



# طراحی و پیاده سازی الگوریتم تطابق اسامی در زبان فارسی به منظور تشخیص ذینفع واحد

لیلا مومنی نسب، [momeninasab.leila@gmail.com](mailto:momeninasab.leila@gmail.com)  
دکتر نیما امیرشکاری، [nima.itpro@gmail.com](mailto:nima.itpro@gmail.com)  
استاد جلال ملکی، [jalal.maleki@liu.se](mailto:jalal.maleki@liu.se)  
پرفسور لارش اهرنبرگ، [lars.ahrenberg@liu.se](mailto:lars.ahrenberg@liu.se)

سومین همایش سالانه بانکداری الکترونیک و نظام های پرداخت

[conf.mbri.ac.ir/ebps3](http://conf.mbri.ac.ir/ebps3)



کاربرد  
مروری بر الگوریتم های موجود  
یک الگوریتم تطابق اسامی برای زبان فارسی  
پی دا  
ارزیابی  
نتیجه  
آینده کاری





## کاربرد

- ❖ در زمینه بانکداری و مالی برای تشخیص تقلب
- ❖ مدیریت ارتباط با مشتری
- ❖ ضد پولشویی
- ❖ رتبه بندی اعتباری
- ❖ تشخیص ذینفع واحد



جمهوری اسلامی ایران



پژوهشگاه ملی و دانش  
پایه فناوری اطلاعات، تهران، ایران



شرکت ملی انفورماتیک



## مروری بر الگوریتم های موجود

- ❖ تغییرات اسامی
- ❖ الگوریتم های تطابق اسامی
- ❖ زبان فارسی



جمهوری اسلامی ایران



پژوهشگاه ملی و دانش  
پایه فناوری اطلاعات، تهران، ایران



شرکت ملی انفورماتیک



سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران



## مروری بر الگوریتم های موجود تغییرات اسامی

در اسامی یکسان اتفاق می افتد

- ❖ تغییرات نوشتاری
  - ✓ خطاهای املائی
  - ✓ نوشتار جایگزین
  - ✓ نویسه گردانی
  - ✓ حروف بی صدا
- ❖ تغییرات فیلد
- ❖ اسامی هم ارز



سازمان اسناد و کتابخانه ملی



پژوهشکده بومی و تاریخی  
سازمان اسناد و کتابخانه ملی



شرکت ملی اسناد و کتابخانه



سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران



## مروری بر الگوریتم های موجود تغییرات اسامی

- ❖ مختصر نویسی
- ❖ قطعه شدن
- ❖ ترجمه
- ❖ اضافه یا حذف اجزاء
- ❖ استفاده از علائم



سازمان اسناد و کتابخانه ملی



پژوهشکده بومی و تاریخی  
سازمان اسناد و کتابخانه ملی



شرکت ملی اسناد و کتابخانه



دفتر نشر و کتابخانه دیجیتال



## مروری بر الگوریتم های موجود الگوریتم های تطابق اسامی

آیا این دو اسم به یک شخص واحد تعلق دارد؟

Mohamedamin و Mohamed Amin

❖ الگوریتم های تطابق صوتی

Soundex ✓

❖ الگوریتم های تفاوت رشته ای

*Levenshtein* ✓

❖ الگوریتم های بر مبنای توکن

Q-gram ✓



پاسداری تهران



پژوهشکده ادبی و انسانی  
دانشگاه تهران



دفترت مانی تصویرخانه



دفتر نشر و کتابخانه دیجیتال



## مروری بر الگوریتم های موجود زبان فارسی

❖ الفبای زبان فارسی

۳۳ حرف ✓

سیستم نوشتاری Perso-Arbic ✓

از چپ به راست ✓

اتصال حروف در نوشتار ✓



پاسداری تهران



پژوهشکده ادبی و انسانی  
دانشگاه تهران



دفترت مانی تصویرخانه



## مروری بر الگوریتم های موجود یک الگوریتم تطابق اسامی برای زبان فارسی

- ❖ Levenshtein → Arabic Edit Distance Algorithm (AEDA)
- PEDA



پاسداری تهران



پژوهشکده مهندسی و فناوری

دانشگاه تهران، تهران، ایران



شرکت ملی صنایع پتروشیمی



پی دا

این دو اسم فارسی چه اندازه با هم شباهت دارند؟

موراد و مراد



پاسداری تهران



پژوهشکده مهندسی و فناوری

دانشگاه تهران، تهران، ایران



شرکت ملی صنایع پتروشیمی



## پی دا

- ❖ سطوح شباهت در زبان فارسی
  - ✓ شباهت فرمی
  - ✓ شباهت صوتی
  - ✓ شباهت کی یوردی
- ❖ هسته اصلی کد پی دا



پاسداری آشنایان



پژوهشکده پی‌دی و دانگی

دکتر سحر کرمی، دکتر سحر کرمی



دفترت ملی تصویر مانتک



## پی دا سطوح شباهت در زبان فارسی شباهت فرمی

Form Similarity in Persian Alphabet (between origin letter forms)		
No.	Similar Groups	Similarity Index
1.	(ی - ی) (ه - ه) (و - و) (ز - ز) (س - س) (ش - ش) (ت - ت) (ث - ث) (ج - ج) (ح - ح) (خ - خ) (د - د) (ذ - ذ) (ر - ر) (ز - ز) (ص - ص) (ض - ض) (ط - ط) (ک - ک)	1
2.	(ع - ع) (ک - ک) (ی - ی) (ب - ب) (ی - ی) (ب - ب) (ی - ی) (ن - ن) (ن - ن) (ف - ف) (ق - ق)	0.8
3.	(ب - ب) (ب - ب) (ج - ج) (ز - ز)	0.54
4.	(ج - ج) (س - س) (ش - ش) (ر - ر) (ت - ت) (ن - ن)	0.6
5.	(ع - ع) (ف - ف)	0.4
6.	(ک - ک) (ل - ل) (ب - ب) (ن - ن) (ب - ب) (ن - ن)	0.27
7.	(ج - ج) (خ - خ)	0.2
8.	(ب - ب) (ن - ن) (ب - ب) (ن - ن) (ی - ی) (ت - ت) (ی - ی) (ن - ن) (ی - ی) (ت - ت) (ی - ی) (ن - ن)	0.14
9.	(ئ - ئ) (ت - ت) (ئ - ئ) (ن - ن)	0.07
10.	(ه - ه) (م - م)	0
	Any other pair of Persian letters	0



پاسداری آشنایان



پژوهشکده پی‌دی و دانگی

دکتر سحر کرمی، دکتر سحر کرمی



دفترت ملی تصویر مانتک



## پی‌دا سطوح شباهت در زبان فارسی شباهت صوتی

Phonetic Similarity in Persian Alphabet (between origin letter forms)		
No.	Similar Groups	Similarity Index
1	(ک - ک) (ی - ی) (ا - ا) (ع - ا) (ا - ا) (ا - ا) (و - و) (ب - ط) (ت - س) (ص - ص) (ز - ض) (ظ - ظ) (ح - ح) (ه - ه) (ع - ع) (ا - ع) (ق - ق) (ه - ه) (ع - ع) (ی - ی)	1
0.8		
0.6		
0.4	(ب - ب) (ت - د) (ث، س، ص - ز، ض، ظ)	
0.2	(ج - ج) (ز - ز) (ش - و) (ف - ک) (ک - گ) (م - ن)	
0.1	(ن - ن)	
0	Any other combination of Persian letters	



پژوهشکده ادبی و انسانی  
دانشگاه گیلان، گیلان، ایران



دفترت مانی انصوریانک



## پی‌دا سطوح شباهت در زبان فارسی شباهت کی بوردی



$$Sim_{kb}(a, b) = 1 - \frac{\sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}}{\psi}$$



پژوهشکده ادبی و انسانی  
دانشگاه گیلان، گیلان، ایران



دفترت مانی انصوریانک



## پی‌دا هسته اصلی کد پی‌دا

### Levenshtein ❖

- ✓ تبدیل اسم مبدأ به اسم مقصد با کمترین هزینه
- ✓ حداقل تعداد عملیات برای تبدیل
  - جایگزینی
  - حذف
  - اضافه



پاسداری گیلان



پژوهشکده بیولوژی و دامپزشکی

دکتر سحر کرمی، دکتر سحر کرمی



دفترت ملی تصویر مانتک



## پی‌دا هسته اصلی کد پی‌دا

### Levenshtein ❖

- ✓ یک ماتریکس می‌سازد
- ✓ سطر اول و ستون اول ماتریکس را پر می‌کند

		t				
		p	u	z	l	e
s		1	2	3	4	5
	p					
	u					
	z					
	l					



پاسداری گیلان



پژوهشکده بیولوژی و دامپزشکی

دکتر سحر کرمی، دکتر سحر کرمی



دفترت ملی تصویر مانتک

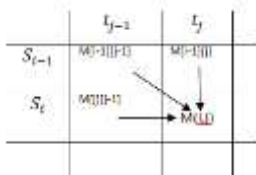




## پی‌دا هسته اصلی کد پی‌دا

### Levenshtein ❖

✓ سلول‌های ماتریکس را با توجه به مقادیر همسایه‌ها پر می‌کند



	p	u	z	z	l	e
0	1	2	3	4	5	6
p	1	0	1	2	3	4
z	2	1	1	1	2	3
z	3	2	2	1	1	2
e	4	3	3	2	2	2
l	5	4	4	3	3	2

$$M(i, j) = \min \begin{cases} M(i-1, j-1) + \text{substitution cost} \\ M(i-1, j) + \text{delete cost} \\ M(i, j-1) + \text{insert cost} \end{cases}$$



پاسداری اسلامبول



پژوهشکده پی‌دا و رایگی

دکتر سحر کرمی، دکتر سحر کرمی



شرکت ملی صنایع پتروشیمی

## پی‌دا هسته اصلی کد پی‌دا

### هزینه عملیات حذف و اضافه ❖

$$f_{\omega}(a_i, a_{i-1}) = \begin{cases} 0 & \text{if } a_i = \text{blank} \\ 0 & \text{if } a_i \neq a_{i-1} \wedge a_i = + \wedge a_{i-1} = \xi = \{\text{alif, ya, waw}\} \\ 0 & \text{if } a_i = a_{i-1} \\ \mu & \text{if } a_i = \xi = \{\text{alif, ya, waw}\} \\ 1 & \text{Otherwise} \end{cases}$$

محمد عرفان

اشیا

محمد

اسماعیل

محمد عرفان

اشیا

محمد

اسماعیل

### هزینه عمل جایگزینی ❖

$$f_{\omega}(a, b) = \begin{cases} \frac{\alpha(a, b) \cdot \omega + \beta(a, b) \cdot \lambda + \gamma(a, b) \cdot \sigma}{\omega + \lambda + \sigma} & \text{if } a \neq b \\ 0 & \text{Otherwise} \end{cases}$$



پاسداری اسلامبول



پژوهشکده پی‌دا و رایگی

دکتر سحر کرمی، دکتر سحر کرمی



شرکت ملی صنایع پتروشیمی



## پی‌دا هسته اصلی کد پی‌دا

❖ مثال

موراد و مراد

	م	و	ز	ا	د	
م	0	1	0	1.15	1.3	2.3
و	1	0	0.77778	0.15	0.3	1.3
ز	2	1	0.92778	0.3	0.15	1.00595
ا	3	1.15	0.77778	0	1.3	0.57778
د	4	2.15	1.87222	1.3	0.99444	0.57778

$$\text{Sim}(\text{موراد، مراد}) = 100 - \frac{0.57778 \times 100}{\max(4,5)} = 88\%$$



پاسداری گیلان



پژوهشکده پی‌دا و دانگی

دکتر سوزان کوزلوی، سوسان لورن



دفترت ملی تصویرماتیک



## ارزیابی

- ❖ انجام سه سری از آزمایشات
- ❖ نتایج تطابق



پاسداری گیلان



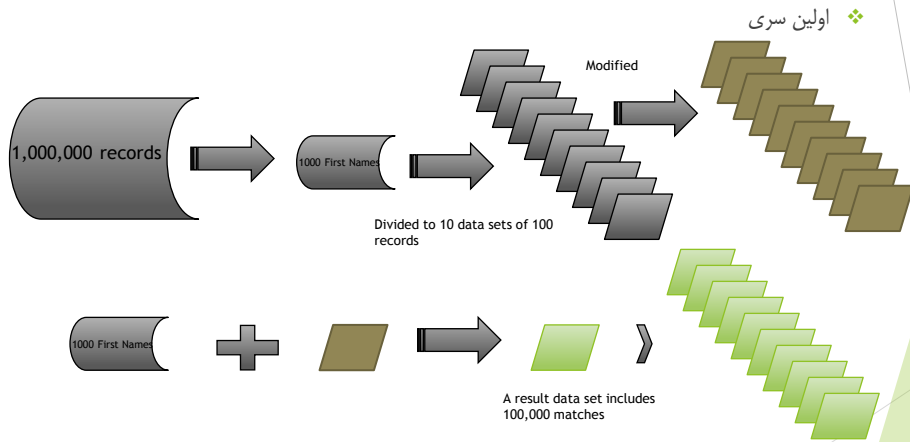
پژوهشکده پی‌دا و دانگی

دکتر سوزان کوزلوی، سوسان لورن



دفترت ملی تصویرماتیک

## ارزیابی انجام سه سری از آزمایشات



پاسدانی آشنایان

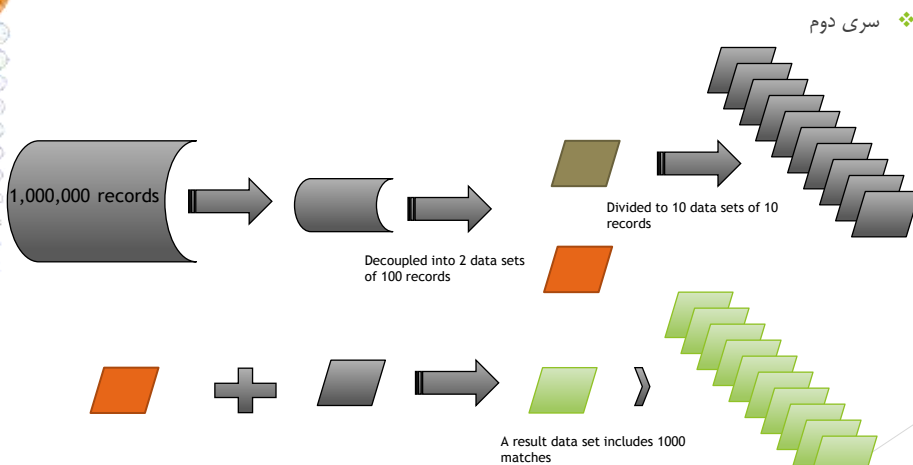


پژوهشکده مهندسی و فناوری  
دانشگاه آزاد اسلامی، واحد تهران غرب



دفتر ارتباط علمی با صنعت

## ارزیابی انجام سه سری از آزمایشات



پاسدانی آشنایان



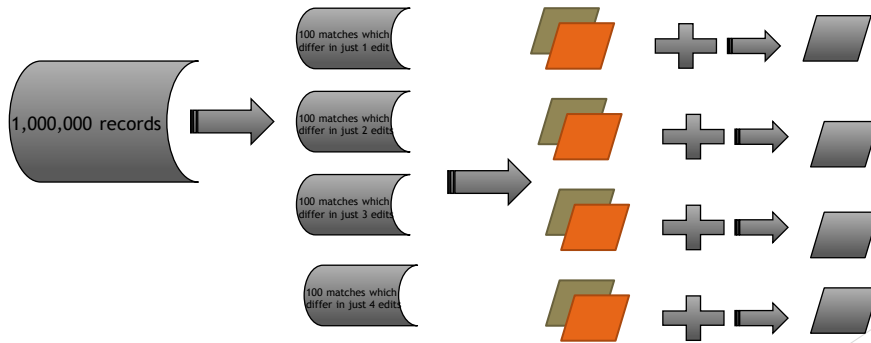
پژوهشکده مهندسی و فناوری  
دانشگاه آزاد اسلامی، واحد تهران غرب



دفتر ارتباط علمی با صنعت

## ارزیابی انجام سه سری از آزمایشات

❖ پی دا برای تعداد مختلفی از عملیات چگونه عمل می کند؟



پاسداری تهران



پژوهشکده پی دی ایت  
دکتر سید محمد حسینی



مرکز سیستم های هوشمند

## ارزیابی نتایج تطابق

❖ اولین سری

PEDA	Precision	Recall	f-measure
DS01	0.8	0.9	0.85
DS02	0.77	0.95	0.85
DS03	0.89	0.95	0.92
DS04	0.81	0.97	0.88
DS05	0.73	0.96	0.83
DS06	0.79	0.95	0.86
DS07	0.79	0.99	0.88
DS08	0.79	0.96	0.87
DS09	0.75	0.95	0.84
DS10	0.7	0.95	0.81
f-measure mean			0.86



پاسداری تهران



پژوهشکده پی دی ایت  
دکتر سید محمد حسینی



مرکز سیستم های هوشمند



## ارزیابی نتایج تطابق

❖ اولین سری: مقایسه با لونشتین

Levenshtein	DS0 1	DS0 2	DS0 3	DS0 4	DS0 5	DS0 6	DS0 7	DS0 8	DS0 9	DS1 0
True positives	95%	100%	94%	99%	99%	96%	97%	96%	97%	98%
False positives	5%	0%	6%	1%	1%	4%	3%	4%	3%	2%
True negatives	59%	65%	65%	74%	67%	65%	69%	68%	62%	67%
False negatives	41%	35%	35%	26%	33%	35%	31%	32%	38%	33%

PEDA	DS0 1	DS0 2	DS0 3	DS0 4	DS0 5	DS0 6	DS0 7	DS0 8	DS0 9	DS1 0
True positives	86%	78%	90%	81%	73%	79%	79%	21%	76%	70%
False positives	14%	22%	10%	19%	27%	21%	21%	79%	24%	30%
True negatives	92%	91%	91%	96%	93%	92%	98%	94%	96%	93%
False negatives	8%	9%	9%	4%	7%	8%	2%	6%	4%	7%



پاسداری گیلان



پژوهشکده پی سی و آی تی  
دانشگاه گیلان، گیلان، ایران



دفترت مانی تصویرهاستک



## ارزیابی نتایج تطابق

❖ دومین سری

	Precision	Recall	f-measure
DS01	1	1	1
DS02	0.64	0.78	0.7
DS03	0.83	0.55	0.66
DS04	0.75	0.43	0.55
DS05	0.71	0.83	0.76
DS06	0.86	1	0.92
DS07	1	1	1
DS08	1	0.75	0.86
DS09	1	0.71	0.83
DS10	0.71	0.83	0.76
f-measure mean			0.80



پاسداری گیلان



پژوهشکده پی سی و آی تی  
دانشگاه گیلان، گیلان، ایران



دفترت مانی تصویرهاستک



## ارزیابی نتایج تطابق

❖ سومین سری

	1 edit	2 edits	3 edits	4 + edits
True positives	99%	81%	69%	42%
False positives	1%	19%	31%	58%



سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران



پژوهشگاه ملی زبان و ادبیات

دکتر سید علی حسینی



دفترت ملی اسناد و کتابخانه ملی



## ارزیابی

- ❖ انواع تغییرات اسامی را در داده های تست آورده ایم
- ❖ نتایج را برای انواع تغییرات اسامی بررسی نمودیم

s	t	Similarity
ح ج بری	بری	45%
زری	زهرا	68%
سید احمد	احمد	70%
نسب آقا	نسب	69%
آقا شهزاد	شهزاد	64%
محبوبه	محبوبه	76%
زینب نه به خانم	زینب ن ن خا	68%
فاطمه	فاطمی	64%



سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران



پژوهشگاه ملی زبان و ادبیات

دکتر سید علی حسینی



دفترت ملی اسناد و کتابخانه ملی



## نتیجه

- ❖ پی‌دا برای تغییرات نوشتاری و قطعه‌شدن خوب عمل می‌کند
- ❖ توصیه می‌شود در کنار الگوریتم‌های دیگر استفاده شود



پایگاه داده‌های علمی



پژوهشگاه ملی و دانش

دانشگاه تهران، تهران، ایران



شرکت ملی صنایع پتروشیمی

29



## آینده کاری

- ❖ توسعه قوانین شباهت
- ❖ هوشمندسازی پی‌دا
- ❖ ترکیب با پایگاه داده‌های موارد خاص (مانند اسامی اشخاص، شرکت‌ها، نام‌نامه و ...)



پایگاه داده‌های علمی



پژوهشگاه ملی و دانش

دانشگاه تهران، تهران، ایران



شرکت ملی صنایع پتروشیمی

30



## با تشکر

سومین همایش سالانه بانکداری الکترونیک و نظام های پرداخت  
۱۶ و ۱۷ دی ماه ۱۳۹۲ - مرکز همایش های برج میلاد

[conf.mbri.ac.ir/ebps3](http://conf.mbri.ac.ir/ebps3)

