

شناسایی هویت نویسنده بر اساس زبان فردی: پژوهشی پیکره‌ای در زبان فارسی

رامین گلشائی

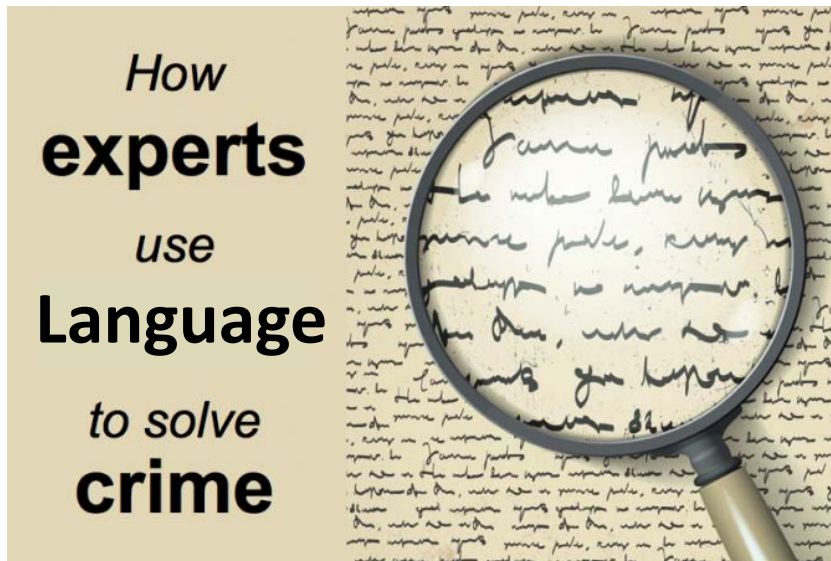
پژوهشگاه علوم و فناوری اطلاعات ایران

I WAS BORN OUTSIDE
IN A SMALL TOWN OUTSIDE
OF CHICAGO. I ALWAYS LIKED MATH.
I WAS PRETTY GOOD AT IT TOO. WHICH I
GUESS IS WHY I ENDED UP PURSUING
FINANCE AS A CAREER. IN MY SENIOR YEAR
AT COLLEGE I TRANSFERRED TO A SCHOOL
ON THE WEST COAST. I WORKED HARD AND
EARNED EVERY BIT OF MY DEGREE IN ACTUARIAL SCIENCES.
IT FEELS LIKE YESTERDAY BUT IT'S BEEN NEARLY 14 YEARS
SINCE I STARTED MY FIRST JOB AS AN ADVISOR. AT FIRST
IT WAS JUST A FEW CLIENTS. I TOOK MY FIRST
AND PUT IN A LOT OF HARD WORK BUT LITTLE TIME
I BUILT MY BOOK. A FEW GOOD REFERRALS REALLY HELPED ME
EXPAND MY BUSINESS AND SOLIDIFY MY REPUTATION AS A TOP PROD-
UCER. I TRIED EVERYTHING FROM SMALL COMPANIES TO WAREHOUSES.
I'VE ALWAYS BEEN INDEPENDENT AND I'VE ALWAYS LIKED THE IDEA
OF WORKING THE WAY YOU LIKE THEM DONE. I HATE THE
FEELING OF A SMALL COMPANY. I MET A LOT OF GREAT PEOPLE
-SOURCES OF A LARGER COMPANY. I FEEL LIKE I'VE BEEN HITTING THE RE-
ALITY BUT WHERE I AM NOW JUST DOESN'T FEEL LIKE THE RIGHT THING FOR ME.
Now I'M LOOKING FOR SOMETHING A LITTLE MORE FREEDOM AND I'M LOOKING
FOR A PLACE WHERE I CAN REALLY MAKE MY MARK.

مقدمه

- زبان فردی (idiolect):

شیوه‌ی منحصر به فرد هر گویشور در به کار گیری عناصر زبانی از جمله واژه‌ها، ساختار دستوری و آواها گفته می‌شود.

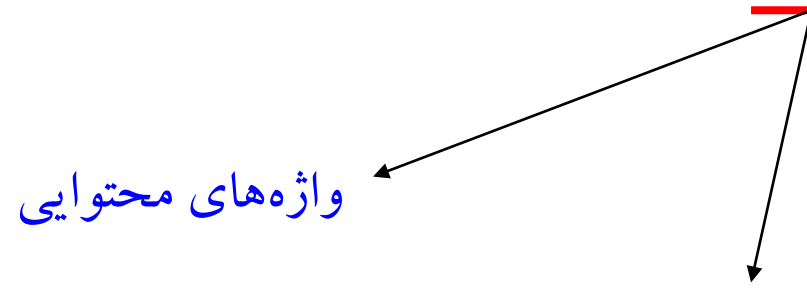


زبان‌شناسی قانونی (حقوقی)

شناسایی مولف (نامه، پیامک تهدید آمیز، ...)
سرقت علمی و مالکیت فکری
آواشناسی قانونی و شناسایی صدا

- زبان فردی (idiolect):

شیوه‌ی منحصر به فرد هر گویشور در به کار گیری عناصر زبانی نظیر واژه‌ها، ساختار دستوری و آواها گفته می‌شود.



واژه‌های محتوایی

واژه‌های دستوری:

حروف تعریف، حروف اضافه، ضمائر، حروف ربط، ...

چرا واژه های دستوری؟

۱. واژه های دستوری **مستقل از موضوع** خاصی در متن به کار می روند؛
۲. واژه های دستوری **پربسامدتر** از واژه های محتوایی هستند؛
۳. تولید و درک واژه های دستوری به طور **ناخود آگاه** صورت می گیرد. (Kestemont, 2014)



احتمال تثبیت چشم بر روی واژه های دستوری = ۳۵٪ | واژه های محتوایی = ۸۵٪
(Carpenter & Just, 1983)

پرسی ها

- آیا واژه های دستوری می توانند نوشتارهای افراد را از یکدیگر متمایز کنند و حکم اثر انگشت نویسنده را داشته باشند؟
- چه توالی هایی از واژه های دستوری (تک نگاشتی، دونگاشتی یا سه نگاشتی) بهتر می تواند متون نویسنده ها را از یکدیگر تفکیک کند؟
- حداقل کلمات موجود در متن برای تفکیک دقیق متون نویسنده ها چقدر است؟

روش پژوهش

۱. انتخاب ۵ نویسنده و گردآوری پیکره ی متنی آثار نویسندگان
 ۲. استانداردسازی شیوه ی نگارش و خطایابی املایی متون پیکره با استفاده از نرم افزار Vafa
 ۳. نمونه گیری از پیکره ها و انجام آزمایش بر روی نمونه های متنی ۴۰۰۰، ۱۰۰۰۰، ۵۰۰۰۰، ۳۰۰۰ و ۲۰۰۰ کلمه ای
 ۴. استفاده از نرم افزار آماری R و بسته ی تخصصی stylo برای انجام تحلیل ها
 ۵. استخراج واژه های پربسامد از متون، حذف واژه های محتوایی و غیردستوری، و انتخاب ۲۰ واژه ی دستوری پربسامد
 ۶. انجام «تحلیل مولفه های اصلی» و «تحلیل خوشه ای» بر روی داده ها برای محاسبه ی شباهت بین متون
- تحلیل خوشه ای بر اساس مقیاس دلتا (Burrows (2002)

جمع آوری متون

- متون از مقالات و خودنوشته‌های ۵ محقق حوزه ی علوم انسانی و عموماً از پایگاه‌های اینترنتی شخصی یا تخصصی آنها جمع آوری شد.

حداقل ۱۵ و حداکثر ۱۰۵ مقاله

- حوزه های فعالیت نویسندگان:

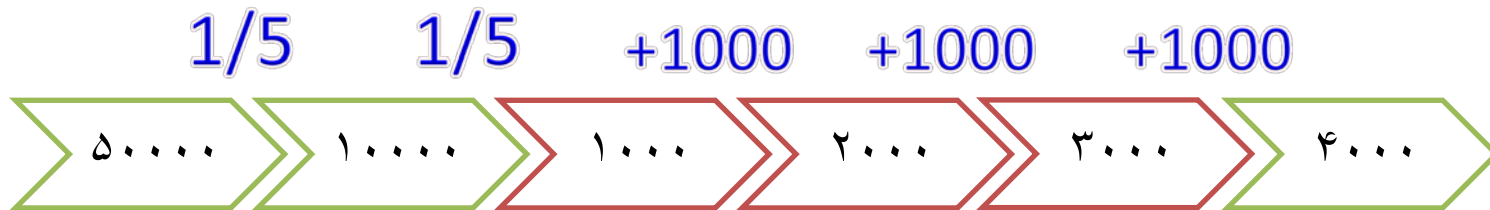
نویسنده	حجم پیکره
A	۱۰۸۶۹۰
B	۱۰۷۴۹۹
C	۱۰۷۱۰۷
D	۴۴۸۵۴
E	۳۴۲۴۶

- علوم سیاسی
- فلسفه و معارف اسلامی
- انسان شناسی
- فلسفه
- ادبیات و سینما

پالایش واژه های پربسامد

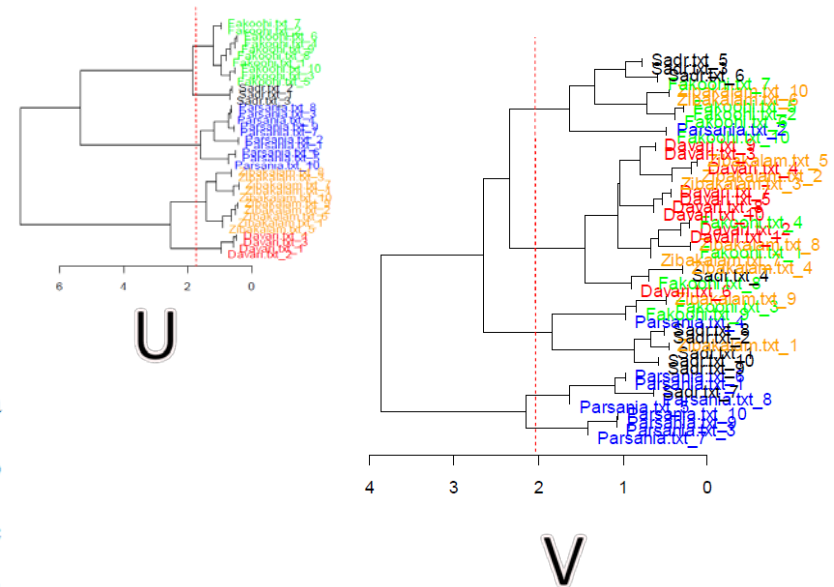
- واژه های غیردستوری از فهرست بسامدی حذف شدند:

واژه های محذوف	مثال
اعداد	یک، دو، سه، ...
علامت جمع بدون نیم فاصله	ها، های
وندهای فاقد نیم فاصله	می، ای
افعال (ربطی و غیرربطی)	است، بود، شد، ...
ضمایر شخصی	او، آنها، ...
واژه های محتوایی	فرهنگ، ایران، علم، ...



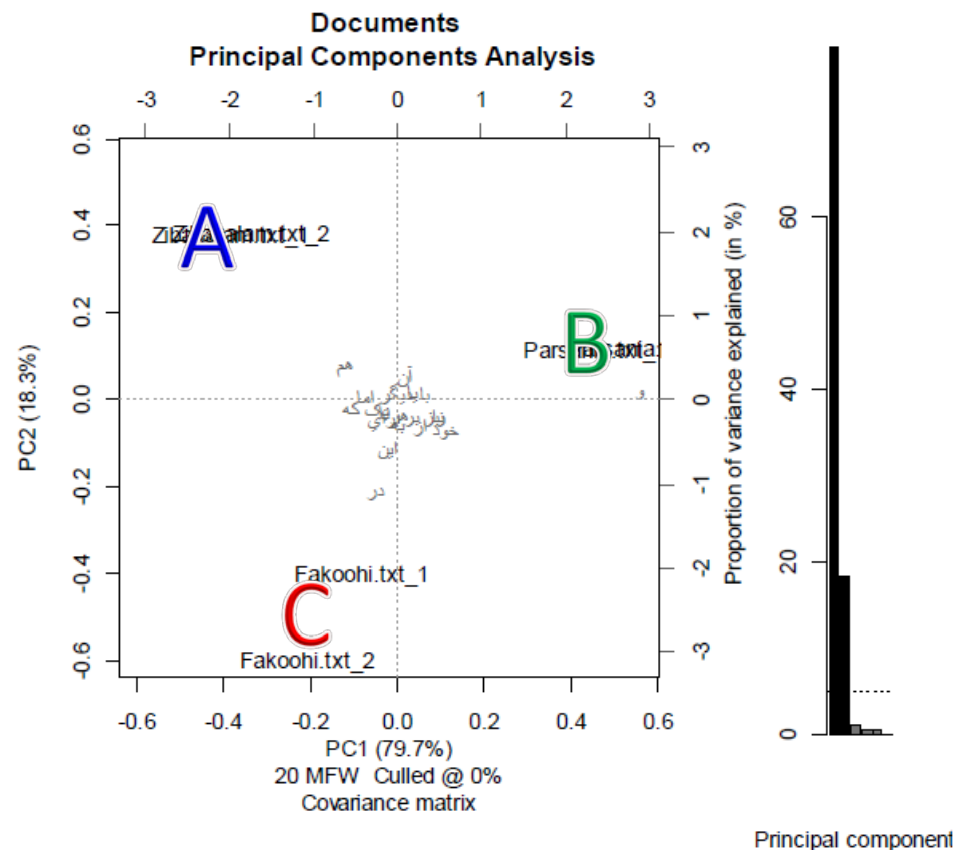
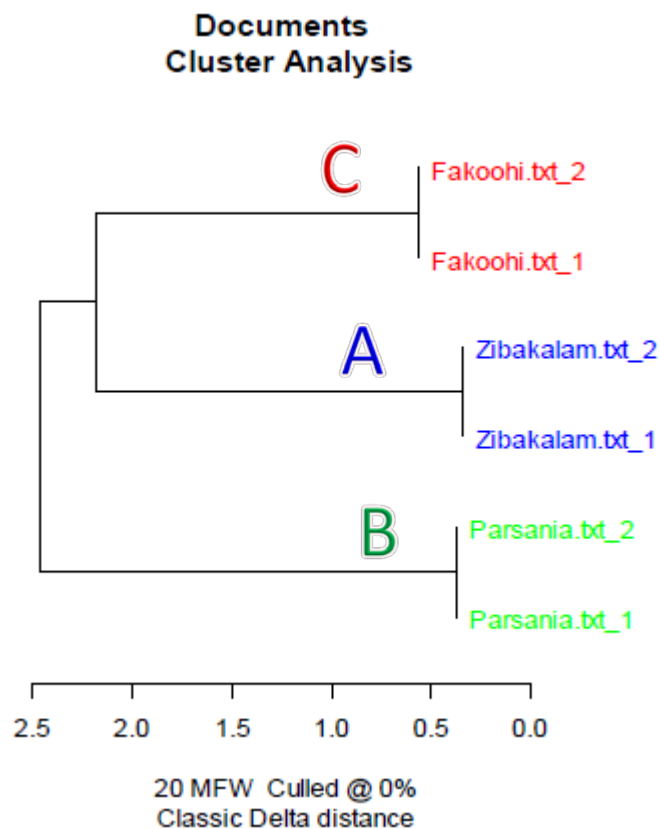
ارزیابی دقت خوشه‌بندی: Adjusted Rand Index

$$RI = \frac{a+d}{a+b+c+d}$$



- a. جفت‌هایی که هم در طبقه‌بندی U و هم در طبقه‌بندی V هم‌گروه هستند؛
- b. جفت‌هایی که در طبقه‌بندی U هم‌گروه ولی در طبقه‌بندی V غیرهم‌گروه هستند؛
- c. جفت‌هایی که در طبقه‌بندی V هم‌گروه ولی در طبقه‌بندی U غیرهم‌گروه هستند؛
- d. جفت‌هایی که هم در طبقه‌بندی U و هم در طبقه‌بندی V غیرهم‌گروه هستند.

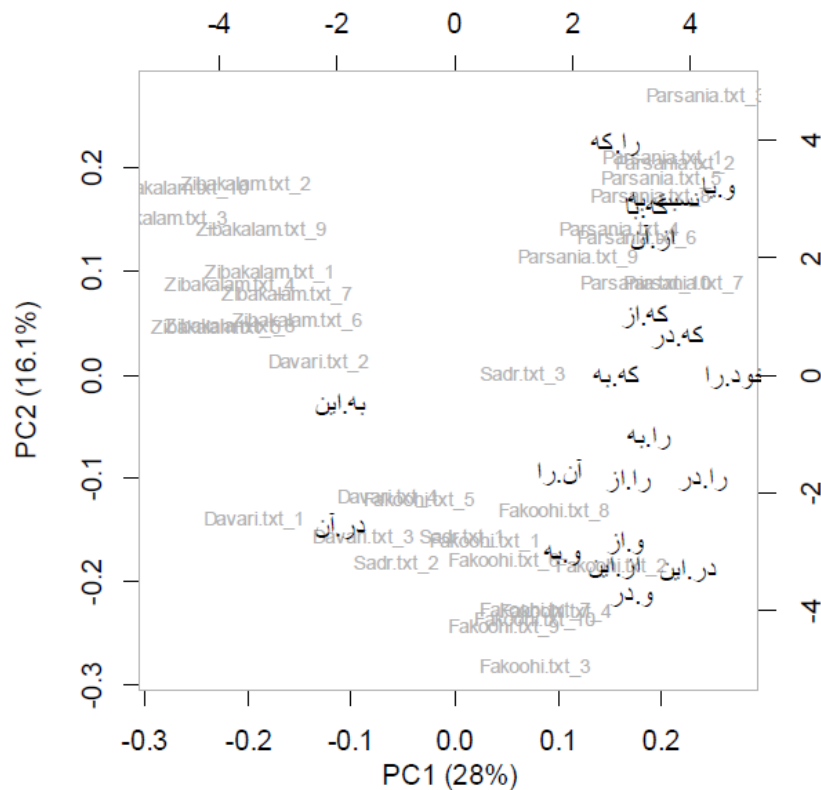
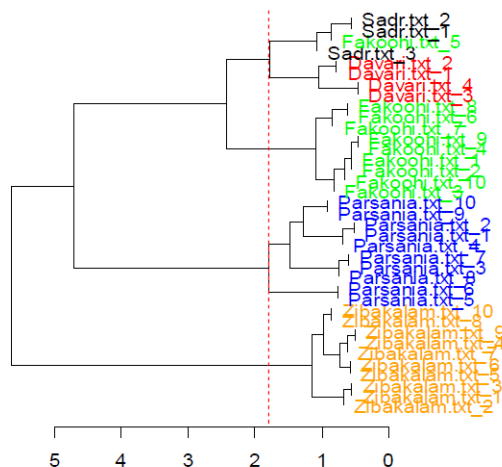
متون ۵۰۰۰ کلمه ای



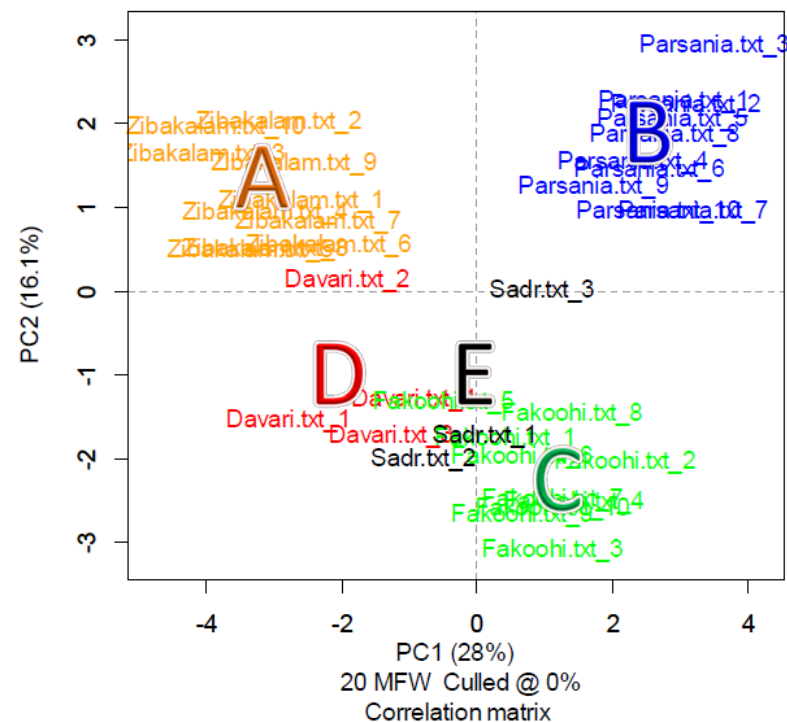
واژه های دستوری تک نگاشتی

متون ۱۰۰۰۰ کلمه ای

Documents
Cluster Analysis

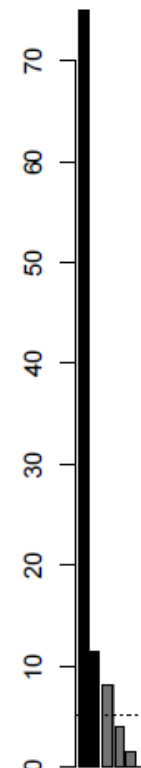
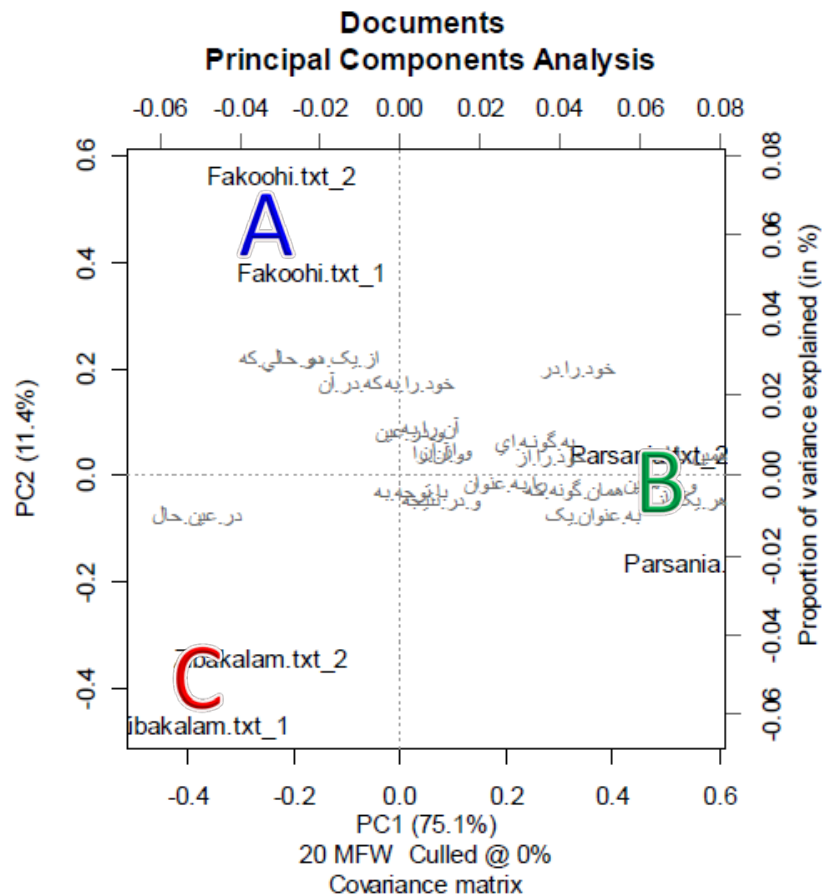
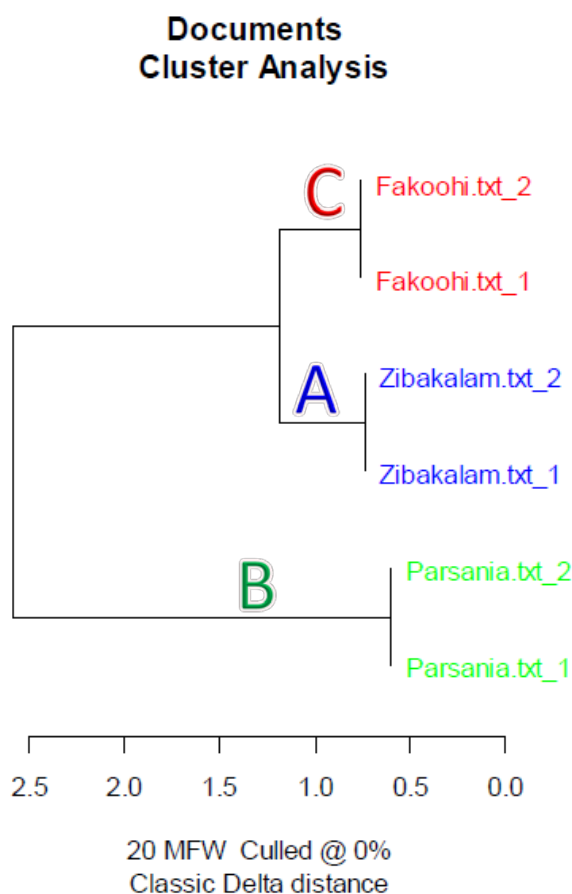


Documents
Principal Components Analysis



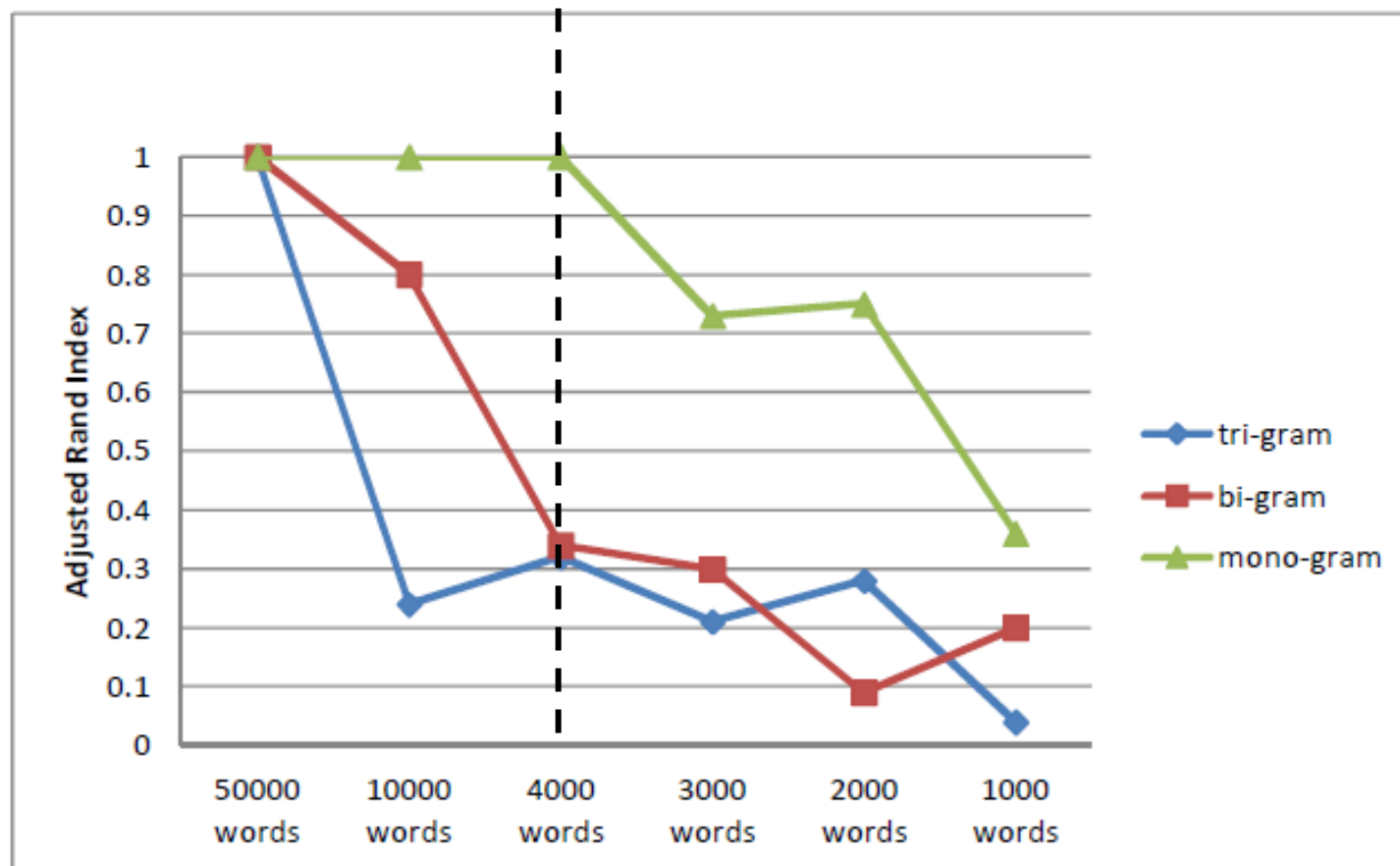
واژه های دستوری دو نگاشتی

متون ۵۰۰۰ کلمه ای



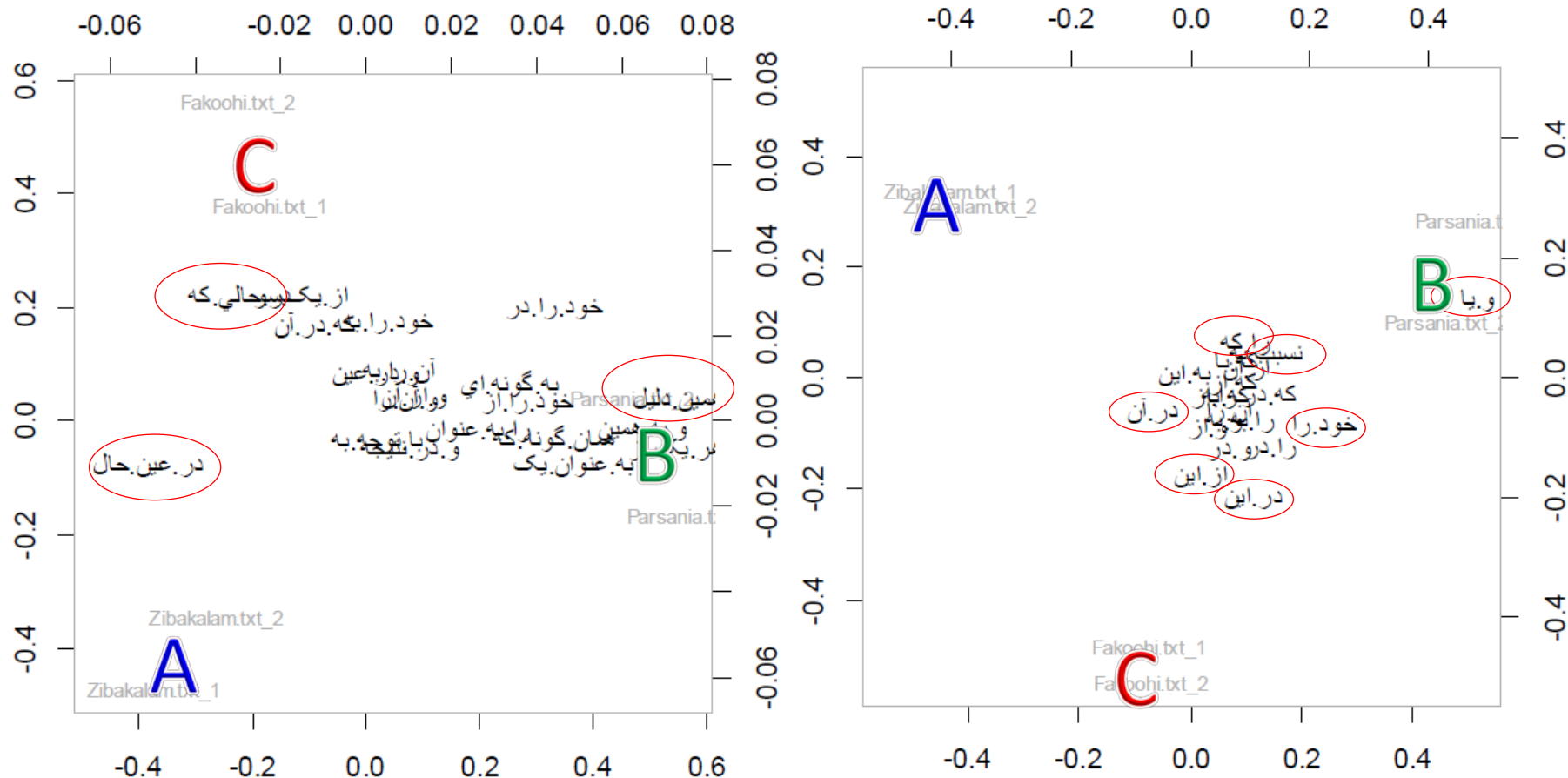
واژه های دستوری سه نگاشتی

Principal component

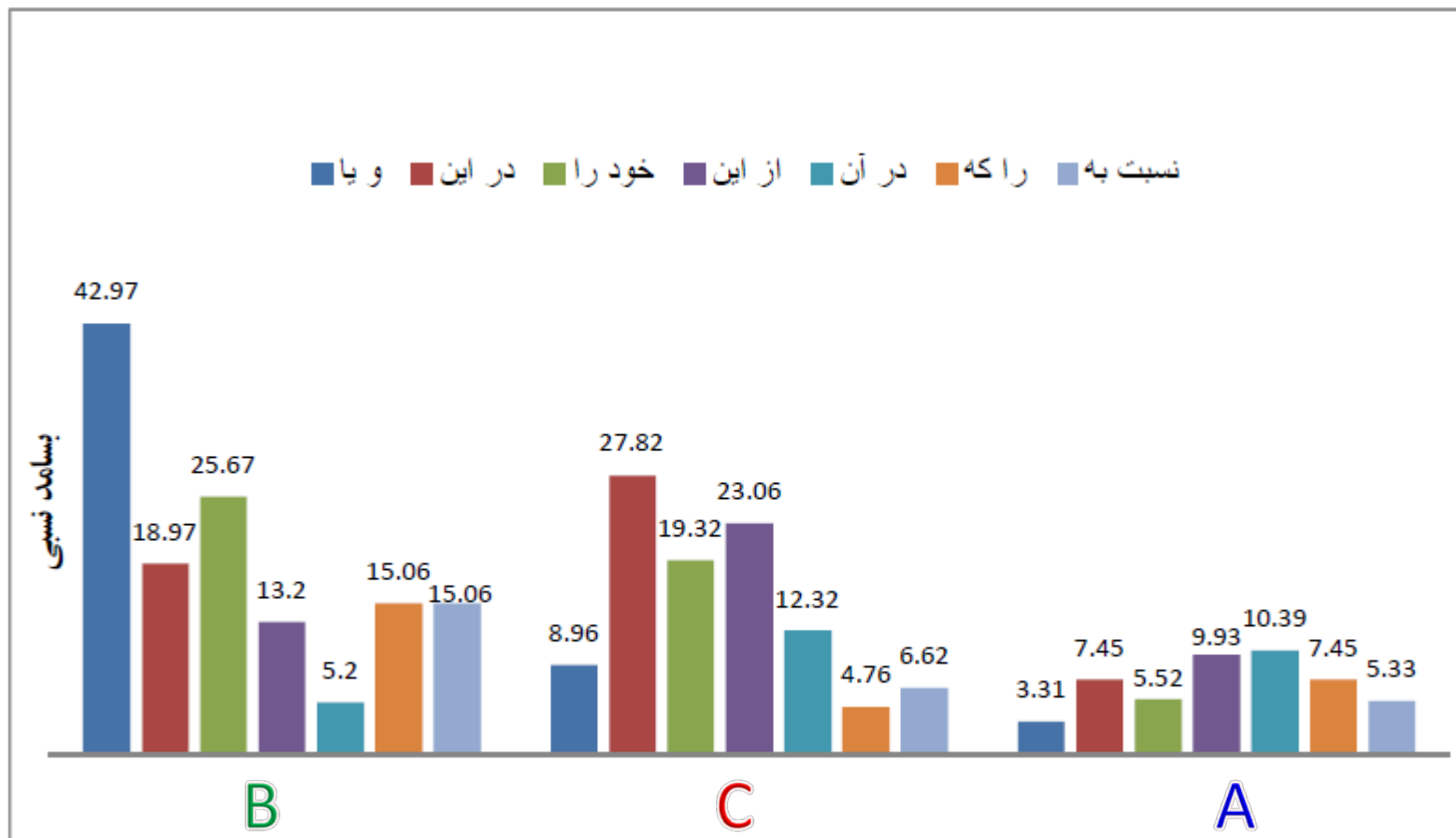


شکل ۴-۵۳. نمودار انگاشتی تغییرات شاخص تعدیل شده‌ی رند بر اساس تعداد کلمات متون

واژه‌های دستوری ممیز سبک

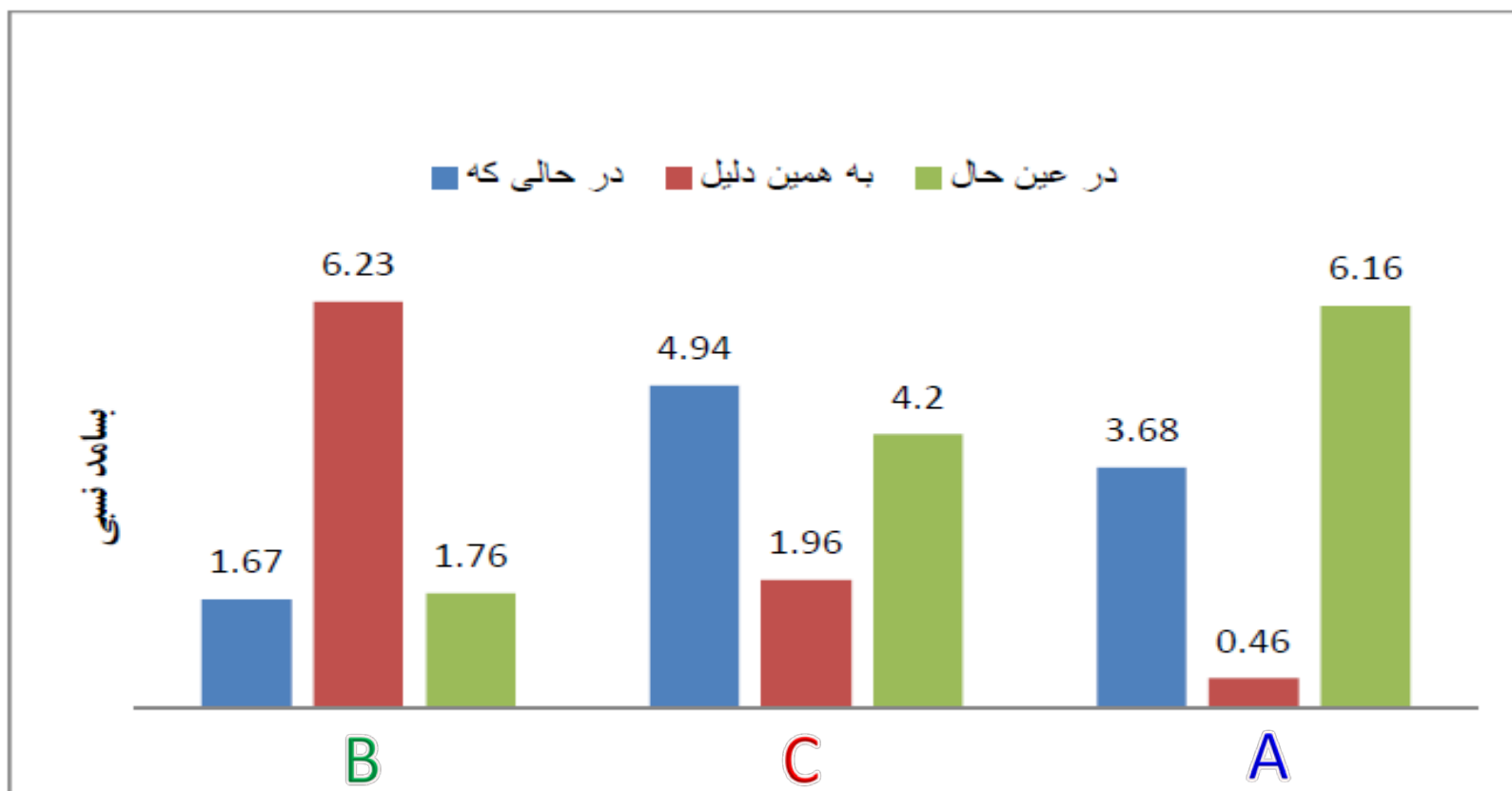


واژه‌های دستوری ممیز سبک



شکل ۴-۵۶. بسامد نسبی برخی دونگاشتی‌های پرسامد ممیز سبکی در پیکره‌ی متنی سه نویسنده

واژه‌های دستوری ممیز سبک



شکل ۴-۵۷. بسامد نسبی برخی سه‌نگاشتهای پرسامد ممیز سبکی در پیکره‌ی متنی سه نویسنده

جمع‌بندی نتایج

- واژه‌های دستوری می‌توانند نوشتارهای نویسنده‌های فاقد سبک ادبی بخصوص را از یکدیگر متمایز کنند.
 - واژه‌های دستوری تک‌نگاشتی نسبت به واژه‌های دونگاشتی و سه‌نگاشتی در تفکیک نویسنده‌ها بهتر عمل می‌کنند.
 - کمینه‌ی واژه‌ها برای تفکیک دقیق متون (تحت شرایط آزمایشی این پژوهش) حدود ۴۰۰۰ واژه است.
- در پژوهش‌های غیرفارسی بین ۱۰۰۰ تا ۵۰۰۰ واژه (Holmes et al. 2001, Eder 2013)
- در نظم و نثر فارسی (نویسندگان صاحب سبک) حدود ۱۰۰۰ واژه (Modaber Dabagh 2007)

منابع

- Burrows, J. F. 2002. Delta: A measure of stylistic difference and a guide to likely authorship, *Literary and Linguistic Computing*, 17, 267–287.
- Carpenter, P. A. & Just, M. A. 1983. What your eyes do while your mind is reading. In K. Rayner (Ed.), *Eye movements in reading: Perceptual and language processes* (pp. 275–307). New York: Academic Press.
- Eder, M. 2013. Does size matter? Authorship attribution, small samples, big problem. *Literary and Linguistic Computing*. DOI: <http://dx.doi.org/10.1093/llc/fgt066>
- Faili, H., Ehsan, N., Montazery, M., and Pilehvar, M.M. 2016. Vafa spell-checker for detecting spelling, grammatical, and real-word errors of Persian language. *Digital Scholarship in Humanities*, 31 (1), 95-117.
- Holmes, D. I., Gordon, L.J., and Wilson, C. 2001. A widow and her soldier: stylometry and the American civil war. *Literary and Linguistic Computing*, 16(4), 403-420.
- Kestemont, M. 2014. Function words in authorship attribution: from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature*, April 27, Gothenburg, Sweden, pp. 59-66.
- Modaber Dabagh, R. 2007. Authorship attribution and statistical text analysis. *Metodološki zvezki*, 4(2), 149-163.