

Deduplicação e Blocagem

Luiz Fernando Böhm
lfbohm@inf.ufrgs.br



Definição

- **Deduplicação**

- Processo de determinar se pares de informações representam uma mesma entidade em uma mesma base ou em diferentes bases de dados.
 - Informação: Tuplas de uma tabela, registros de um arquivo.
 - Entidade: Pessoa, paciente, empresa, produto, etc.
- Ou seja: identificar quando dois dados quaisquer são na verdade o mesmo!

Definição

- **Deduplicação**

- Principais aplicações:

- Fazer um *merge* dos dados que representam a mesma entidade
 - Remover os dados duplicados

- Problema:

- Comparar todos x todos = $O(n^2)$

Termos Relacionados

- **Record Linkage**
- **Data Linkage**
- **Data Matching**
- **Entity Resolution**
- **Entity Disambiguation (SBBD 2011)**
- **Merge and Purge**

Etapas da Deduplicação

- **Pré-processamento dos dados**
- **Montagem de uma base de dados de testes avaliada**
 - **Ferramenta Febrl**
- **Escolha dos métodos de matching**
 - **Determinísticos x Probabilísticos**
- **Execução da deduplicação**
- **Avaliação dos resultados**
 - **Precisão e Revocação, F-Measure**
 - **Falsos Positivos e Falsos Negativos**

Pré-processamento

- **Preparação dos dados para execução do algoritmo de deduplicação**
- **Representa em torno de 90% do esforço de deduplicação**
- **É um processo cíclico que pode ser reiniciado várias vezes conforme são executadas as etapas posteriores**
- **Pode sofrer alterações para melhorar o desempenho da deduplicação ou para atender exigências dos algoritmos**

Pré-processamento

- **Normalização de atributos numéricos**
 - Valores entre 0 e 1
- **Padronização dos dados**
 - Alterar campos texto inteiramente para maiúsculas ou minúsculas
 - Remover acentos, cedilha e caracteres inválidos
- **Remoção de ruídos**
 - Validar CPF, CEP, Datas, etc.
- **Criação de novos campos**
- **Criação de índices simples e compostos**

Base de Dados de Testes

- **Auxiliar na escolha do método de deduplicação mais adequado**
- **Avaliar qualidade da deduplicação**
- **Estimar tempo de execução da deduplicação**

Base de Dados de Testes

- **Como montá-la?**
 - Base de dados de testes deve manter as mesmas características da base de dados completa (conceito de resampling de KDD)
 - Profiling dos dados
 - Média e desvio padrão dos valores
 - Proporção entre as classes
 - Proporção de erros: erros de digitação, fonéticos e OCR

Base de Dados de Testes

- **Não há um tamanho ideal**
 - Quanto maior o número de registros, melhor!
- **Idealmente, é um especialista no domínio que deve fazer a avaliação dos dados da base de testes**

- **Freely Extensible Biomedical Record Linkage**
 - Open Source
 - Python e GUI
 - Padronização e limpeza dos dados
 - Geração automática de bases de dados de testes

FEBRL

▼ Febri - (None)*

File Tools Help

Execute New Open Save Quit

Standardisation Deduplication
Geocoding Linkage

Data Explore Index Compare Classify Output/Run Evaluate Log

First data set type: ☒ CSV ☐ COL ☐ TAB ☐ SQL Missing values: View Data Edit Data

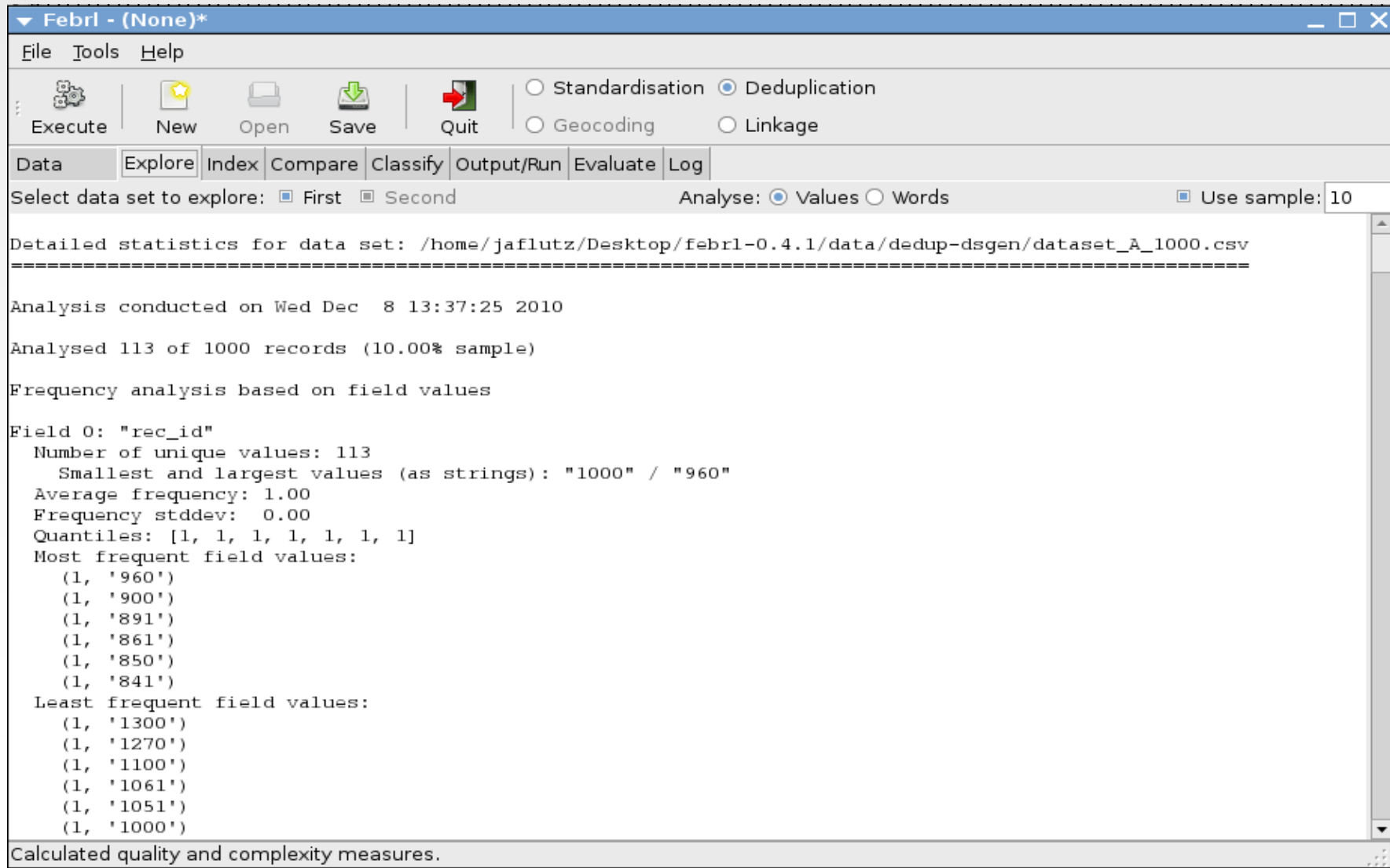
Filename: dataset_A_1... Delimiter: Use headerline Strip fields

Record identifier field: rec_id

rec_id	given_name	surname	street_number	address_1	address_2	suburb	postcode	state	da
1491	elle	lee	2	ellerston avenue	rosedale	epping	7000	nsw	191
2551	zachary	kanhanouvong	2	healy place	belmont cottage	northam	4133	nsw	197
2661	cooper	gillick	23	templestowe avenue		tully	2560	vic	198
3501	joshua	dodson	16	templeton street		alexander heights	2264	nsw	192
3451	tiaza	karandakis	1	cockle street	locn 6357	turramurra	2010	vic	
3201	mitchell	lawrence	7	stace place		lindfield	2142	nsw	193
620	lachlan	monteleone	6	kingsbury street		torrensville	2042		196
1420	lachlan-john	corby	17	astelia place	millamurra	vincentia	2153	qld	193
2581	jasmine	mchenry	8	tristania street		tarragijdi	6530	vic	190
3180	teegan	notley	4	clark close		charters towers	2827	nsw	193
1761	alissa	drumgoon	159	pridham street		phillip bay	2484	vc	190
330	leah	tarrant	33	gellibrand street	laura downs	griffith	4066	sa	191
3220	nacoya	rowe	5	colvin street		teneriffe	2482	nsw	199
3201	mitchell	lawrence	7	stace place		lindfield	2142	nsw	193

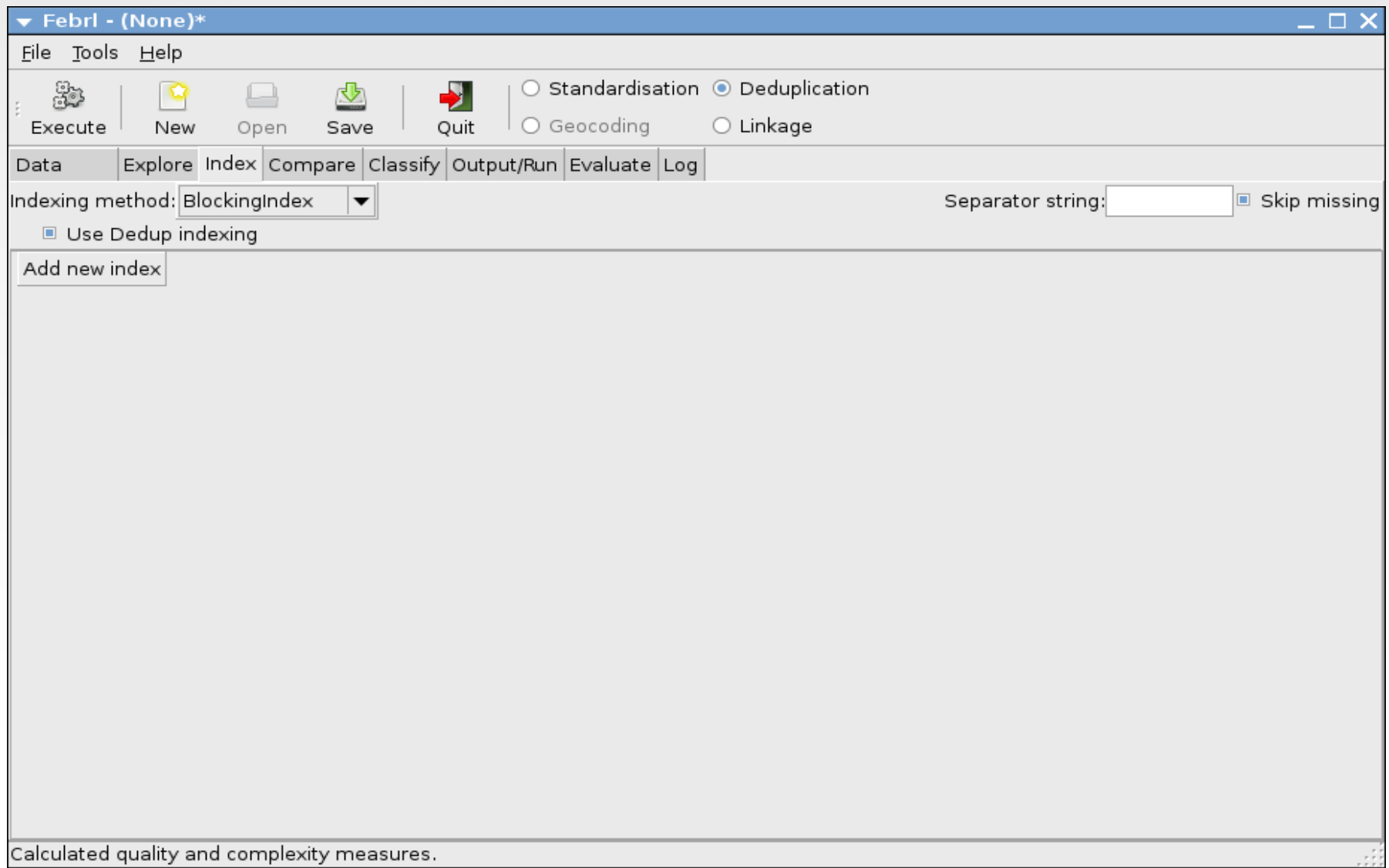
Calculated quality and complexity measures.

Carregando a Base de Dados



Visualização de Estatísticas da Base de Dados (Profiling)

FEBRL



Esquema de Indexação da Base de Dados

▼ Febrl - (None)*

File Tools Help

Execute New Open Save Quit

☐ Standardisation ☒ Deduplication
☐ Geocoding ☐ Linkage

Data Explore Index Compare Classify Output/Run Evaluate Log

Field comparison function: Edit-Dist ▼

Field name A: given_name ▼ Field name B: given_name ▼ ☐ Cache comparisons Maximum cache size: None

Missing value weight: 0.0 Agreeing value weight: 1.0 Disagreeing value weight: 0.0

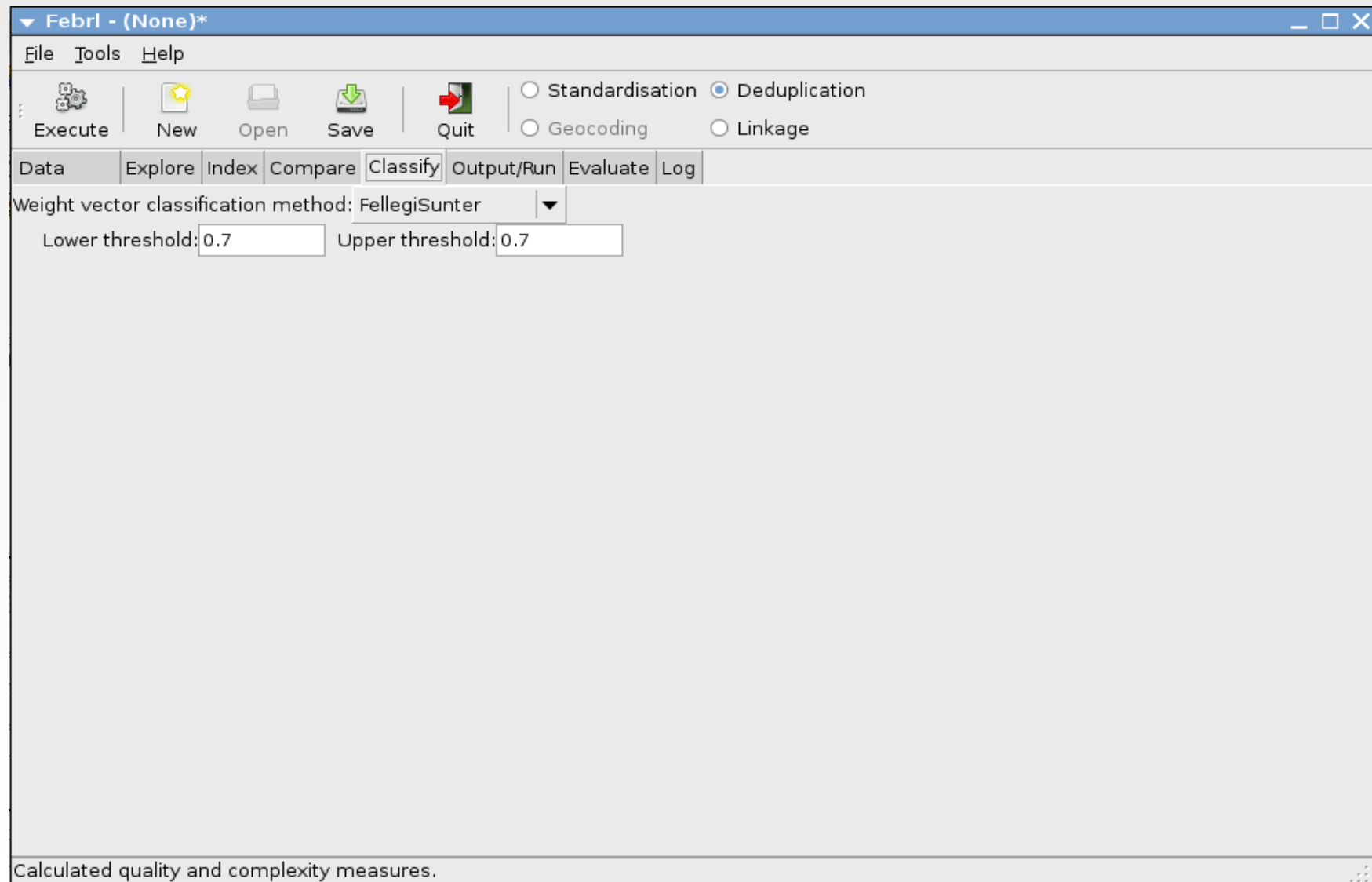
Threshold: 0.7

Add new comparison function

Calculated quality and complexity measures.

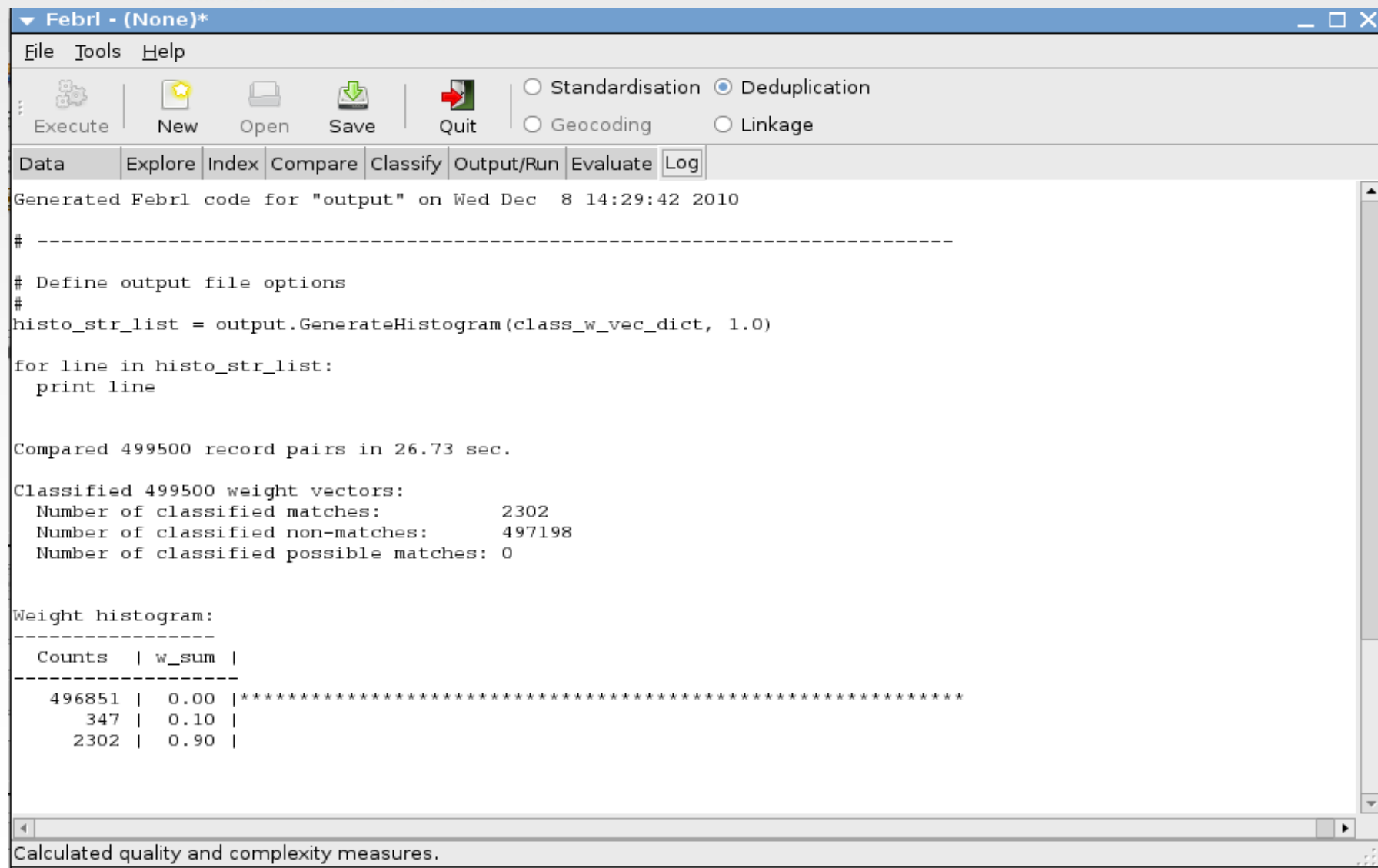
Algoritmo de Comparação dos registros

FEBRL



Algoritmo de Classificação do Matching

FEBRL



The screenshot shows the FEBRL application window titled "Febri - (None)*". The menu bar includes "File", "Tools", and "Help". The toolbar contains icons for "Execute", "New", "Open", "Save", and "Quit", along with radio buttons for "Standardisation", "Deduplication" (selected), "Geocoding", and "Linkage". The "Data" menu is open, showing options: "Explore", "Index", "Compare", "Classify", "Output/Run", "Evaluate", and "Log". The main text area displays the following log output:

```
Generated Febri code for "output" on Wed Dec 8 14:29:42 2010

# -----
# Define output file options
#
histo_str_list = output.GenerateHistogram(class_w_vec_dict, 1.0)

for line in histo_str_list:
    print line

Compared 499500 record pairs in 26.73 sec.

Classified 499500 weight vectors:
  Number of classified matches:      2302
  Number of classified non-matches:  497198
  Number of classified possible matches: 0

Weight histogram:
-----
Counts | w_sum |
-----
496851 | 0.00 | *****
  347  | 0.10 |
  2302 | 0.90 |

Calculated quality and complexity measures.
```

Log com os Resultados

- **Dsgen**
 - Script python para gerar bases de teste
 - Utiliza pequenas bases de dados e gera duplicatas a partir delas
 - Pode ser modificado
 - Insere erros fonéticos, de digitação, OCR

- **Parâmetros do Dsgen:**
 - Nro de Originais
 - Nro de Duplicatas
 - Nro máximo de duplicados por registro
 - Nro máximo de modificações por campo
 - Nro máximo de modificações por registro
 - Tipo de modificação (erro de digitação, fonética, OCR)

Métodos de Matching

- **Determinísticos**
 - Resultado é absoluto
 - Par de registros REPRESENTA a mesma entidade OU
 - Par de registros NÃO REPRESENTA a mesma entidade

Métodos de Matching

- **Probabilísticos**

- Resultado mais flexível

- Define-se um threshold: Valor entre 0 e 1 que corresponde à probabilidade mínima para que um par seja considerado duplicado

- Alternativamente, pode-se definir 3 intervalos de aceitação:

- ✗ Probabilidade $< X$: Par não é duplicado

- ? $X \leq \text{Probabilidade} < Y$: Indefinido

- ✓ Probabilidade $\geq Y$: Par é duplicado

Modelo de Fellegi-Sunter

Métodos de Matching

- **Determinísticos**

- Validam um matching quando um campo (ou parte de um campo) é exatamente igual nos dois registros.

- Ex.: Matching sobre um campo CEP, utilizando os 5 primeiros números.

- **Probabilísticos**

- Utiliza-se uma função de matching que gera valores entre 0 e 1 e define-se o threshold que dê o melhor resultado na base de teste.

- Ex.: n-grams, métodos de distância entre strings.

Execução da Deduplicação

- **Tempo de Execução:**

- Pode ser estimado através de um cálculo da média de tempo que é gasto para comparar 2 registros na base de dados de teste.
- Para uma base com 1.000 registros, o número de comparações N é:

$$N = \frac{1.000 \times (1.000 - 1)}{2} = 499.500$$

Execução da Deduplicação

- **Tempo de Execução:**
 - Caso o tempo gasto para comparar 2 registros seja de 1 centésimo de segundo – o tempo para fazer as 499.500 comparações será de aproximadamente 1 hora e 23 minutos.
 - Para 100.000 registros o tempo seria superior a 500 dias!
 - Para reduzir o tempo de comparação, regras mais restritivas devem ser testadas com prioridade.

Avaliação dos Resultados

- **Precisão, Revocação e F-Measure:**

Precisão (precision): $\frac{\text{Número de relevantes recuperados}}{\text{Número total de recuperados}}$

Revocação (recall): $\frac{\text{Número de relevantes recuperados}}{\text{Número total de relevantes}}$

$$F = \frac{(\beta^2 + 1) \times P \times R}{(\beta^2 \times P) + R}$$

- Se $\beta = 1$, a mesma ênfase é dada a P e a R
- Se $\beta = 2$, Revocação é enfatizada 2 vezes em relação à Precisão
- Se $\beta = 0.5$, enfatiza a Precisão 2 vezes em relação à Revocação

Avaliação dos Resultados

- **Falso Positivo:** Não-relevante que foi recuperado (pares que não são duplicados, mas foram identificados como sendo)
- **Falso Negativo:** Relevante que não foi recuperado (pares que deveriam, mas não foram detectados como duplicados)

Definição

- **Blocagem**

- Técnica utilizada para reduzir o número de comparações entre os registros, fazendo a segmentação dos registros em blocos menores.
 - Somente registros do mesmo bloco são comparados.
 - Idealmente, todos os registros que referenciam a mesma entidade, e somente estes, devem ficar no mesmo bloco.
 - Reduz-se o número de comparações sem reduzir a precisão.

Etapas da Blocagem

- Pré-processamento dos dados
- Montagem de uma base de dados de testes avaliada
- Escolha dos algoritmos de matching
- **Escolha da chave de blocagem**
 - **Método Manual x Método Automático**
- Execução da deduplicação nos blocos
- **Avaliação dos resultados**
 - **Pair Completeness (PC), Reduction Rate (RR) e F-Score**

Escolha da Chave de Blocação

- **Manual**
 - Especialista no domínio pode definir qual a “melhor” chave de blocação.
 - Como existe uma base de testes avaliada, podem ser feitos testes para determinar qual a melhor chave entre as candidatas.

Escolha da Chave de Blocação

- **Automática**

- Artigo “Learning Blocking Schemes for Record Linkage” de Matthew Michelson e Craig A. Knoblock (AAAI-06)
- Iterativamente identifica qual a melhor chave de bloqueio
- Cria conjunções entre regras até cobrir todos os registros
- Interrompe as iterações quando atinge threshold pré-estabelecido ou quando adição de regras não cobre mais nenhum registro
- Problema: É preciso definir quais são as regras que serão testadas

Deduplicação com Blocagem

- **Tempo de Execução:**

- Em uma base com 1.000 registros o número de comparações N, sem blocagem é:

$$N = \frac{1.000 \times (1.000 - 1)}{2} = 499.500$$

- A mesma base com 1.000 registros, divididos em 20 blocos de 50 registros, tem o número de comparações M de:

$$M = \frac{50 \times (50 - 1)}{2} \times 20 = 24.500$$

Deduplicação com Blocação

- **Tempo de Execução:**
 - Redução de 95% no número de comparações
 - Tempo gasto na deduplicação sem blocação: 1 hora e 23 minutos
 - Tempo gasto na deduplicação com blocação: 4 min e 5 seg

Avaliação dos Resultados

- **Pair Completeness (PC)**
 - Indica qual a taxa dos pares duplicados que ficaram no mesmo bloco

$$PC = \frac{\text{Pares Corretos nos Blocos}}{\text{Total de Pares Corretos}}$$

Avaliação dos Resultados

- **Reduction Rate (RR)**

- Indica a redução na quantidade de comparações que o processo de blocagem vai proporcionar

$$RR = 1 - \frac{\text{Total de Pares Gerados nos Blocos}}{\text{Total de Pares Possíveis}}$$

Avaliação dos Resultados

- **F-Score**
 - Média harmônica entre o PC e o RR
 - Referencial no processo de decisão

$$\text{F - Score} = \frac{2 \times \text{RR} \times \text{PC}}{\text{RR} + \text{PC}}$$

Referências

- **Deduplicação**

- Jin, L.; Li, C.; Mehrotra, S. Efficient Record Linkage in Large Data Sets, 2003.
- Lifang, G.; Baxter, R.; Deane, V.; Chris, R. Record Linkage: Current Practice and Future Directions. Technical Report 03/83, CSIRO Mathematical and Information Sciences, April 2003.
- N. Koudas, A. Marathe, and D. Srivastava. Flexible string matching against large databases in practice. Proceedings of VLDB, 2004.
- Böhm, L. F. Elaboração de uma estratégia de deduplicação de dados utilizando técnicas de blocagem em um cadastro hospitalar de pacientes. Lume: <http://hdl.handle.net/10183/26350>

- **Blocagem**

- M. Michelson and C. A. Knoblock. Learning blocking schemes for record linkage. In National Conference on Artificial Intelligence (AAAI-06), Boston, 2006.
- Christen, P. Towards Parameter-free Blocking for Scalable Record Linkage. ANU Joint Computer Science Technical Report Series, Agosto/2007
- Baxter, R.; Christen, P.; Churches, T. A comparison of fast blocking methods for record linkage. ACM SIGKDD workshop on Data Cleaning, Record Linkage and Object Consolidation, Washington DC, 2003. p. 25-27.

- **Febrl**

<http://sourceforge.net/projects/febrl/>



OBRIGADO!
PERGUNTAS?

Deduplicação e Blocagem

Luiz Fernando Böhm
lfbohm@inf.ufrgs.br

