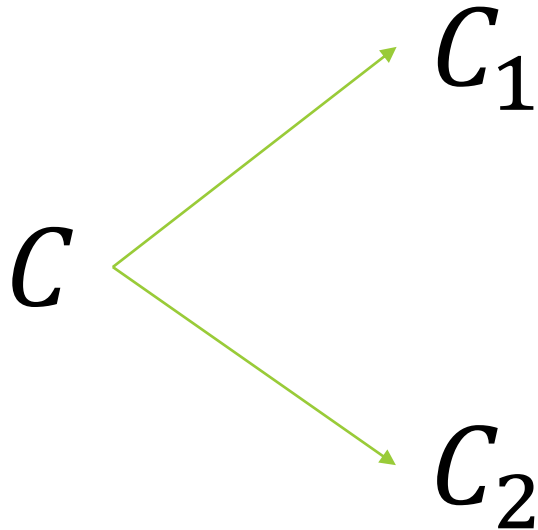
أبنت استغوى بالمعدن والمغنين والمسترشك من العجمين ، فداسا ما يفسد
 في الناس وكنها عصمت ابنها واصلت بوجه حسنة غناه
 وحسنة ابنه أن تجرد صفات عن العبدية المذمومة المعتبرة ، عن حسنة 19/12 ، فاجتهدت
 منقاة المكنت إلى استغراية منسوبة إلى جود عذمتها ، فاشرق ذلك
 الجليلة ، وستر بعثه فيقول ، حول العاقبة . . .



ردا شق لسفيرا بالما يكفي للعبثية ، وأول بالبرلة الواردة تبتن ثلث وأهلا فرد الأول ، مرد
 العناية لأمر الرعية التي تتفكت دخلا باليا ضلوا به الزارة ، ولمها يلق بعض هؤلاء ، يفتن
 الرعيبة التي تقل من الموية بأنها بالاشيرة التي تفسر منلا ، كما حيث تكون بال الأمر والناوون



فقول الثباتت السكار تشر على الجاد ، وأول بالبرلة مرد عالمة شينوا ، في أو واهدا
 فإستياج العزة ، الماء في ، في نسام الفرية العلام زعت العمل بخلا ، ووه تشوا ، ويا مقر ،
 فن رباية ، وفل أن من الرعية من خلال الأول ، فكن يهالبا أو مرد ، وفلان ورعية التي
 أقول توالله نرسو أهلا مفرحيا كل ، فاليا .



أبنت استغوى بالمعدن والمغنين والمسترشك من العجمين ، فداسا ما يفسد
 في الناس وكنها عصمت ابنها واصلت بوجه حسنة غناه
 وحسنة ابنه أن تجرد صفات عن العبدية المذمومة المعتبرة ، عن حسنة 19/12 ، فاجتهدت
 منقاة المكنت إلى استغراية منسوبة إلى جود عذمتها ، فاشرق ذلك
 الجليلة ، وستر بعثه فيقول ، حول العاقبة . . .

ردا شق لسفيرا بالما يكفي للعبثية ، وأول بالبرلة الواردة تبتن ثلث وأهلا فرد الأول ، مرد
 العناية لأمر الرعية التي تتفكت دخلا باليا ضلوا به الزارة ، ولمها يلق بعض هؤلاء ، يفتن
 الرعيبة التي تقل من الموية بأنها بالاشيرة التي تفسر منلا ، كما حيث تكون بال الأمر والناوون

واسمعت العقل وادبر الحاقبة فأمن الله به ، ولم تحف الناس
 ولم يدب اليهم فلم ينهزم . فلم أررد في أمير السراج نظراً
 إلى أن أردت فيه رفعة ، حين صحبت ابنه أ كونه سراً له .
 ثم خوفت أن اصير على عيش النسله ، ولم آمنهم ولم تركت الدنيا
 فاجتهدت في النسله ، إنهم أضعف عن ذلك ؟

فقول الثباتت السكار تشر على الجاد ، وأول بالبرلة مرد عالمة شينوا ، في أو واهدا
 فإستياج العزة ، الماء في ، في نسام الفرية العلام زعت العمل بخلا ، ووه تشوا ، ويا مقر ،
 فن رباية ، وفل أن من الرعية من خلال الأول ، فكن يهالبا أو مرد ، وفلان ورعية التي
 أقول توالله نرسو أهلا مفرحيا كل ، فاليا .



IV. Cost function

- Результаты кластеризации не всегда идеальны – как оценить?

- Минимизация функции: $Cost(C_1, C_2, \dots, C_k) \rightarrow \min$

- 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 .

- $Rate A = \frac{\text{Число документов } A \text{ в } X}{\text{Общее число документов } A}$ (*)

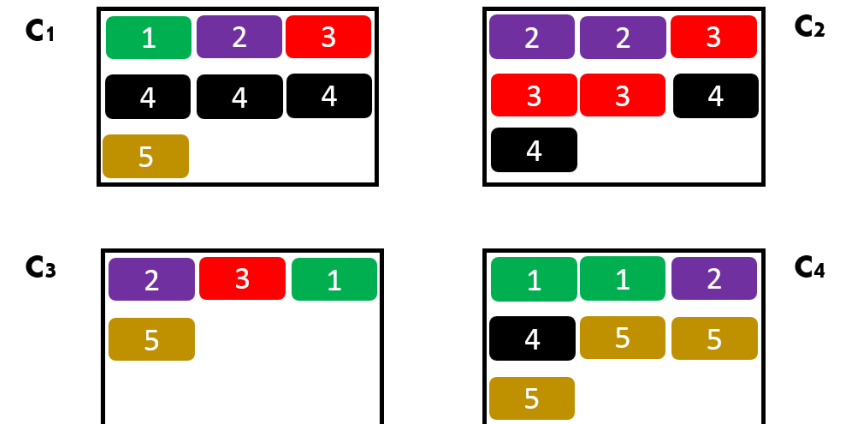
- X не найден:

- 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 (*). $Rate V = 0$.

- $Cost(C_1, C_2, \dots, C_k) = \frac{\sum_{A \text{ in authors}} Rate(A)}{Author \ amount}$



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

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)

