به نام خدا

An n-gram based method for nearly copy detection in plagiarism systems

PAN FIRE 2016

دکتر بهروز مینایی

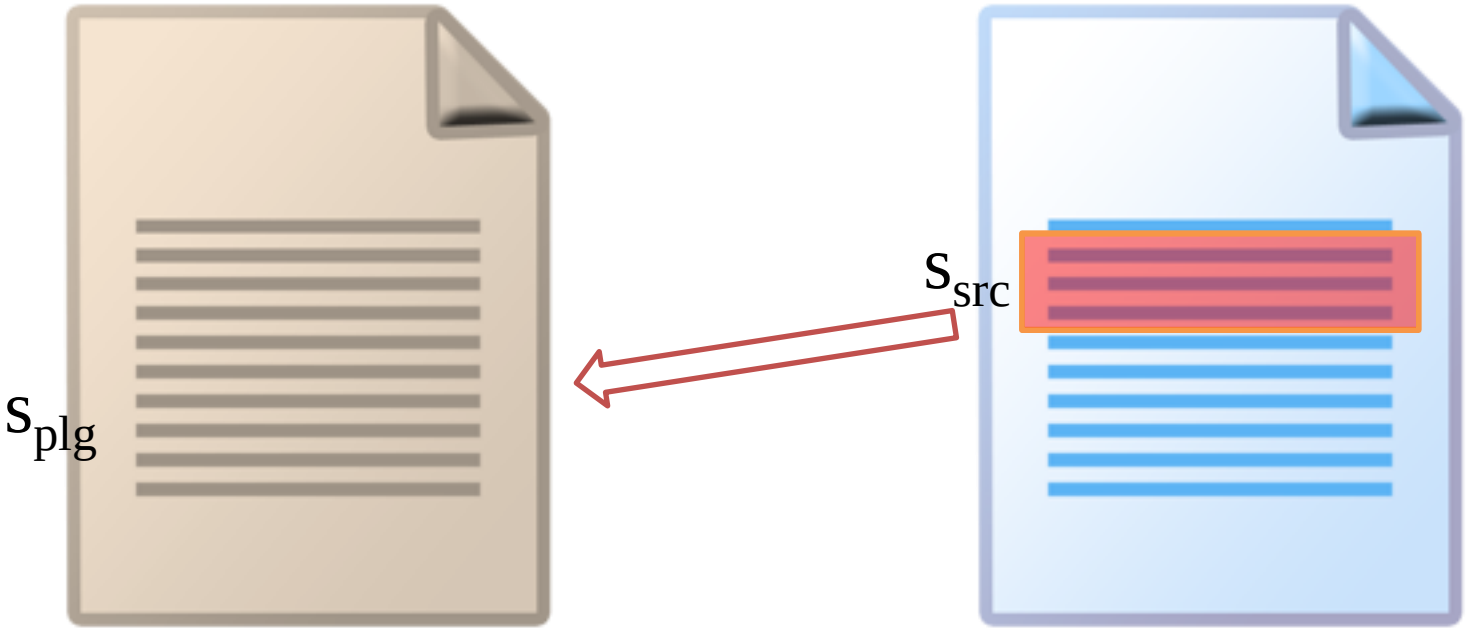
مهدی نیکنام

تعریف

- تصاحب غیرقانونی، دزدیدن و انتشار ایده، تجربه، تفکر و کلام فرد دیگر و عرضه آن به عنوان کار اصلی خود.

- تخصیص دادن خلاقیت ادبی یا پژوهشی دیگری یا بخشی از آن یا متن ناشی از آن به خود، گویی که خود شخص آن را خلق کرده است.

نمونه ای از سرقت علمی

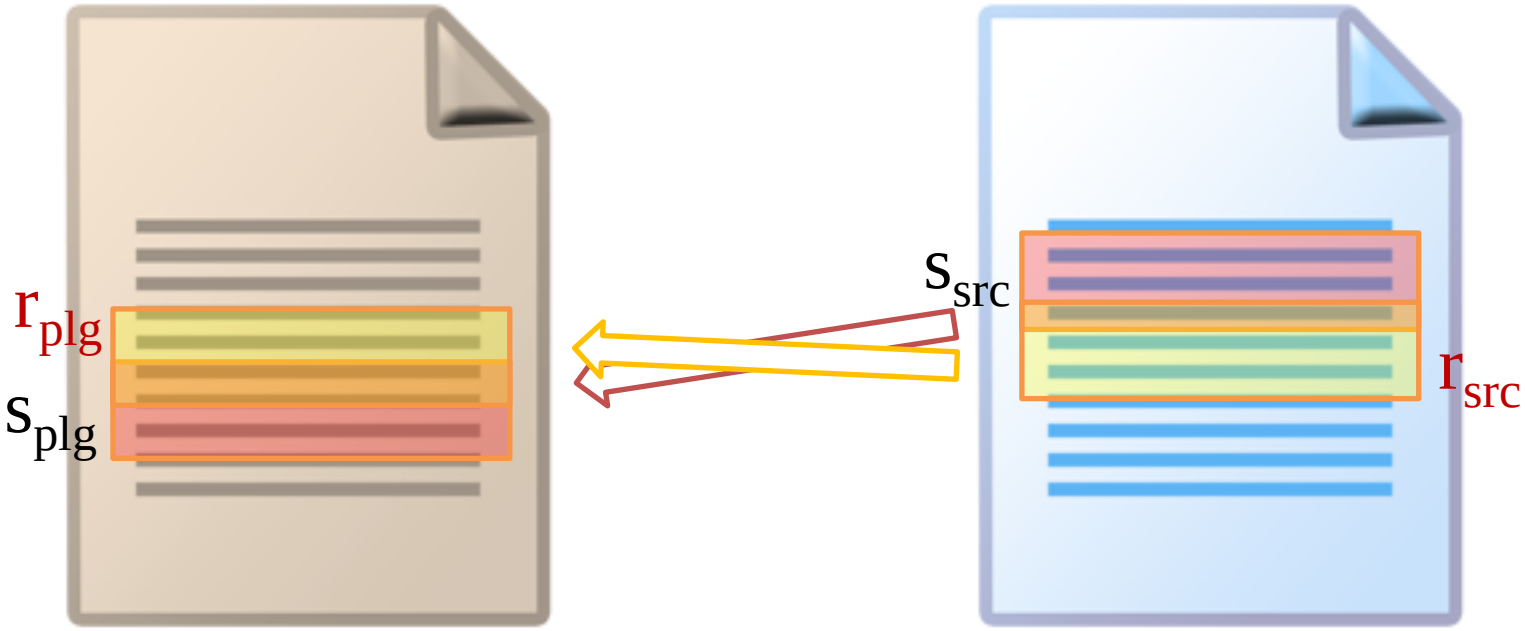


Plagiarism document (d_{plg})

Source document (d_{src})

$$s = \langle s_{plg}, d_{plg}, s_{src}, d_{src} \rangle$$

نمونه از شناسایی سرقت علمی



Plagiarism document (d_{plg})

Source document (d_{src})

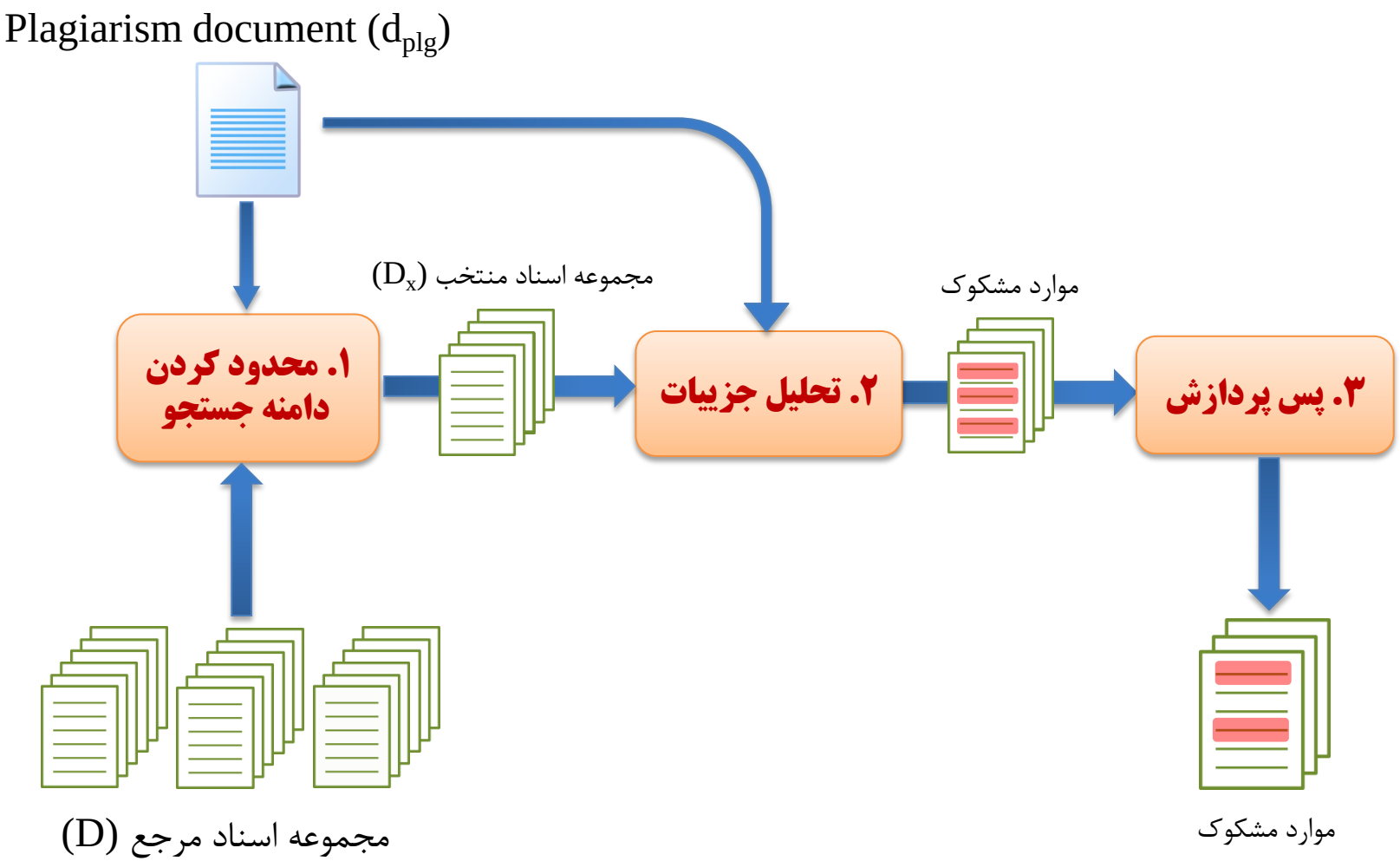
Source document (d'_{src})

$r = \langle r_{plg}, d_{plg}, r_{src}, d'_{src} \rangle$

$s = \langle s_{plg}, d_{plg}, s_{src}, d_{src} \rangle$

$s_{plg} \cap r_{plg} \neq \emptyset \, , s_{src} \cap r_{src} \neq \emptyset \, , d_{src} = \acute{d}_{src}$

استراتژی کلی شناسایی سرقت علمی



الگوریتم پیشنهادی

مراحل الگوریتم پیشنهادی

۱. پیش پردازش

۲. محدود کردن دامنه جستجو

۳. مقایسه دقیق بین نتایج

۴. پس پردازش

- یک دست کردن حروف
- حذف کلمات بی‌بار
- حذف حروف و کلمات غیر فارسی و اعداد

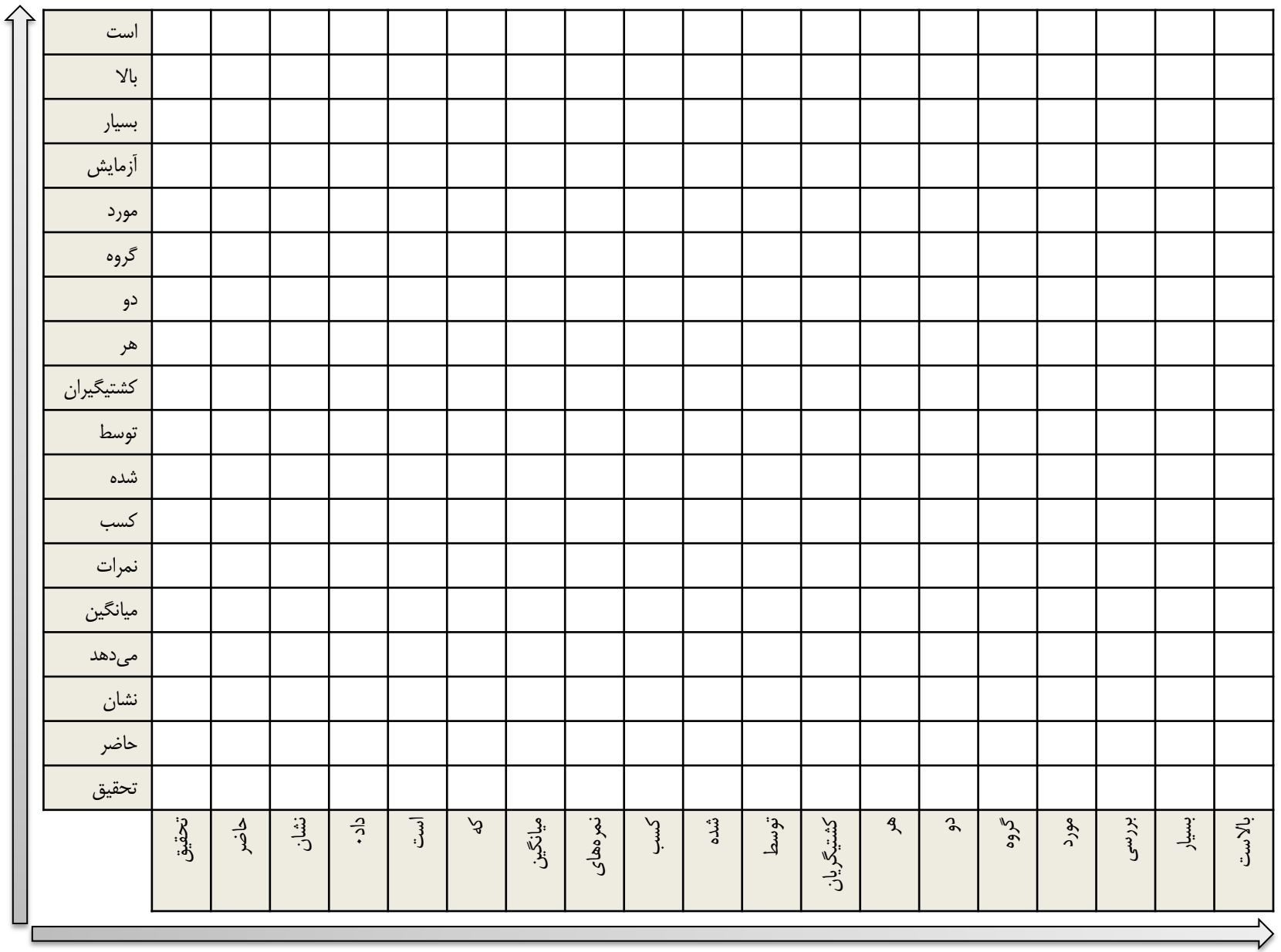
- ایجاد جدول معکوس از اسناد منبع
- تبدیل سند مشکوک به سرقت علمی به مجموعه ای از چندتایی های کلمه محور (word-base ngram)
- جستجوی هر چندتایی در جدول معکوس و بازیابی اسناد مرتبط با معیار tf-idf
- انتخاب ۲۵ مقاله با بیشترین درجه شباهت به عنوان اسناد کاندیدا

پیش پردازش

محدود کردن دامنه جستجو

مقایسه دقیق بین نتایج

پس پردازش



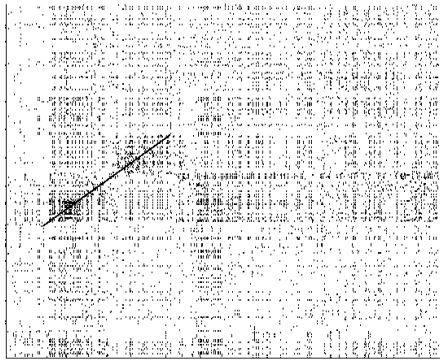
تحقیق حاضر نشان داده است که میانگین نمره‌های کسب شده توسط کشتی‌گیران هر دو گروه مورد بررسی بسیار بالاست

تحقیق حاضر نشان می‌دهد میانگین نمرات کسب شده توسط کشتی‌گیران هر دو گروه مورد آزمایش بسیار بالا است.

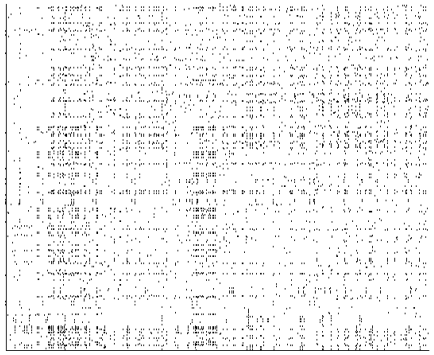
تحقیق حاضر نشان می‌دهد میانگین نمرات کسب شده توسط کشتی‌گیران هر دو گروه مورد آزمایش بسیار بالا است.

تحقیق حاضر نشان داده است که میانگین نمره‌های کسب شده توسط کشتی‌گیران هر دو گروه مورد بررسی بسیار بالاست

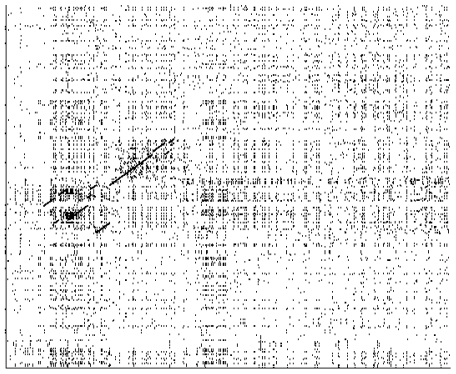
نمونه هایی از مقایسه دو سند بر روی نمودار



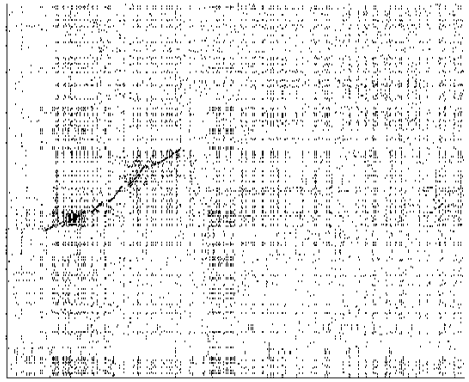
نمایش تصویری از استفاده بدون تغییر از یک سند در سند دیگر



نمایش تصویری از شباهت‌های دو سند بدون سرقت علمی



نمایش تصویری کپی متن با جابه‌جایی جملات

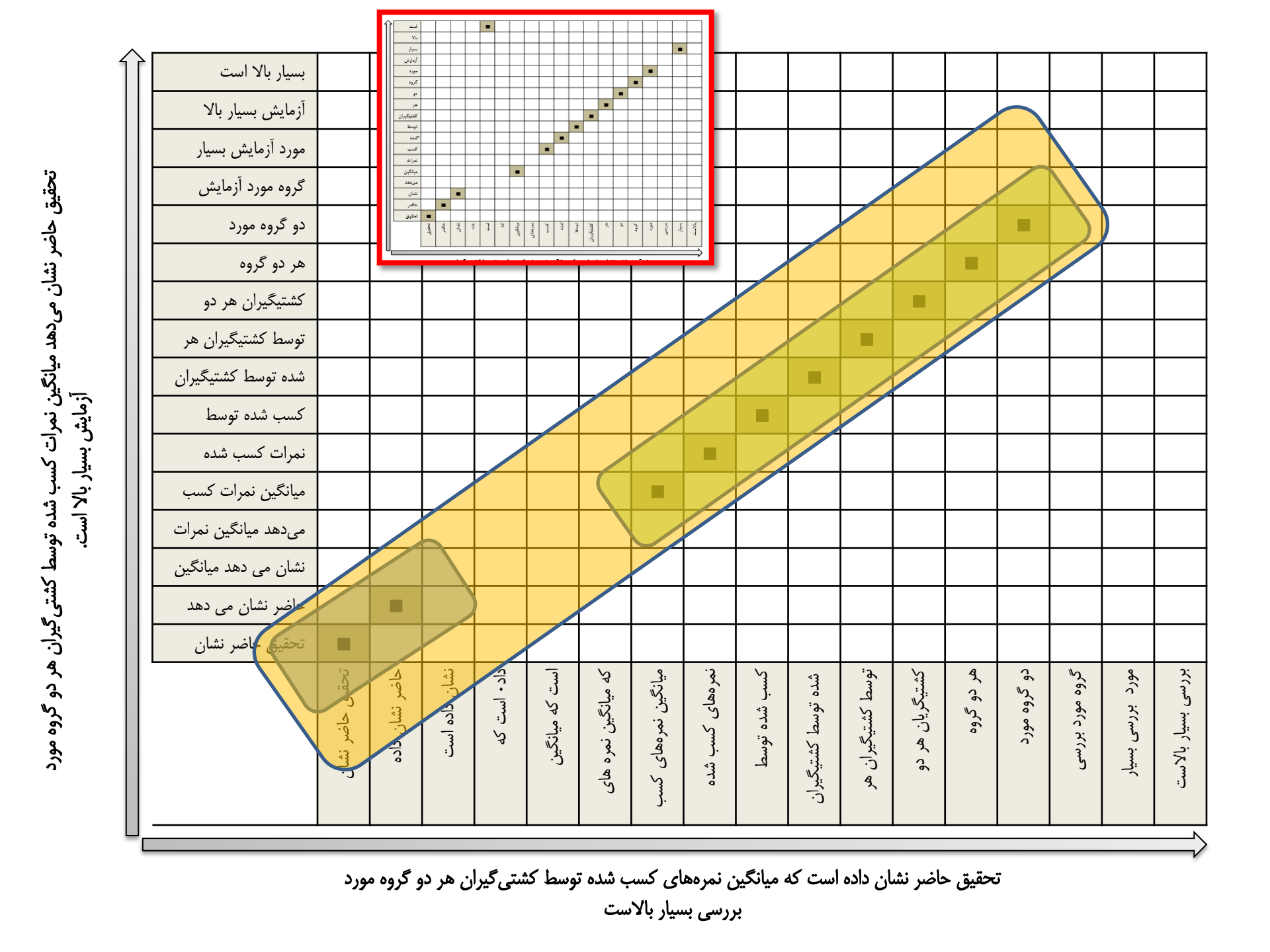


نمایش تصویری کپی متن با جایگزینی برخی کلمات

تحقیق حاضر نشان می‌دهد میانگین نمرات کسب شده توسط کشتی‌گیران هر دو گروه مورد آزمایش بسیار بالا است.

	بسیار بالا است	آزمایش بسیار بالا	مورد آزمایش بسیار	گروه مورد آزمایش	دو گروه مورد	هر دو گروه	کشتیگیران هر دو	توسط کشتیگیران هر	شده توسط کشتیگیران	کسب شده توسط	نمرات کسب شده	میانگین نمرات کسب	می‌دهد میانگین نمرات	نشان می دهد میانگین	حاضر نشان می دهد	تحقیق حاضر نشان
تحقیق حاضر نشان																
حاضر نشان داده																
نشان داده است																
داد٠ است که																
است که میانگین																
که میانگین نمره های																
میانگین نمره‌های کسب																
نمره‌های کسب شده																
کسب شده توسط																
شده توسط کشتیگیران																
توسط کشتیگیران هر																
کشتیگریان هر دو																
هر دو گروه																
دو گروه مورد																
گروه مورد بررسی																
مورد بررسی بسیار																
بررسی بسیار بالاست																

تحقیق حاضر نشان داده است که میانگین نمره‌های کسب شده توسط کشتی‌گیران هر دو گروه مورد بررسی بسیار بالاست


تحقیق حاضر نشان می‌دهد میانگین نمرات کسب شده توسط کشتی‌گیران هر دو گروه مورد آزمایش بسیار بالا است.
بسیار بالا است
آزمایش بسیار بالا
مورد آزمایش بسیار
گروه مورد آزمایش
دو گروه مورد
هر دو گروه
کشتیگیران هر دو
توسط کشتیگیران هر
شده توسط کشتیگیران
کسب شده توسط
نمرات کسب شده
میانگین نمرات کسب
می‌دهد میانگین نمرات
نشان می دهد میانگین
حاضر نشان می دهد
تحقیق حاضر نشان
تحقیق حاضر نشان داده
نشان داده است
داده است که
است که میانگین
که میانگین نمره های
میانگین نمره‌های کسب
نمره‌های کسب شده
کسب شده توسط
شده توسط کشتیگیران
توسط کشتیگیران هر
کشتیگیران هر دو
هر دو گروه
دو گروه مورد
گروه مورد بررسی
مورد بررسی بسیار
بررسی بسیار بالاست
تحقیق حاضر نشان داده است که میانگین نمره‌های کسب شده توسط کشتی‌گیران هر دو گروه مورد
بررسی بسیار بالاست

- حذف نمونه های شناسایی شده که طول آن ها از میزان آستانه β کمتر باشد.

ارزیابی الگوریتم

Rank / Team	Runtime (h:m:s)	Recall	Precision	Granularity	F-Measure	PlagDet
1 Mashhadirajab	02:22:48	**0.9191**	0.9268	1.0014	**0.9230**	**0.9220**
2 Gharavi	**00:01:03**	0.8582	0.9592	**1**	0.9059	0.9059
3 Momtaz	00:16:08	0.8504	0.8925	**1**	0.8710	0.8710
4 Minaei	00:01:33	0.7960	0.9203	1.0396	0.8536	0.8301
5 Esteki	00:44:03	0.7012	0.9333	**1**	0.8008	0.8008
6 Talebpour	02:24:19	0.8361	**0.9638**	1.2275	0.8954	0.7749
7 Ehsan	00:24:08	0.7049	0.7496	**1**	0.7266	0.7266
8 Gillam	21:08:54	0.4140	0.7548	1.5280	0.5347	0.3996
9 Mansourizadeh	00:02:38	0.8065	0.9000	3.5369	0.8507	0.3899

ویژگی های الگوریتم ارائه شده

- سادگی الگوریتم
- سرعت مناسب اجرای الگوریتم
- قابلیت شناسایی کپی با وجود ایجاد تغییرات در متن
- قابلیت استفاده از الگوریتم برای سایر کاربردها

منابع

- M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, “An evaluation framework for plagiarism detection,” in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 2010, pp. 997–1005.
- B. Stein, S. M. zu Eissen, and M. Potthast, “Strategies for retrieving plagiarized documents,” in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 825–826.
- C. Grozea, C. Gehl, and M. Popescu, “ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection,” in *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse*, 2009, p. 10.
- J. Kasprzak, M. Brandejs, and M. Kripac, “Finding plagiarism by evaluating document similarities,” in *Proc. SEPLN*, 2009, vol. 9, pp. 24–28.
- C. Basile, D. Benedetto, E. Caglioti, G. Cristadoro, and M. D. Esposti, “A plagiarism detection procedure in three steps: Selection, matches and ‘squares’,” in *Proc. SEPLN*, 2009, pp. 19–23.
- M. Zechner, M. Muhr, R. Kern, and M. Granitzer, “External and intrinsic plagiarism detection using vector space models,” in *Proc. of 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection*, 2009, pp. 47–55.

با تشکر از توجه شما