



وزارت علوم، تحقیقات و فناوری
پژوهشگاه علوم و فناوری اطلاعات ایران

طرح پژوهشی

شناسایی هویت نویسنده بر اساس زبان فردی:

پژوهشی پیکره‌ای در زبان فارسی

مجری:

رامین گلشائی

تیر ۱۳۹۵

به نام خدا

صفحه تایید پژوهشگاه

حقوق: پژوهشگاه علوم و فناوری اطلاعات ایران

طرح پژوهشی «شناسایی هویت نویسنده بر اساس زبان فردی: پژوهشی پیکره‌ای در زبان فارسی» پیرو قرارداد شماره ۳۳/۴۸۵۳ مورخ ۹۴/۴/۱۳ میان پژوهشگاه علوم و فناوری اطلاعات ایران (کارفرما) و رامین گلشائی (مجری) اجرا شده است. این گزارش و تمامی حقوق مادی آن بر اساس «قانون حمایت از حقوق مولفان و مصنفان و هنرمندان» مصوب سال ۱۳۴۸ و اصلاحیه‌های بعدی آن و همچنین آیین‌نامه‌های اجرایی این قانون متعلق به پژوهشگاه علوم و فناوری اطلاعات ایران است و هر گونه استفاده از تمامی یا پاره‌ای از آن شامل نقل قول، تکثیر، انتشار، کاربرد نتایج، تکمیل و مانند آنها به صورت چاپی، الکترونیکی یا وسایل دیگر فقط با اجازه کتبی پژوهشگاه امکان‌پذیر است. نقل قول در حد هزار واژه در انتشارات علمی مانند کتاب و مقاله با درج اطلاعات کامل کتاب‌شناختی، نیازی به مجوز پژوهشگاه ندارد. صحت مندرجات گزارش بر عهده مجری طرح پژوهشی است.

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

مجری طرح پژوهشی: رامین گلشائی، پژوهشگاه علوم و فناوری اطلاعات ایران

نشانی: تهران، خیابان انقلاب، چهارراه فلسطین، شماره ۱۰۹۰، طبقه ۳، اتاق ۳۰۹

صندوق پستی: ۱۳۷۱۳۱۸۵ تلفن: ۰۲۱-۶۶۴۹۴۹۸۰

وبگاه: www.irandoc.ac.ir

رایانامه: golshaie@irandoc.ac.ir

چکیده

زبان فردی به شیوه‌ی منحصر به فرد هر گویشور در به کارگیری عناصر زبانی از جمله واژه‌ها، ساختار دستوری و آواها گفته می‌شود که اخیراً موضوع پژوهش‌های رایانشی و زبان‌شناختی در شناسایی نویسنده‌ی متون فاقد هویت واقع شده است. یکی از مولفه‌های زبان‌شناختی که گفته می‌شود تجلی گاه زبان فردی واقع می‌شود، واژه‌های دستوری در زبان است. واژه‌های دستوری از آن جهت که به طور ناخودآگاه در تولید زبان به کار گرفته می‌شوند، مستقل از موضوع متن به کار می‌روند و بسامد بالایی در متون کوتاه دارند، همواره مورد توجه پژوهشگران در حوزه‌ی شناسایی سبک نویسنده بوده‌اند. در این پژوهش امکان تفکیک متون متعلق به یک نویسنده از متون دیگر با استفاده از واژه‌های دستوری زبان فارسی بررسی شده است. ابتدا پیکره‌هایی متنی از نوشته‌های ۵ محقق و نویسنده‌ی معاصر جمع‌آوری و سپس با استفاده از نرم‌افزار خطایاب وفا، استانداردسازی شدند. با استفاده از بسته‌ی تحلیل آماری stylo در نرم‌افزار R، واژه‌های پربسامد دستوری با توالی‌های یک تا سه‌نگاشتی از متون استخراج شدند و سپس قابلیت تفکیک متون بر اساس این واژه‌ها و با استفاده از تحلیل مولفه‌های اصلی و همچنین تحلیل خوشه‌ای بر اساس مقیاس فاصله‌ای دلتا بررسی شد. نتایج نشان می‌دهد که واژه‌های دستوری در زبان فارسی قابلیت تفکیک متون متعلق به یک نویسنده را دارند و عملکرد واژه‌های تک‌نگاشتی بهتر از دونگاشتی و سه‌نگاشتی‌ها در متون کم‌حجم است. همچنین نتایج پژوهش نشان می‌دهد که حجم کمینه‌ی متن برای شناسایی موفقیت‌آمیز نویسنده در متون فارسی حدود ۴۰۰۰ واژه بر اساس ۲۰ واژه‌ی دستوری پربسامد است.

کلیدواژه‌ها: شناسایی نویسنده، تحلیل پیکره‌ای، تحلیل خوشه‌ای، روش دلتا، تحلیل مولفه‌های اصلی، سبک‌سنجی

فهرست مطالب

فصل اول کلیات پژوهش	۱
۱-۱. مقدمه	۲
۱-۲. بیان مسئله	۳
۱-۳. ضرورت پژوهش	۳
۱-۴. هدف پژوهش	۴
۱-۵. سوالات پژوهش	۴
۱-۶. محدودیت‌های پژوهش	۴
۱-۷. اصطلاحات تخصصی	۴
فصل دوم پیشینه و مبانی نظری پژوهش	۶
۲-۱. مقدمه	۷
۲-۲. نشانگرهای سبکی در شناسایی نویسنده	۷
۲-۲-۱. نشانگرهای واژگانی	۷
۲-۲-۲. نشانگرهای کاراکتری	۸
۲-۲-۳. نشانگرهای نحوی	۹
۲-۲-۴. نشانگرهای معنایی	۱۱
۲-۳. روش‌های رایانشی متداول	۱۲
۲-۳-۱. روش‌های بی‌نظارت	۱۲
۲-۳-۲. روش‌های بانظارت	۱۶
۲-۳. زبان فردی به مثابه‌ی اثرانگشت گویشور	۱۹
۲-۳-۱. زبان فردی در گفتار	۱۹
۲-۳-۲. زبان فردی در نوشتار	۲۲

فصل سوم روش پژوهش	۲۴
۳-۱. مقدمه	۲۵
۳-۲. نشانگرهای سبکی	۲۵
۳-۳. پیکره‌ی پژوهش	۲۶
۳-۴. روش تحلیل	۲۸
۳-۵. نرم‌افزار جستجوی پیکره	۲۹
۳-۶. روال انجام پژوهش	۳۰
فصل چهارم تحلیل داده‌ها	۳۳
۴-۱. مقدمه	۳۴
۴-۲. آزمایش ۱ (متون ۵۰۰۰ کلمه‌ای)	۳۴
۴-۳. آزمایش ۲ (متون ۱۰۰۰۰ کلمه‌ای)	۴۰
۴-۴. آزمایش ۳ (متون ۱۰۰۰ کلمه‌ای)	۴۸
۴-۵. آزمایش ۴ (متون ۲۰۰۰ کلمه‌ای)	۵۴
۴-۶. آزمایش ۵ (متون ۳۰۰۰ کلمه‌ای)	۶۲
۴-۷. آزمایش ۶ (متون ۴۰۰۰ کلمه‌ای)	۶۸
۴-۷. جمع‌بندی آزمایش‌ها	۷۵
۴-۸. تحلیل‌های پیکره‌ای	۷۷
فصل پنجم نتیجه‌گیری	۸۲
۵-۱. مقدمه	۸۳
۵-۲. پاسخ به پرسش‌های پژوهش	۸۳
۵-۳. پیشنهاداتی برای پژوهش‌های آتی	۸۶
منابع	۸۸

فهرست جداول

- جدول ۱-۳. تعداد کلمات موجود در پیکره به تفکیک نویسندگان..... ۲۷
- جدول ۲-۳. حجم تقریبی پیکره مورد استفاده در هر آزمایش..... ۳۰
- جدول ۳-۳. فهرست واژه‌های محذوف از جدول واژه‌های پربسامد..... ۳۱

فهرست اشکال

- شکل ۱-۲. تحلیل خوشه‌ای از شباهت برخی همخوان‌های انگلیسی ۱۳
- شکل ۲-۲. تحلیل مولفه‌های اصلی نه رمان از چهار نویسنده بر اساس ۱۰۰ واژه‌ی پرسامد ۱۶
- شکل ۱-۳. برخی امکانات رابط کاربردی در بسته‌ی استایلو (Eder et al., 2013) ۲۹
- شکل ۲-۳. محیط نرم‌افزار پیکره‌ای AntConc ۳۰
- شکل ۱-۴. تحلیل مولفه‌های اصلی تک‌نگاشتی در متون ۵۰۰۰۰ کلمه‌ای ۳۵
- شکل ۲-۴. تحلیل مولفه‌های اصلی تک‌نگاشتی به‌همراه بارهای عاملی در متون ۵۰۰۰۰ کلمه‌ای ۳۶
- شکل ۳-۴. تحلیل خوشه‌ای واژه‌های دستوری تک‌نگاشتی در متون ۵۰۰۰۰ کلمه‌ای ($ARI=1$) ۳۷
- شکل ۴-۴. تحلیل مولفه‌های اصلی دونگاشتی‌ها به‌همراه بار عاملی در متون ۵۰۰۰۰ کلمه‌ای ۳۸
- شکل ۵-۴. تحلیل خوشه‌ای بر اساس واژه‌های دستوری دونگاشتی در متون ۵۰۰۰۰ کلمه‌ای ($ARI=1$) ۳۹
- شکل ۶-۴. تحلیل مولفه‌های اصلی واژه‌های سه‌نگاشتی به‌همراه بار عاملی در متون ۵۰۰۰۰ کلمه‌ای ۴۰
- شکل ۷-۴. تحلیل خوشه‌ای واژه‌های دستوری سه‌نگاشتی در متون ۵۰۰۰۰ کلمه‌ای ($ARI=1$) ۴۰
- شکل ۸-۴. تحلیل مولفه‌های اصلی تک‌نگاشتی در متون ۱۰۰۰۰ کلمه‌ای ۴۱
- شکل ۹-۴. تحلیل مولفه‌های اصلی تک‌نگاشتی به‌همراه بار عاملی در متون ۱۰۰۰۰ کلمه‌ای ۴۲
- شکل ۱۰-۴. تحلیل خوشه‌ای واژه‌های دستوری تک‌نگاشتی در متون ۱۰۰۰۰ کلمه‌ای ($ARI=1$) ۴۳
- شکل ۱۱-۴. تحلیل مولفه‌های اصلی دونگاشتی در متون ۱۰۰۰۰ کلمه‌ای ۴۴
- شکل ۱۲-۴. تحلیل مولفه‌های اصلی دونگاشتی به‌همراه بار عاملی در متون ۱۰۰۰۰ کلمه‌ای ۴۴
- شکل ۱۳-۴. تحلیل خوشه‌ای دونگاشتی‌ها در متون ۱۰۰۰۰ کلمه‌ای ($ARI=0.8$) ۴۵
- شکل ۱۴-۴. تحلیل مولفه‌های اصلی سه‌نگاشتی در متون ۱۰۰۰۰ کلمه‌ای ۴۶
- شکل ۱۵-۴. تحلیل مولفه‌های اصلی سه‌نگاشتی به‌همراه بار عاملی در متون ۱۰۰۰۰ کلمه‌ای ۴۶
- شکل ۱۶-۴. تحلیل خوشه‌ای سه‌نگاشتی‌ها در متون ۱۰۰۰۰ کلمه‌ای ($ARI=0.24$) ۴۷

- شکل ۴-۱۷. تحلیل مولفه‌های اصلی تک‌نگاشتی در متون ۱۰۰۰ کلمه‌ای..... ۴۹
- شکل ۴-۱۸. تحلیل مولفه‌های اصلی تک‌نگاشتی به‌همراه بار عاملی در متون ۱۰۰۰ کلمه‌ای..... ۴۹
- شکل ۴-۱۹. تحلیل خوشه‌ای تک‌نگاشتی‌ها در متون ۱۰۰۰ کلمه‌ای ($ARI=0.36$)..... ۵۰
- شکل ۴-۲۰. تحلیل مولفه‌های اصلی دونگاشتی در متون ۱۰۰۰ کلمه‌ای..... ۵۱
- شکل ۴-۲۱. تحلیل مولفه‌های اصلی دونگاشتی به‌همراه بار عاملی در متون ۱۰۰۰ کلمه‌ای..... ۵۱
- شکل ۴-۲۲. تحلیل خوشه‌ای دونگاشتی‌ها در متون ۱۰۰۰ کلمه‌ای ($ARI=0.2$)..... ۵۲
- شکل ۴-۲۳. تحلیل مولفه‌های اصلی سه‌نگاشتی در متون ۱۰۰۰ کلمه‌ای..... ۵۳
- شکل ۴-۲۴. تحلیل مولفه‌های اصلی سه‌نگاشتی به‌همراه بار عاملی در متون ۱۰۰۰ کلمه‌ای..... ۵۳
- شکل ۴-۲۵. تحلیل خوشه‌ای سه‌نگاشتی‌ها در متون ۱۰۰۰ کلمه‌ای ($ARI=0.04$)..... ۵۴
- شکل ۴-۲۶. تحلیل مولفه‌های اصلی تک‌نگاشتی در متون ۲۰۰۰ کلمه‌ای..... ۵۵
- شکل ۴-۲۷. تحلیل مولفه‌های اصلی تک‌نگاشتی به‌همراه بار عاملی در متون ۲۰۰۰ کلمه‌ای..... ۵۶
- شکل ۴-۲۸. تحلیل خوشه‌ای تک‌نگاشتی‌ها در متون ۲۰۰۰ کلمه‌ای ($ARI=0.75$)..... ۵۷
- شکل ۴-۲۹. تحلیل مولفه‌های اصلی دونگاشتی در متون ۲۰۰۰ کلمه‌ای..... ۵۸
- شکل ۴-۳۰. تحلیل مولفه‌های اصلی دونگاشتی به‌همراه بار عاملی در متون ۲۰۰۰ کلمه‌ای..... ۵۸
- شکل ۴-۳۱. تحلیل خوشه‌ای دونگاشتی‌ها در متون ۲۰۰۰ کلمه‌ای ($ARI=0.09$)..... ۵۹
- شکل ۴-۳۲. تحلیل مولفه‌های اصلی سه‌نگاشتی در متون ۲۰۰۰ کلمه‌ای..... ۶۰
- شکل ۴-۳۳. تحلیل مولفه‌های اصلی سه‌نگاشتی به‌همراه بار عاملی در متون ۲۰۰۰ کلمه‌ای..... ۶۰
- شکل ۴-۳۴. تحلیل خوشه‌ای سه‌نگاشتی‌ها در متون ۲۰۰۰ کلمه‌ای ($ARI=0.28$)..... ۶۱
- شکل ۴-۳۵. تحلیل مولفه‌های اصلی تک‌نگاشتی در متون ۳۰۰۰ کلمه‌ای..... ۶۲
- شکل ۴-۳۶. تحلیل مولفه‌های اصلی تک‌نگاشتی به‌همراه بار عاملی در متون ۳۰۰۰ کلمه‌ای..... ۶۳
- شکل ۴-۳۷. تحلیل خوشه‌ای تک‌نگاشتی در متون ۳۰۰۰ کلمه‌ای ($ARI=0.73$)..... ۶۴
- شکل ۴-۳۸. تحلیل مولفه‌های اصلی دونگاشتی در متون ۳۰۰۰ کلمه‌ای..... ۶۵
- شکل ۴-۳۹. تحلیل مولفه‌های اصلی دونگاشتی به‌همراه بار عاملی در متون ۳۰۰۰ کلمه‌ای..... ۶۵
- شکل ۴-۴۰. تحلیل خوشه‌ای دونگاشتی در متون ۳۰۰۰ کلمه‌ای ($ARI=0.3$)..... ۶۶
- شکل ۴-۴۱. تحلیل مولفه‌های اصلی سه‌نگاشتی در متون ۳۰۰۰ کلمه‌ای..... ۶۷
- شکل ۴-۴۲. تحلیل مولفه‌های اصلی سه‌نگاشتی به‌همراه بار عاملی در متون ۳۰۰۰ کلمه‌ای..... ۶۷
- شکل ۴-۴۳. تحلیل خوشه‌ای سه‌نگاشتی در متون ۳۰۰۰ کلمه‌ای ($ARI=0.21$)..... ۶۸
- شکل ۴-۴۴. تحلیل مولفه‌های اصلی تک‌نگاشتی در متون ۴۰۰۰ کلمه‌ای..... ۶۹
- شکل ۴-۴۵. تحلیل مولفه‌های اصلی تک‌نگاشتی به‌همراه بار عاملی در متون ۴۰۰۰ کلمه‌ای..... ۷۰
- شکل ۴-۴۶. تحلیل خوشه‌ای تک‌نگاشتی در متون ۴۰۰۰ کلمه‌ای ($ARI=1$)..... ۷۰
- شکل ۴-۴۷. تحلیل مولفه‌های اصلی دونگاشتی در متون ۴۰۰۰ کلمه‌ای..... ۷۱
- شکل ۴-۴۸. تحلیل مولفه‌های اصلی دونگاشتی به‌همراه بار عاملی در متون ۴۰۰۰ کلمه‌ای..... ۷۲

- شکل ۴-۴۹. تحلیل خوشه‌ای دنگاشتی در متون ۴۰۰۰ کلمه‌ای (ARI=0.34) ۷۳
- شکل ۴-۵۰. تحلیل مولفه‌های اصلی سه‌نگاشتی در متون ۴۰۰۰ کلمه‌ای ۷۴
- شکل ۴-۵۱. تحلیل مولفه‌های اصلی سه‌نگاشتی به‌همراه بار عاملی در متون ۴۰۰۰ کلمه‌ای ۷۴
- شکل ۴-۵۲. تحلیل خوشه‌ای سه‌نگاشتی در متون ۴۰۰۰ کلمه‌ای (ARI=0.32) ۷۵
- شکل ۴-۵۳. نمودار Π ننگاشتی تغییرات شاخص تعدیل‌شده‌ی رند بر اساس تعداد کلمات متون ۷۶
- شکل ۴-۵۴. نمودار بار عاملی. سمت راست: تک‌نگاشتی (۵۰۰۰۰ کلمه)؛ سمت چپ: تک‌نگاشتی (۱۰۰۰۰ کلمه) .. ۷۷
- شکل ۴-۵۵. نمودار بار عاملی. سمت راست: دنگاشتی (۵۰۰۰۰ کلمه)؛ سمت چپ: سه‌نگاشتی (۵۰۰۰۰ کلمه) ۷۹
- شکل ۴-۵۶. بسامد نسبی برخی دنگاشتی‌های پرسامد ممیز سبکی در پیکره‌ی متنی سه نویسنده ۷۹
- شکل ۴-۵۷. بسامد نسبی برخی سه‌نگاشتی‌های پرسامد ممیز سبکی در پیکره‌ی متنی سه نویسنده ۸۰

فصل اول

کلیات پژوهش

۱-۱. مقدمه

شناسایی اصالت نوشته‌ها و کشف هویت نویسنده‌ی متون از دیرباز مورد توجه انسان بوده است. یکی از قدیمی‌ترین نمونه‌های آن، اثبات جعلی بودن سند «بخشش کنستانتین» توسط لورنزو والا در سال ۱۴۳۹ میلادی است؛ طبق این سند جعلی، کنستانتین (امپراتور روم) حکومت روم و قسمت غربی امپراطوری روم را به پاپ سپرده بود. پژوهش‌هایی که امروزه در این حوزه انجام می‌شوند، عمدتاً با عنوان سبک‌سنجی^۱ شناخته می‌شوند. در عصر کنونی با تحولی که رایانه در پردازش متون حجیم ایجاد کرده و کاربردهایی که پژوهش‌های سبک‌سنجی در زبان‌شناسی قانونی^۲ (Toty et al., 1987) و کشف سرقت علمی و ادبی (Stein et al., 2010) پیدا کرده است، بررسی نظام‌مند متون با روش‌های علمی دقیق بیشتر مورد توجه قرار گرفته است.

پژوهش‌های نوین سبک‌سنجی که مبتنی بر تحلیل‌های رایانشی هستند، بر مولفه‌های مختلفی از متون تمرکز می‌کنند. برای نمونه یکی از پژوهش‌های مهم و اولیه در این باره اختصاص به شناسایی مولف «مقالات فدرالیست»^۳ داشت (Mosteller and Wallace, 1964) که به تحلیل آماری بسامد واژه‌های دستوری (نظیر حروف اضافه) پرداخته بود. مولفه‌های دیگر می‌تواند در سطح واژگانی (غناي واژه‌ها، بسامد واژه‌ها، غیره)، نحوی (ساخت جمله، بسامد قواعد نحوی، خطاها، اجزای کلام)، معنایی (هم‌معناها، وابستگی‌های معنایی) باشد (Stamatatos, 2009). اما برخی ویژگی‌های منحصر به فرد در واژه‌های دستوری وجود دارد که موجب می‌شود نشانگرهای سبکی خوبی برای تحلیل سبک نویسنده باشند. برخی از این ویژگی‌ها عبارتند از: بسامد بالای آنها در متن که دست پژوهشگر را در تحلیل متون کوتاه‌تر باز می‌گذارد، عدم تاثیرپذیری واژه‌های نقشی از موضوع یا ژانر نوشته، و عدم کنترل خود آگاه نویسنده بر کاربرد واژه‌های دستوری (Binongo, 2003).

در این طرح پژوهشی، کارایی واژه‌های دستوری زبان فارسی در تفکیک سبک زبانی نویسنده‌های مختلف بررسی خواهد شد. مبنای این فرضیه که افراد مختلف دارای سبک زبانی مختلف هستند، مفهومی موسوم به زبان فردی یا گویش فردی در علم زبان‌شناسی است. بر اساس این مفهوم، زبان هر فردی (به لحاظ واژگانی، دستوری، و غیره) به دلیل برخورداری از عادات اکتسابی و مشخصه‌های

¹ stylometry

² forensic linguistics

³ Federalist Papers

۳- شناسایی هویت نویسنده بر اساس زبان فردی

سبکی منحصر به فرد، متفاوت از زبان دیگر افراد است (Wardhaugh, 2010). این پژوهش با رویکردی زبان‌شناختی و آماری به آزمایش امکان تمایز سبک نویسندگان بر اساس زبان فردی آنها خواهد پرداخت.

۲-۱. بیان مسئله

واژه‌های دستوری علاوه بر اینکه در زبان‌های مختلف به شکل‌های مختلفی نمود می‌یابند، ممکن است دارای تنوع نوعی و نقشی نیز باشند. مثلاً در زبان انگلیسی و فارسی نمود این واژه‌ها عمدتاً به صورت واژه‌های مستقل نظیر حروف اضافه، حروف ربط و غیره است. اما مثلاً در زبان ترکی، واژه‌های دستوری به واژه‌های محتوایی متصل می‌شوند و نمود مستقل ندارند. از سوی دیگر، در زبان‌هایی که واژه‌های دستوری عمدتاً به شکل مشابهی نمود می‌یابند (انگلیسی و فارسی)، برخی ویژگی‌های منحصر به فرد نیز ملاحظه می‌شود. مثلاً در زبان فارسی حرف اضافه‌ی «را» نشانگر مفعول بی‌واسطه است که مابه‌ازایی در زبان انگلیسی ندارد. یا برخی واژه‌های دستوری در زبان فارسی، صورتی ترکیبی دارند (نظیر «با این حال») که مابه‌ازای نقشی آنها در زبان انگلیسی (مثلاً «however») ممکن است صورت ساده یا تک‌واژه‌ای داشته باشد. بنابراین، پژوهش‌هایی که نقش موثر واژه‌های دستوری در تفکیک سبک‌های نوشتاری در یک زبان خاص را تایید می‌کنند، لزوماً به معنای تایید آن در زبان‌های دیگر نمی‌تواند باشد.

در این پژوهش، امکان‌سنجی تفکیک سبک‌های نوشتاری نویسنده‌های مختلف بر اساس واژه‌های دستوری در زبان فارسی بررسی شده است. مسئله‌ی این پژوهش در قالب سه زیرمسئله قابل تعریف است که عبارتند از: امکان‌پذیری تفکیک سبک‌های نویسنده‌ها با استفاده از واژه‌های دستوری فارسی، بررسی توالی‌های (نگاشتی) پربسامد واژه‌های دستوری در تفکیک سبک‌ها، و برآورد حداقل کلمات مورد نیاز برای تفکیک موفقیت‌آمیز سبک‌های نوشتاری بر اساس واژه‌های دستوری.

۳-۱. ضرورت پژوهش

زبان‌شناسی قانونی، شاخه‌ای نسبتاً جدید از علم زبان‌شناسی به طور عام و زبان‌شناسی کاربردی به طور خاص است که به بررسی کاربرد تحلیل‌ها، روش‌ها، و دانش زبان‌شناختی در مسائل حقوقی و قضایی از جمله کشف هویت نویسنده و سرقت‌های ادبی/علمی می‌پردازد. تحلیل دقیق و علمی از زبانی که حوزه‌ی عملکرد این روش‌ها و مبنای قضاوت‌های رسمی واقع خواهد شد از اهمیت برخوردار است.

شناسایی هویت نویسنده بر اساس زبان فردی □ ۴

بر همین اساس، با توجه به نوپا بودن رشته‌ی زبان‌شناسی قانونی در ایران و کاربردهای سبک‌شناختی آن، به پژوهش‌های بنیادینی نیاز است که به توصیف و تحلیل رفتار مولفه‌های مختلف زبان فارسی بپردازند. این پژوهش، تحلیلی زبان‌شناختی-آماري از رفتار واژه‌های دستوری در زبان فارسی به‌دست خواهد داد.

۴-۱. هدف پژوهش

اهداف کلی این پژوهش را می‌توان در قالب موارد زیر خلاصه کرد:

- بررسی وجود زبان فردی در نوشتار نویسندگان فارسی با تحلیل n-نگاشتی واژه‌های دستوری پربسامد
- بررسی کارایی مقادیر مختلف n در n-نگاشت‌ها برای شناسایی زبان فردی
- ارائه‌ی الگویی عملیاتی برای کاربرد نتایج طرح در شناسایی هویت نویسنده

۵-۱. سوالات پژوهش

۱. واژه‌های دستوری چگونه می‌توانند سرنخی از زبان فردی به دست دهند؟
۲. در تحلیل n-نگاشتی زبان فارسی، کدام مقادیر n بهتر می‌تواند زبان فردی نویسندگان مختلف را از یکدیگر تفکیک کند؟
۳. چگونه می‌توان هویت نویسنده را از بررسی نوشته‌ها شناسایی کرد؟

۶-۱. محدودیت‌های پژوهش

۱. عدم دسترسی کافی به متون الکترونیکی نویسندگان فارسی
۲. مشکل استانداردسازی تمام و کمال املاي متون فارسی

۷-۱. اصطلاحات تخصصی

n-نگاشت^۴: توالی تعداد n کلمه‌ی مجاور. مقدار کمینه برای n عبارت از $n=1$ است که همان کلمات منفرد بدون توالی هستند.
زبان فردی^۵: عادات و الگوهای کاربردی زبان که برای هر شخصی منحصر به فرد است و در سطوحی نظیر واژگان و نحو نمود می‌یابند.

^۴ n-gram

شناسایی هویت نویسندگان بر اساس زبان فردی □۵

استانداردسازی متن^۶: فرآیندی که طی آن شیوه‌های متنوع نگارش و خطاهای املایی احتمالی تصحیح می‌شوند و به شکل استاندارد درمی‌آیند.

^۵ idiolect

^۶ text normalization

فصل دوم

پیشینه و مبانی نظری پژوهش

۱-۲. مقدمه

حوزه‌ی سبک‌سنجی رایانشی به‌طور عام و تشخیص هویت نویسنده به‌طور خاص، موضوع تحقیقات فراوان با رویکردهای رایانشی و شاخص‌های تحلیلی متنوع بوده است. تمرکز این فصل بر دو موضوع شناسایی رایانشی نویسنده و زبان فردی به‌مثابه‌ی اثر انگشت نویسنده خواهد بود و پیشینه‌ای مختصر از پژوهش‌های انجام‌گرفته در این ارتباط مرور خواهد شد. در موضوع شناسایی رایانشی نویسنده، که دربردارنده‌ی رویکردهای مختلف از جمله رویکرد مهندسی به سبک‌سنجی رایانشی است، به مرور نشانگرهای سبکی مورد استفاده در تحقیقات مختلف و نیز روش‌های رایانشی متداول پرداخته خواهد شد هر کدام در بخش‌های مجزایی بحث خواهند شد. بخش مربوط به زبان فردی، حوزه‌ی سبک‌سنجی رایانشی را به‌طور خاص از منظر تحقیقات زبان‌شناختی و در قالب دو زیربخش گونه‌ی گفتاری و نوشتاری مرور خواهد کرد.

۲-۲. نشانگرهای سبکی در شناسایی نویسنده

شناسایی نویسنده را می‌توان فرآیندی تعریف کرد که طی آن با اندازه‌گیری برخی مولفه‌های متنی می‌توان متونی را که توسط نویسنده‌های مختلف نوشته شده، از یکدیگر متمایز کرد (Stamatatos, 2009). پیش‌فرض مهمی که در تحقیقات مربوط به شناسایی نویسنده وجود دارد این است که نویسنده‌ی یک متن مورد مناقشه، از بین چند نویسنده‌ی مظنون انتخاب می‌شود و بنابراین بدیهی است که فرآیند «شناسایی» به جستجوی فرد از میان نویسندگان ناشناخته و بی‌شمار دلالت ندارد. برای مقایسه‌ی متنی که هویت نویسنده‌ی آن نامعلوم است با متونی که نویسندگان مشخصی دارند، نیاز به شاخص‌هایی است که ویژگی‌های سبکی متون را بتوان بر اساس آن شاخص‌ها استخراج و با یکدیگر مقایسه کرد. این شاخص‌ها طیف وسیعی از مشخصه‌های زبانی را شامل می‌شوند؛ در این پژوهش به چهار مقوله‌ی عمده، یعنی نشانگرهای کاراکتری، واژگانی، نحوی و معنایی، که کمابیش منطبق بر حوزه‌های نظری زبان‌شناختی هستند تمرکز خواهیم کرد.

۱-۲-۲. نشانگرهای واژگانی

واژه‌ها را شاید بتوان متعارف‌ترین و طبیعی‌ترین واحدهای زبانی در اکثر زبان‌ها به حساب آورد که در یک توالی زنجیروار، متن را می‌سازند. اگرچه نشانگرهای واژگانی همگی مبتنی بر بسامد واژه‌ها برای تمایز میان سبک‌های نوشتاری‌اند، اما ممکن است در قالب روش‌های مختلف عملیاتی شوند.

برای نمونه، یکی از نشانگرهای واژگانی در سبک سنجی، تعیین طول جملات^۷ بر حسب تعداد کلمات است. افراد مختلف بسته به سطح تحصیلات و مهارت زبانی ممکن است جملاتی متفاوت از یکدیگر به لحاظ طول واژگانی تولید کنند که می تواند به عنوان نشانگر سبکی مورد بهره برداری واقع شود. برای نمونه، معروف است که ارنست همینگوی، نویسنده آمریکایی، از جملات کوتاه و بی تکلف استفاده می کرد و این یکی از مشخصه های بارز سبک نوشتاری اش بود. این امر در مورد تولید زبانی سایر افراد جامعه نیز صادق است. برای به دست شاخص طول جملات برای افراد مختلف، کافی است بتوانیم مرز جملات در متن را به طور خودکار تشخیص دهیم و میانگینی از طول جملات به کار رفته به دست دهیم.

نشانگر واژگانی دیگری که مبتنی بر بسامد واژگانی عمل می کند، نشانگر غنای واژگانی^۸ است. غنای واژگانی نشان می دهد که واژه های مورد استفاده توسط فرد تا چه حد متنوع هستند. در مورد غنای واژگانی نیز، تحقیقات صورت گرفته در حوزه روان شناسی زبان نشان می دهد افراد مختلف غنای واژگانی متفاوتی دارند (Carroll, 2008). متداول ترین روش برای محاسبه ی غنای واژگانی یک متن این است که تعداد واژه های منحصر به فرد (گونه) آن را بر تعداد کل کلمات (نمونه) به کار رفته تقسیم می کنند (Johansson, 2008). شاخص به دست آمده نسبت گونه به نمونه^۹ را به دست خواهد داد که نشانگر غنای واژگانی متن مورد نظر است. اگرچه این شاخص می تواند تصویری از تنوع واژگانی افراد به دست دهد اما به دلیل تحت تاثیر قرار گرفتن این شاخص از اندازه ی متن، در به کارگیری آن باید جانب احتیاط نگه داشته شود و حجم متون مورد مقایسه حداقل امکان باید یکسان باشد (همان).

۲-۲-۲. نشانگرهای کاراکتری

یک روش دیگر در بازنمایی متن این است که به جای تقطیع متن به کلمات مجزا، آن را مجموعه ای سلسله وار از کاراکترها بدانیم. در این رویکرد، هر نشانه ای که در متن (نوشتاری) واقع می شود (نظیر حروف الفبا، علائم نگارشی، و فاصله های بین کاراکترها) به طور بالقوه یک نشانگر سبکی خواهد بود. بنابراین، کمی سازی متن می تواند در قالب بسامد کاراکترهای متنی انجام شود و در حالت های

⁷ sentence length

⁸ lexical richness

⁹ type-token ratio

گوناگون n-نگاشتی^{۱۰} (تک-نگاشتی، دونگاشتی، سه-نگاشتی و غیره) نمود یابد. در حالت تک-نگاشتی، بسامد کارکترها به صورت تکی محاسبه می شود. در حالت دونگاشتی، بسامد دو کاراکتر متوالی از متن استخراج می شود، و در حالت سه-نگاشتی نیز توالی های سه کاراکتری به دست داده می شود. برای مثال، بازنمایی کاراکتری یک تا سه-نگاشتی از عبارت «سبک فردی» به صورت زیر خواهد بود:

تک-نگاشتی: "س"، "ب"، "ک"، "Ø"، "ف"، "ر"، "د"، "ی"

دونگاشتی: "سب"، "بک"، "کØ"، "Øف"، "فر"، "رد"، "دی"

سه-نگاشتی: "سبک"، "بکØ"، "کØف"، "Øفر"، "فرد"، "ردی"

اگرچه بازنمایی متن در قالب کاراکترها یا n-نگاشت های کاراکتری بی معنی و غیرشمی به نظر می رسد، اما استفاده از این روش از دهه ی ۹۰ میلادی به این سو موفقیت های چشمگیری را در شناسایی سبک نویسنده به دست آورده است (Kestemont, 2014). یکی از نقاط قوت رویکرد کاراکترمحور به تحلیل سبک زبانی این است که مستقل از زبان عمل می کند و به همین دلیل در اکثر زبان ها قابل پیاده سازی است. همچنین در تحلیل دلایل موفقیت این روش در شناسایی سبک نویسنده چنین استدلال شده که به لحاظ اطلاعاتی، کاراکترها قادر محتوای اطلاعاتی بیشتری را نسبت به واژه ها بازنمایی کنند (همان). برای مثال در زبان های تصریفی (همانند لاتین) که وندها پررنگی در تعیین نقش کلمات دارند، نمی توان در سطح کلمات به اطلاعات نقشی دسترسی یافت. در این موارد، بازنمایی کاراکتری (n-نگاشتی) متن به دلیل اینکه در سطحی کوچکتر از واژه عمل می کند، می تواند به اطلاعات نقشی وندها دسترسی یابد. این مسئله تا حدودی در زبان های غیرتصریفی مثل زبان انگلیسی نیز کاربردهایی دارد؛ مثلاً برای دسترسی به اطلاعات زمان گذشته ی فعلی نظیر walked، بازنمایی کاراکتری سه-نگاشتی به صورت "edØ" می تواند این اطلاعات را در اختیار پژوهشگر قرار دهد.

۳-۲-۲. نشانگرهای نحوی

نشانگرهای نحوی همانطور که از نامش پیداست، بر مبنای ساختار نحوی و وابستگی های دستوری بین عبارات زبانی عمل می کند. پیش فرض استفاده از نشانگرهای نحوی در تشخیص هویت نویسنده این

¹⁰ n-gram

شناسایی هویت نویسنده بر اساس زبان فردی □ ۱۰

است که افراد در تولید زبان از الگوهای نحوی کمابیش مشابهی استفاده می‌کنند. نکته‌ای که استفاده از رویکردی نحوبنیاد را با چالش مواجه می‌کند، استفاده از ابزارهای خودکار تحلیل نحوی است. در واقع برای اینکه بتوان در تحلیل متنی از ساختار نحوی بهره جست، نیاز است که با استفاده از نرم‌افزارهای رایانشی خاص (نظیر تقطیع گر نحوی^{۱۱} یا برچسب‌زن نحوی^{۱۲}) بازنمایی نحوی را جمله به‌دست دهیم. این پیش‌پردازش در بسیاری از زبان‌ها به‌دلیل محدودیت در دسترسی به نرم‌افزارهای لازم امکانپذیر نیست.

اولین و ساده‌ترین سطح در بازنمایی نحوی جمله، بازنمایی متن بر اساس اجزای کلام^{۱۳} است. در واقع به‌جای اینکه بازنمایی متن بر اساس واژه‌های ملموس باشد، مقوله‌ی دستوری واژه‌ها مبنای بازنمایی قرار می‌گیرد که می‌توان همانند واژه‌ها، بسامد آنها و توالی‌شان را به‌صورت n-نگاشتی به‌دست داد. برای این منظور، ابتدا متن مورد نظر توسط برچسب‌زن نحوی برچسب زده می‌شود، سپس بسامد مقوله‌های نحوی محاسبه می‌شود. مثلاً برای جمله‌ی the boy ate the apple در زبان انگلیسی می‌توان بازنمایی نحوی زیر را ارائه داد که مبتنی بر اجزای کلام است:

ART_N_V_ART_N

در این جمله‌ی انگلیسی دو اسم apple و boy به کار رفته و قبل از آنها نیز حرف تعریف the آمده است. اگر بخواهیم بازنمایی واژگانیِ دونگاشتی از این جمله به‌دست دهیم، توالی the boy و the apple دو نمونه‌ی بسامدی مجزا تلقی خواهند شد. در صورتی که اگر بازنمایی متن بر اساس مقوله‌های نحوی یا اجزای کلام باشد، این دو گروه اسمی (the boy و the apple) در قالب الگوی انتزاعی تر ART_N در یک گروه قرار خواهند گرفت.

استفاده از اجزای کلام در بسیاری از پژوهش‌های شناسایی سبک نویسنده موفقیت‌آمیز بوده است (Gamon, 2004; Koppel and Schler, 2003)، اما با این حال مقوله‌های نحوی روابط ساختاری متن را به‌دست نمی‌دهد و برای ردگیری سبک نویسنده در قالب روابط نحوی پیچیده‌تر نیاز به تحلیل‌های سلسله‌مراتبی پیچیده‌تر است.

¹¹ syntactic parser

¹² POS tagger

¹³ Parts of Speech (POS)

۴-۲-۲. نشانگرهای معنایی

مشخصه‌های معنایی متن یکی دیگر از مولفه‌هایی هستند که این پتانسیل را دارند، برخی مشخصه‌های سبکی نویسنده‌ی یک متن را فاش کنند. این نشانگرهای معنایی می‌توانند طیف وسیعی از مشخصه‌ها نظیر شمار دستوری اسامی، زمان و نمود دستوری فعل و مشخصه‌های زیرمقوله‌ای فعل را شامل شود (Gamon, 2004). برخی پژوهش‌ها همچنین از رویکردهای نظام‌مندتر و دستوربنیان‌تری به این مسئله استفاده کرده‌اند. مثلاً در برخی پژوهش‌ها (Whitelaw & Argamon, 2004; Whitelaw & Patrick, 2004)، مجموعه‌ای از مشخصه‌های معنایی بر اساس دستور نظام‌مند نقشگرا^{۱۴} تعریف شده‌اند که کلمات یا عبارات زبانی بخصوصی را با اطلاعات معنایی مرتبط می‌کنند. دستور نظام‌مند نقشگرا متن را به مثابه‌ی تجلی انتخاب‌هایی می‌داند که در درون شبکه‌ای از معانی مرتبط شکل گرفته‌اند و این معانی عمدتاً غیرارجاعی (به جهان خارج) هستند. بر اساس این رویکرد، وایتلا و آرگامون معانی سبکی را همان جوانبی از معنی در نظر می‌گیرند که غیرارجاعی بوده و مستقل از اشیاء و رویدادهای ارجاع‌داده‌شده توسط متن شکل گرفته‌اند.

برخی مطالعات اخیر از معناشناسی قاب‌بنیان^{۱۵} نیز در شناسایی نویسنده، به‌ویژه در آثار ترجمه‌شده، استفاده کرده‌اند. معناشناسی قاب‌بنیان نظریه‌ای جدید در معناشناسی است که معنای یک واژه را در دایره‌ی دانش دایره‌المعارفی آن واژه و روابط معنایی با واژه‌های دیگر تعریف می‌کند. برای نمونه، طبق این نظریه، معنای واژه‌ای نظیر «فروختن» بدون در نظر گرفتن دانش دایره‌المعارفی قاب دادوستد و ارتباط آن با دیگر واژه‌ها نظیر «خریدن»، «خریدار»، «کالای مورد معامله» و غیره قابل درک نخواهد بود. در پژوهشی اخیر (Hedegaard & Simonsen, 2011) که برای شناسایی نویسنده‌ی یک متن ترجمه‌شده انجام شده، پژوهشگران با استفاده از روابط معنایی تعریف‌شده بر اساس شبکه‌ی قاب‌ها^{۱۶}، دقت شناسایی نویسنده‌ی اصلی متن را افزایش دادند. فرض بنیادین آنها در استفاده از معناشناسی قاب‌بنیان این بوده که ترجمه‌ی یک متن بسته به سبک مترجم می‌تواند نشانگرهای نحوی و واژگانی نویسنده‌ی اصلی را دگرگون کند، اما قاب‌های معنایی به کاررفته در متن اصلی در متن ترجمه‌شده دگرگون نمی‌شوند.

¹⁴ Systemic functional grammar

¹⁵ Frame Semantics

¹⁶ FameNet

تحلیل متن بر اساس نشانگرهای معنایی نیاز به تحلیل عمیق متون و استفاده از ابزارهای رایانشی خودکار دارد. با توجه به اینکه ابزارهای خودکار پردازش معنایی زبان در ابتدای مسیر هستند، چالش‌های استفاده از نشانگرهای معنایی بیشتر از نشانگرهای صوری واژگانی و نحوی است.

۲-۳. روش‌های رایانشی متداول

در بخش قبل برخی مشخصه‌های مهم سبکی در قالب نشانگرهای واژگانی، نحوی، کاراکتری و معنایی معرفی شدند. این نشانگرها در واقع حکم شاخص‌هایی را دارند که جوانب مختلف تولید زبانی نویسنده یا گوینده را به صورت متغیرهای کمی بازنمایی می‌کنند تا در نهایت بتوان سبک گویشور را از دیگر سبک‌ها متمایز کرد. برای تحلیل نشانگرهای سبکی، روش‌ها و تکنیک‌های مختلفی توسط پژوهشگران مختلف به کار گرفته شده است. در ادامه به بررسی برخی از مهم‌ترین تکنیک‌ها که ذیل دو رویکرد عمده بی‌نظارت^{۱۷} و بانظارت^{۱۸} قرار می‌گیرند، خواهیم پرداخت.

۲-۳-۱. روش‌های بی‌نظارت

روش‌های بی‌نظارت به مجموعه روش‌هایی گفته می‌شود که به طور کلی برای کاوش و کشف الگو در داده‌ها به کار گرفته می‌شوند و هیچ معیار از پیش تعیین شده‌ای برای مطابقت نتایج با آن وجود ندارد. در حوزه شناسایی سبک زبانی این امر می‌تواند اینگونه نمود یابد که مثلاً سه نویسنده هر کدام ۴ متن را تولید کرده‌اند و ما می‌خواهیم میزان شباهت سبکی را بین کل این ۱۲ متن نامعلوم کشف کنیم بدون اینکه مشخصات سبکی هر نویسنده را برای مطابقت این ۱۲ متن با آن مشخصات از قبل بدانیم. برخی از رایج‌ترین روش‌های بی‌نظارت در حوزه سبک‌سنجی عبارتند از تحلیل خوشه‌ای^{۱۹} و تحلیل مولفه‌های اصلی^{۲۰} که در ادامه به معرفی هر یک می‌پردازیم.

۲-۳-۱-۱. تحلیل خوشه‌ای

تحلیل خوشه‌ای یکی از روش‌های بی‌نظارت است که داده‌ها را بر اساس اطلاعات موجود در آنها و روابطی که بین آنها وجود دارد، دسته‌بندی می‌کند. این دسته‌بندی‌ها معمولاً به صورت یک نمودار درختی نمایش داده می‌شوند. داده‌هایی که در یک مقوله قرار می‌گیرند نسبت به داده‌های

¹⁷ unsupervised

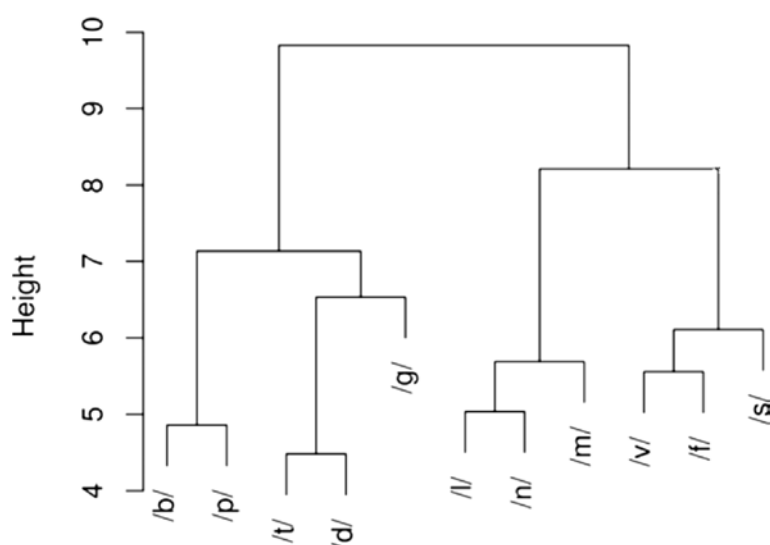
¹⁸ supervised

¹⁹ Cluster Analysis

²⁰ Principal Component Analysis

شناسایی هویت نویسنده بر اساس زبان فردی □ ۱۳

برون مقوله‌ای شباهت بیشتری به یکدیگر دارند. به عنوان مثالی زبان شناختی، پژوهشگری برای تعیین شباهت بین همخوان‌های /b/، /d/، /f/، /g/، /l/، /m/، /n/، /p/، /s/، /t/ و /v/ از ۲۰ نفر خواسته که شباهت هر یک از این همخوان‌ها را به صورت دوبه‌دو در مقیاس بین ۰ (کاملاً متفاوت) تا ۱ (کاملاً مشابه) ارزیابی کنند. داده‌های به دست آمده از این ارزیابی را می‌توان با روش تحلیل خوشه‌ای سلسله‌مراتبی تحلیل کرد و شباهت بین همخوان‌ها را به دست آورد. نمودار ۱-۲ نتیجه‌ی فرضی چنین تحلیلی را نشان می‌دهد (Gries, 2009: 337):



شکل ۱-۲. تحلیل خوشه‌ای از شباهت برخی همخوان‌های انگلیسی

در شکل ۱-۲، همخوان‌های انگلیسی ابتدا به دو شاخه‌ی عمده تقسیم شده‌اند که با بررسی همخوان‌های موجود در آن مشخص می‌شود زیرشاخه‌ی عمده‌ی سمت چپ دربردارنده‌ی همخوان‌های انسدادی است در حالی که زیرشاخه‌ی عمده‌ی سمت راست همخوان‌هایی را شامل می‌شود که ماهیت غیرانسدادی یا تداومی (مسیر حرکت هوا به طور کامل مسدود نمی‌شود) دارند. در نمودار تحلیل خوشه‌ای، مقیاسی در محور عمودی نمودار وجود دارد که ارتفاع (height) را نشان می‌دهد و شاخصی برای سنجش میزان دوری یا نزدیکی خوشه‌ها نسبت به هم است. هر چه دو خوشه در ارتفاع بالاتری از یکدیگر متمایز شود، نشان‌دهنده‌ی شباهت کمتر آنها نسبت به یکدیگر است. با بررسی زیرخوشه‌ها می‌توان مشاهده کرد همخوان‌های /p/ و /b/ و نیز همخوان‌های /t/ و /d/ به دلیل شباهت بالایشان در ارزیابی در یک خوشه قرار گرفته‌اند که به ترتیب در مشخصه‌ی لبی و لثوی مشترک هستند.

تحلیل خوشه‌ای در تحقیقات زبان‌شناختی کاربرد فراوان داشته است (Sandra & Rice, 1995; Gries, 2006) و در مطالعات شناسایی سبک نویسنده نیز می‌تواند مورد استفاده واقع شود. در کاربردهای سبک‌سنجی می‌توان نوشته‌ها یا متون زبانی را به‌مثابه‌ی متغیرهایی فرض کرد که قرار است میزان شباهت‌شان با یکدیگر ارزیابی شود، و نشانگرهای سبکی را به‌مثابه‌ی مشخصه‌هایی لحاظ کرد که نوشته‌ها بر اساس آنها با یکدیگر مقایسه می‌شوند. در زبان فارسی می‌توان به پژوهش مدبر دباغ (Modaber Dabagh, 2007) اشاره کرد که با استفاده از تحلیل خوشه‌ای نمونه آثار شعرای فارسی بر اساس واژه‌های دستوری پربسامد، به بررسی تفکیک متون مربوط به هر شاعر پرداخته است.

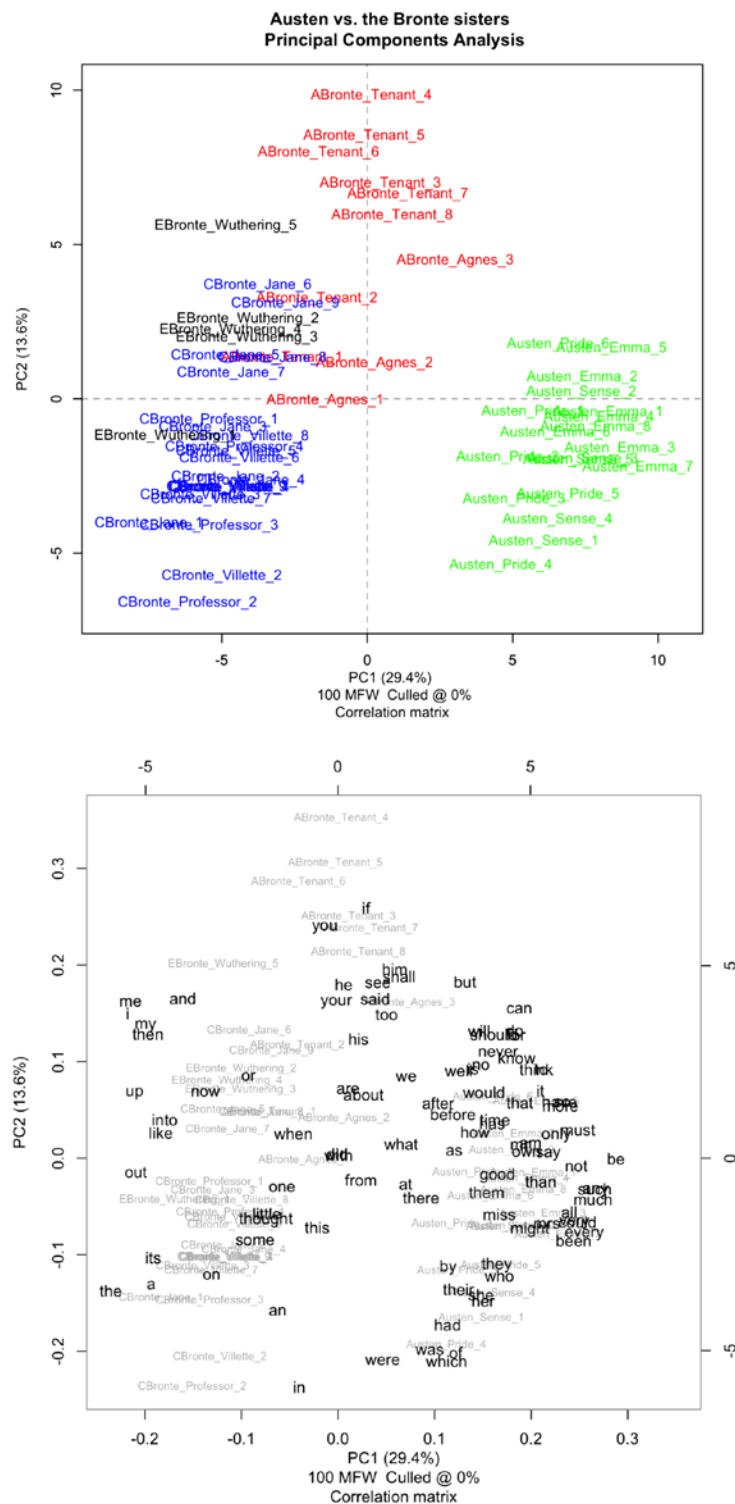
۲-۳-۱-۲. تحلیل مولفه‌های اصلی

تحلیل مولفه‌های اصلی روشی چندمتغیری و بی‌نظارت است که برای کاهش ابعاد داده‌ها به‌کار می‌رود. برای نمونه، در مطالعات سبک‌سنجی معمولاً از نشانگرهای سبکی فراوانی به‌عنوان متغیر استفاده می‌شود تا سبک نویسندگان بر اساس آنها توصیف شود. تمام این نشانگرها حاوی اطلاعات ارزشمند نیستند و اغلب بین این متغیرها همبستگی وجود دارد. برای یافتن نشانگرهای موثری که فاقد همبستگی باشند و بیشترین پراکنش را در سبک نویسنده‌ها نشان دهند، نیاز است با استفاده از روش تحلیل مولفه‌های اصلی، تعداد این نشانگرهای فراوان را کاهش دهیم و در قالب چند متغیر یا مولفه جدید با تعداد کمتر (معمولاً دو مولفه) بازنمایی کنیم. مولفه‌های جدید فاقد اطلاعات حشو هستند و حداکثر پراکنش در نمونه‌های مورد تحلیل را توصیف می‌کنند. مولفه‌ی اول (PC1) بر روی محور X نشان داده می‌شود که بیشترین پراکنش را نشان می‌دهد. مولفه‌ی دوم (PC2) دومین مولفه‌ای است که بیشترین پراکنش را در محور Y نشان خواهد داد.

شکل ۲-۲ مثالی فرضی از تحلیل مولفه‌های اصلی نه رمان از چهار نویسنده را نشان می‌دهد (Eder et al., accepted). این نمودار با تحلیل ۱۰۰ واژه‌ی پربسامد در هر یک از این رمان‌ها به‌دست آمده است. قسمت بالایی نمودار، شباهت سبکی نویسنده‌ها را بر اساس دو مولفه‌ی اصلی نشان می‌دهد. مولفه‌ی اصلی اول (PC1) با داشتن مقدار ۲۹/۴٪ مسئول بیشترین پراکنش در تمایز سبکی متون بوده است. مولفه‌ی اصلی دوم (PC2) نیز مسئول ۱۳/۶٪ پراکنش بوده است. این دو مولفه در مجموع ۴۲٪ از کل پراکنش متون را تبیین می‌کنند. در قسمت پایین نمودار، سهم هر یک از متغیرها (نشانگرهای واژگانی) در تمایز بین سبک نویسندگان مشاهده می‌شود. مثلاً نمودار پایینی نشان

شناسایی هویت نویسنده بر اساس زبان فردی □ ۱۵

می‌دهد که در مولفه‌ی اول، واژه‌هایی نظیر be, much و the بیشترین ویژگی متمایزکننده‌ی دو نویسنده‌ی CBronte/EBronte و Austen بوده‌اند و در مولفه‌ی دوم، واژه‌هایی نظیر you, if و in متمایزکننده‌ی سبک Austen/CBronte و ABronte بوده‌اند.



شناسایی هویت نویسنده بر اساس زبان فردی □ ۱۶

شکل ۲-۲. تحلیل مولفه‌های اصلی نه رمان از چهار نویسنده بر اساس ۱۰۰ واژه‌ی پربسامد. نمودار بالا، شباهت سبکی نویسنده‌های رمان‌ها را در مختصات دو مولفه‌ی اصلی به‌دست می‌دهد. نمودار پایین، نقش نشانگرهای سبکی (که در اینجا واژه‌ها هستند) در میزان تمایز سبک‌های نویسندگان را نشان می‌دهد.

روش‌های تحلیل خوشه‌ای و تحلیل مولفه‌های اصلی در حوزه‌ی سبک‌سنجی، جزو روش‌های کاوشگرانه‌ای هستند که عمدتاً بدون داشتن پیشداوری نسبت به درستی یا نادرستی تحلیل، سعی در توصیف و یافتن الگو در داده‌های زبانی دارند. اما مواقعی وجود دارد که هدف پژوهشگر کاوش، توصیف و کشف الگوها و روابط در داخل مجموعه‌ای از متون نیست، بلکه هدفش این است که معین کند یک متن با نویسنده‌ای نامعلوم به کدام یک از متونی که نویسندگان مشخص است تعلق دارد. روش‌های مورد استفاده برای پاسخ به این پرسش تحت عنوان روش‌های بانظارت شناخته می‌شوند که در بخش بعد به برخی از آنها اشاره می‌شود.

۲-۳-۲. روش‌های بانظارت

روش‌های بانظارت ویژگی‌های سبکی یک متن ناشناخته را با معیارهایی از قبل مشخص شده می‌سنجند. این روش‌ها طیف وسیعی از تکنیک‌ها و الگوریتم‌هایی را شامل می‌شوند که گاهی تحت عنوان یادگیری ماشینی نیز از آنها یاد می‌شود. برخی از این روش‌ها عبارتند از شبکه‌های عصبی^{۲۱}، دسته‌بندی کننده‌ی نایو بیز^{۲۲} و روش دلتا (Juola, 2006).

۲-۳-۲-۱. شبکه‌های عصبی

شبکه‌های عصبی مدل‌های رایانشی الهام گرفته از ساختار مغز انسان هستند که متشکل از واحدهای ساده اما به شدت درهم تنیده‌اند که اتصالات نورونی را شبیه‌سازی می‌کنند. این شبکه‌ها که معمولاً متشکل از سه لایه‌ی درون‌داد، لایه‌ی مخفی، و لایه‌ی برون‌داد هستند، ابتدا آموزش می‌بینند و پس از آموزش می‌توان از آنها در پردازش داده‌های جدید استفاده کرد. هنگام آموزش، واحدهای لایه‌ی درون‌داد، اطلاعات ورودی را دریافت می‌کنند و پس از پردازش و بازنمایی الگو، این اطلاعات به لایه‌ی برون‌داد فرستاده می‌شود. در لایه‌ی برون‌داد نتیجه‌ی به‌دست آمده و نتیجه‌ی مطلوب با یکدیگر

²¹ neural networks

²² naive Bayes classifier

شناسایی هویت نویسنده بر اساس زبان فردی □ ۱۷

مقایسه می‌شوند و میزان خطا مجدداً طی فرایندی به نام پس‌انتشار خطا^{۲۳} به لایه‌ی قبل ارسال می‌شود. این چرخه همچنان ادامه می‌یابد تا خطای سیستم کاهش و سیستم بتواند نتیجه‌ی مطلوب را به دست دهد.

شبکه‌های عصبی در شناسایی نویسنده‌ی متون عملکرد موفقیت‌آمیزی داشته‌اند. برای نمونه، در پژوهشی که در آن با استفاده از شبکه‌های عصبی مقالات فدرالیست (Mosteller and Wallace, 1964) مجدداً مورد تحلیل قرار گرفته (Tweedie et al., 1996)، عملکرد این روش با عملکرد روش پژوهش اولیه در باب مقالات فدرالیست مقایسه شده و نتایج آن مطلوب ارزیابی شده است. در زبان فارسی نیز از این روش برای شناسایی مولف استفاده شده (Farahmandpour nad Nikmehr, 2015) که عملکرد آن مطلوب ارزیابی شده اما بهترین روش در میان روش‌های به کار گرفته شده توسط پژوهشگران نبوده است. با وجود دقتی که شبکه‌های عصبی در دسته‌بندی داده‌ها و بازشناسایی الگو دارند، اما عملکرد آنها عمدتاً به جعبه‌ای سیاه تشبیه می‌شود که فرآیند حل مسئله در آن چندان روشن نیست.

۲-۳-۲. دسته‌بندی کننده‌ی نایو بیز

دسته‌بندی کننده‌های نایو بیز به مجموعه‌ای از دسته‌بندی کننده‌های احتمالاتی گفته می‌شود که بر اساس قانون بیز عمل می‌کنند، با این تفاوت که قائل به استقلال رخدادی بین مولفه‌های مورد دسته‌بندی هستند. قانون بیز یک احتمال شرطی است که احتمال رخداد یک رویداد را بر اساس رویدادی دیگر تخمین می‌زند و در حالت نایو، احتمالات شرطی جزئی به صورت مستقل از یکدیگر برای مولفه‌های رویداد تشکیل شده و این احتمالات در یکدیگر ضرب می‌شوند. در باب شناسایی نویسنده‌ی یک متن نیز چنین فرض می‌شود که رخداد مشخصه‌های متنی مستقل از یکدیگر هستند. بر اساس این فرض، هدف پژوهشگر یافتن مولف m به شرط رخداد مجموعه مشخصه‌های $\{a_1, a_2, \dots, a_n\}$ است که از متن استخراج شده‌اند. فرمول محاسبه‌ی این احتمال شرطی به صورت زیر خواهد بود:

$$P(m|a) = P(a_1|m) \times P(a_2|m) \times \dots \times P(a_n|m) \times P(m)$$

²³ error backpropogation

شناسایی هویت نویسنده بر اساس زبان فردی □ ۱۸

در فرمول فوق، محاسبه‌ی احتمال مولف m به شرط مشخصه‌ی a در قالب حاصل ضرب احتمالات جزئی رخداد هر یک از مشخصه‌های متنی $a_1, a_2, \dots, a_n$ به شرط مولف m ضرب در احتمال رخداد مولف m محاسبه می‌شود.

الگوریتم نایو بیز در مطالعات شناسایی مولف و طبقه‌بندی متون عملکرد خوبی داشته است. برای نمونه، از این الگوریتم در شناسایی مولف بر اساس واژه‌های دستوری (Zhao and Zobel, 2005) در زبان انگلیسی استفاده شده است. پژوهشگران در آزمایش‌های خود که بر روی متونی با حجم‌های مختلف انجام شده‌اند، نشان می‌دهند که الگوریتم‌های احتمالاتی بیز عملکرد بالاتری داشته‌اند؛ به‌طور مشخص، الگوریتم نایو بیز بهتر از الگوریتم‌های دیگر نظیر نزدیک‌ترین همسایه^{۲۴} و درخت تصمیم^{۲۵} عمل کرده اما عملکرد آن پایین‌تر از الگوریتم شبکه‌های بیزی (نوع دیگری از الگوریتم بیز) بوده است.

۳-۲-۳-۲. روش دلتا

روش دلتا، روشی است که توسط باروز (Burrows, 2002) برای شناسایی نویسنده معرفی شده است. این روش مبنای آماری داشته و بر اساس محاسبه‌ی نمره‌ی معیار (Z-score) عمل می‌کند. عملکرد این روش بدین صورت است که ابتدا واژه‌های پربسامد متونی از یک نویسنده‌ی مشخص استخراج می‌شود (مثلاً ۱۰۰ واژه). سپس برای هر یک از واژه‌ها، توزیع معیار محاسبه می‌شود. منظور از توزیع معیار، بسامد میانگین آن کلمه به همراه پراکنش آن است. در مرحله‌ی بعد، بسامد واژه‌های هر یک از نوشته‌هایی که قرار است نویسنده‌ی آنها شناسایی شود، در مقایسه با ۱۰۰ واژه‌ی استخراج‌شده بر اساس نمره‌ی معیار امتیازدهی می‌شوند تا معلوم شود بسامد واژه‌های این نوشته‌ها به چه میزانی بالا یا پایین از حد نرمال است. نمره‌ی معیار مثبت نشانگر فراوانی بیش از میانگین، نمره‌ی معیار منفی نشانگر فراوانی کمتر از میانگین و نمره‌ی معیار صفر نشانگر فراوانی میانگین آن واژه است. مقیاس دلتا در واقع میانگین قدرمطلق تفاضل نمرات معیار مشخصه‌های مجموعه متون اصلی (متونی که الگوریتم بر اساس آن آموزش می‌بیند) و نمرات معیار همان مشخصه‌ها در متن هدف (فاقد هویت) است. هر چه مقدار دلتا کوچکتر باشد، نشانگر شباهت بیشتر متن هدف به متن اصلی است.

²⁴ nearest neighbor

²⁵ decision tree

روش دلتا در میان روش‌های شناسایی نویسنده کم‌رقیب است. دقت و کارآمدی این روش در شناسایی نویسنده‌ی متون نشر و نظم موفقیت‌آمیز ارزیابی شده است (Hoover, 2004). در پژوهشی که به آزمون عملکرد روش دلتا به همراه روش‌های دیگر و برگرفته از آن در سه زبان انگلیسی، آلمانی و فرانسه پرداخته، عملکرد روش دلتا بسیار قابل قبول توصیف شده است (Jannidis et al., 2015). این روش به دلیل داشتن منطقی ساده و شهودی، می‌تواند روش جذابی برای پیاده‌سازی و آزمون در زبان‌های دیگر محسوب تلقی گردد.

۲-۳. زبان فردی به مثابه‌ی اثر انگشت گویشور

زبان فردی^{۲۶} مفهومی آشنا در حوزه‌ی مطالعات زبان‌شناختی است. اگر بخواهیم تعریفی ساده از این مفهوم به دست دهیم، «زبان فردی شیوه‌ی منحصر بفرد [هر فرد در] صحبت کردن، شامل به کارگیری صداها، کلمات، دستور زبان و سبک زبانی است» (Wardhaugh and Fuller, 2015: 9). بنابراین اگر بخواهیم از استعاره‌ی اثر انگشت استفاده کنیم، می‌توان این فرضیه را مطرح کرد که زبان فردی هر کس از آنجا که منحصر به خود اوست، حکم اثر انگشت زبانی‌اش را خواهد داشت. پیامد این فرضیه در حوزه‌ی شناسایی نویسنده‌ی یک متن این خواهد بود که اگر بتوان مشخصه‌های سبکی و مبتنی بر زبان فردی متون نگاشته شده توسط یک نویسنده را استخراج کرد، می‌توان با استخراج مشخصه‌های سبکی یک متن فاقد هویت و مقایسه‌ی آن را سبک‌های استخراج شده از متون دارای هویت، مشخص کرد که سبک متن فاقد هویت به سبک کدام نویسنده نزدیک‌تر است.

پیش از آنکه به شواهد موجود در تایید زبان فردی بپردازیم، این نکته را باید مدنظر داشت که آنچه از تعریف فوق برمی‌آید این است که زبان فردی بیشتر در حوزه‌ی گفتاری نمود می‌یابد. به همین دلیل این مفهوم اکثراً به «گویش فردی» در فارسی ترجمه می‌شود. اما بر اساس شواهدی که در ادامه آنها را مرور خواهیم کرد، نشان داده خواهد شد که مفهوم idiolect هم در زبان گفتاری و هم در زبان نوشتاری وجود دارد. از این رو در این پژوهش به جای «گویش فردی» از معادل «زبان فردی» که جامع‌تر است استفاده شده است.

۲-۳-۱. زبان فردی در گفتار

²⁶ idiolect

شناسایی هویت نویسنده بر اساس زبان فردی □ ۲۰

تحقیقات بسیاری درباره‌ی سبک منحصر بفرد افراد در استعمال زبان انجام شده است. در اینجا به سه نمونه اشاره خواهد شد که مورد نخست به دلیل بُعد قانونی و اهمیت تاریخی آن و دو مورد بعدی نیز به دلیل استفاده از روش‌های پیکره‌ای نوین حائز اهمیت است. نکته‌ی قابل توجه دیگر در باب این مطالعات این است که گفتارهای مورد تحلیل، صحبت‌های افرادی متعارف جامعه هستند و همانند نویسندگان هیچگونه سبک ادبی متمایزی برای آنها قابل تصور نیست.

مورد نخست مربوط به پرونده‌ی قتل منتسب به نوجوانی ۱۹ ساله به نام درِک بنتلی^{۲۷} است که در سال ۱۹۵۲ به اتهام مشارکت در قتل یک مامور پلیس در بریتانیا محاکمه و اعدام شد. ماجرا از این قرار بود که بنتلی و همدستش کریس کریگ (در آن هنگام ۱۶ سال داشت) به قصد دستبرد از دیوار یک انبار بالا می‌روند و وارد پشت‌بام انبار می‌شوند. زنی با مشاهده‌ی این صحنه با پلیس تماس گرفته و ماموران پلیس در محل حاضر می‌شوند. بنتلی خود را تسلیم پلیس می‌کند اما کریگ مقاومت کرده و مامور پلیس را با شلیک گلوله از پای درمی‌آورد. بنتلی که گفته می‌شود به معلولیت ذهنی نیز دچار بوده و هنگام کشته‌شدن پلیس در بازداشت وی بوده، متهم به مشارکت در قتل می‌شود. با وجود اینکه اتهاماتش را در دادگاه نمی‌پذیرد و ادعا می‌کند اعترافات ارائه‌شده به دادگاه گفته‌های او نیست، محاکمه و اعدام می‌شود. در حرکتی نمادین، بنتلی ۴۰ سال پس از مرگش در سال ۱۹۹۳ مورد عفو سلطنتی واقع شده و در سال ۱۹۹۸ حکم اعدام وی ابطال می‌شود.

سی و شش سال پس از مرگ بنتلی، کولتھارد (Coulthard, 1995, 1998) شواهدی زبان‌شناختی را به دادگاه ارائه می‌دهد که نشان می‌دهد اعترافات منسوب به بنتلی، با مشارکت و دستکاری پلیس نگاشته شده است. یکی از شواهد، بسامد بالا و غیرعادی کلمه‌ی then (برای بیان توالی رخدادی) و همچنین ساختار دستوری این واژه در متن اعترافات منسوب به بنتلی است. کولتھارد دو پیکره تهیه کرد که در یکی گزارشات شاهدان دیگر و در پیکره‌ی دیگر گزارشات سه مامور پلیس را گردآوری کرد. با مقایسه‌ی این دو پیکره متوجه شد در پیکره‌ی شاهدان، کلمه‌ی then یک بار در هر ۹۳۰ کلمه و در پیکره‌ی ماموران پلیس یک بار در هر ۷۸ کلمه به کار رفته است. تعداد کلمه‌ی then در اظهارات بنتلی (یک بار در هر ۵۸ کلمه) بسیار نزدیک به پیکره‌ی اظهارات پلیس بود. کولتھارد همچنین در بررسی ساختار دستوری کلمه‌ی then متوجه شد که این کلمه در متن اظهارات بنتلی به صورت پس‌اضافه (مثلاً به شکل

²⁷ Derek Bentley

then I ran after the... به جای then I ran after them... به کار رفته است. وی با مقایسه متن بنتلی با پیکره‌ی گفتاری کوبیلد^{۲۸} دریافت که در کل این پیکره توالی I then با بسامد بسیار کم و به تعداد ۹ بار در هر ۱۶۵۰۰۰ کلمه (یک بار در هر ۱۸۳۳۳ کلمه) به کار رفته است، در حالی که در متن بنتلی این توالی یک بار در هر ۱۹۴ کلمه به کار رفته است. در پیکره‌ی اظهارات پلیس نیز این توالی یک بار در هر ۱۱۹ کلمه به کار رفته بود که نشانگر ردپای سبک منحصر به فرد زبان پلیس در متن اعتراف منتسب به بنتلی است.

پژوهش‌های اخیر وجود زبان فردی در گفتمان سیاسی را نیز نشان داده است. مثلاً در پژوهشی که به تحلیل گفتار تونی بلر (نخست‌وزیر اسبق بریتانیا) پرداخته (Mollin, 2009)، نشان داده شده که سخنوران مختلف تمایلات متمایزکننده‌ای در به کارگیری قیده‌های بیشینه‌ساز (fully و entirely و absolutely در انگلیسی) دارند. در این پژوهش پیکره‌ای ۳ میلیون کلمه‌ای از سخنرانی‌های تونی بلر جمع‌آوری شده و باهم‌آیی‌های مربوط به قیده‌های بیشینه‌ساز از آن استخراج شده است. در نهایت با استفاده از مقیاس‌های آماری و رتبه‌دهی قیده‌های بیشینه‌ساز، خاص بودن یا متعارف بودن این عبارات با پیکره‌ی ملی بریتانیا مقایسه شده است. تحلیل‌های بعدی نشان می‌دهند که از میان ۲۵ باهم‌آیی موثر انتخاب شده، ۱۶ مورد به‌طور خاص مربوط به زبان فردی تونی بلر می‌شود و بقیه‌ی موارد یا در گفتار عمومی متداول هستند، یا ویژه‌ی سبک روزنامه‌ای هستند، و یا خاص سبک گفتار پارلمانی‌اند. نتایج این پژوهش به‌روشنی نشان می‌دهد که زبان فردی اشخاص در قالب ترجیحات و ازگانی متفاوتی نمود می‌یابد و این ترجیحات متفاوت نیز خود می‌تواند تابعی از گونه‌ی کاربردی بخصوصی باشد.

پژوهش دیگری بر روی گفتار پنج سخنگوی خبری کاخ سفید انجام شده (Barlow, 2010) که شواهد محکمی از نقش پررنگ ساخت‌های دستوری در زبان فردی به‌دست می‌دهد. فرضیه‌ی مورد آزمون این پژوهش این است که تنوع در زبان فردی گویشوران در قالب مولفه‌های اصلی واژگانی دستوری نمود می‌یابد و محدود به استعمال برخی واژه‌ها و عبارات خاص و منحصر به فرد نمی‌شود. برای آزمون فرضیه‌ی پژوهش، گفتار پنج سخنگوی خبری کاخ سفید که پیکره‌ای مشتمل بر ۳ میلیون کلمه را می‌سازند، به‌صورت فایل‌های نوشتاری ۲۰۰۰۰۰ کلمه‌ای آماده شدند. سپس کلمات موجود در متون هم‌آیند یابی شدند تا جفت‌های دونگاشت (توالی‌های دو کلمه‌ای) استخراج

شوند. بعد ۱۰ دونگاشت پرسامد از متون نمونه استخراج شدند و پس از تلفیق با یکدیگر و حذف دونگاشت‌های تکراری، در نهایت ۱۵ دونگاشت باقی ماندند که نمایانگر دو نداشت‌های پرسامد متون پیکره بودند. در مرحله‌ی بعد، متون سخنگوها بر اساس این ۱۵ دونگاشت پرسامد نمرده‌دهی شدند، به این ترتیب که رتبه‌ی هر کدام از دونگاشت‌ها (اینکه یک دونگاشت خاص چندمین دونگاشت پرسامد در متن مربوطه است) در متون معین شد. یافته‌های اولیه‌ی این پژوهش نشان می‌دهد که الگوی رتبه‌بندی دونگاشت‌ها برای سخنگوهای مختلف متفاوت و در گفتارهای مختلف یک سخنگوی واحد مشابه است. پس از تحلیل دونگاشت‌ها و حتی سه‌نگاشت‌ها، تحلیل برچسب‌های نحوی نیز نشان داد که گویشوران در به‌کارگیری ساخت‌های نحوی نیز تمایلات متفاوتی دارند. مهم‌ترین نتیجه‌ای که از این پژوهش به‌دست داده شده این است که تفاوت‌های موجود در زبان فردی صرفاً محدود به استفاده‌ی افراد از کلمات منحصر‌به‌فرد و کلیشه‌ای نمی‌شود، بلکه ساختارهای نحوی را نیز در بر می‌گیرد.

۲-۳-۲. زبان فردی در نوشتار

از آنجا که مفهوم زبان فردی (گویش فردی) در زبان‌شناسی عمدتاً برای زبان گفتاری به‌کار می‌رود، پژوهش‌های سبک‌سنجی نشان داده‌اند که این مفهوم در زبان نوشتار نیز صادق است. در این بخش به دو پژوهش اشاره خواهد شد که رد پای زبان فردی را در زبان نوشتاری نشان می‌دهد.

در یکی از پژوهش‌های اخیر (Johnson and Wright, 2014) که بر روی شناسایی نویسنده‌ی رایانامه‌های شرکت انرژی انرون انجام گرفته، محققان با استفاده از توالی‌های واژگانی n-نگاشتی، سعی می‌کنند مصادیقی از زبان فردی را به‌دست دهند. در این پژوهش پیکره‌ای ۲/۵ میلیون کلمه‌ای متشکل از ۶۳۰۰۰ ایمیل که توسط ۱۷۶ کارمند شرکت آمریکایی انرون نوشته شده، آماده شده و به‌لحاظ سبکی مورد تحلیل واقع می‌شوند. نتایج تحلیل نشان می‌دهد که هر یک از کارمندان شرکت الگوهای سبکی متمایز از الگوی دیگران را در نگارش خود به‌کار می‌گیرند. در ادامه‌ی تحلیل‌ها پژوهشگران همچنین با استفاده از یک آزمایش آماری نشان می‌دهند که با استفاده از n-نگاشت‌ها می‌توان نویسندگان رایانامه‌های مجهول‌الهویت را با دقت ۱۰۰ درصد شناسایی کرد.

در پژوهشی دیگر (Panicheva et al., 2010) از مفهوم زبان فردی در تحلیل عواطف^{۲۹} استفاده شده است. تحلیل عواطف فرآیندی است که با استفاده از تحلیل رایانشی متن مشخص می‌کند، نویسنده‌ی یک متن یا نظر، چه نگرشی (مثبت، منفی، خنثی) به یک محصول یا ایده داشته است. در پژوهش مذکور، محققان با تلفیق مفهوم زبان فردی و حوزه‌ی تحلیل عواطف، رویکرد نوینی را به تحلیل عواطف یک متن اتخاذ کرده‌اند. در واقع فرضیه‌ی محققان این بوده است که نویسندگانی که نظرات و احساساتشان را در متنی نسبت به یک محصول یا ایده بیان می‌کنند به گونه‌ای منحصر به فرد و متفاوت از دیگران این کار را انجام می‌دهند و این تفاوت در زبان فردی آنها مستتر است. بنابراین با شناسایی نویسنده‌ی متن می‌توان تا حد زیادی فرآیند تشخیص بار عاطفی نظر ثبت شده را تسهیل کرد.

پژوهش‌های انجام شده در حوزه‌ی شناسایی نویسنده در زبان نوشتاری عمدتاً با رویکرد مهندسی به موضوع نزدیک شده‌اند و مسئله را چندان از دید زبان فردی بررسی نکرده‌اند. یکی از پیش فرض‌هایی که در مورد زبان فردی در این فصل مطرح شد، تحقیقاتی بود که نشان می‌داد زبان فردی بیش از اینکه به کاربرد منحصر به فرد و عادت‌وار کلمات توسط افراد محدود شود، ریشه در ساختارهای نحوی و روابط واژگی-دستوری دارد. در این پژوهش قصد بر این است که با رویکردی مبتنی بر زبان فردی، تحلیلی پیکره‌ای از سبک نویسندگان مختلف در زبان فارسی به دست داده شود و قابلیت تفکیک سبک نویسنده‌ها بر اساس واژه‌های دستوری بررسی شود. واژه‌های دستوری به دلیل اینکه مجموعه‌ای بسته از عناصر زبانی را تشکیل می‌دهند، مستقل از محتوا و موضوع متن به کار گرفته می‌شوند، و سرنخ‌هایی از ساختارهای کلان نحوی به دست می‌دهند، حائز اهمیت هستند. در فصل بعد، توضیحات مربوط به روش شناسی پژوهش، پیکره و الگوریتم‌های مورد استفاده ارائه شده است.

²⁹ sentiment analysis

فصل سوم

روش پژوهش

۱-۳. مقدمه

این پژوهش به صورت توصیفی-تحلیلی انجام شده است. در این فصل روش‌شناسی پژوهش شامل نشانگرهای سبکی، نحوه‌ی گردآوری پیکره‌ی پژوهش، روش‌های مورد استفاده در تحلیل سبک نویسنده، نرم‌افزار جستجوی پیکره و روال انجام پژوهش شرح داده خواهند شد.

۲-۳. نشانگرهای سبکی

نشانگر سبکی مورد استفاده در این پژوهش، نشانگرهای واژگانی و از نوع واژه‌های دستوری است. دلیل انتخاب نشانگرهای واژگانی (و نه مثلاً نشانگرهای کاراکتری) در این پژوهش، هم‌راستایی با بنیادی‌ترین تلقی معناداری در زبان‌شناسی است چرا که واژه‌ها بنیادی‌ترین واحدهای معنادار قابل تشخیص در نوشتار را تشکیل می‌دهند. همچنین واژه‌ها به عنوان نشانگر سبکی، در مقابل نشانگرهای کاراکتری، نقش معنادارتری در یافتن الگوهای مبتنی بر زبان فردی گویشوران ایفا می‌کنند.

واژه‌ها طبق تقسیم‌بندی‌های زبان‌شناختی به دو دسته‌ی عمده‌ی واژه‌های محتوایی و واژه‌های دستوری تقسیم می‌شوند. واژه‌های محتوایی تمام واژه‌هایی را در می‌گیرند که محتوای معنایی پررنگ‌تری دارند، تعدادشان نامحدود است و حجم عمده‌ی فرهنگ‌های لغت را به خود اختصاص می‌دهند. واژه‌های دستوری (شامل حروف اضافه، حروف ربط، غیره)، در مقابل، محتوای معنایی پررنگی ندارند، تعدادشان در هر زبان محدود است و نقشی دستوری در جملات ایفا می‌کنند. در این پژوهش، به دلایل زیر نشانگرهای واژگانی از مقوله‌ی واژه‌های دستوری انتخاب شده‌اند:

- واژه‌های دستوری بسامد بالایی نسبت به واژه‌های محتوایی در متون دارند. این ویژگی امتیاز بزرگی در حوزه‌ی تشخیص سبک نویسنده محسوب می‌شود، چرا که در دنیای واقعی و در موقعیت‌های کاربردی، حجم متون ایده‌آل نیست و گاهی محقق ناچار است با داده‌های اندکی که در اختیار دارد، سبک نویسنده‌ی متن را مشخص یا نویسنده را شناسایی کند.
- واژه‌های دستوری اساساً مستقل از عنوان و موضوع متن به کار برده می‌شوند. از آنجا که واژه‌های دستوری نقش دستوری در جمله ایفا می‌کنند، مورد استعمال آن‌ها متکی به موضوع متن نیست. به همین دلیل مزیت مهم این واژه این است که متون مورد تحلیل در شرایط آزمایشگاهی نیازی به داشتن موضوع یکسان ندارند که بتوان آنها را با یکدیگر مقایسه کرد.

- واژه‌های دستوری به‌طور ناخودآگاه به‌کار گرفته می‌شوند و نویسنده برخلاف واژه‌های محتوایی کنترل خودآگاه بر استفاده یا استفاده نکردن از آنها ندارد.

تعداد نشانگرهای دستوری مورد استفاده در این پژوهش که بر اساس آن‌ها سبک نویسنده‌ها با یکدیگر مقایسه و متمایز می‌شدند، ۲۰ واژه‌ی دستوری پربسامد تعیین شد.

۳-۳. پیکره‌ی پژوهش

۱-۳-۳. گردآوری پیکره

پیکره‌ی مورد استفاده در این پژوهش، از مجموعه مقالات نوشتاری پنج نویسنده که به‌صورت الکترونیکی در دسترس بوده‌اند، گردآوری شده است. این متون از مقالات قابل دسترس ناصر فکوهی و نیوشا صدر در وبگاه اینترنتی «انسان‌شناسی و فرهنگ»^{۳۰}، وبگاه شخصی صادق زیباکلام^{۳۱}، وبگاه شخصی رضا داوری اردکانی^{۳۲}، و وبگاه شخصی حمید پارسانیا^{۳۳} جمع‌آوری شده‌اند. در جمع‌آوری پیکره، ملاحظات زیر مدنظر قرار گرفت:

- سعی بر آن بود موضوعاتی که نویسندگان به آنها پرداخته‌اند تا حد امکان مشابه انتخاب شوند یا حداقل در یک حوزه (در اینجا علوم انسانی) باشند. هرچند استفاده از نشانگرهای واژه‌های دستوری در این پژوهش که مستقل از موضوع به‌کارگرفته می‌شوند، ضرورت تشابه متون را برطرف می‌کند، اما این نکته برای بالابردن روایی پژوهش مدنظر بوده است.
- در جمع‌آوری متون بر این نکته بسیار تاکید شد که اطمینان حاصل شود متون به قلم خود نویسنده نوشته شده باشد. از آنجا که این پژوهش، به‌صورت آزمایشگاهی انجام می‌شود و هدف از آن آزمودن کارایی واژه‌های دستوری در متمایزکردن سبک نوشته‌های فارسی است، اطمینان از اصالت نوشته‌ها بسیار حائز اهمیت است. این دغدغه موجب شد نتوان از مقالات یا رساله‌های تحصیلات تکمیلی به‌عنوان پیکره‌های قابل دسترس در ایرانداک استفاده شود، چرا که

³⁰ www.anthropology.ir

³¹ www.zibakalam.com

³² www.rezadavari.ir

³³ www.parsania.ir

اصول ارجاع‌دهی در بسیاری موارد در نگارش این نوشتجات رعایت نمی‌شود. در نهایت تصمیم بر این شد که مقالات موجود در وبلاگ‌های شخصی برخی محققان و اندیشمندان، می‌تواند نسبت به سایر منابع از اصالت بیشتری برخوردار باشد.

مقالات پس از گردآوری به‌طور مجزا برای هر نویسنده در فایل‌های متنی ذخیره شدند.

۲-۳-۳. پیش‌پردازش

قبل از اینکه مطالب گردآوری‌شده در آزمایش‌های سبک‌سنجی استفاده شوند، لازم بود که برخی عملیات پیش‌پردازشی بر روی آنها انجام شود. در ابتدا هر یک از فایل‌ها به‌صورت دستی مرور شدند تا برخی مطالب غیرمرتبط با پژوهش پالایش شوند. مثلاً نام نویسندگان مقالات از ابتدای آنها حذف شد. همچنین در مواردی مشاهده شد که متن مقاله به روایت شخص دیگری غیر از خود نویسنده است که این مقالات کنار گذاشته شدند. همچنین سعی شد مقالاتی که به‌طور فراوان از منابع دیگر نقل قول مستقیم کرده بودند، از پیکره پژوهش کنار گذاشته شوند.

پس از پیش‌پردازش دستی اولیه، کل پیکره به‌طور خودکار توسط نرم‌افزار خطایاب وفا^{۳۴} (Faili et al., 2016) استانداردسازی شد. این خطایاب به‌صورت افزونه‌ای در نرم‌افزار مایکروسافت ورد نصب می‌شود و متن را به‌لحاظ یکدست‌سازی حروف (جایگزینی حروف عربی با فارسی)، رعایت نیم‌فاصله‌ها، نحوی و معنایی خطایابی می‌کند. مهم‌ترین انگیزه از به‌کارگیری نرم‌افزار خطایاب در این پژوهش، یکدست‌سازی نگارش و کشف خطاهای املایی در متون بوده است.

پس از مرحله‌ی پیش‌پردازش، پیکره آماده‌ی استفاده است. تعداد کلمات پیکره‌ی اصلی پژوهش به تفکیک نویسندگان آن به‌صورت جدول ۳-۱ است.

جدول ۳-۱. تعداد کلمات موجود در پیکره به تفکیک نویسندگان

نویسنده	تعداد کلمات پیکره
صادق زیباکلام	۱۰۸۶۹۰
حمید پارسا‌نیا	۱۰۷۴۹۹
ناصر فکوهی	۱۰۷۱۰۷

35 stylo

شکل ۳-۱. برخی امکانات رابط کاربردی در بسته‌ی استایلو (Eder et al., 2013)

۳-۵. نرم‌افزار جستجوی پیکره

در این پژوهش از نرم‌افزار AntConc (Anthony, 2011) برای استخراج تحلیل بافتی واژه‌های دستوری استفاده شده است. این نرم‌افزار قادر است پیکره‌های بزرگی را در قالب فایل‌های متنی پردازش کند. برخی از مهم‌ترین قابلیت‌های این نرم‌افزار عبارتند از:

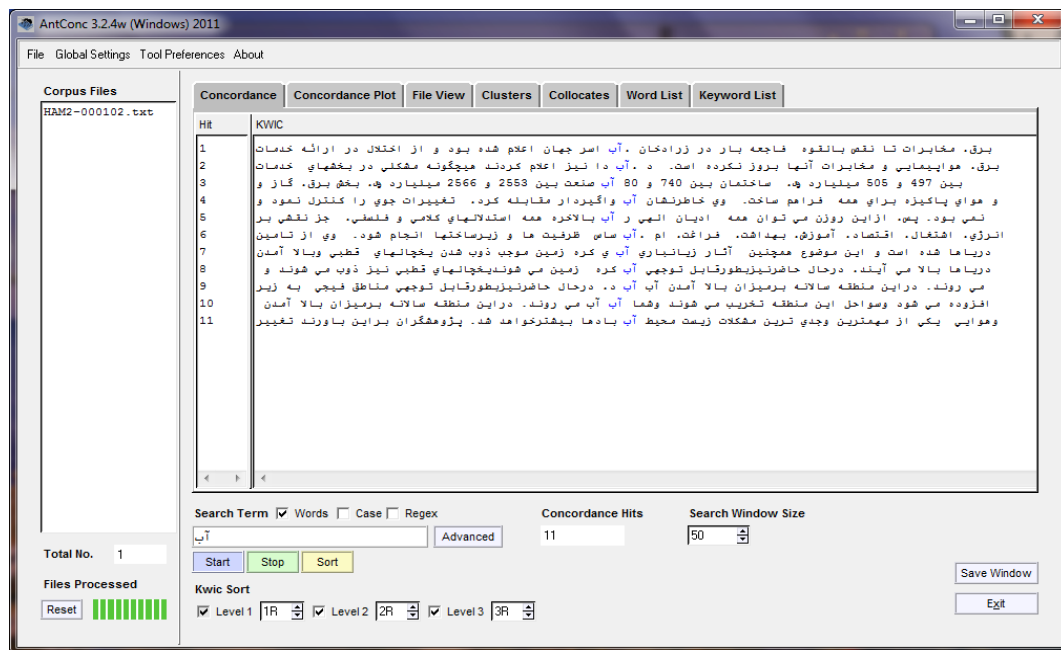
- پشتیبانی از استاندارد یونی‌کد برای پردازش خط فارسی.
- قابلیت جستجوی کلیدواژه‌های موردنظر به صورت فهرست واژه‌نما. در فهرست واژه‌نما، قابلیت مرتب‌سازی الفبایی همسایگی‌های راست و چپ کلیدواژه وجود دارد.
- قابلیت جستجوی یک کلیدواژه در بافت واژگانی کلمه‌ای دیگر، با امکان تعیین بازه‌ی جستجو نسبت به چپ و راست کلیدواژه.
- استخراج باهم‌آیی‌های واژگانی با استفاده از دو مقیاس آماری نمره‌تی^{۳۶} و اطلاعات متقابل^{۳۷}. باهم‌آیی‌ها، واژه‌هایی هستند که هم‌وقوعی آنها تصادفی نیست. مثلاً دو کلمه «نان» و «پنیر» در عبارت «نان و پنیر» هم‌آیند یکدیگر محسوب می‌شوند و یک کل معنایی را می‌سازند.
- قابلیت محاسبه‌ی بسامد واژه‌های پیکره و مرتب‌سازی آنها بر اساس بسامد.
- قابلیت استخراج خوشه‌های واژگانی و عبارات n-نگاشتی^{۳۸} از پیکره.

^{۳۶} T-score

^{۳۷} Mutual Information (MI)

^{۳۸} N-gram

شناسایی هویت نویسنده بر اساس زبان فردی □۳۰



شکل ۳-۲ محیط نرم افزار پیکره‌ای AntConc

شکل ۳-۲ محیط نرم افزار AntConc را نشان می دهد که در آن فهرست واژه‌نما برای کلیدواژه «آب» نشان داده شده است.

۳-۶. روال انجام پژوهش

این پژوهش در قالب ۶ آزمایش انجام شده و هر آزمایش با هدف بررسی کارایی واژه‌های دستوری تک‌نگاشتی، دونگاشتی و سه‌نگاشتی انجام شده است. همچنین در هر آزمایش مقدار متغیری از تعداد واژه‌ها به عنوان پیکره از کل پیکره‌ی پژوهش نمونه‌گیری شده‌اند. جدول ۳-۲ میزان حجم پیکره برای هر آزمایش به همراه نویسندگان نمونه متن‌ها نشان می‌دهد.

جدول ۳-۲. حجم تقریبی پیکره مورد استفاده در هر آزمایش

آزمایش	حجم پیکره برای هر نمونه متن	نویسنده متون
۱	۵۰۰۰۰	فکوهی، پارسانیا، زیباکلام
۲	۱۰۰۰۰	فکوهی، پارسانیا، زیباکلام، صدر، اردکانی
۳	۱۰۰۰	فکوهی، پارسانیا، زیباکلام، صدر، اردکانی
۴	۲۰۰۰	فکوهی، پارسانیا، زیباکلام، صدر، اردکانی

۵	۳۰۰۰	فکوهی، پارسانی، زیباکلام، صدر، اردکانی
۶	۴۰۰۰	فکوهی، پارسانی، زیباکلام، صدر، اردکانی

روال انجام آزمایش‌ها به این صورت بوده که ابتدا در آزمایش اول متون ۵۰۰۰۰ کلمه‌ای از نویسندگان تحلیل شدند. برای این منظور، متون سه نویسنده دارای تعداد واژه‌ی کافی یعنی ۱۰۰۰۰۰ واژه بودند که برای هر نویسنده دو پیکره‌ی ۵۰۰۰۰ کلمه‌ای به‌طور تصادفی نمونه‌گیری شدند. در آزمایش دوم، حجم نمونه‌متن‌ها به ۱۰۰۰۰ کاهش داده شد و متون همه‌ی پنج نویسنده در تحلیل قرار گرفت. در آزمایش سوم، حجم نمونه‌متن‌ها به ۱۰۰۰ کاهش یافت. هدف از اعمال این کاهش‌ها ارزیابی عملکرد واژه‌های دستوری در تمایز سبک نویسندگان در حجم داده‌های مختلف است. در سطح ۱۰۰۰ واژه چون عملکرد تشخیص نویسندگان دچار افت دقت شد، در هر یک از آزمایش‌های بعدی ۱۰۰۰ واژه به حجم متون اضافه شد تا حداقل شناسایی نویسندگان در سطح واژه‌های دستوری تک‌نگاشتی با دقت ۱۰۰ درصد صورت گیرد.

پس از مشخص شدن پیکره‌های تحلیل، اولین گام در تحلیل سبک‌های نویسندگان، استخراج واژه‌های دستوری پربسامد بود که نقش نشانگر واژگانی را در متمایز کردن سبک نویسندگان ایفا می‌کردند. واژه‌های دستوری پربسامد با استفاده از فرامین موجود در بسته‌ی استایلو استخراج و سپس به‌صورت دستی بررسی شدند تا واژه‌های غیردستوری از آن پالایش شوند. واژه‌های غیردستوری که از فهرست واژه‌های پربسامد حذف شدند عمدتاً شامل مواردی می‌شدند که در جدول ۳-۳ آمده است.

جدول ۳-۳. فهرست واژه‌های محذوف از جدول واژه‌های پربسامد

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

در طی فرآیند شمارش خودکار واژه‌های پربسامد و قبل از حذف واژه‌های غیردستوری، گزینش^{۳۹} درصدی بر روی واژه‌های پربسامد اعمال شد. طبق این گزینش، واژه‌هایی که تا درصد خاصی در همه‌ی متون به کار نرفته بودند، کنار گذاشته می‌شدند. حد کمینه‌ی این گزینش ۲۰٪ و حد بیشینه‌ی آن ۸۰٪ در نظر گرفته شد. گزینش واژه‌های پربسامد با حد ۲۰٪ بدین معناست که واژه‌هایی که در حداقل ۲۰٪ از متن‌ها به کار رفته باشند، در تحلیل لحاظ خواهند شد.

پس از محاسبه‌ی تعداد واژه‌های پربسامد و حذف واژه‌های غیردستوری، جدول حاصله آماده‌ی تحلیل سبک‌سنجی است. در مرحله‌ی بعد، نحوه‌ی تمایز نویسندگان و سهم هر کدام واژه‌ها در تمایز سبک‌ها به واسطه‌ی تحلیل مولفه‌های اصلی بررسی شدند. سپس با استفاده از تحلیل خوشه‌ای و بر اساس مقیاس فاصله‌ای دلتا، خوشه‌بندی متون مشابه به لحاظ سبکی به دست داده شد.

برای ارزیابی دقت خوشه‌بندی، از شاخصی به نام شاخص تعدیل‌شده‌ی رند^{۴۰} (Hubert and Arabie, 1985) استفاده شد. این شاخص، نسخه‌ی بهبودیافته‌ی شاخص رند (Rand, 1971) است که خوشه‌بندی تولیدشده توسط الگوریتم را با خوشه‌بندی ایده‌آل (که آثار هر نویسنده باید در خوشه‌های مجزا قرار گیرند) مقایسه می‌کند و میزان مطابقت آن را به دست می‌دهد. اگر U و V را دو طبقه‌بندی متفاوت از اشیاء مشخص در نظر بگیریم، شاخص رند را می‌توان طبق فرمول زیر برای مقایسه‌ی این دو طبقه‌بندی محاسبه کرد:

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

- 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.

³⁹ culling

⁴⁰ Adjusted Rand Index

فصل چهارم

تحليل داده‌ها

۱-۴. مقدمه

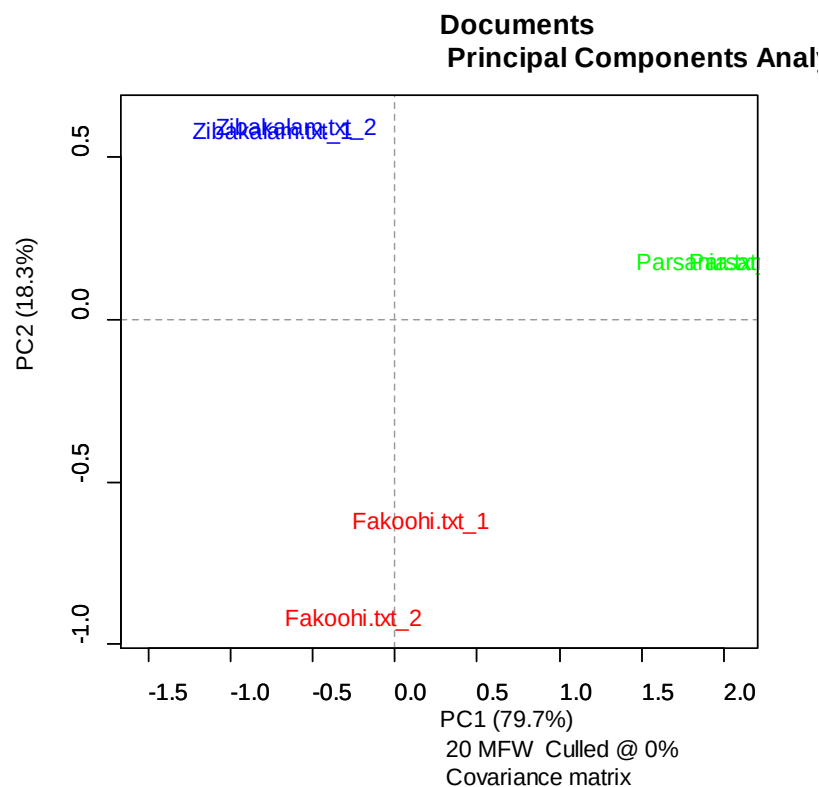
این فصل به تحلیل متون نویسنده‌ها بر اساس واژه‌های دستوری پربسامد می‌پردازد و داده‌ها را بر اساس دو تحلیل آماری یعنی تحلیل مولفه‌های اصلی و تحلیل خوشه‌ای تجزیه و تحلیل می‌کند. تحلیل‌های انجام‌شده به ۶ آزمایش و سپس تحلیل‌های پیکره‌ای تقسیم شده‌اند. آزمایش‌ها ابتدا با حجم متنی ۵۰۰۰۰ کلمه‌ای آغاز شده و با کاهش حجم متون به یک‌پنجم تا حجم متنی ۱۰۰۰ واژه پیش رفته‌اند. از آنجا که در سطح ۱۰۰۰ واژه دقت طبقه‌بندی متون افت چشمگیری داشته است، حجم متون با فواصل ۱۰۰۰ کلمه‌ای افزایش یافته و در سطح ۴۰۰۰ کلمه (که نتیجه‌ی مطلوب حاصل شده) متوقف شده است. پس از آزمایش‌ها، بررسی‌هایی پیکره‌ای از واژه‌های دستوری پربسامد در برخی متون انجام شده و نتایج مربوطه ارائه شده‌اند.

۲-۴. آزمایش ۱ (متون ۵۰۰۰۰ کلمه‌ای)

در آزمایش نخست، متن‌های ۵۰۰۰۰ کلمه‌ای از نویسندگان به‌طور تصادفی نمونه‌گیری شد. از آنجا که فقط سه نویسنده دارای پیکره‌ی ۱۰۰۰۰۰ کلمه‌ای بودند، تحلیل سبکی فقط بر روی این سه نویسنده انجام شد. به این ترتیب، پیکره‌ی هر نویسنده به دو نمونه‌متن ۵۰۰۰۰ کلمه‌ای تقسیم شد و واژه‌های دستوری آنها به‌لحاظ تک‌نگاشتی، دونگاشتی و سه‌نگاشتی تحلیل شدند.

۱-۲-۴. تحلیل تک‌نگاشتی

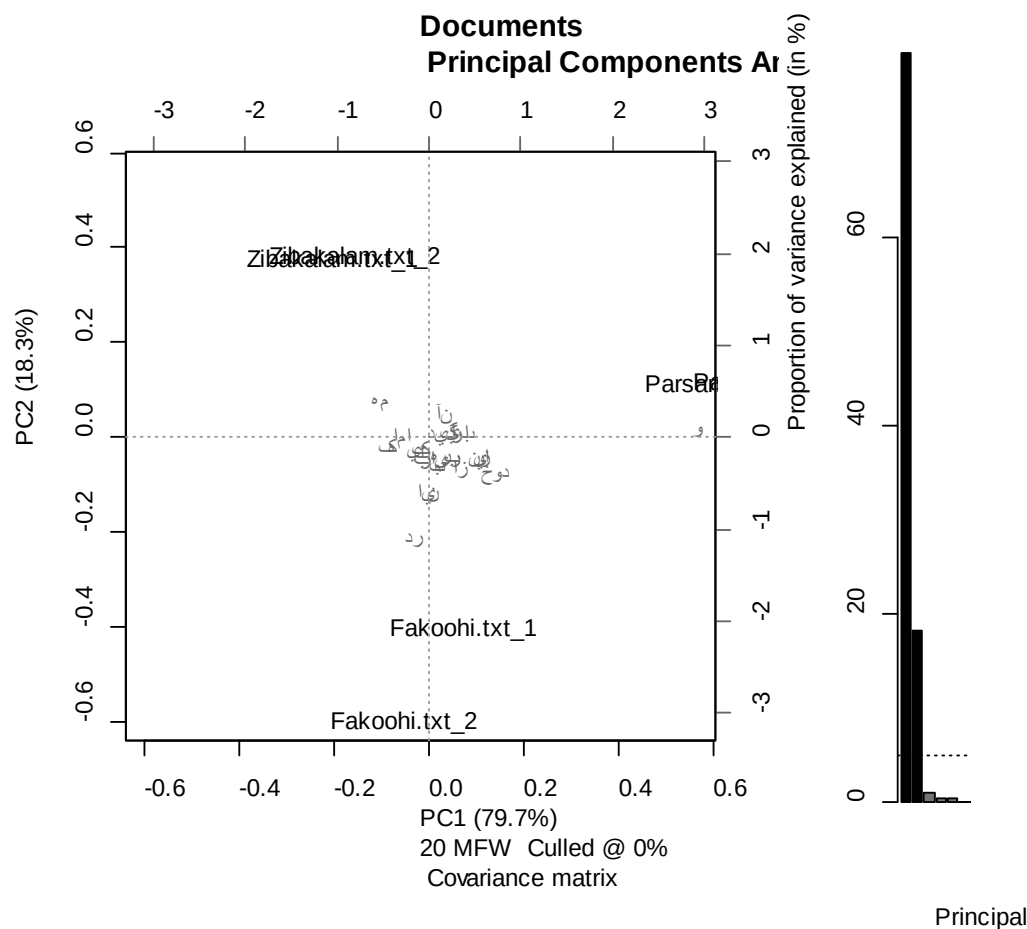
در سطح تحلیل تک‌نگاشتی، ابتدا ۲۰ واژه‌ی دستوری پربسامد از پیکره‌ها استخراج شدند و سپس با اعمال گزینش ۸۰٪ بر روی واژه‌ها، واژه‌هایی که حداقل در ۸۰٪ از متون به‌کار رفته بودند در تحلیل گنجانده شدند. شکل ۴-۱ نتیجه‌ی حاصل از تحلیل مولفه‌های اصلی را بر روی متون نویسنده‌ها نشان می‌دهد.



شکل ۱-۴. تحلیل مولفه‌های اصلی تک‌نگاشتی در متون ۵۰۰۰۰ کلمه‌ای

شکل ۱-۴ نشان می‌دهد که ۲۰ واژه‌ی پربسامد به‌خوبی می‌تواند نوشته‌های سه نویسنده را از یکدیگر متمایز کند. طبق این نمودار، مولفه‌ی اصلی نخست (PC1) حدود ۸۰٪ از پراکنش داده‌ها و مولفه‌ی اصلی دوم (PC2) حدود ۱۸٪ از پراکنش را توضیح می‌دهد که در مجموع ۹۸٪ از پراکنش داده‌ها را می‌توان توسط این دو مولفه توضیح داد. برای مشاهده‌ی اینکه کدام واژه‌های دستوری بیشترین تمایز را میان نوشته‌های نویسندگان به‌وجود آورده‌اند، باید به اطلاعات همبستگی بین متون و واژه‌های دستوری دسترسی داشت که در شکل ۲-۴ در قالب بارهای عاملی^{۴۱} نشان داده شده است.

⁴¹ Factor loadings



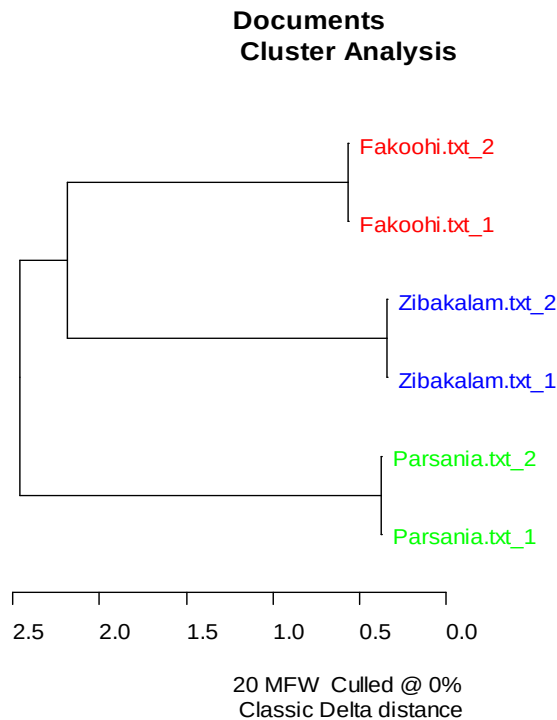
شکل ۴-۲. تحلیل مولفه‌های اصلی تک‌نگاشتی به همراه بارهای عاملی در متون ۵۰۰۰۰ کلمه‌ای

طبق شکل ۴-۲، می‌توان مشاهده کرد که از میان ۲۰ واژه‌ی دستوری، سه واژه‌ی «و»، «در» و «هر» نسبت به دیگر واژه‌ها نقش بیشتری در تمایز متون نویسنده‌ها داشته‌اند. در واقع واژه‌های دستوری «و» در متون پارسانیا، «در» در متون فکوهی و «هم» در متون زیباکلام نقش مشخصه‌ی سبکی را ایفا کرده‌اند. همچنین، واژه‌های «خود» در نوشته‌های پارسانیا و «این» در نوشته‌های فکوهی بیشتر به کار رفته‌اند.

در مرحله‌ی بعد بر اساس مقیاس فاصله‌ای دلتا تحلیل خوشه‌ای بر روی داده‌ها انجام شده تا طبقه‌بندی دقیق هر یک از متون مشخص شود. همچنین برای برآورد دقت خوشه‌بندی، شاخص تعدیل‌شده‌ی

شناسایی هویت نویسنده بر اساس زبان فردی □ ۳۷

رَند (ARI) نیز برای تحلیل خوشه‌ای محاسبه شده است. نمودار شکل ۴-۳، نتایج حاصل از تحلیل خوشه‌ای را نشان می‌دهد.



شکل ۴-۳. تحلیل خوشه‌ای واژه‌های دستوری تک‌نگاشتی در متون ۵۰۰۰۰ کلمه‌ای (ARI=1)

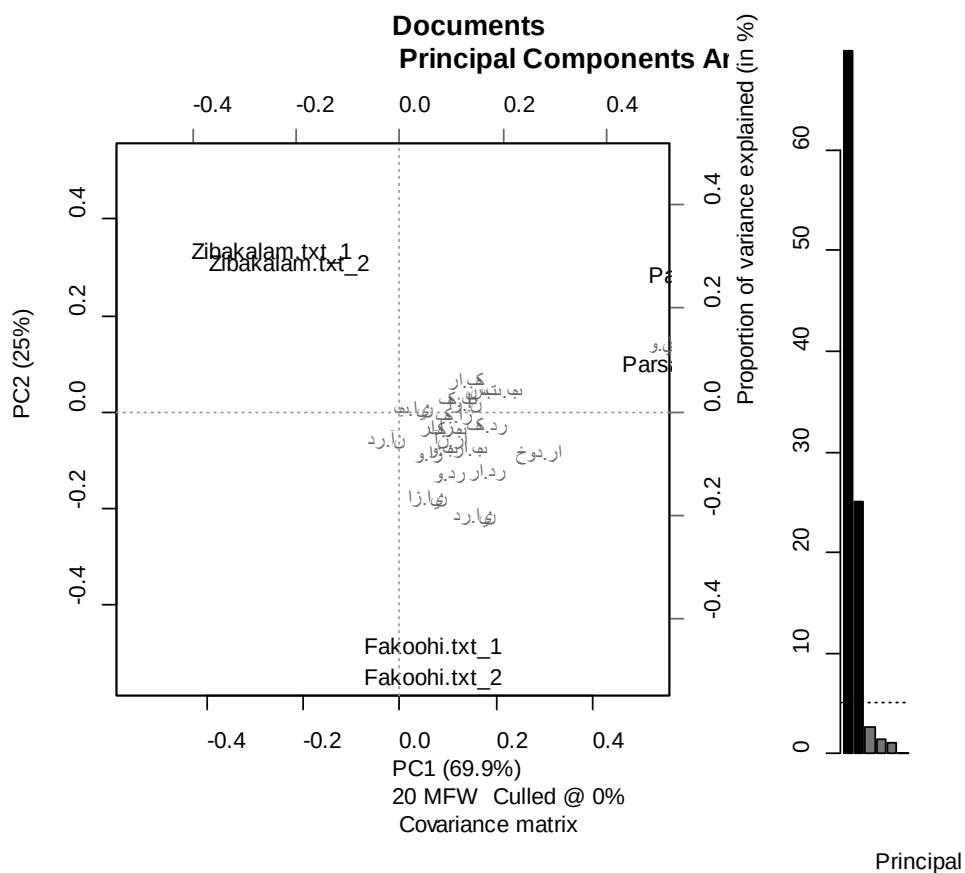
طبق شکل ۴-۳، متون نویسنده‌ها هر کدام در خوشه‌ی مربوط به خود قرار گرفته و طبقه‌بندی با دقت ۱۰۰٪ انجام شده است.

۴-۲-۲. تحلیل دونگاشتی

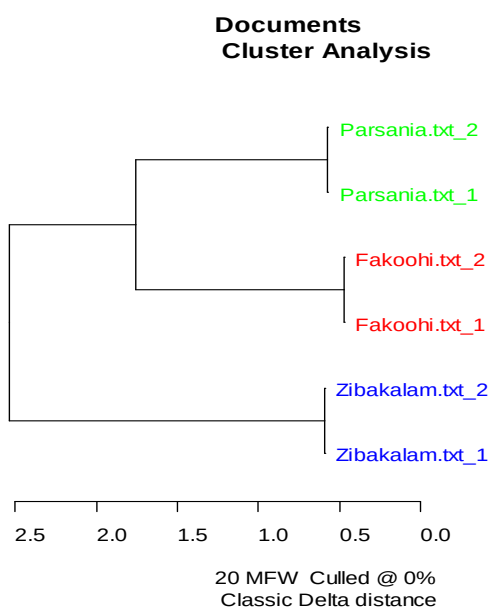
در تحلیل دونگاشتی، توالی‌های دوتاییِ واژه‌های دستوری با گزینش ۸۰٪ استخراج شدند. نمودارهای ۴-۴ و ۵-۴ نتایج تحلیل مولفه‌های اصلی و تحلیل خوشه‌ای را نشان می‌دهند. بر اساس تحلیل مولفه‌های اصلی، مشاهده می‌شود مولفه‌های اصلی ۱ و ۲ در مجموع ۹۵٪ از پراکنش داده‌ها را توضیح می‌دهند. در واژه‌های دستوری دونگاشتی، توالی‌های «و یا» و «خود را» بیشترین پراکنش را در سطح مولفه‌ی نخست و توالی‌های «در این» و «از این» بیشترین پراکنش را در سطح مولفه‌ی دوم توضیح می‌دهند.

بررسی نتایج تحلیل خوشه‌ای متون در شکل ۴-۵ نیز نشان می‌دهد که دسته‌بندی متون با دقت ۱۰۰٪ انجام گرفته است و شاخص تعدیل‌شده‌ی رَند (ARI) مقدار ۱ را نشان می‌دهد.

شناسایی هویت نویسنده بر اساس زبان فردی □ ۳۸



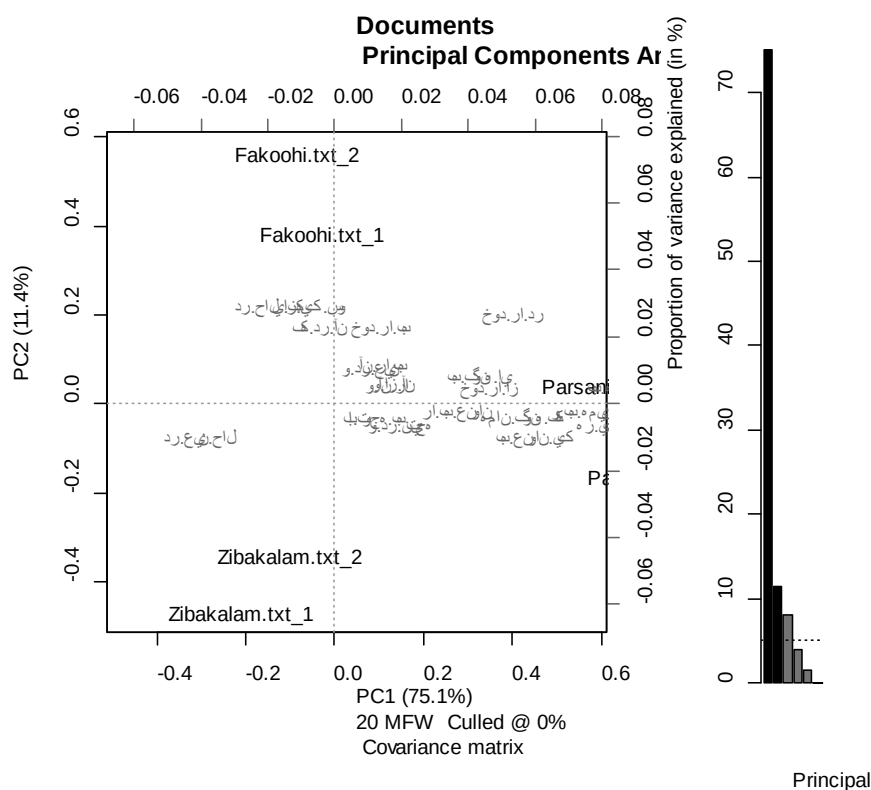
شکل ۴-۴. تحلیل مولفه‌های اصلی دنگاشتی‌ها به همراه بار عاملی در متون ۵۰۰۰۰ کلمه‌ای



شکل ۴-۵. تحلیل خوشه‌ای بر اساس واژه‌های دستوری دونگاشتی در متون ۵۰۰۰۰ کلمه‌ای (ARI=1)

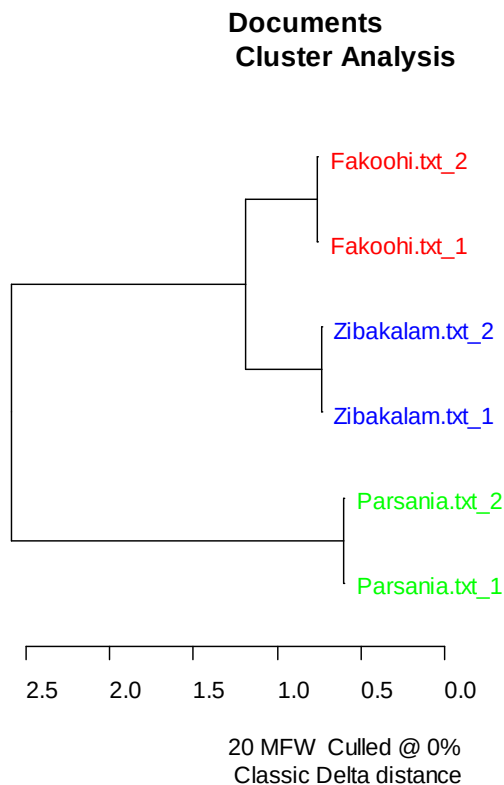
۳-۲-۴. تحلیل سه‌نگاشتی

تحلیل واژه‌های دستوری با توالی سه‌تایی (سه‌نگاشتی) نیز با گزینش ۸۰٪ واژه‌ها انجام شد. نتایج به‌دست آمده در شکل ۴-۶ نشان می‌دهد که دو مولفه‌ی اصلی ۱ و ۲ در مجموع بیش از ۸۶٪ پراکنش داده‌ها را توضیح می‌دهند. بررسی بار عاملی در نمودار نشان می‌دهد که واژه‌های دستوری سه‌نگاشتی به‌طور موثرتر و معنادارتری سبک نویسنده‌ها را از یکدیگر متمایز کرده‌اند. مثلاً سه‌نگاشتی‌هایی نظیر «خود را در»، «به عنوان یک»، «هر یک از»، «خود را از» و «به همین دلیل» در متون پارسانیا بیشتر به چشم می‌خورند. سه‌نگاشتی‌های «از یک سو» و «در حالی که» بیشتر در متون فکوهی دیده می‌شوند و سه‌نگاشتی «در عین حال» در سبک زیبا کلام پررنگ‌تر است. نتایج تحلیل خوشه‌ای واژه‌های سه‌نگاشتی (شکل ۴-۷) نشان می‌دهد متون مربوط به هر نویسنده با دقت ۱۰۰٪ طبقه‌بندی شده است (ARI=1).



شناسایی هویت نویسندگان بر اساس زبان فردی □ ۴۰

شکل ۴-۶. تحلیل مولفه‌های اصلی واژه‌های سه‌نگاشتی به همراه بار عاملی در متون ۵۰۰۰۰ کلمه‌ای



شکل ۴-۷. تحلیل خوشه‌ای واژه‌های دستوری سه‌نگاشتی در متون ۵۰۰۰۰ کلمه‌ای (ARI=1)

۴-۳. آزمایش ۲ (متون ۱۰۰۰۰ کلمه‌ای)

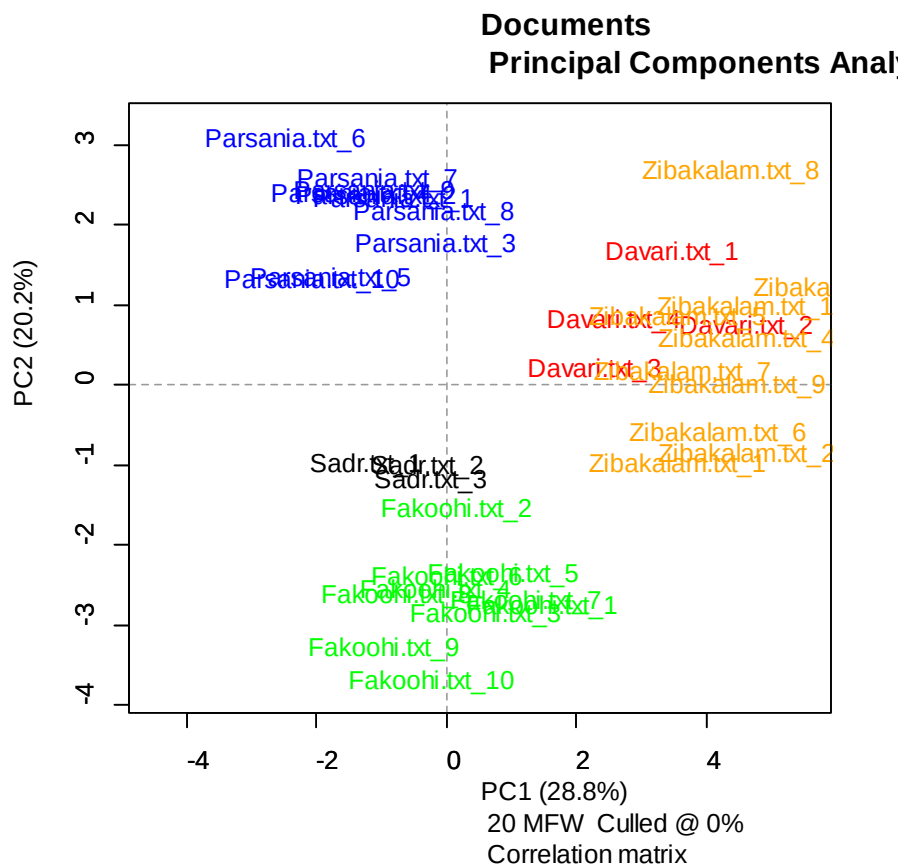
در آزمایش دوم، حجم کلمات نمونه متن‌ها به یک‌پنجم آزمایش نخست کاهش داده شد. هر یک از متون نویسنده‌ها ۱۰۰۰۰ واژه تعیین شد که به صورت تصادفی از پیکره‌ی مادر هر نویسنده استخراج شد. به این ترتیب، ۱۰ نمونه‌متن ۱۰۰۰۰ کلمه‌ای برای فکوهی، پارسانیا و زیباکلام، ۴ نمونه‌متن ۱۰۰۰۰ کلمه‌ای برای داوری و ۳ نمونه‌متن ۱۰۰۰۰ کلمه‌ای برای صدر به دست آمد.

۴-۳-۱. تحلیل تک‌نگاشتی

واژه‌های دستوری تک‌نگاشتی پرسامد با گزینش ۸۰٪ در متون استخراج شدند. شکل ۴-۸ و ۴-۹ تحلیل مولفه‌های اصلی و بار عاملی این متون را نشان می‌دهد. به نظر می‌رسد متون پارسانیا و فکوهی دسته‌های متراکم‌تری را نسبت به متون زیباکلام تشکیل داده‌اند. همچنین متون داوری و صدر نیز در

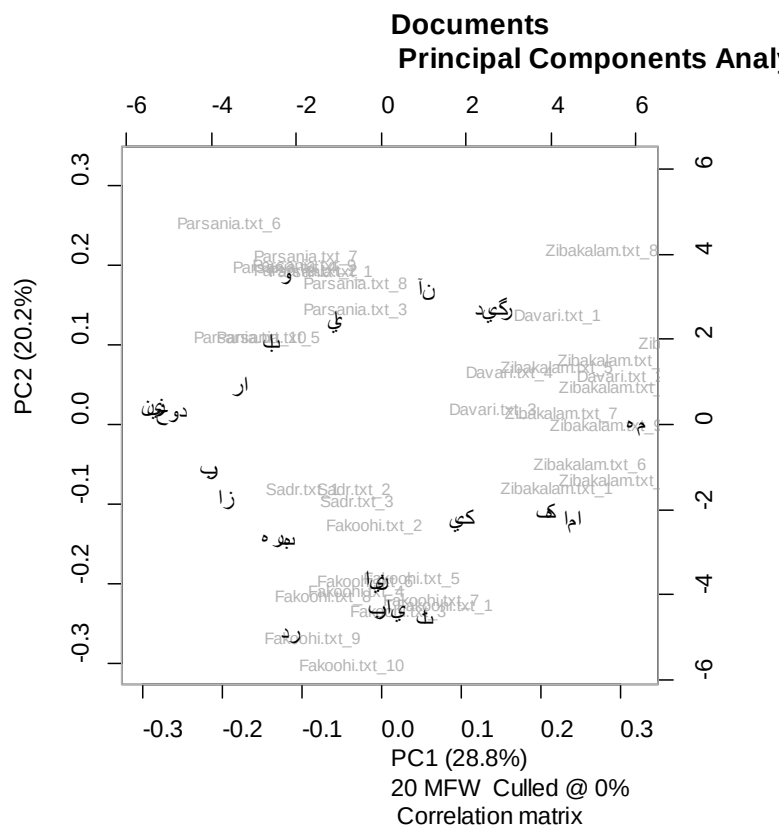
شناسایی هویت نویسندگان بر اساس زبان فردی □ ۴۱

سطح دو مولفه‌ی اصلی به ترتیب چندان از متون زیباکلام و فکوهی متمایز نشده‌اند. طبق اطلاعات شکل، دو مولفه‌ی اصلی ۱ و ۲ در مجموع ۴۹٪ از پراکنش داده‌ها را توضیح می‌دهند.

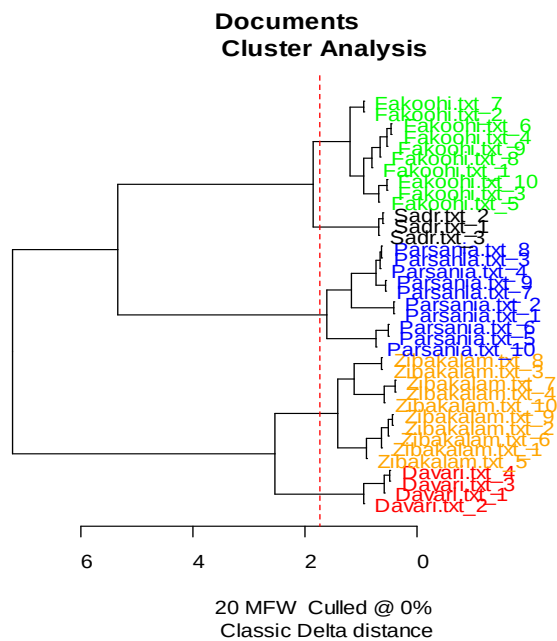


شکل ۴-۸. تحلیل مولفه‌های اصلی تک‌نگاشتی در متون ۱۰۰۰۰ کلمه‌ای

در نهایت شکل ۴-۱۰، تحلیل خوشه‌ای متون مربوطه را بر اساس مقیاس فاصله‌ای دلتا نشان می‌دهد. طبق این شکل، در تحلیل خوشه‌ای تمام متون نویسندگان در خوشه‌های مربوط به خود جای گرفته‌اند و این خوشه‌های ۵ گانه (که ۵ نویسنده را متمایز می‌کنند) در مقیاس فاصله‌ای حدوداً ۱/۷ متمایز می‌شوند. از آنجا که دقت خوشه‌بندی ۱۰۰٪ بوده است، شاخص تعدیل‌شده‌ی رند مقدار ۱ را نشان می‌دهد.



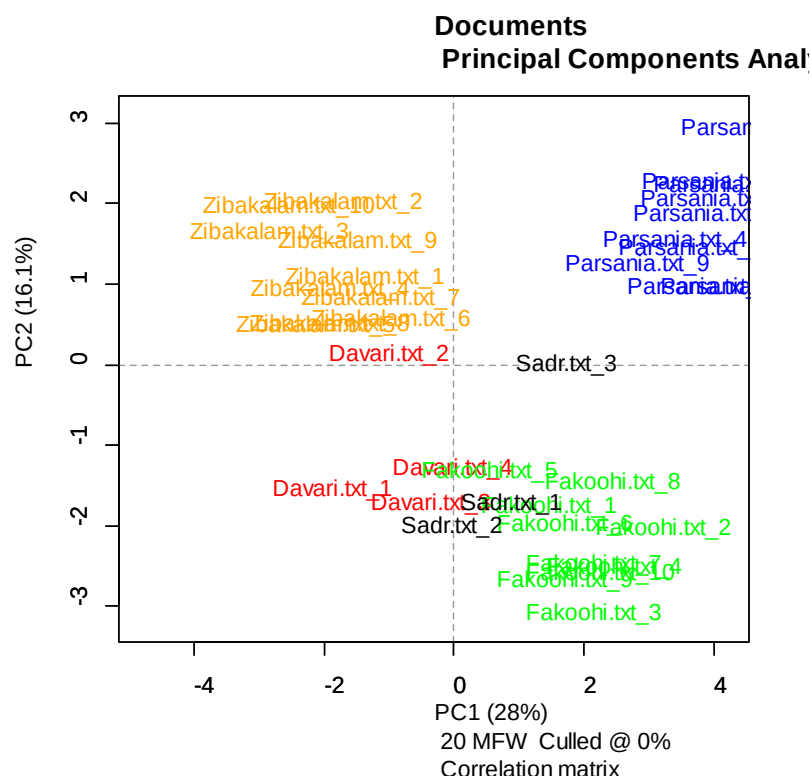
شکل ۹-۴. تحلیل مولفه‌های اصلی تک‌نگاشتی به همراه بار عاملی در متون ۱۰۰۰۰ کلمه‌ای



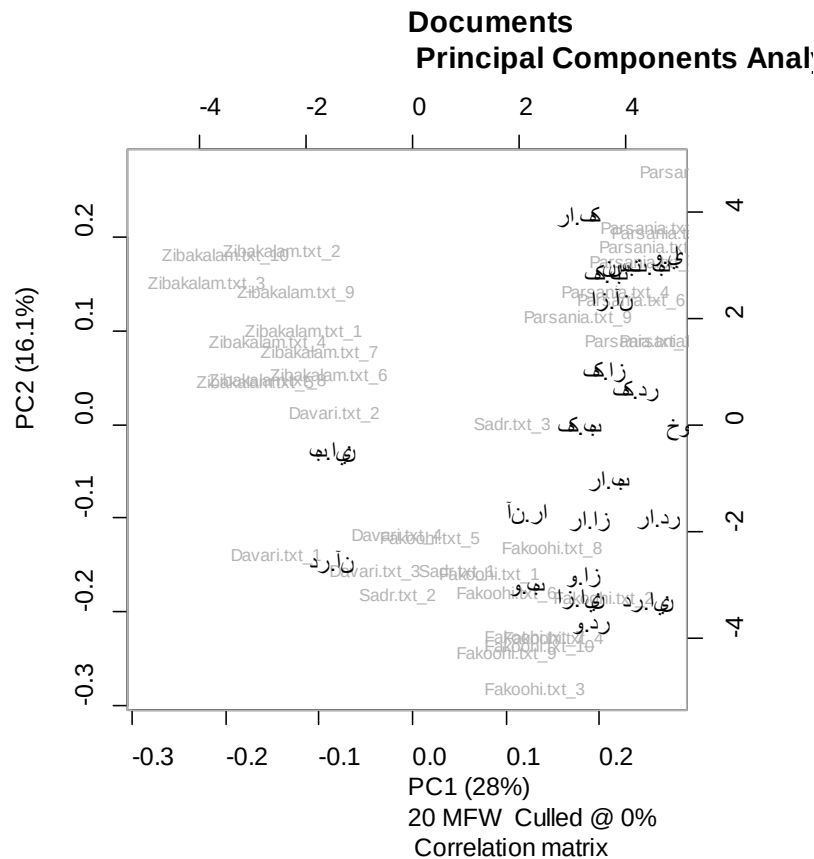
شکل ۴-۱۰. تحلیل خوشه‌ای واژه‌های دستوری تک‌نگاشتی در متون ۱۰۰۰۰ کلمه‌ای (ARI=1)

۲-۳-۴. تحلیل دونگاشتی

برای تحلیل واژه‌های دستوری دونگاشتی، پربسامدترین توالی‌های دوتایی واژه‌های دستوری متون استخراج و تحت‌گزینش ۸۰٪ قرار گرفتند. شکل ۴-۱۱ تحلیل مولفه‌ای اصلی واژه‌های دونگاشتی در این متون را نشان می‌دهد. نتایج تحلیل نشان می‌دهد که متون پارسانیا، زیباکلام و فکوهی خوشه‌های متراکم‌تری را تشکیل می‌دهند و متون صدر و داوری نسبتاً در بین این دسته‌ها پراکنده هستند. دو مولفه اصلی ۱ و ۲ در مجموع بیش از ۴۴٪ پراکنش داده‌ها را توضیح می‌دهد. شکل ۴-۱۲ واژه‌های دونگاشتی را نیز در نمودار تحلیل مولفه‌های اصلی نمایش می‌دهد. بارهای عاملی در این نمودار نشان می‌دهند که دونگاشتی‌هایی نظیر «خود را»، «و یا»، «به این»، «را در»، «در این» و «در آن» بیشترین اثر را در مولفه ۱ و دونگاشتی‌هایی نظیر «را که» و «و در» بیشترین تاثیر را در مولفه ۲ ایفا کرده‌اند.

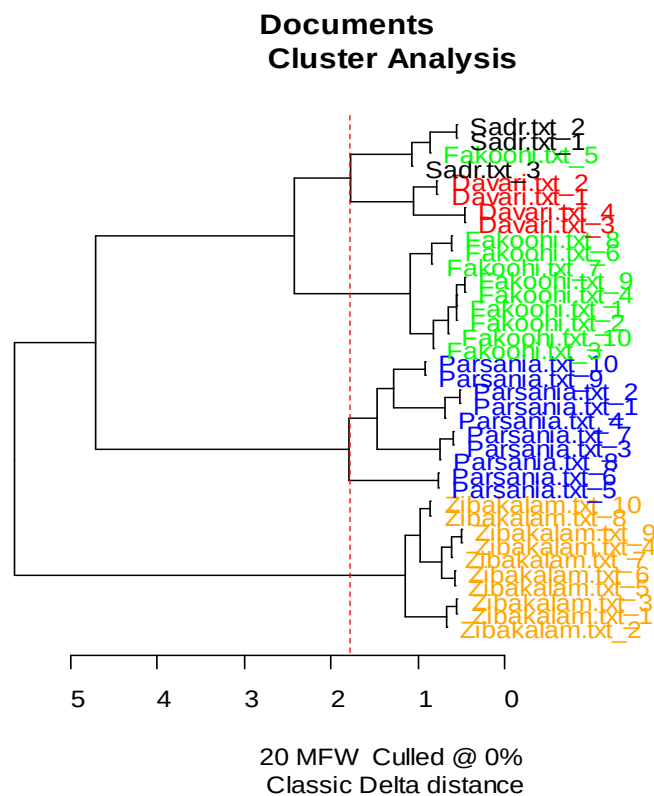


شکل ۴-۱۱. تحلیل مولفه‌های اصلی دونگاشتی در متون ۱۰۰۰۰ کلمه‌ای



شکل ۴-۱۲. تحلیل مولفه‌های اصلی دونگاشتی به همراه بار عاملی در متون ۱۰۰۰۰ کلمه‌ای

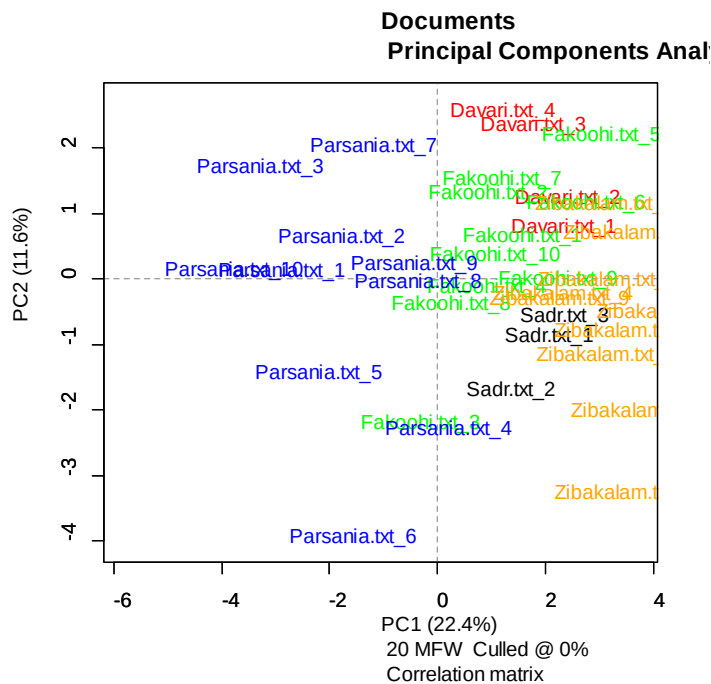
شکل ۴-۱۳ تحلیل خوشه‌ای واژه‌های دستوری دونگاشتی را در متون ۱۰۰۰۰ کلمه‌ای نشان می‌دهد. در این نمودار، تقسیم‌بندی متون به دسته‌هایی با نویسنده‌های واحد با دقت ۱۰۰٪ صورت نگرفته است. این اولین شاهد از افت دقت خوشه‌بندی در دونگاشتی‌های متون ۱۰۰۰۰ کلمه‌ای است. طبق نمودار، متون ۵ و ۶ پارسانیا خوشه‌ای مجزا را می‌سازند که در محل نقطه‌چین جداکننده‌ی ۵ خوشه از یکدیگر متمایز می‌شوند. برای برآورد دقت خوشه‌بندی از شاخص تعدیل‌شده‌ی رند استفاده شد که مقدار آن ۰/۸ به دست آمد.



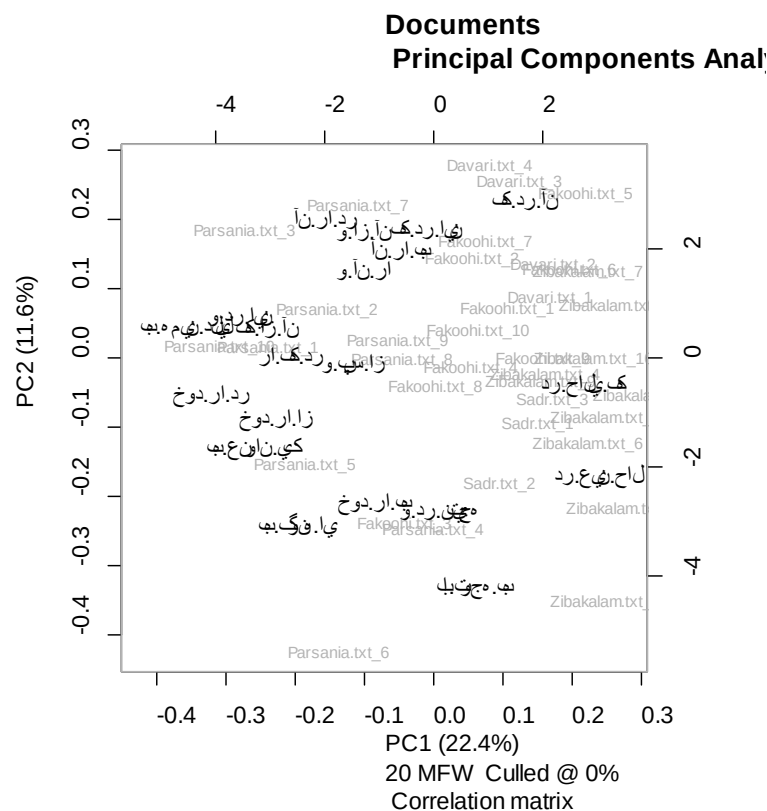
شکل ۴-۱۳. تحلیل خوشه‌ای دنگاشتی‌ها در متون ۱۰۰۰۰ کلمه‌ای (ARI=0.8)

۳-۴-۳. تحلیل سه‌نگاشتی

پس از استخراج توالی‌های سه‌تاییِ واژه‌های دستوری پرسامد، گزینش ۶۰٪ بر روی آنها اعمال شد. به دلیل کاهش حجم متون، بسامد توالی‌های سه‌نگاشتی در متون پایین می‌آید و احتمال وجود توالی‌هایی با بسامد صفر در هر یک از متون بالا می‌رود. بر همین اساس، اعمال گزینش ۸۰٪ موجب شد که تعداد واژه‌های دستوری پرسامد به ۲۰ عدد نرسد. لذا ارزش آستانه‌ی گزینش تا حد ۶۰٪ پایین آورده شد. تحلیل مولفه‌های اصلی سه‌نگاشتی‌ها در شکل ۴-۱۴ نشان می‌دهد که دو مولفه‌ی اصلی ۱ و ۲ در مجموع ۳۴٪ از پراکنش داده‌ها را توضیح می‌دهند. همچنین اطلاعات نمودار نشان می‌دهد که متون مربوط به هر نویسنده نسبت به قبل پراکنده‌تر هستند و تراکم فضایی خود را از دست داده‌اند. شکل ۴-۱۵، بارهای عاملی را در نمودار تحلیل مولفه‌های اصلی نشان می‌دهد. به طور کلی، تمایز نویسندگان در مولفه‌ی اصلی ۱ بارزتر است و در این راستا سه‌نگاشتی‌هایی نظیر «در حالی که»، «در عین حال»، «به همین دلیل» و «خود را در» بیشترین اثر را در این مولفه داشته‌اند.

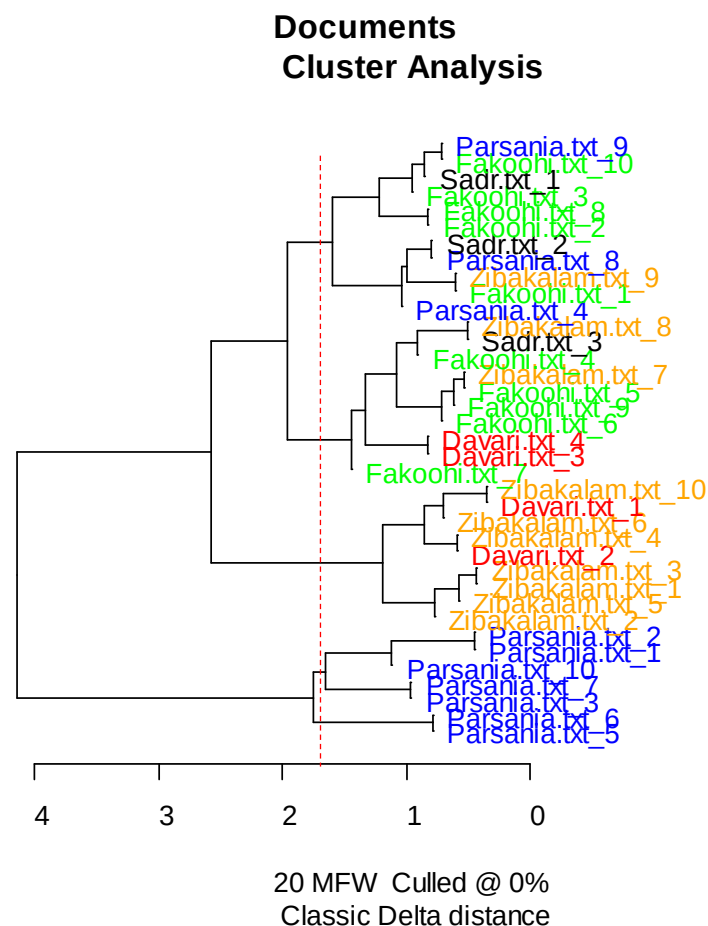


شکل ۴-۱۴. تحلیل مولفه‌های اصلی سه‌نگاشتی در متون ۱۰۰۰۰ کلمه‌ای



شکل ۴-۱۵. تحلیل مولفه‌های اصلی سه‌نگاشتی به همراه بار عاملی در متون ۱۰۰۰۰ کلمه‌ای

شکل ۴-۱۶، نمودار تحلیل خوشه‌ای واژه‌های دستوری سه‌نگاشتی در متون ۱۰۰۰۰ کلمه‌ای را نشان می‌دهد. بررسی خوشه‌ها نشان می‌دهد که عملکرد الگوریتم در دسته‌بندی متون نویسنده‌های واحد در خوشه‌های یکسان مطلوب نیست. ارزیابی دقت خوشه‌بندی بر اساس شاخص تعدیل‌شده‌ی رند، مقدار ۰/۲۴ را به‌دست می‌دهد.



شکل ۴-۱۶. تحلیل خوشه‌ای سه‌نگاشتی‌ها در متون ۱۰۰۰۰ کلمه‌ای (ARI=0.24)

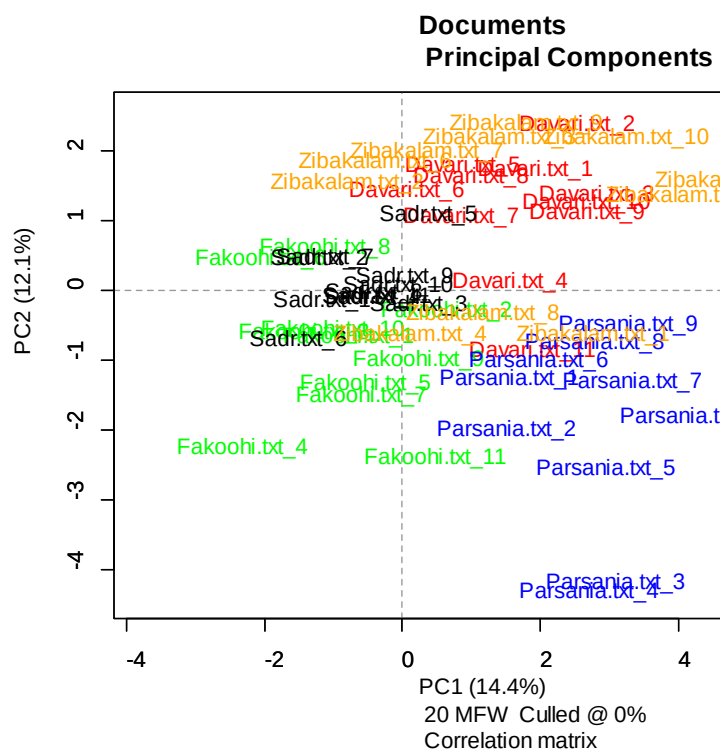
نتایج این آزمایش نشان داد، با کاهش حجم متون به ۱۰۰۰۰ کلمه، عملکرد واژه‌های دستوری تک‌نگاشتی در تمایز نویسنده‌ها همچنان ۱۰۰٪ باقی ماند اما عملکرد واژه‌های دونگاشتی و سه‌نگاشتی با افت مواجه شد. در آزمایش بعدی، روند کاهش حجم متون ادامه می‌یابد و از ۱۰۰۰۰ کلمه به یک‌پنجم کاهش داده می‌شود تا به ۲۰۰۰ کلمه برسد و رفتار الگوریتم بررسی شود. برای ارزیابی کارکرد الگوریتم، تحلیل‌ها در سطح ۱۰۰۰ واژه نیز انجام شد که قبل از آزمایش ۲۰۰۰ کلمه‌ای به گزارش نتایج آن می‌پردازیم.

۴-۴. آزمایش ۳ (متون ۱۰۰۰ کلمه‌ای)

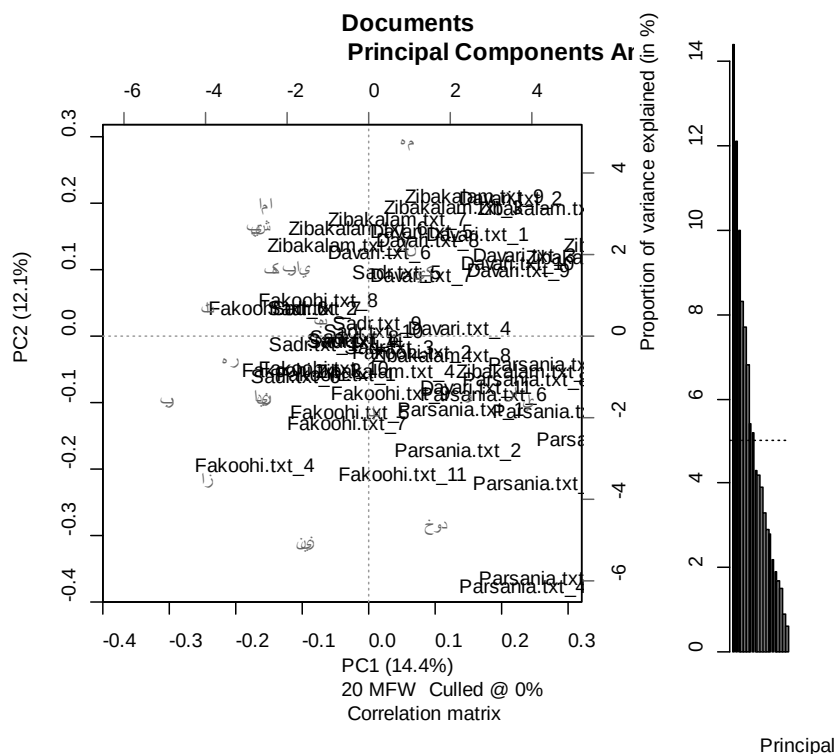
در این آزمایش، حجم نمونه‌متن‌های نویسنده‌ها ۱۰۰۰ کلمه تعیین شد. مشکلی که در اینجا وجود داشت این بود که اگر پیکره‌های مادر به نمونه‌متن‌های ۱۰۰۰ کلمه‌ای تقسیم می‌شدند (مثلاً در مورد پیکره‌های ۱۰۰۰۰۰ کلمه‌ای ۳ نویسنده)، نمونه‌متن‌های برخی نویسنده‌ها ۱۰۰ عدد می‌شد که تحلیل و بازنمایی تصویری آنها دشوار می‌شد. از این رو، پیکره‌هایی حدوداً ۱۰۰۰۰ کلمه‌ای از پیکره‌های مادر به‌طور تصادفی انتخاب شد و این پیکره‌های حدوداً ۱۰۰۰۰ کلمه‌ای به نمونه‌متن‌های ۱۰۰۰ کلمه‌ای تقسیم شدند که این امر منجر به شکل‌گیری ۱۱ نمونه‌متن برای هر نویسنده شد.

۴-۴-۱. تحلیل تک‌نگاشتی

در این سطح از تحلیل، واژه‌های پربسامد استخراج و از میان آنها ۲۰ واژه‌ی دستوری با گزینش ۷۰٪ انتخاب شدند. شکل ۴-۱۷ و ۴-۱۸ نتایج تحلیل مولفه‌های اصلی را نشان می‌دهد. دو مولفه‌ی اصلی ۱ و ۲ در مجموع بیش از ۲۶٪ پراکنش داده‌ها را توضیح می‌دهند.



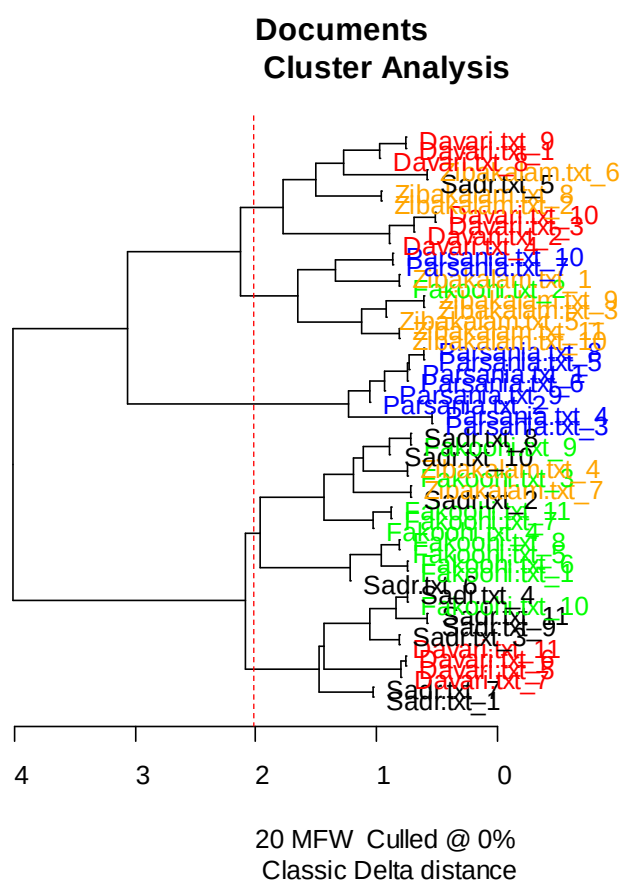
شکل ۴-۱۷. تحلیل مولفه‌های اصلی تک‌نگاشتی در متون ۱۰۰۰ کلمه‌ای



شکل ۴-۱۸. تحلیل مولفه‌های اصلی تک‌نگاشتی به همراه بار عاملی در متون ۱۰۰۰ کلمه‌ای

بررسی نمودار تحلیل مولفه‌های اصلی، تفکیک واضحی از نویسندگان به دست نمی‌دهد. همچنین بررسی مولفه‌های دیگر در شکل ۴-۱۸ نشان می‌دهد که ۸ مولفه‌ی اصلی هر کدام بیش از ۵٪ پراکنش داده‌ها را توضیح می‌دهند.

در مرحله‌ی بعد، تحلیل خوشه‌ای بر روی داده‌ها انجام شد. بررسی نمودار خوشه‌ای در شکل ۴-۱۹ نشان می‌دهد به نظر خوشه‌هایی قابل متمایز در نمودار شکل گرفته‌اند که البته با پراکنندگی دیگر داده‌ها مواجه هستند. برای نمونه، اکثریت متن‌های پارسانیا و فکوهی بهتر از نویسندگان دیگر به خوشه‌های مجزا تعلق یافته‌اند. برای محاسبه‌ی دقت خوشه‌بندی، شاخص تعدیل شده‌ی رند محاسبه شد و مقدار آن ۰/۳۶ به دست آمد.

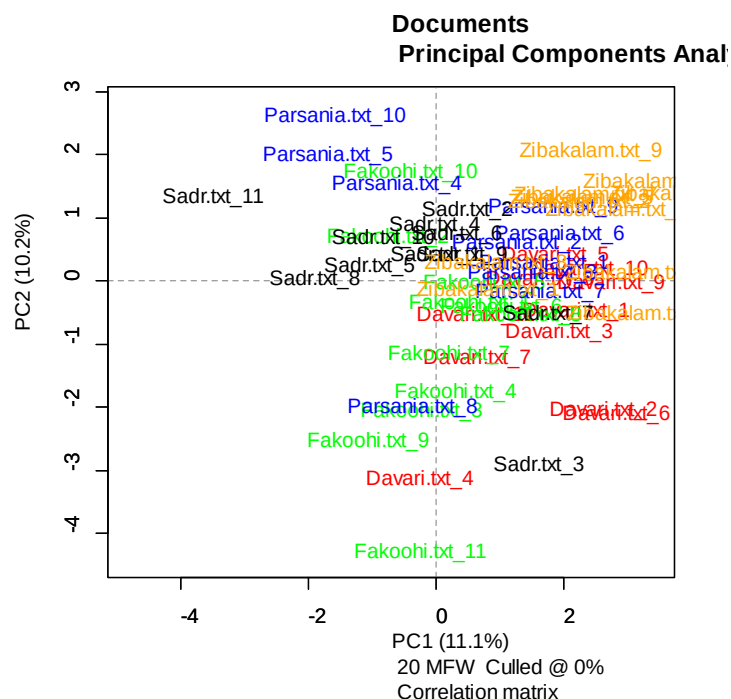


شکل ۴-۱۹. تحلیل خوشه‌ای تک‌نگاشتی‌ها در متون ۱۰۰۰ کلمه‌ای (ARI=0.36)

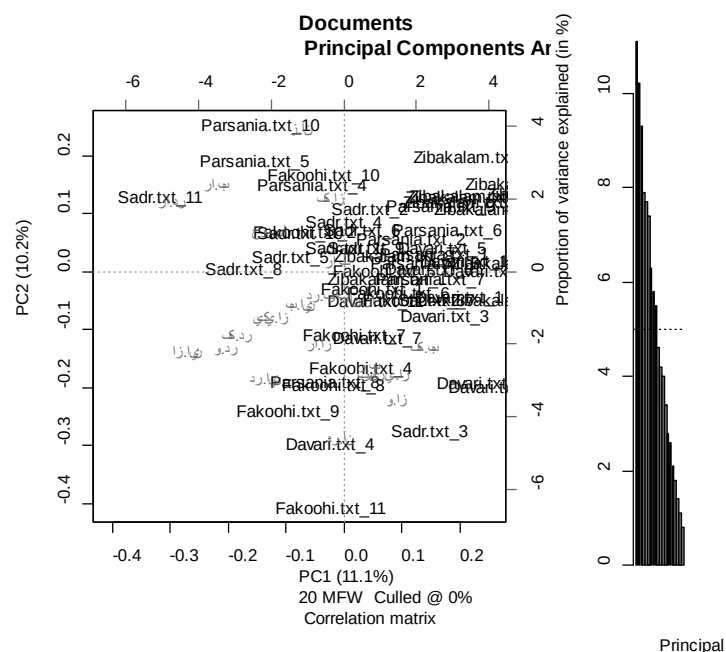
۴-۴-۲. تحلیل دونگاشتی

در این سطح از تحلیل، واژه‌های پربسامد استخراج و از میان آنها ۲۰ واژه‌ی دستوری با گزینش ۴۰٪ انتخاب شدند. شکل ۴-۲۰ و ۴-۲۱ نتایج تحلیل مولفه‌های اصلی را نشان می‌دهد. دو مولفه‌ی اصلی ۱ و ۲ در مجموع بیش از ۳۰٪ پراکنش داده‌ها را توضیح می‌دهند. بررسی نمودار تحلیل مولفه‌های اصلی در سطح واژه‌های دستوری دونگاشتی تفکیک واضحی از نویسندگان به‌دست نمی‌دهد. همچنین بررسی مولفه‌های دیگر در شکل ۴-۲۱ نشان می‌دهد که ۹ مولفه‌ی اصلی هر کدام بیش از ۵٪ پراکنش داده‌ها را توضیح می‌دهند. به‌نظر می‌رسد به‌دلیل کم‌بودن حجم متون، دو مولفه‌ی اصلی ۱ و ۲ قادر نیستند تفکیک واضحی از متون نویسنده‌ها به‌دست دهند. برای روشن‌تر شدن دسته‌بندی متون، تحلیل خوشه‌ای بر روی داده‌ها انجام شد.

شناسایی هویت نویسنده بر اساس زبان فردی □ ۵۱



شکل ۴-۲۰. تحلیل مولفه‌های اصلی دونگاشتی در متون ۱۰۰۰ کلمه‌ای

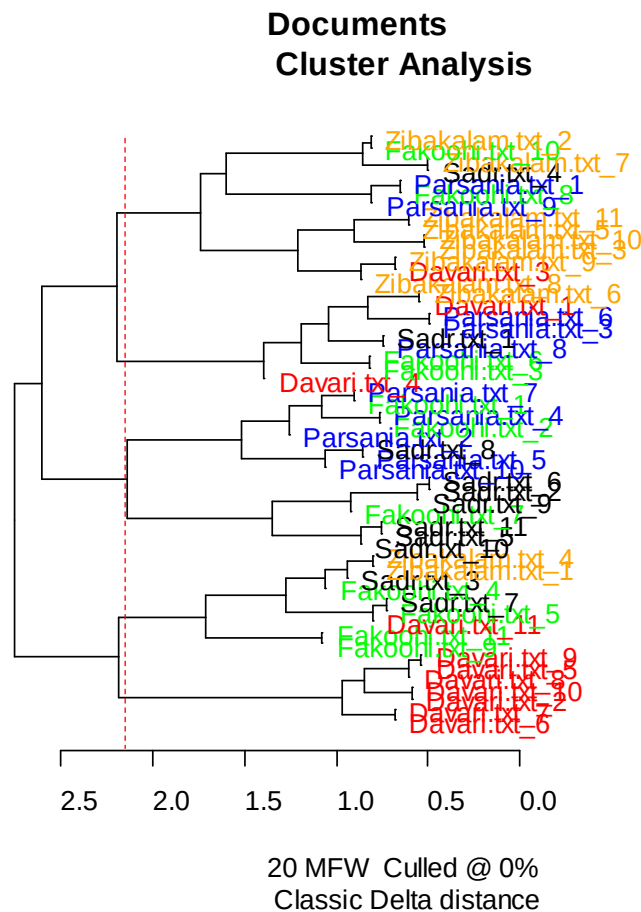


شکل ۴-۲۱. تحلیل مولفه‌های اصلی دونگاشتی به همراه بار عاملی در متون ۱۰۰۰ کلمه‌ای

شکل ۴-۲۲ نمودار تحلیل خوشه‌ای واژه‌های دستوری دونگاشتی در متون ۱۰۰۰ کلمه‌ای را نشان می‌دهد. با بررسی نمودار می‌توان دریافت که به نظر می‌رسد در مورد دو نویسنده یعنی صدر و

شناسایی هویت نویسنده بر اساس زبان فردی □ ۵۲

زیباکلام اکثریت نمونه‌متن‌ها در خوشه‌های متمایز جای گرفته‌اند. محاسبه‌ی شاخص تعدیل‌شده‌ی رند برای ارزیابی دقت خوشه‌بندی عدد ۰/۲ را به‌دست می‌دهد.

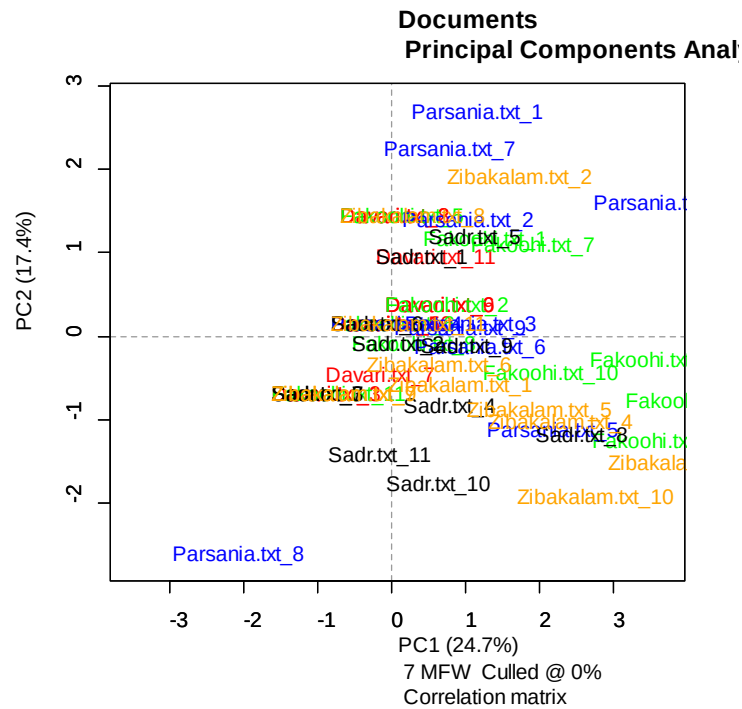


شکل ۴-۲۲. تحلیل خوشه‌ای دونگاشتی‌ها در متون ۱۰۰۰ کلمه‌ای (ARI=0.2)

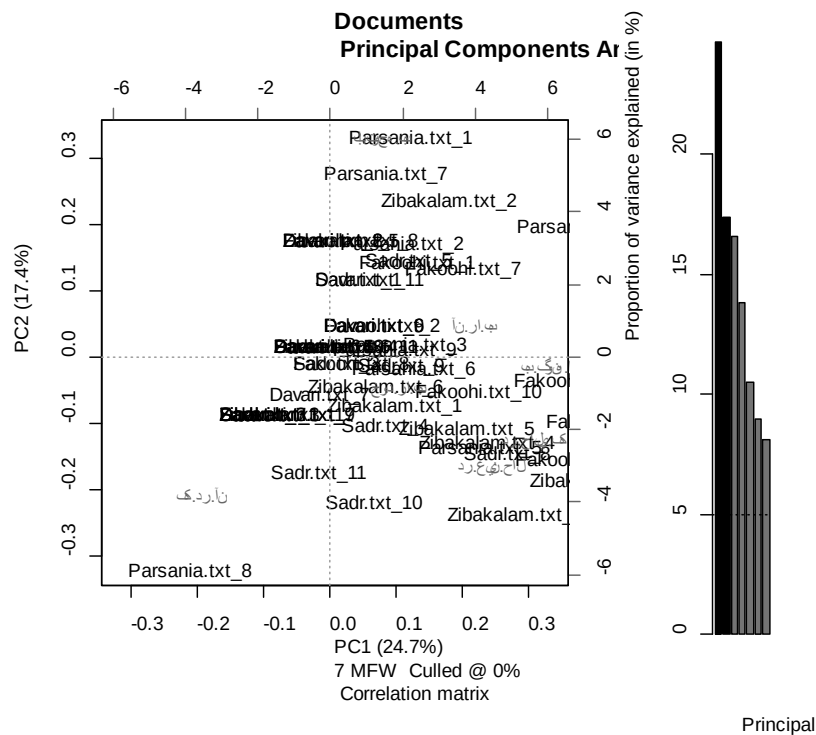
۴-۴-۳. تحلیل سه‌نگاشتی

در تحلیل سه‌نگاشتی به‌دلیل حجم پایین متون، گزینش واژه‌های دستوری غیرصفر در متون نویسندگان با حداقل مقدار یعنی ۲۰٪ اعمال شد که موجب شد تنها ۷ واژه‌ی دستوری برای تحلیل در دسترس باشد. بنابراین تمام تحلیل‌های انجام شده با تعداد ۷ واژه‌ی دستوری سه‌نگاشتی انجام شده است. شکل ۴-۲۳ و ۴-۲۴ تحلیل مولفه‌های اصلی داده‌ها را نشان می‌دهند. دو مولفه‌ی اصلی ۱ و ۲ روی هم رفته، حدود ۳۲٪ از پراکنش داده‌ها را توضیح می‌دهند.

شناسایی هویت نویسنده بر اساس زبان فردی □ ۵۳



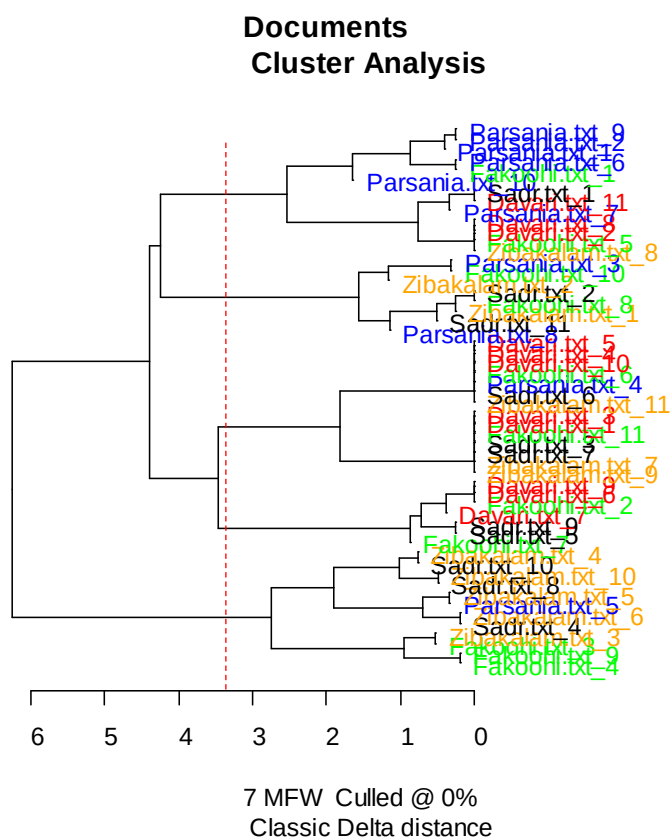
شکل ۴-۲۳. تحلیل مولفه‌های اصلی سه‌نگاشتی در متون ۱۰۰۰ کلمه‌ای



شکل ۴-۲۴. تحلیل مولفه‌های اصلی سه‌نگاشتی به همراه بار عاملی در متون ۱۰۰۰ کلمه‌ای

شناسایی هویت نویسنده بر اساس زبان فردی □ ۵۴

برای بررسی دقیق‌تر نحوه‌ی تفکیک متون، تحلیل خوشه‌ای با مقیاس فاصله‌ای دلتا انجام شد. شکل ۴-۲۵ نمودار درختی این تحلیل را نشان می‌دهد. خوشه‌ها به‌طور غیرمتراکم و با فاصله‌های زیاد از یکدیگر قرار دارند و تفکیک آشکاری نمی‌توان از نویسنده‌ها به‌دست داد. محاسبه‌ی شاخص تعدیل‌شده‌ی رند، عدد ۰/۰۴ را به‌دست می‌دهد که بسیار نامطلوب است.



شکل ۴-۲۵. تحلیل خوشه‌ای سه‌نگاشتی‌ها در متون ۱۰۰۰ کلمه‌ای (ARI=0.04)

۵-۴. آزمایش ۴ (متون ۲۰۰۰ کلمه‌ای)

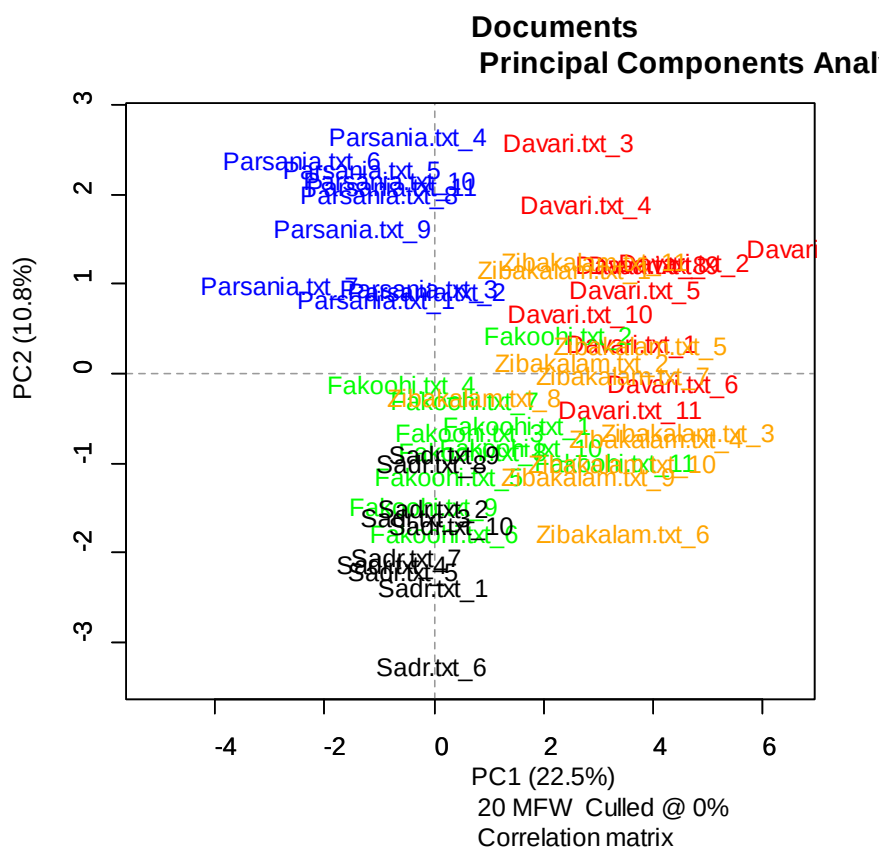
در این آزمایش، حجم نمونه‌متن‌های نویسنده‌ها ۲۰۰۰ کلمه تعیین شد. مشکل آزمایش قبل در این آزمایش نیز وجود داشت: اینکه که اگر پیکره‌های مادر به نمونه‌متن‌های ۲۰۰۰ کلمه‌ای تقسیم می‌شدند (مثلاً در مورد پیکره‌های ۱۰۰۰۰۰ کلمه‌ای ۳ نویسنده)، نمونه‌متن‌های برخی نویسنده‌ها ۵۰ عدد می‌شد که تحلیل و بازنمایی تصویری آنها دشوار می‌شد. از این‌رو، پیکره‌هایی حدوداً ۲۰۰۰۰ کلمه‌ای از پیکره‌های مادر به‌طور تصادفی انتخاب شد و این پیکره‌های حدوداً ۲۰۰۰۰ کلمه‌ای به

شناسایی هویت نویسنده بر اساس زبان فردی □ ۵۵

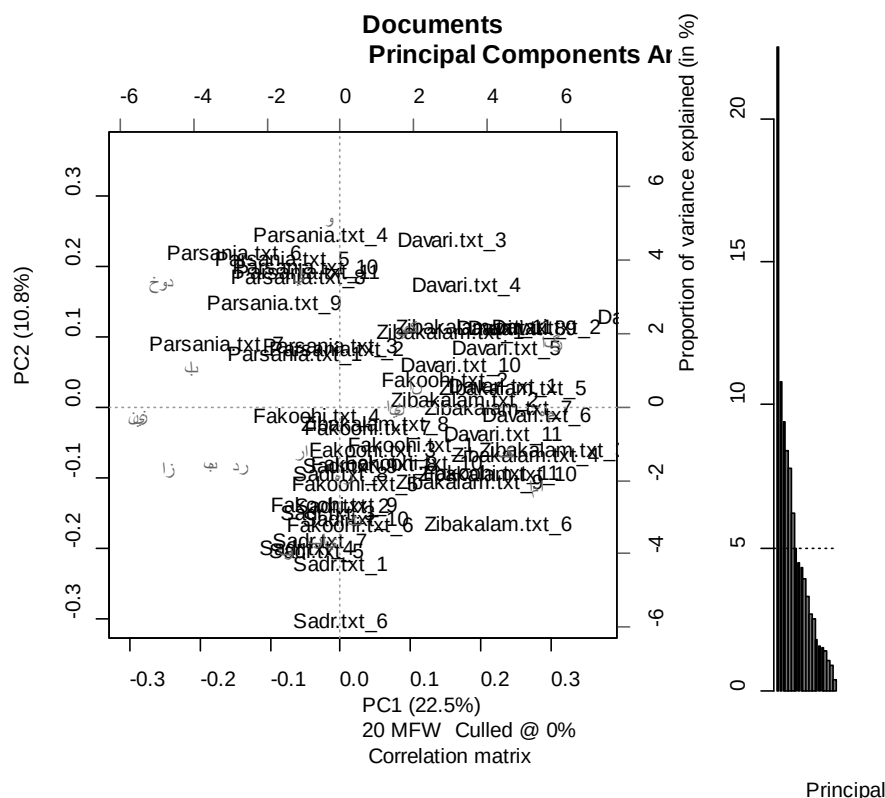
نمونه‌متن‌های ۲۰۰۰ کلمه‌ای تقسیم شدند که این امر منجر به شکل‌گیری ۱۱ نمونه‌متن برای هر نویسنده شد.

۱-۵-۴. تحلیل تک‌نگاشتی

در تحلیل واژه‌های دستوری تک‌نگاشتی در این آزمایش نیز همانند آزمایش‌های قبلی، واژه‌های پربسامد استخراج و از میان آنها ۲۰ واژه‌ی دستوری با گزینش ۸۰٪ انتخاب شدند. شکل ۴-۲۶ نتایج تحلیل مولفه‌های اصلی را نشان می‌دهد. دو مولفه‌ی اصلی ۱ و ۲ در مجموع بیش از ۳۲٪ پراکنش داده‌ها را توضیح می‌دهند.



شکل ۴-۲۶. تحلیل مولفه‌های اصلی تک‌نگاشتی در متون ۲۰۰۰ کلمه‌ای



شکل ۴-۲۷. تحلیل مولفه‌های اصلی تک‌نگاشتی به همراه بار عاملی در متون ۲۰۰۰ کلمه‌ای

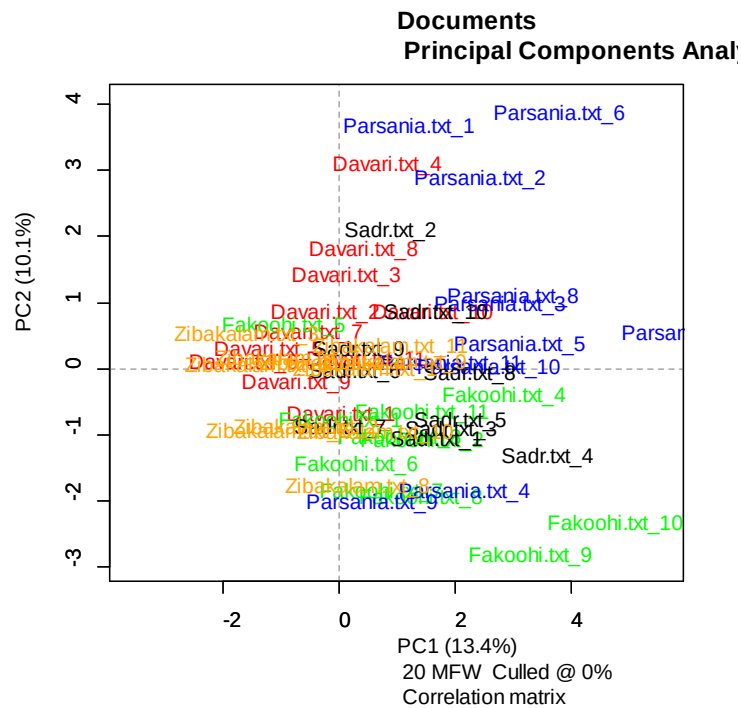
در قسمت راست شکل ۴-۲۷، که میزان مشارکت مولفه‌های اصلی را در توضیح پراکنش داده‌ها نشان می‌دهد، مشاهده می‌شود که مولفه‌های اصلی ۲ به بعد، با فاصله‌ی اندکی از یکدیگر قرار گرفته‌اند و تعداد ۶ مولفه (با احتساب مولفه‌های ۱ و ۲) بیش از ۵٪ از پراکنش داده‌ها را توضیح می‌دهند. این امر نشان می‌دهد که تمایز سبک نویسندگان در قالب دو مولفه‌ی ۱ و ۲ کافی نیست و در هم‌رفتگی متون نویسندگان در نمودار ۴-۲۶ نشانگر عدم کفایت دو مولفه در تفکیک این متون است.

با اینکه تحلیل مولفه‌های اصلی تصویر روشنی از دسته‌بندی متون ارائه نمی‌کند، تحلیل خوشه‌ای می‌تواند اطلاعات بیشتری از شباهت بین متون در اختیارمان قرار دهد. شکل ۴-۲۸ نتایج این تحلیل را نشان می‌دهد. با توجه به داده‌های نمودار درختی، مشاهده می‌شود که اکثر متون در خوشه‌های متمایز خود قرار گرفته‌اند و برخی از متون در خوشه‌هایی غیر از خوشه‌های نویسندگان خود قرار گرفته‌اند. محاسبه‌ی شاخص تعدیل شده‌ی رند دقتی معادل ۰/۷۵ را برای خوشه‌بندی نشان می‌دهد.

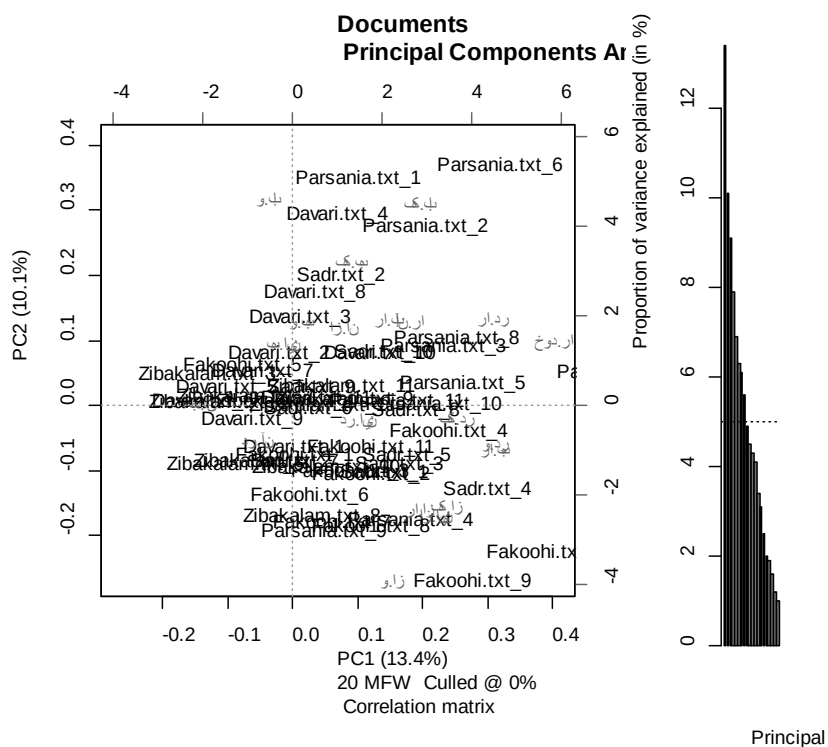


۲-۵-۴. تحلیل دوانگاشتی

برای تحلیل توالی‌های دوناگشتی واژه‌های دستوری، واژه‌های پربسامد با گزینش ۶۰٪ لحاظ شدند. شکل ۴-۲۹ نتایج تحلیل مولفه‌های اصلی را نشان می‌دهد. در شکل ۴-۳۰ همان تحلیل به همراه بار عاملی و مشارکت عامل‌های اصلی بیش از ۲ نیز مشاهده می‌شود. نتایج این تحلیل نشان می‌دهد که دو مولفه‌ی اصلی ۱ و ۲ در مجموع بیش از ۲۳٪ از پراکنش داده‌ها را توضیح می‌دهند. مشارکت مولفه‌های اصلی دیگر در شکل ۴-۳۰ نشان می‌دهد که بیش از دو مولفه‌ی اصلی مشارکت بیش از ۵٪ در پراکنش داده‌ها داشته‌اند. از آنجا که تفکیک‌پذیری متون نویسنده‌ها در سطح دو مولفه‌ی اصلی نامطلوب است، مطالعه‌ی بارهای عاملی (واژه‌های دستوری) و نسبت‌دادن آنها به نویسنده‌ی خاصی دشوار است.

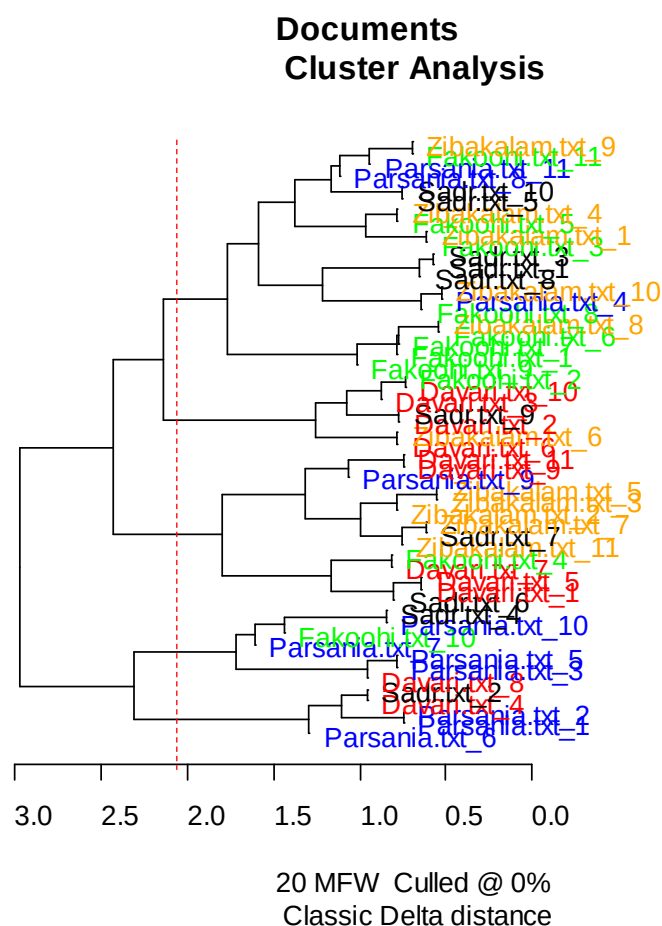


شکل ۴-۲۹. تحلیل مولفه‌های اصلی دونگاشتی در متون ۲۰۰۰ کلمه‌ای



شکل ۴-۳۰. تحلیل مولفه‌های اصلی دونگاشتی به همراه بار عاملی در متون ۲۰۰۰ کلمه‌ای

نتایج حاصل از تحلیل خوشه‌ای واژه‌های دستوری دونگاشتی در متون ۲۰۰۰ کلمه‌ای در شکل ۴-۳۱ آمده است. با توجه به نمودار درختی مشاهده می‌شود که خوشه‌های ۵ گانه به‌طور آشکار تراکم درونی نداشته و تمایز آشکاری نیز از یکدیگر ندارند. محاسبه‌ی شاخص تعدیل‌شده‌ی رند، عدد ۰/۰۹ را برای این خوشه‌بندی به‌دست می‌دهد.

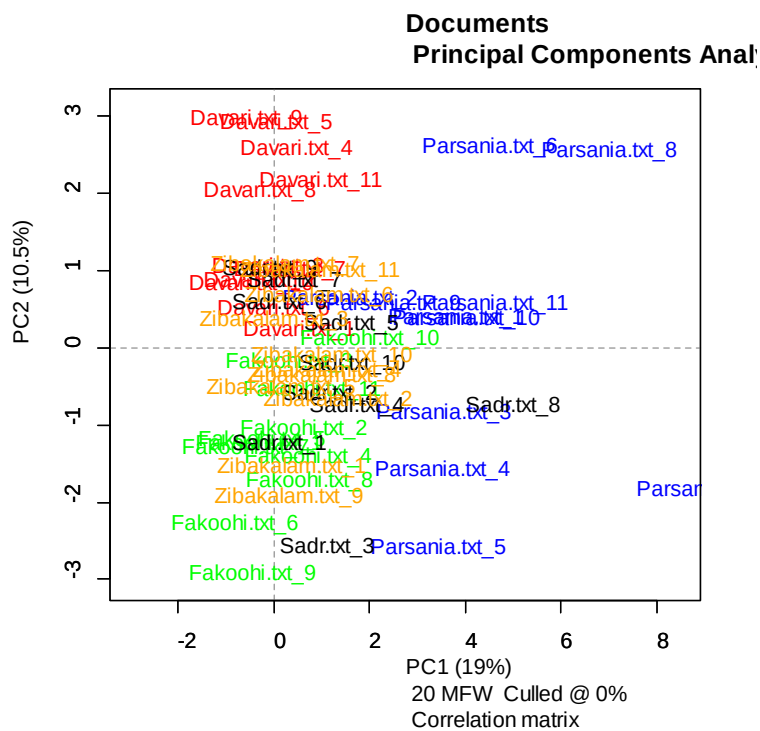


شکل ۴-۳۱. تحلیل خوشه‌ای دونگاشتی‌ها در متون ۲۰۰۰ کلمه‌ای (ARI=0.09)

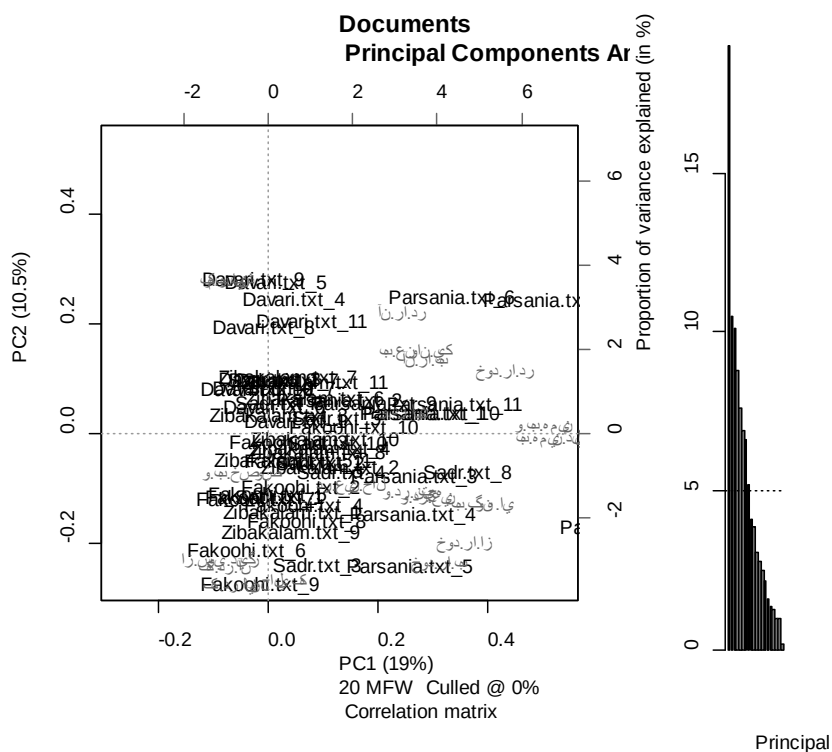
۳-۵-۴. تحلیل سه‌نگاشتی

در استخراج توالی‌های سه‌تایی واژه‌های دستوری، از آنجا که بسامد این توالی‌ها به‌دلیل حجم کم متون در بسیاری از متون صفر بود، گزینش ۲۰٪ بر توالی‌های پربسامد اعمال شد. شکل ۴-۳۲ و ۴-۳۳ نتایج حاصل از تحلیل مولفه‌های اصلی بر اساس واژه‌های دستوری سه‌نگاشتی را نشان می‌دهد.

شناسایی هویت نویسندگان بر اساس زبان فردی □ ۶۰



شکل ۴-۳۲. تحلیل مولفه‌های اصلی سه‌نگاشتی در متون ۲۰۰۰ کلمه‌ای

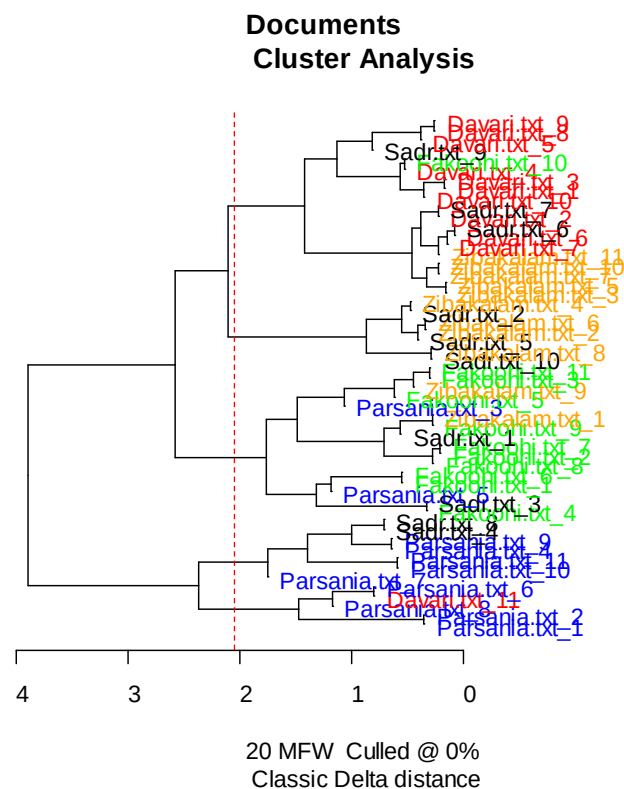


شکل ۴-۳۳. تحلیل مولفه‌های اصلی سه‌نگاشتی به همراه بار عاملی در متون ۲۰۰۰ کلمه‌ای

شناسایی هویت نویسنده بر اساس زبان فردی □ ۶۱

بررسی شکل ۴-۳۲ نشان می‌دهد که دو مولفه‌ی اصلی ۱ و ۲ در مجموع حدود ۲۰٪ از پراکنش داده‌ها را توضیح می‌دهند. همچنین با توجه به داده‌های ۴-۳۳، حدود ۸ مولفه‌ی اصلی بیش از ۵٪ در پراکنش داده‌ها مشارکت داشته‌اند. از این‌رو متمایز کردن نویسنده‌ها بر اساس دو مولفه‌ی اصلی چندان نتیجه‌بخش نیست.

شکل ۴-۳۴ نتایج تحلیل خوشه‌های واژه‌های دستوری سه‌نگاشتی را نشان می‌دهد. همانند تحلیل دونگاشتی‌ها، خوشه‌های شکل گرفته چندان متراکم و متمایز نیستند. اما به‌نظر می‌رسد عملکرد واژه‌های سه‌نگاشتی اندکی بهتر از دونگاشتی بوده است. محاسبه‌ی شاخص تعدیل‌شده‌ی رند نشان می‌دهد که خوشه‌بندی متون بر اساس واژه‌های سه‌نگاشتی با دقت ۰/۲۸ انجام گرفته است.



شکل ۴-۳۴. تحلیل خوشه‌ای سه‌نگاشتی‌ها در متون ۲۰۰۰ کلمه‌ای (ARI=0.28)

نتایج آزمایشات این بخش نشان داد که تمایز متون نویسنده‌ها بر اساس ۲۰۰۰ واژه در هر سه سطح واژه‌های تک‌نگاشتی، دونگاشتی و سه‌نگاشتی دقتی کمتر از ۱۰۰٪ دارد. به‌همین دلیل در آزمایشات

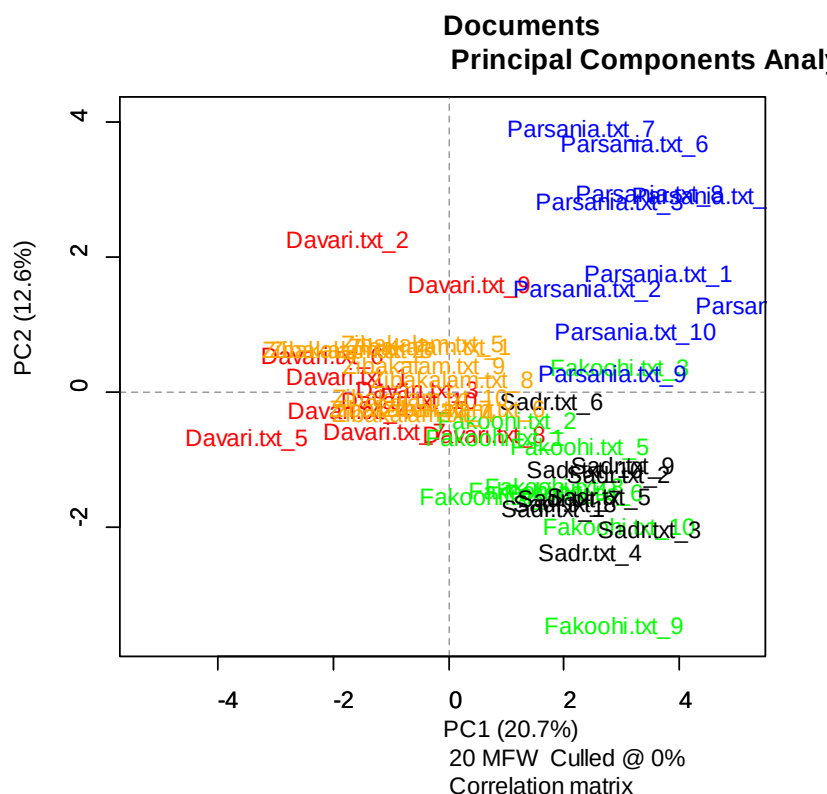
بعدی، به میزان ۱۰۰۰ واژه به متون افزوده شد تا حد بهینه برای تفکیک ۱۰۰٪ نویسنده‌ها از یکدیگر (حداقل در یکی از سه سطوح تک‌نگاشتی، دونگاشتی و سه‌نگاشتی) معین شود.

۶-۴. آزمایش ۵ (متون ۳۰۰۰ کلمه‌ای)

برای انجام این آزمایش، پیکره‌ای حدوداً ۳۰۰۰۰ کلمه‌ای برای هر نویسنده از پیکره‌ی مادر به‌طور تصادفی انتخاب شد و سپس این پیکره به نمونه‌متن‌های ۳۰۰۰ کلمه‌ای برای هر نویسنده تقسیم شد. تعداد نمونه‌متن‌ها برای هر نویسنده حدوداً ۱۰ متن بود.

۱-۶-۴. تحلیل تک‌نگاشتی

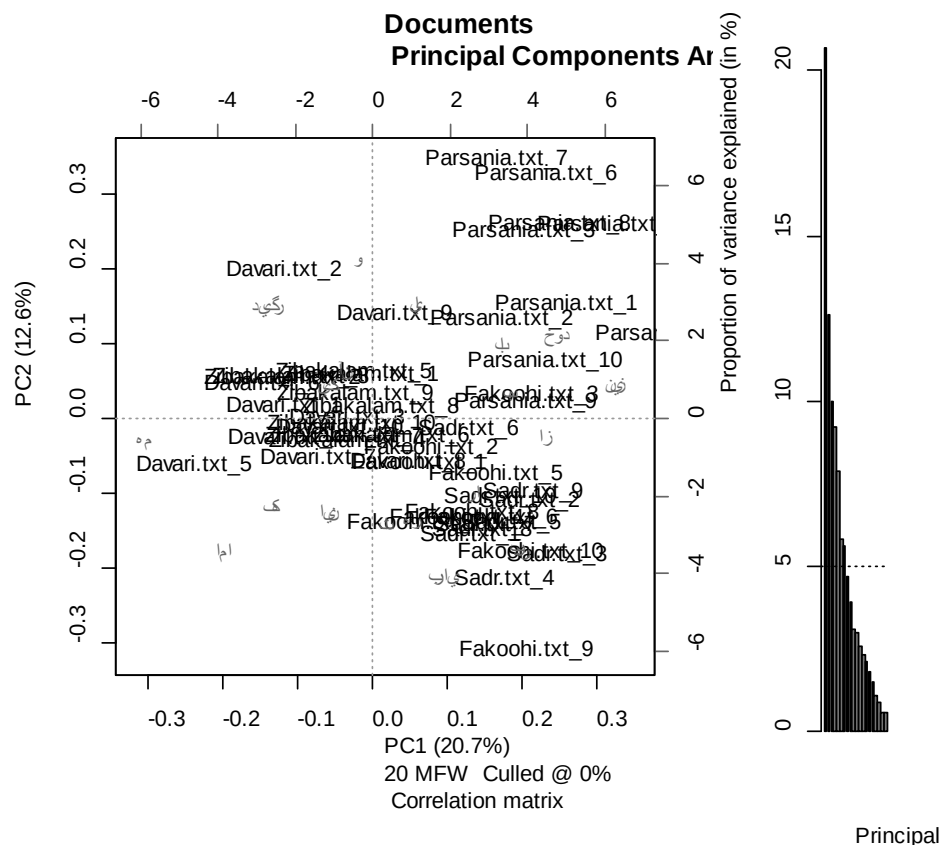
در تحلیل تک‌نگاشتی، ۲۰ واژه‌ی پربسامد نمونه‌متن‌ها استخراج شدند و با مقیاس ۸۰٪ گزینش شدند. شکل ۴-۳۵ تحلیل مولفه‌های اصلی متون مورد تحلیل را نشان می‌دهد. دو مولفه‌ی اصلی ۱ و ۲ روی هم رفته، حدود ۳۳٪ از پراکنش داده‌ها را توضیح می‌دهند.



شکل ۴-۳۵. تحلیل مولفه‌های اصلی تک‌نگاشتی در متون ۳۰۰۰ کلمه‌ای

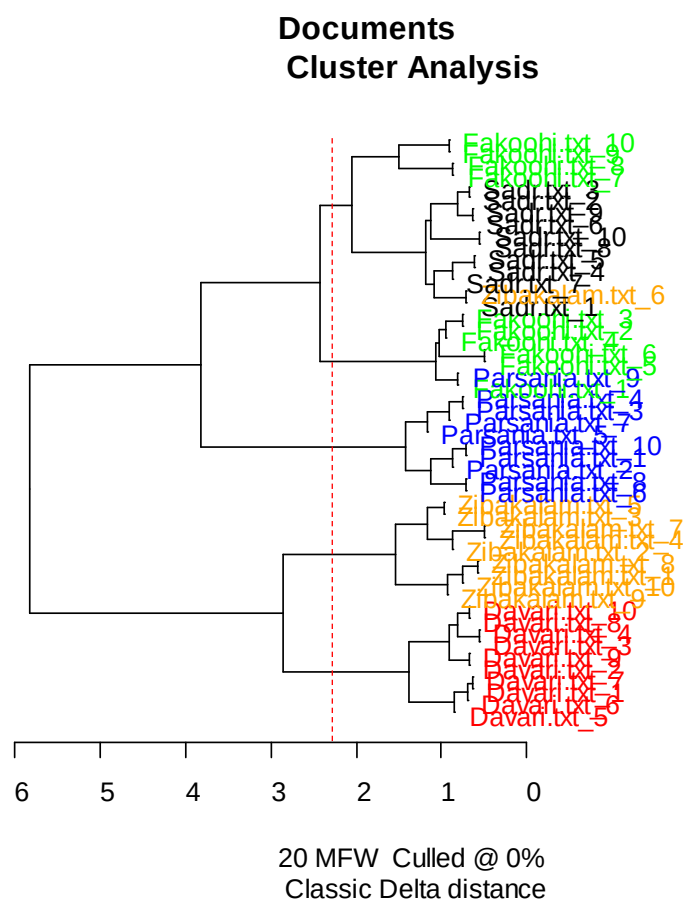
شناسایی هویت نویسنده بر اساس زبان فردی □ ۶۳

شکل ۴-۳۶، تحلیل مولفه‌های اصلی را به همراه بار عاملی و نمودار مولفه‌های دیگر نشان می‌دهد. طبق این شکل، مولفه‌هایی که هر کدام بیش از ۵٪ پراکنش داده‌ها را توضیح می‌دهند، تعدادشان ۷ مولفه است. بنابراین، دو مولفه‌ای اصلی که داده‌های ما را توضیح می‌دهند نمی‌توانند پراکنش داده‌ها را به خوبی نشان دهند.



شکل ۴-۳۶. تحلیل مولفه‌های اصلی تکنگاشتی به همراه بار عاملی در متون ۳۰۰۰ کلمه‌ای

برای طبقه‌بندی آثار نویسنده‌ها و ارزیابی کارایی مقیاس دلتا، تحلیل خوشه‌ای نیز بر روی داده‌ها انجام گرفت که نتایج آن در شکل ۴-۳۷ آمده است. خط جداکننده خوشه‌ها که در نقطه‌ی تفکیک ۵ خوشه به مثابه‌ی ۵ نویسنده‌ی آثار قرار گرفته، آثار داوری، صدر و پارسانیا را کاملاً در خوشه‌های مجزا قرار داده اما متون زیباکلام و فکوهی در خوشه‌های دیگر پراکنده شده‌اند. محاسبه‌ی شاخص تعدیل شده‌ی رند، دقت ۰/۷۳ را برای این خوشه‌بندی به دست می‌دهد.

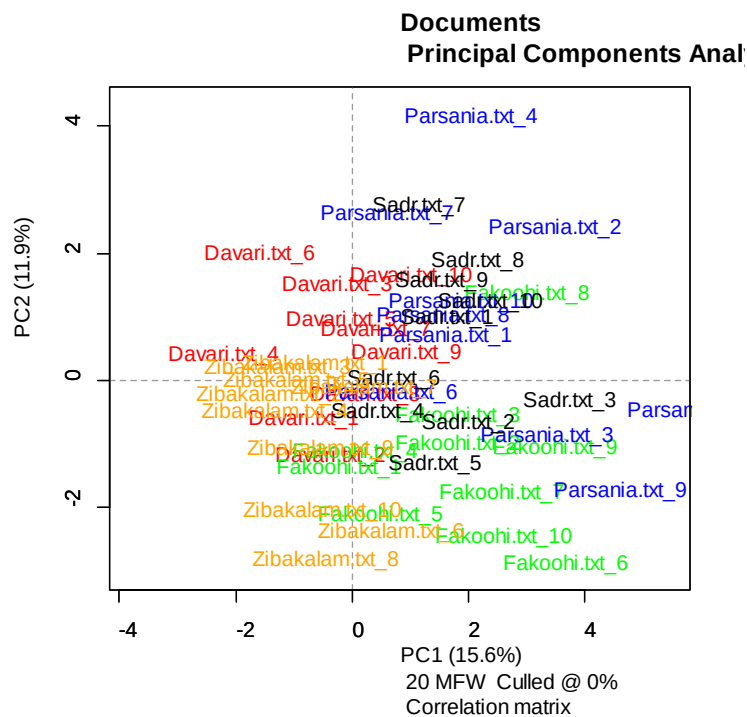


شکل ۴-۳. تحلیل خوشه‌ای تک‌نگاشتی در متون ۳۰۰۰ کلمه‌ای (ARI=0.73)

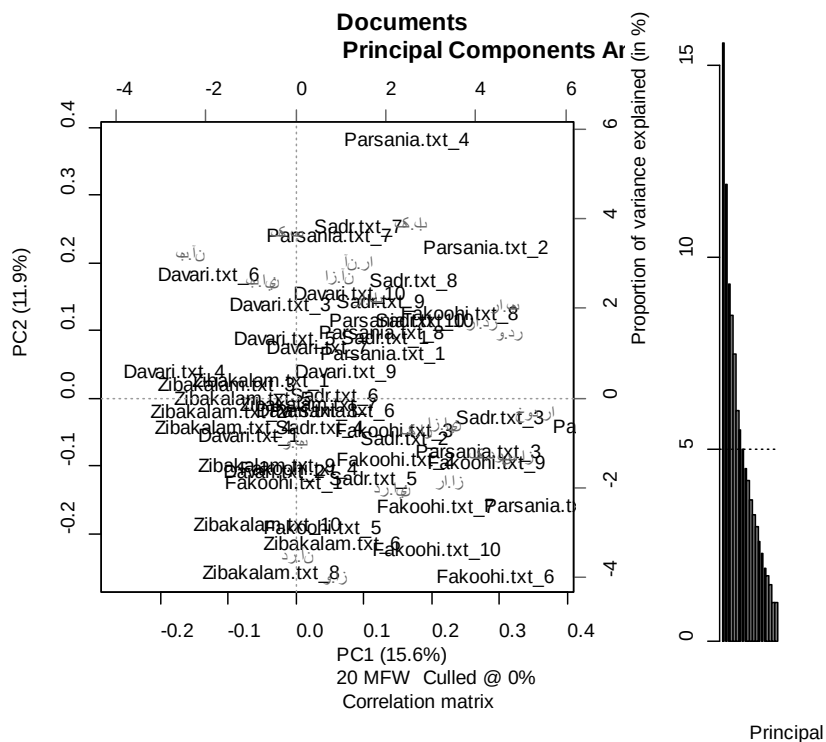
۲-۵-۴. تحلیل دونگاشتی

در تحلیل دونگاشتی همانند آزمایش‌های پیشین، توالی‌های دوتایی پربسامد از واژه‌های دستوری استخراج شدند و تعداد ۲۰ عدد با گزینش ۷۰٪ انتخاب شدند. شکل ۴-۳۸، نمودار تحلیل مولفه‌های اصلی واژه‌های دونگاشتی را نشان می‌دهد. در مولفه‌ی اصلی ۱ و ۲ در مجموع قادرند حدود ۲۷٪ از پراکنش داده‌ها را توضیح دهند. عدم تفکیک واضح نویسنده‌ها در این نمودار احتمالاً به دلیل ضعف دو مولفه‌ی اصلی در تبیین عمده‌ی پراکنش در داده‌هاست. با بررسی شکل ۴-۳۹ این مسئله تایید می‌شود. در شکل ۴-۳۹ که مولفه‌های دیگر نیز در سمت راست نمودار نشان داده شده‌اند، سطح مشارکت ۷ مولفه در پراکنش داده‌ها، هر کدام بیش از ۵٪ بوده است.

شناسایی هویت نویسنده بر اساس زبان فردی □ ۶۵



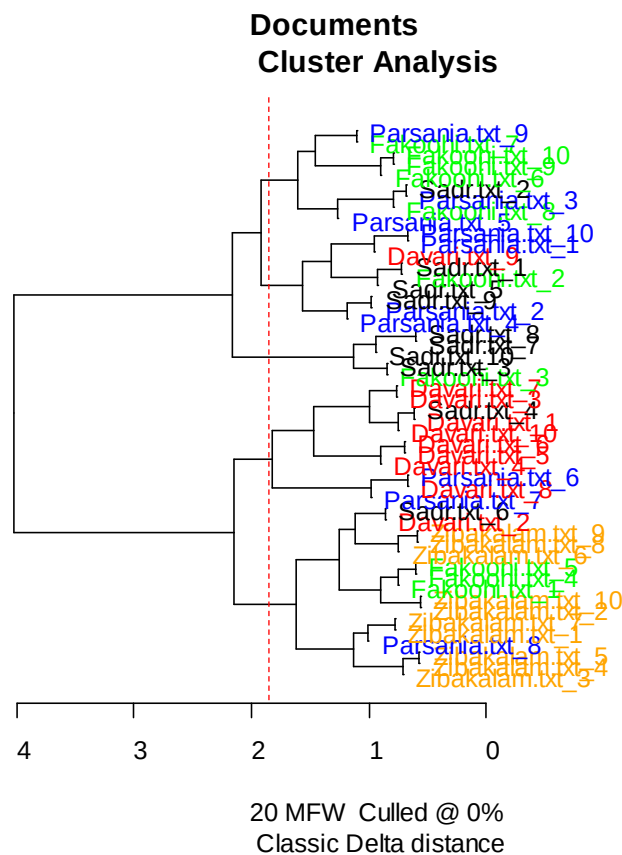
شکل ۴-۳۸. تحلیل مولفه‌های اصلی دونگاشتی در متون ۳۰۰۰ کلمه‌ای



شکل ۴-۳۹. تحلیل مولفه‌های اصلی دونگاشتی به همراه بار عاملی در متون ۳۰۰۰ کلمه‌ای

شناسایی هویت نویسنده بر اساس زبان فردی □ ۶۶

برای دستیابی به تصویری روشن تر از شباهت متون تحلیل شده، تحلیل خوشه‌ای بر روی نمونه‌نوشته انجام شد. شکل ۴-۴۰ نتایج این تحلیل را بر اساس مقیاس دلتا نشان می‌دهد. طبق شکل به نظر می‌رسد متون مربوط به داوری و زیباکلام اکثریت متون مربوط به دو خوشه‌ی پایین را تشکیل می‌دهند و پراکندگی در باقی خوشه‌ها بیشتر به نظر می‌رسد. محاسبه‌ی شاخص تعدیل شده‌ی رند، عدد ۰/۳ را به عنوان دقت این خوشه‌بندی به دست می‌دهد.

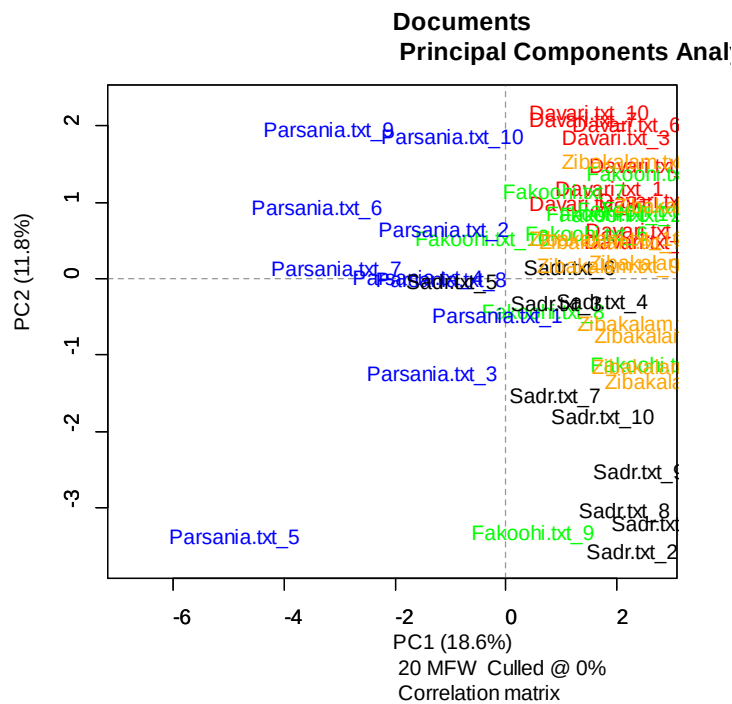


شکل ۴-۴۰. تحلیل خوشه‌ای دوناگشتی در متون ۳۰۰۰ کلمه‌ای (ARI=0.3)

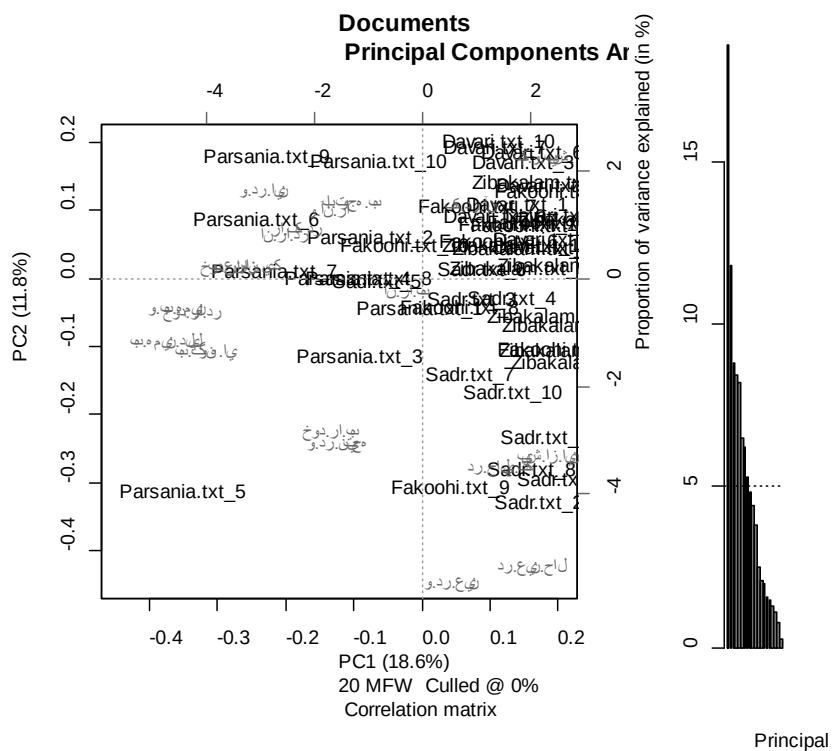
۳-۶-۴. تحلیل سه‌نگاشتی

تحلیل توالی‌های سه‌تایی از واژه‌های دستوری همانند آزمایش‌های پیشین در متون ۳۰۰۰ کلمه‌ای انجام شد. به دلیل پایین بودن حجم متون، گزینش ۳۰٪ بر روی واژه‌های دستوری پربسامد اعمال شد. شکل ۴-۴۱ نتایج تحلیل مولفه‌های اصلی متون با توالی‌های سه‌تایی واژه‌های دستوری را نشان می‌دهد. همچنین در شکل ۴-۴۲ مشارکت مولفه‌های دیگر در مقدار پراکنش و بار عاملی نشان داده شده است.

شناسایی هویت نویسنده بر اساس زبان فردی □ ۶۷



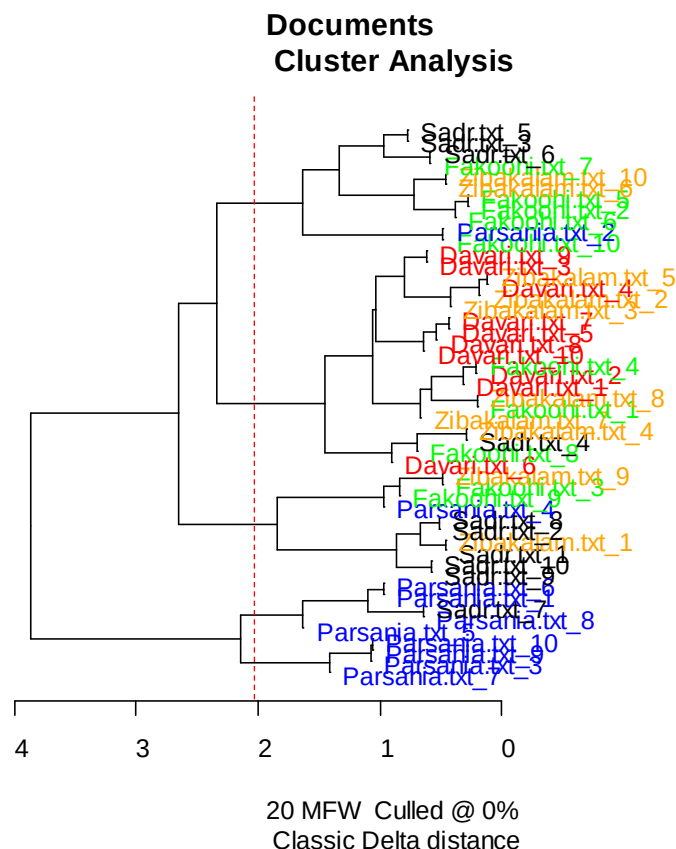
شکل ۴-۴۱. تحلیل مولفه‌های اصلی سه‌نگاشتی در متون ۳۰۰۰ کلمه‌ای



شکل ۴-۴۲. تحلیل مولفه‌های اصلی سه‌نگاشتی به همراه بار عاملی در متون ۳۰۰۰ کلمه‌ای

شناسایی هویت نویسنده بر اساس زبان فردی □ ۶۸

طبق نمودارهای فوق، مولفه‌های اصلی ۱ و ۲ در مجموع حدود ۳۰٪ از پراکنش داده‌ها را توضیح می‌دهند. در نهایت، تحلیل خوشه‌های متون نویسنده‌ها ساختار درختی شکل ۴-۴۳ را به دست می‌دهد. محاسبه‌ی شاخص تعدیل‌شده‌ی رند، مقدار ۰/۲۱ دقت را برای تحلیل خوشه‌ای واژه‌های دستوری سه‌نگاشتی به دست می‌دهد.



شکل ۴-۴۳. تحلیل خوشه‌ای سه‌نگاشتی در متون ۳۰۰۰ کلمه‌ای (ARI=0.21)

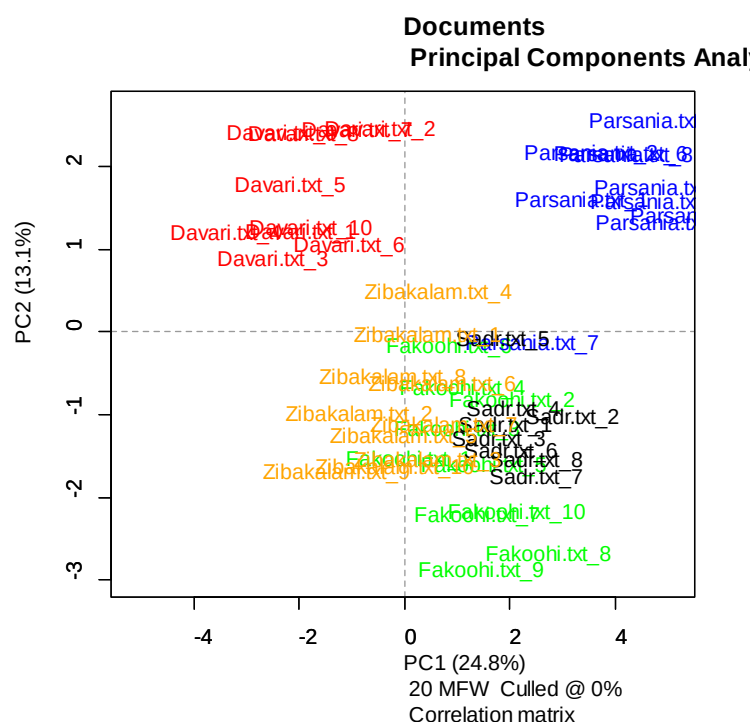
۴-۷. آزمایش ۶ (متون ۴۰۰۰ کلمه‌ای)

از آنجایی که در آزمایش چهارم، دقت ۱۰۰٪ در طبقه‌بندی متون نویسنده‌ها و در هریک از سطوح N-نگاشتی تحلیل به دست نیامد، مجدداً ۱۰۰۰ کلمه به حجم نمونه‌متن‌ها اضافه شد تا تعداد کلمات در متون به ۴۰۰۰ واژه برسد. هر یک از نمونه‌متن‌های ۴۰۰۰ واژه‌ای از پیکره‌ای ۴۰۰۰۰ کلمه‌ای استخراج شدند که خود این پیکره‌ی اخیر یا به‌طور تصادفی از پیکره‌های مادر نمونه‌گیری شده بودند (در مورد

متون زیباکلام، پارسانیا و فکوهی) یا کل متون نویسنده را شامل می‌شدند (در مورد صدر و داوری که پیکره‌هایشان کم حجم بود).

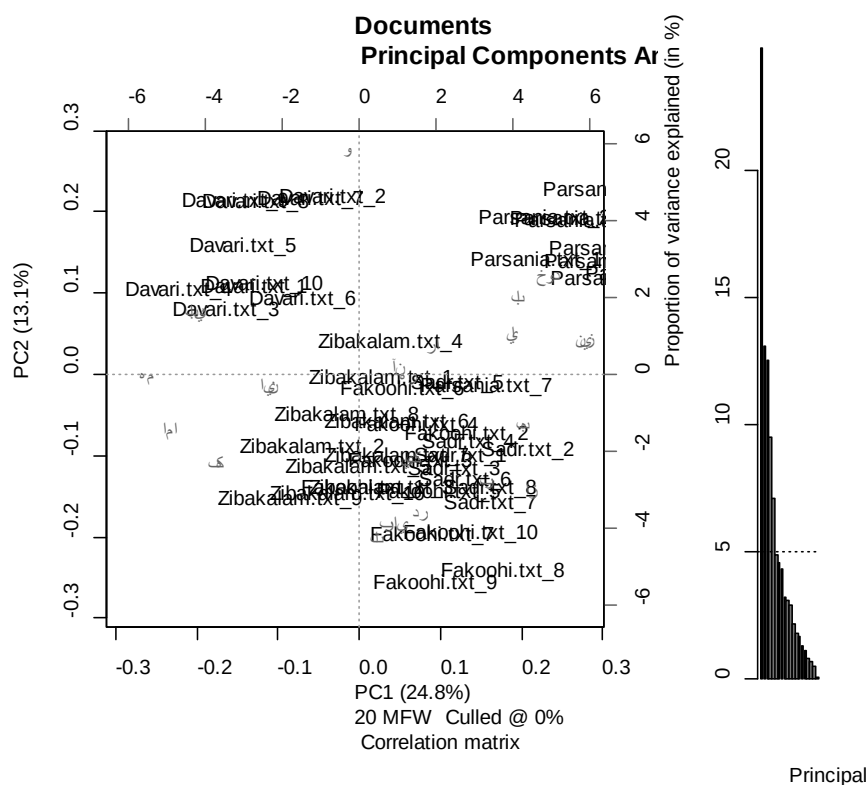
۱-۶-۴. تحلیل تک‌نگاشتی

در تحلیل تک‌نگاشتی، همانند آزمایش‌های پیشین پربسامدترین واژه‌های دستوری استخراج و با گزینش ۸۰٪ انتخاب شدند. شکل ۴-۴۴ نمودار تحلیل مولفه‌های اصلی متون را نشان می‌دهد. دو مولفه اصلی ۱ و ۲ در مجموع حدود ۳۸٪ از پراکنش داده‌ها را توضیح می‌دهند. بررسی نمودار نشان می‌دهد که در سطح ۴۰۰۰ کلمه، واژه‌های دستوری تفکیک نسبتاً بهتری از متون به‌دست داده‌اند و عمده‌ی متون نویسنده‌ها در مناطق نزدیک به هم قرار گرفته‌اند. در شکل ۴-۴۵ نیز که اطلاعات مربوط به بار عاملی و مولفه‌های دیگر نشان داده شده، تعداد ۵ مولفه هر کدام بیش از ۵٪ مشارکت در پراکنش داده‌ها داشته‌اند. تحلیل خوشه‌ای متون نویسنده‌ها در سطح ۴۰۰۰ واژه و بر اساس مقیاس دلتا، شکل ۴-۴۶ را به‌دست می‌دهد. بررسی این نمودار نشان می‌دهد که نمونه‌متن‌های هر نویسنده در خوشه‌ی مربوط به خود جای گرفته و دقت خوشه‌بندی ۱۰۰٪ است.

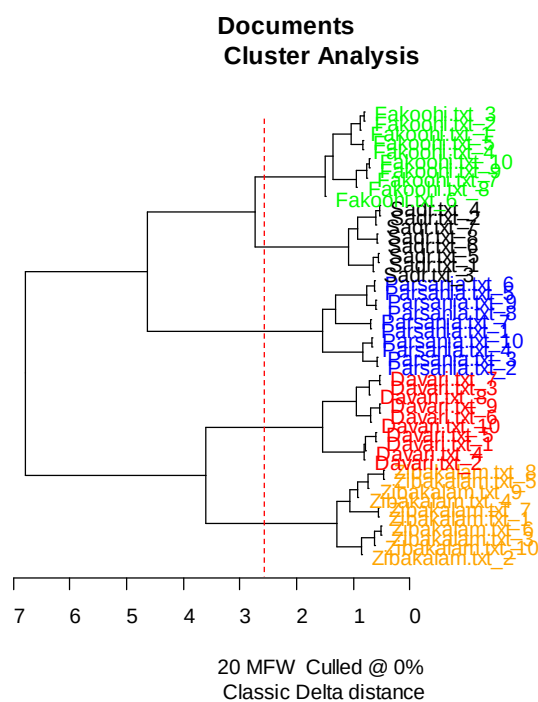


شکل ۴-۴۴. تحلیل مولفه‌های اصلی تک‌نگاشتی در متون ۴۰۰۰ کلمه‌ای

شناسایی هویت نویسندگان بر اساس زبان فردی □۷۰

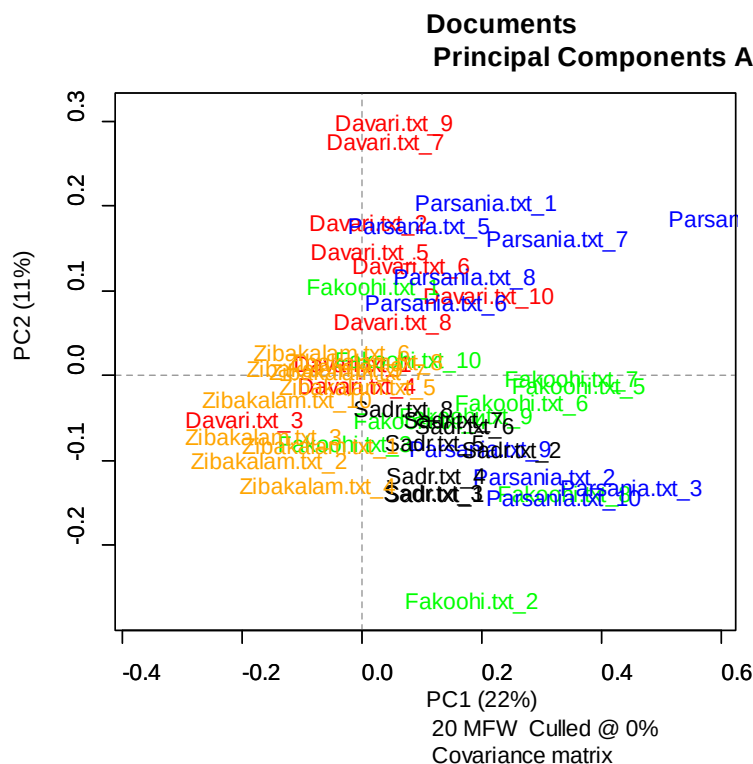


شکل ۴-۴. تحلیل مولفه‌های اصلی تک‌نگاشتی به همراه بار عاملی در متون ۴۰۰۰ کلمه‌ای

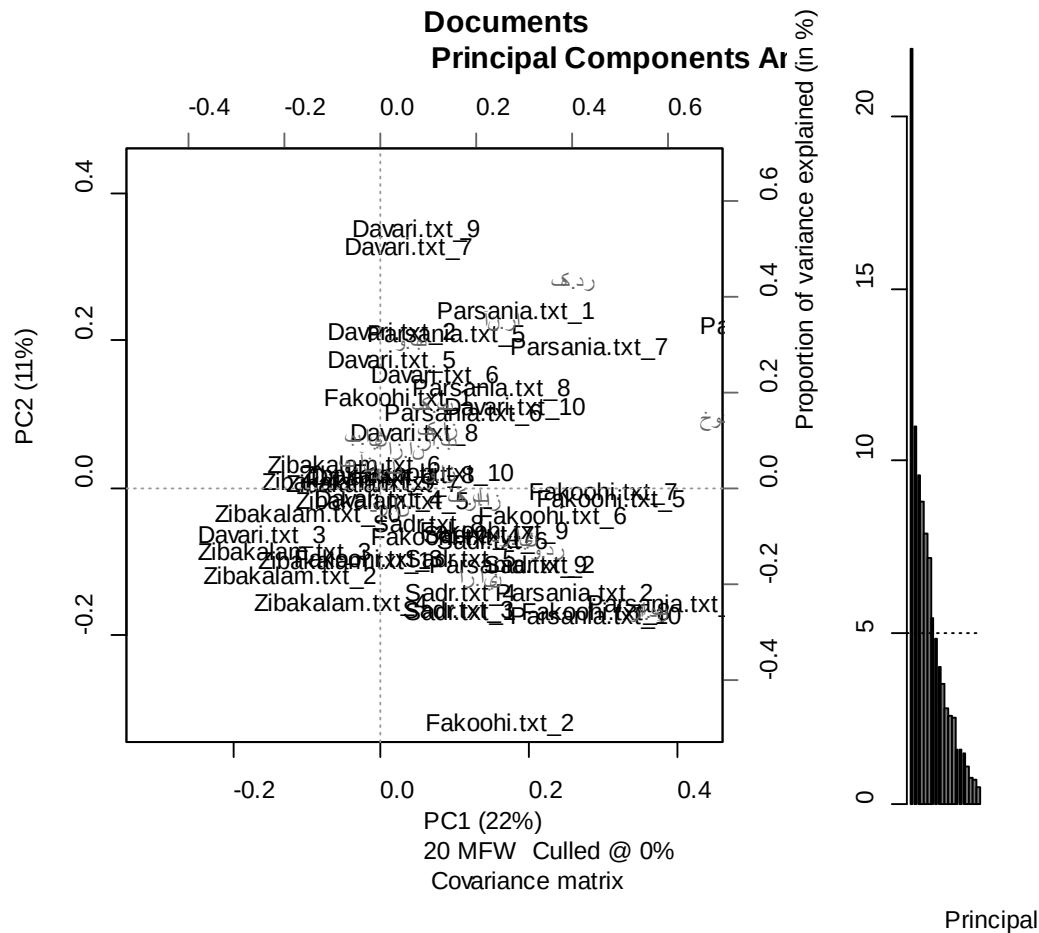


شکل ۴-۴. تحلیل خوشه‌ای تک‌نگاشتی در متون ۴۰۰۰ کلمه‌ای (ARI=1)

در تحلیل دونگاشتی نیز با توجه به افزایش نسبی حجم متون نسبت به آزمایش قبلی، از گزینش ۸۰٪ برای حذف واژه‌های دستوری دارای بسامد صفر در متون استفاده شد. شکل ۴-۴۷ نمودار تحلیل مولفه‌های اصلی برای واژه‌های دستوری دونگاشتی را نشان می‌دهد. دو مولفه اصلی ۱ و ۲ روی هم رفته ۳۳٪ از پراکنش داده‌ها را توضیح می‌دهند. بررسی نمودار نشان می‌دهد که ظاهراً در سطح دو مولفه، تفکیک متون چندان دارای تمایز مشهود نیست. در شکل ۴-۴۸، اطلاعات مولفه‌های اصلی دیگر و همچنین بار عاملی مشاهده می‌شود. طبق این شکل، تعداد ۷ مولفه اصلی هر کدام مشارکتی بیش از ۵٪ در توضیح پراکنش داده‌ها داشته‌اند.

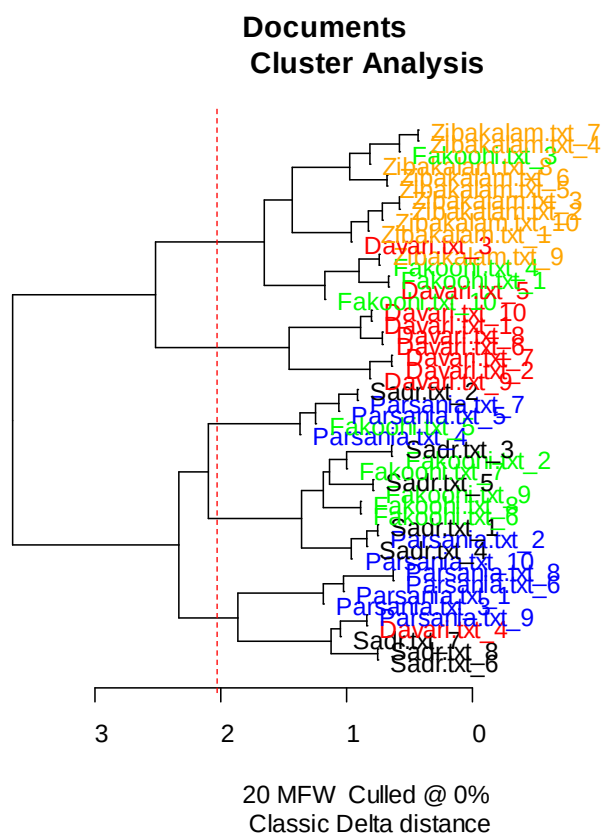


شکل ۴-۴۷. تحلیل مولفه‌های اصلی دونگاشتی در متون ۴۰۰۰ کلمه‌ای



شکل ۴-۴. تحلیل مولفه‌های اصلی دونگاشتی به همراه بار عاملی در متون ۴۰۰۰ کلمه‌ای

در نهایت، شکل ۴-۴۹ نمودار درختی حاصل از تحلیل خوشه‌ای داده‌ها را بر اساس مقیاس فاصله‌ای دلتا نشان می‌دهد. بررسی نمودار درختی نشان می‌دهد که اکثر متون زیباکلام، پارسانیا و داوری در خوشه‌های مجزا قرار گرفته‌اند و متون دیگر نویسنده‌ها پراکندگی بیشتری از خود نشان می‌دهند. محاسبه‌ی شاخص تعدیل‌شده‌ی رند، دقت ۰/۳۴ را برای این خوشه‌بندی به دست می‌دهد.

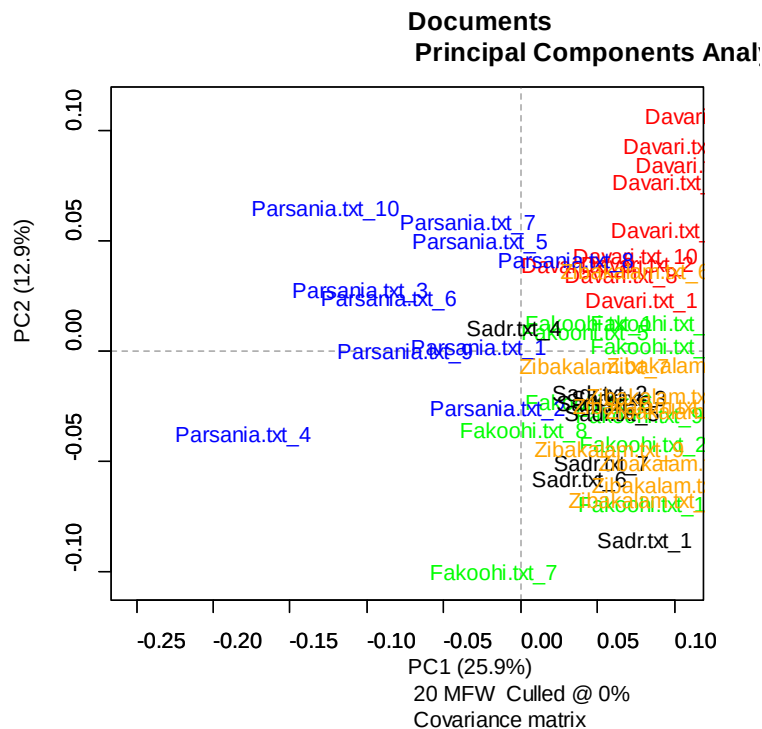


شکل ۴-۴. تحلیل خوشه‌ای دوناگشتی در متون ۴۰۰۰ کلمه‌ای (ARI=0.34)

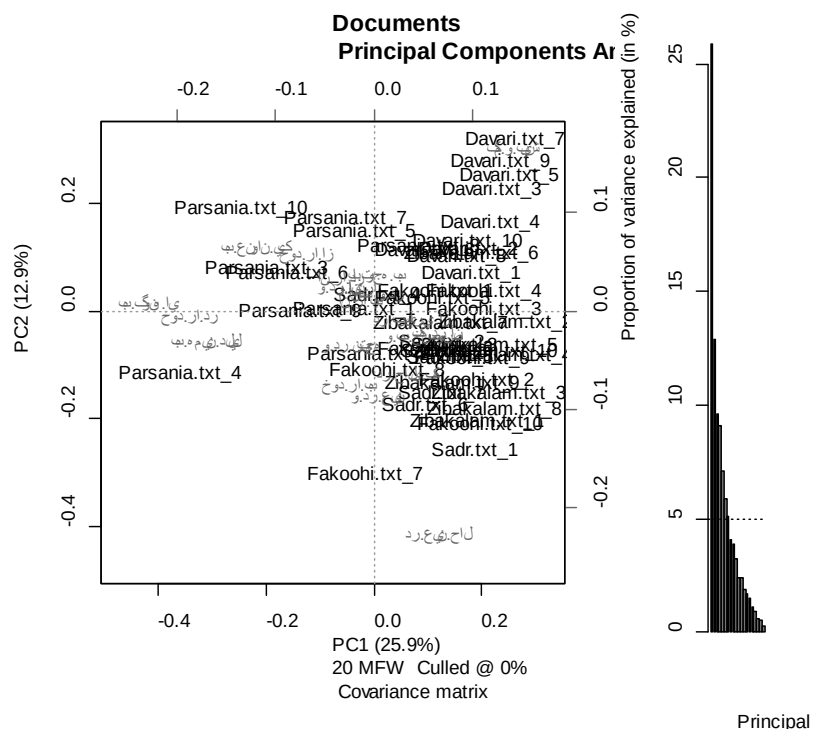
۳-۶-۴. تحلیل سه‌نگاشتی

در تحلیل توالی‌های سه‌تایی واژه‌های دستوری، از گزینش ۳۰٪ استفاده شد. شکل ۴-۵۰ نمودار تحلیل مولفه‌های اصلی بر اساس واژه‌های دستوری سه‌نگاشتی را نشان می‌دهد. دو مولفه اصلی ۱ و ۲ در مجموع قادرند حدود ۳۹٪ از پراکنش داده‌ها را توضیح دهند. بر اساس اطلاعات شکل ۴-۵۱ که تحلیل مولفه‌های اصلی را به همراه بار عاملی و مولفه‌های اصلی دیگر نشان می‌دهد، ۷ مولفه اصلی هر کدام بیش از ۵٪ از پراکنش داده‌ها را توضیح می‌دهند.

شناسایی هویت نویسندگان بر اساس زبان فردی



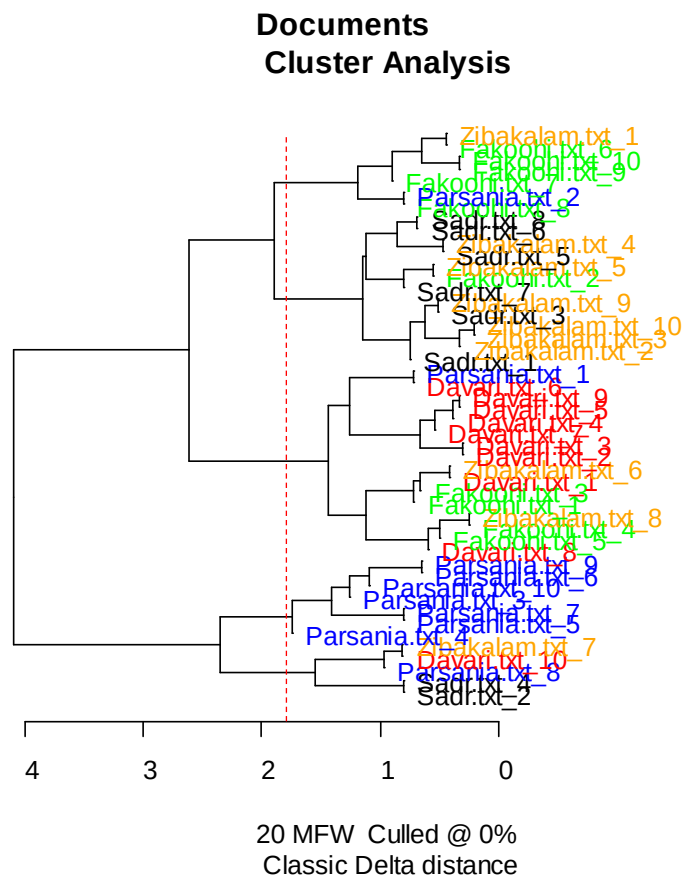
شکل ۴-۵. تحلیل مولفه‌های اصلی سه‌نگاشتی در متون ۴۰۰۰ کلمه‌ای



شکل ۴-۵۱. تحلیل مولفه‌های اصلی سه‌نگاشتی به همراه بار عاملی در متون ۴۰۰۰ کلمه‌ای

شناسایی هویت نویسنده بر اساس زبان فردی □ ۷۵

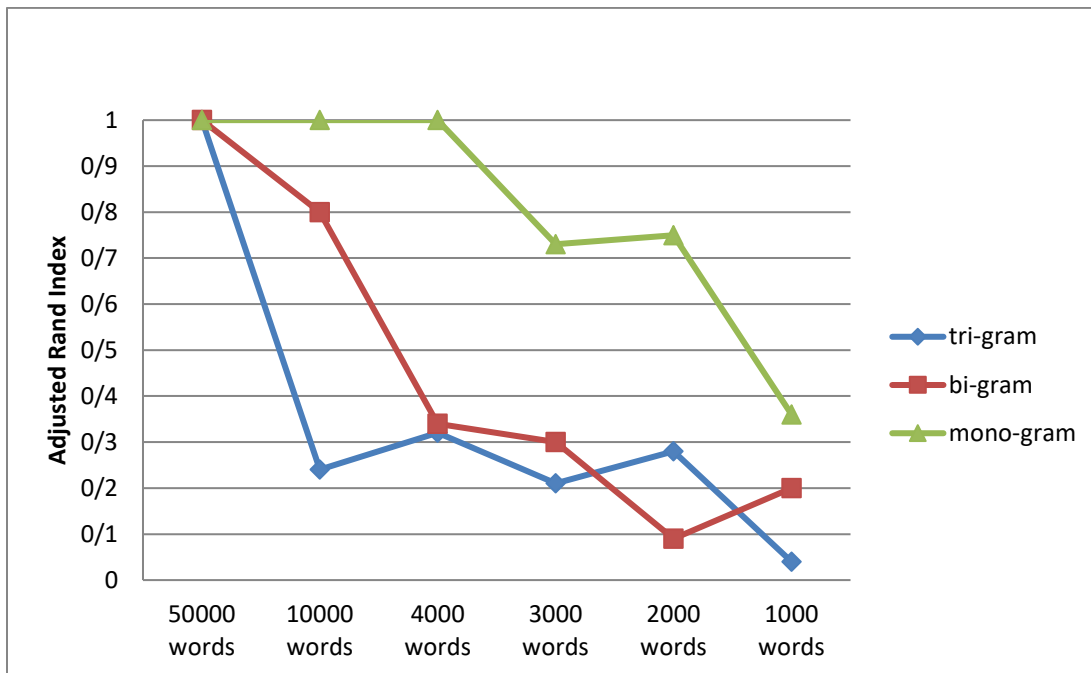
به منظور دسته بندی متون نویسندگان، تحلیل خوشه ای بر اساس مقیاس فاصله ای دلتا بر روی داده ها انجام شد. نتایج خوشه بندی در شکل ۴-۵۲ نشان داده شده است. محاسبه ی شاخص تعدیل شده ی رند برای این خوشه بندی، مقدار ۰/۳۲ را به دست می دهد که مقدار مطلوبی نیست.



شکل ۴-۵۲. تحلیل خوشه ای سه نگاشتی در متون ۴۰۰۰ کلمه ای (ARI=0.32)

۴-۷. جمع بندی آزمایش ها

در این بخش قصد داریم نتایج آزمایش ها را ادغام کرده و جمع بندی ای از آنها ارائه کنیم. با کنار هم قراردادن آزمایش هایی که هر کدام بر روی مقدار متفاوتی از داده های متنی انجام شد می توان به تصویری روشن تر از کارایی مقیاس دلتا در طبقه بندی متون نویسندگان این پژوهش به دست آورد. شکل ۴-۵۳، نمودار تغییرات شاخص تعدیل شده ی رند را برای واژه های N نگاشتی در حجم های متنی آزمایش شده در این پژوهش نشان می دهد.



شکل ۴-۵۳. نمودار n.نگاشتی تغییرات شاخص تعدیل‌شده‌ی ژند بر اساس تعداد کلمات متون

طبق شکل ۴-۵۳، مشاهده می‌شود که به‌طور کلی با کاهش تعداد کلمات متون، دقت دسته‌بندی متون متعلق به یک نویسنده در گروه مربوطه که ما آن را دقت شناسایی نویسنده می‌نامیم کاهش می‌یابد. اما به‌نظر می‌رسد واژه‌های تک‌نگاشتی، دونگاشتی و سه‌نگاشتی در این میان اندکی متفاوت عمل می‌کنند. نخست اینکه عملکرد واژه‌های تک‌نگاشتی (یا همان واژه‌های متعارف متنی) بهترین عملکرد را نسبت به واژه‌های دونگاشتی و سه‌نگاشتی داشته است. در سه سطح واژه‌ای ۵۰۰۰۰ کلمه، ۱۰۰۰۰ کلمه و ۴۰۰۰ کلمه، دقت واژه‌های تک‌نگاشتی در شناسایی متون ۱۰۰٪ است و تغییر نمی‌کند. نتایج نشان می‌دهد که متون ۴۰۰۰ کلمه‌ای احتمالاً حد کمینه‌ی متونی است که تماماً به‌درستی شناسایی می‌شوند. البته چنانچه آستانه‌ی دقت را تا حد ۰٫۷ پایین بیاوریم، آنگاه می‌توان گفت که تا حد ۲۰۰۰ کلمه نیز عملکرد الگوریتم با استفاده از واژه‌های تک‌نگاشتی پذیرفتنی است.

نکته‌ی دوم اینکه، در میان واژه‌های دونگاشتی و سه‌نگاشتی، به‌نظر می‌رسد به‌طور کلی عملکرد واژه‌های دونگاشتی (به استثناء متون ۲۰۰۰ کلمه‌ای) در همه‌ی اندکی بهتر از واژه‌های سه‌نگاشتی بوده است. نکته‌ی دیگری که در عملکرد واژه‌های دستوری سه‌نگاشتی بیشتر مشهود است این است که با کاهش تعداد متون از ۵۰۰۰۰ به ۱۰۰۰ واژه، شاهد نوسان در دقت شناسایی نویسنده هستیم. مثلاً در

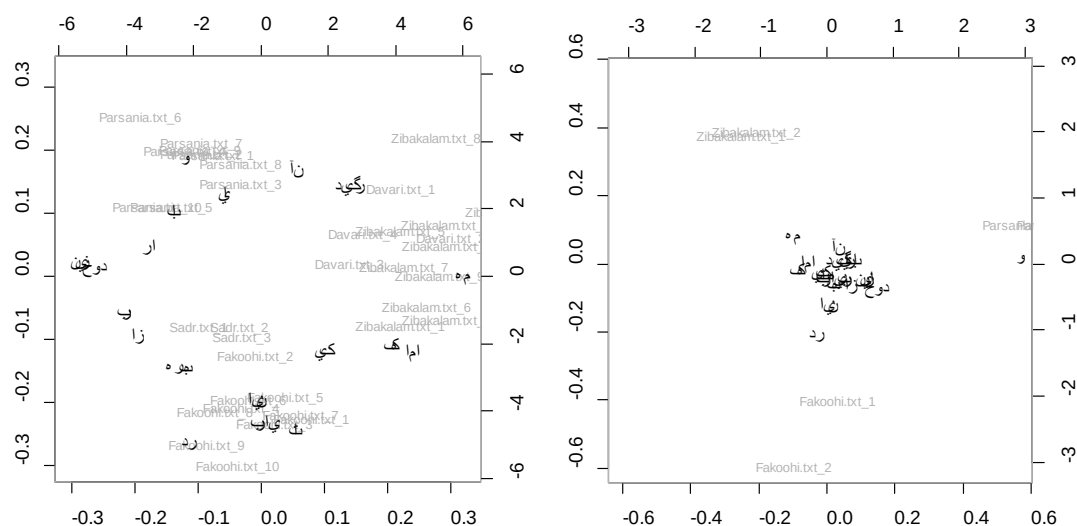
شناسایی هویت نویسنده بر اساس زبان فردی □ ۷۷

سطح متون ۴۰۰۰ کلمه‌ای، با آنکه تعداد متون نسبت به سطح ۱۰۰۰۰ کلمه‌ای بیش از نصف کاهش یافته است، اما دقت شناسایی اندکی بیشتر بوده است که البته به نظر می‌رسد معنی‌دار نیست.

یکی از دلایلی که می‌توان بر رفتار متفاوت واژه‌های n -نگاشتی و افت عملکرد شناسایی نویسنده با حرکت از واژه‌های تک‌نگاشتی تا سه‌نگاشتی عنوان کرد، کاهش تعداد کلمات متون و به دنبال آن کاهش رخداد توالی‌های واژگانی در پیکره است. در واقع با کاهش تعداد کلمات در پیکره‌ی متون، احتمال اینکه توالی‌های چندگانه در متن تکرار شوند، کاهش بیشتری می‌یابد. اگرچه توالی‌های چندگانه‌ی واژگانی، نشانگرهای دقیق‌تری از روال‌های خودکارشده‌ی واژگانی می‌توانند باشند، اما کاهش تعداد کلمات متن شناسایی آنها را با مشکل مواجه می‌سازد. از این رو، عملکرد مطلوب توالی‌های واژگانی در متون حجیم می‌تواند بروز کند.

۸-۴. تحلیل‌های پیکره‌ای

در بخش آزمایشات، نقش واژه‌های دستوری پربسامد را در تفکیک نویسنده‌های متون فارسی بررسی کردیم. در تحلیل‌هایی که انجام شد، تحلیل مولفه‌های اصلی، تصویری روشن از واژه‌هایی که بیشترین تاثیر را در تمایز سبک‌ها داشته‌اند ارائه کرد. در این بخش قصد داریم، با مرور برخی نمودارهای مولفه‌های اصلی و بررسی‌های پیکره‌ای نگاهی عمیق‌تر به واژه‌های دستوری بیافکنیم که سبک نویسندگان را از یکدیگر جدا کرده است.



شکل ۴-۵۴. نمودار بار عاملی. سمت راست: تک‌نگاشتی (۵۰۰۰۰ کلمه)؛ سمت چپ: تک‌نگاشتی (۱۰۰۰۰ کلمه)

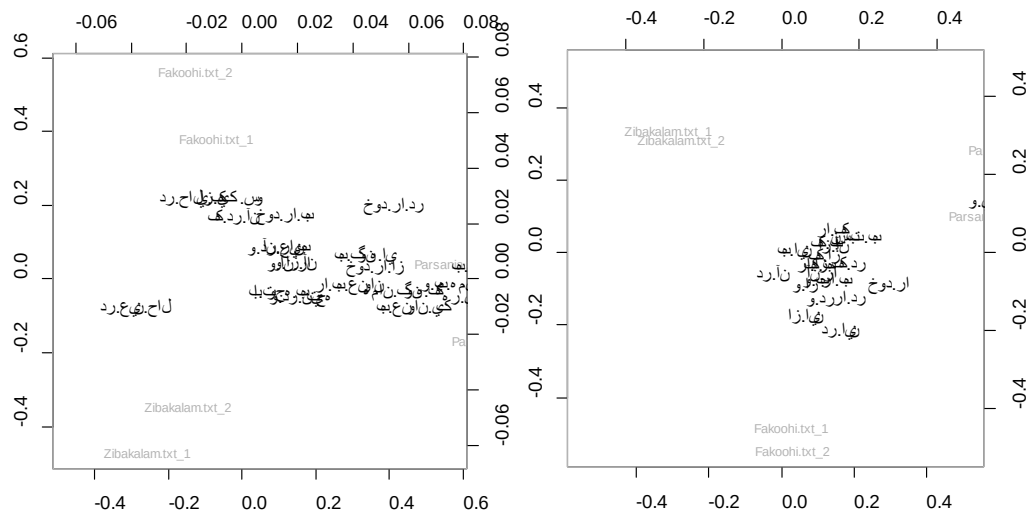
در شکل ۴-۵۴، دو نمودار بار عاملی که در بخش‌های قبل تحلیل شدند، نشان داده شده است. نمودار سمت راست مربوط به واژه‌های دستوری تک‌نگاشتی در پیکره‌های ۵۰۰۰۰ کلمه‌ای ۳ نویسنده است. در این نمودار ملاحظه می‌شود که واژه‌ی دستوری «و» در متون پارسانیا بیش از دیگران به کار رفته است. بررسی نزدیکی سایر واژه‌های دستوری به نویسندگان دیگر نشان می‌دهد که واژه‌ی «هم» در متون زیباکلام بیشتر به کار رفته و دو واژه‌ی «در» و «این» به متون فکوهی (به‌ویژه متن شماره ۱) نزدیک‌تر است.

در نمودار سمت چپ شکل ۴-۵۴، هر یک از واژه‌های دستوری پربسامد در متون ۱۰۰۰۰ کلمه‌ای نشان داده شده‌اند و بهتر از یکدیگر قابل تفکیک هستند. در این نمودار نیز مشاهده می‌شود که واژه‌ی «و» در مرکز متون پارسانیا قرار گرفته است. به‌علاوه واژه‌های «یا»، «با»، «را»، «خود» و «نیز» هم در نزدیکی متون این نویسنده قرار دارند. در قسمت پایین، متون صدر و فکوهی قرار دارند. مجدداً واژه‌های «در» و «این» در میان متون فکوهی به چشم می‌خورد که واژه‌های دیگری نظیر «برای» و «تا» نیز آن را همراهی می‌کنند. در نهایت متون زیباکلام در سمت راست نمودار دیده می‌شود که از متون داوری به‌وضوح قابل تفکیک نیست. با این حال، واژه‌ی «هم» مجدداً در نزدیکی متون زیباکلام مشاهده می‌شود.

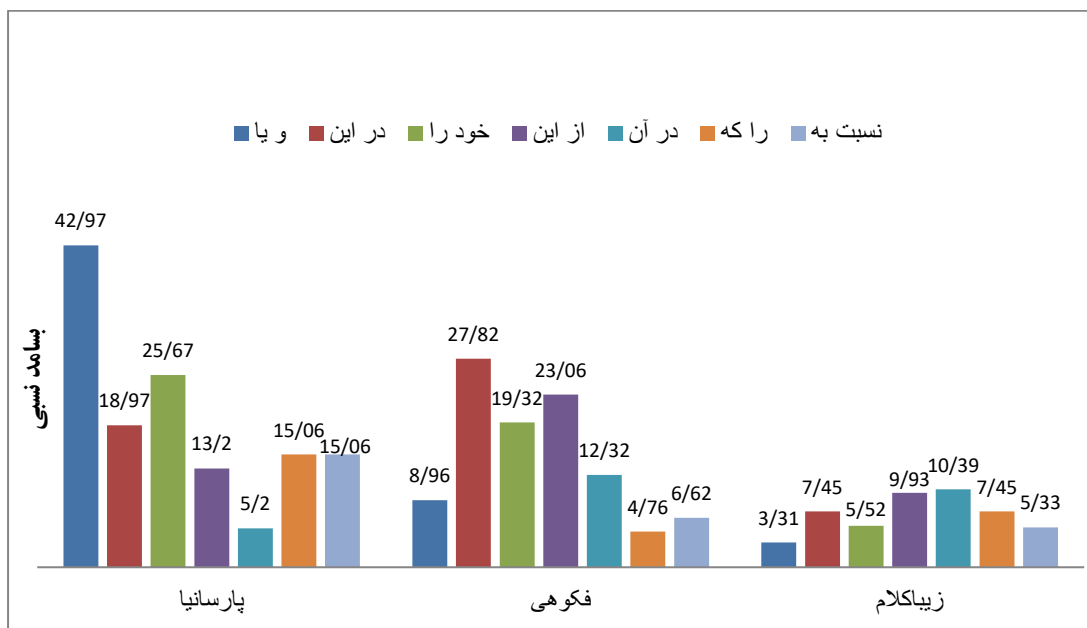
از آنجایی که واژه‌های دستوری متعارف (یا تک‌نگاشتی) اطلاعات زیادی را از نحوه‌ی به کارگیری این واژه‌ها در جملات و بافت دستوری آن‌ها به دست نمی‌دهند، بررسی توالی‌های دونگاشتی و سه‌نگاشتی پربسامد در مرحله‌ی بعد می‌تواند اطلاعات بیشتری در این راستا به دست دهد. بررسی شکل ۵۵-۴ که نمودار بار عاملی واژه‌های دونگاشتی و سه‌نگاشتی در متون ۵۰۰۰۰ کلمه‌ای است، اطلاعات جالب توجهی را در اختیارمان قرار می‌دهد. نمودار واژه‌های دونگاشتی نسبت به نمودار سه‌نگاشتی متراکم‌تر است اما با این حال می‌توان دونگاشتی‌های متمایزکننده‌ی سبک را تا حدودی مشخص کرد. مثلاً توالی «و یا» بیشترین پراکنش را در سطح مولفه‌ی نخست (محور افقی) توضیح می‌دهد و در متون پارسانیا به‌طور چشمگیر به کار رفته است. بعد از این توالی، دونگاشتی «خود را»، «را در» و «در آن» بیشترین پراکنش را در سطح مولفه‌ی نخست توضیح می‌دهد. توالی‌های «در این» و «از این» بیشترین پراکنش را در سطح مولفه‌ی اصلی دوم (محور عمودی) توضیح می‌دهد و به نظر می‌رسند عمده‌تاً مشخصه‌ی متون فکوهی بوده‌اند. بررسی پیکره‌ای این دو دونگاشتی و جستجوی هم‌آیندهای موثر آنها نشان می‌دهد که این دو دونگاشتی در عباراتی نظیر «از این رو» و «از این لحاظ» و نیز در عبارات «در این زمینه» و «در این حالت» بیشترین بسامد را داشته‌اند. همچنین توالی‌های «را که» و «نسبت به» در

شناسایی هویت نویسنده بر اساس زبان فردی □ ۷۹

متون پارسانیا بیشتر از فکوهی و زیباکلام به کار رفته‌اند. شکل ۴-۵۶ بسامدهای نسبی برخی از دونگاشتی‌های فوق را نشان می‌دهند که از پیکره‌ی متنی این نویسندگان استخراج شده‌اند.



شکل ۴-۵۵. نمودار بار عاملی. سمت راست: دونگاشتی (۵۰۰۰۰ کلمه)؛ سمت چپ: سه‌نگاشتی (۵۰۰۰۰ کلمه)

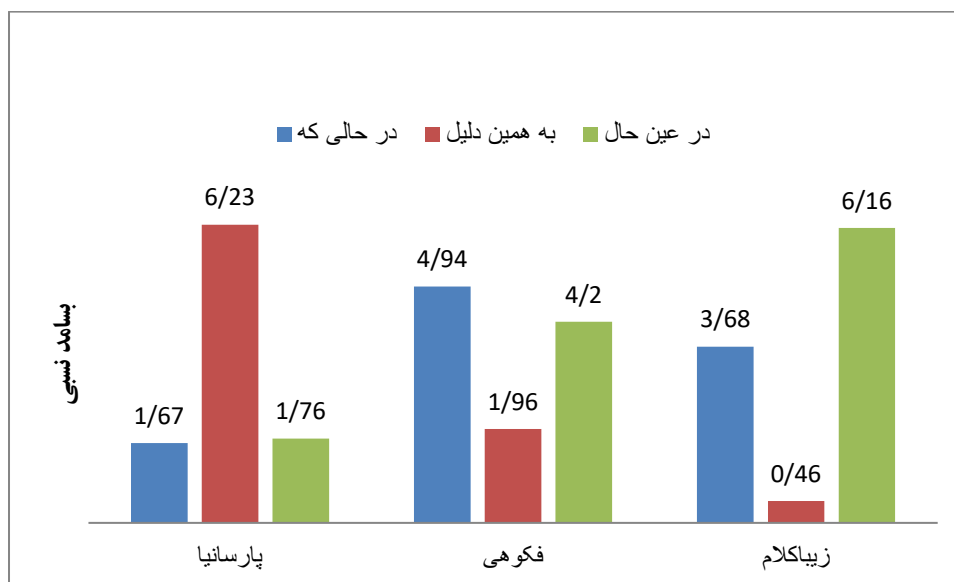


شکل ۴-۵۶. بسامد نسبی برخی دونگاشتی‌های پرسامد ممیز سبکی در پیکره‌ی متنی سه نویسنده

نمودار سمت چپ در شکل ۴-۵۵، بارهای عاملی مربوط به واژه‌های دستوری سه‌نگاشتی را نشان می‌دهد که الگوهای دستوری روشن‌تری را نسبت به واژه‌های دونگاشتی در اختیار قرار می‌دهند. بر

شناسایی هویت نویسنده بر اساس زبان فردی □ ۸۰

اساس این نمودار می‌توان دریافت که در سطح مولفه‌ی اصلی اول (محور افقی) سه‌نگاشتی‌هایی نظیر «در عین حال»، «به همین دلیل»، «و به همین»، «هر یک از»، «همان گونه که»، «در حالی که»، «خود را در» و «به عنوان یک» بیشترین تاثیر را پراکندگی متون در طول این محور داشته‌اند. از میان این سه‌نگاشتی‌ها، برخی نظیر «در عین حال»، «در حالی که» و «به همین دلیل» را می‌توان جزو حروف ربطی مرکب دانست که علاوه بر صورت چند کلمه‌ای‌شان، ماهیت هم‌آیندی دارند و معنای واحدی را می‌رسانند. شکل ۴-۵۷ بسامدهای این سه حرف ربطی مرکب را در متون سه نویسنده نشان می‌دهد که از پیکره‌های مربوطه‌شان استخراج شده‌اند. حرف ربط «در عین حال» بیشترین بسامد را در متون زیباکلام و سپس در متون فکوهی دارد. حرف ربط «به همین دلیل» ممیزه سبکی متون پارسانیا است. همچنین حرف ربط «در حالی که» بیشتر در متون فکوهی، سپس در متون زیباکلام و کمتر در متون پارسانیا به چشم می‌خورد.



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

بر اساس اطلاعات شکل ۴-۵۷ و همچنین مشاهدات پیکره‌ای، ساختار گفتمانی علت و معلولی که به وسیله‌ی حرف ربط «به همین دلیل» بیان شده، در متون پارسانیا غلبه داشته است. در حالی که سبک گفتمانی در متون زیباکلام و فکوهی بیشتر در راستای تقابل آراء و تضارب ایده‌های ناسازگار پیش رفته و در قالب حرف ربط «در حالی که» بیان شده است. این معنای گفتمانی تا حدودی در حرف ربط «در عین حال» نیز دیده می‌شود که مشخصه‌ی بارز متون زیباکلام است. در متون زیباکلام این

شناسایی هویت نویسنده بر اساس زبان فردی □ ۸۱

حرف ربط گاهی در معنای «همزمان» یا «در همین حال» ظاهر شده و گاهی ایده‌های ناسازگار با مطالب پیش گفته را نشانگذاری کرده است.

فصل پنجم

نتیجه گیری

۵-۱. مقدمه

در این پژوهش موضوع شناسایی نویسنده بر اساس زبان فردی و بر اساس تحلیل پیکره‌ای مورد تحلیل قرار گرفت. زبان فردی بر اساس واژه‌های دستوری عملیاتی شد و سپس با استخراج واژه‌های دستوری پربسامد با توالی‌های یک تا سه‌نگاشتی، کارایی آنها در تفکیک متون نویسنده‌ها بررسی شد. در این بخش، به جمع‌بندی نتایج پژوهش پرداخته و به سوالات و فرضیه‌های مطرح‌شده در تحقیق خواهیم پرداخت. در انتها نیز پیشنهاداتی برای پژوهش‌های آتی ارائه خواهد شد.

۵-۲. پاسخ به پرسش‌های پژوهش

در ابتدای این طرح سه پرسش به عنوان مسئله محوری این پژوهش طرح شد که عبارت بودند از:

الف) واژه‌های دستوری چگونه می‌توانند سرنخی از زبان فردی به دست دهند؟

ب) در تحلیل n.نگاشتی زبان فارسی، کدام مقادیر n بهتر می‌تواند زبان فردی نویسندگان مختلف را از یکدیگر تفکیک کند؟
ج) چگونه می‌توان هویت نویسنده را از بررسی نوشته‌ها شناسایی کرد؟

• پاسخ پرسش الف:

نتایج این پژوهش نشان می‌دهد که نویسندگان مختلف واژه‌های دستوری را به‌طور مشابه به کار نمی‌گیرند. در واقع با وجود اینکه برخی واژه‌های دستوری پربسامد توسط همه‌ی نویسندگان به کار گرفته می‌شوند، اما اولویت به کارگیری آنها توسط نویسندگان متفاوت است. همانطور که در بخش تحلیل‌های پیکره‌ای نیز به آن پرداخته شد، برخی واژه‌های دستوری تقریباً به‌طور کمابیش یکسانی توسط همه نویسندگان استفاده می‌شود، اما برخی دیگر با بسامد بیشتری در متون یک نویسنده نسبت به متون دیگر نویسندگان به کار گرفته می‌شوند.

در توضیح علت اولویت‌دهی نویسندگان مختلف به واژه‌های دستوری متفاوت، می‌توان دو دلیل را مد نظر قرار داد. دلیل نخست ریشه در زبان فردی گویشوران دارد. زبان فردی هر گویشوری در بردارنده‌های عادات زبانی وی است که در سطوح مختلف آوایی، واژگانی و نحوی زبان می‌تواند نمود پیدا کند. بنابراین، به کارگیری و اولویت‌دهی به برخی واژه‌های دستوری خاص می‌تواند مبتنی بر

الگوهایی خودکار شده و عادت‌وار باشد که شخص در طول سالیان در مواجهه با درونداد زبانی محیط خود آنها را درونی کرده است. نتایج به‌دست آمده در پژوهش کنونی در راستای این تبیین است.

دلیل محتمل دوم در اولویت‌دهی نویسنده‌های مختلف به واژه‌های دستوری متفاوت می‌تواند به دلیل نگارش متن در یک ژانر بخصوص باشد. در واقع می‌توان این احتمال را مدنظر داشت که ساختار گفتگمانی یک ژانر بخصوص بر گزینش واژه‌های دستوری خاص تاثیر بگذارد. برای مثال، بسیاری از پژوهش‌های سبک‌سنجی نشان داده است که به‌کارگیری ضمائر شخصی همبستگی معنی‌داری با برخی عوامل متنی و غیرمتنی نظیر ژانر، چشم‌انداز روایی نویسنده، جنسیت نویسنده و حتی موضوع متن دارد (Kestemont, 2014). با توجه به اینکه در پژوهش فعلی ضمائر شخصی و هر واژه‌ای که احتمال نوعی رابطه بین موضوع متن و واژه‌های مربوط وجود داشت از فهرست واژه‌های دستوری پربسامد حذف شدند، تفاوت موضوعی و ژانری نمی‌تواند تبیین مناسبی برای نتایج این پژوهش ارائه دهد. قضاوت درباره‌ی همبستگی احتمالی برخی واژه‌های دستوری (غیر از ضمائر شخصی) با ژانر و عوامل متنی و غیرمتنی نیاز بررسی بیشتر و دقیق‌تر را می‌طلبد.

• پاسخ پرسش ب:

با توجه به نتایجی که از خوشه‌بندی واژه‌های دستوری تک‌نگاشتی، دونگاشتی و سه‌نگاشتی به‌دست آمد، واژه‌های تک‌نگاشتی یا همان واژه‌های دستوری متعارف، بهترین کارایی را با دقت ۱۰۰٪ تا سطح ۴۰۰۰ واژه از خود نشان دادند (رجوع کنید به شکل ۴-۵۳). دقت واژه‌های دونگاشتی و سه‌نگاشتی با کاهش حجم متون، بلافاصله با افت کارایی در تفکیک صحیح متون مواجه شد.

با آنکه متونی با کمینه حجم ۴۰۰۰ واژه به‌عنوان آستانه‌ی بهترین عملکرد الگوریتم دلتا در متون فارسی معین شد، اما به‌نظر می‌رسد این آستانه ماهیت حدی ندارد و اتفاق نظری در مورد کمینه‌ی متون برای تحلیل مطمئن سبک‌شناختی وجود ندارد. برای مثال، برخی پژوهشگران ۵۰۰۰ واژه را به‌عنوان حد کمینه‌ی قابل اعتماد برای متون داستانی در انگلیسی و دیگر زبان‌ها تعیین کرده‌اند (Eder, 2013)، در حالی که برخی دیگر قائل به کمینه‌ی ۱۰۰۰ واژه (Holmes et al., 2001) به‌عنوان حد قابل اعتماد برای پژوهش‌های سبک‌سنجی هستند. با این حال، در غیاب شواهد مستدل برای زبان فارسی، فعلاً می‌توان کمینه‌ی ۴۰۰۰ واژه را به‌عنوان حد مطلوب برای الگوریتم دلتا و با مدنظر قراردادن تعداد ۲۰ واژه‌ی دستوری برای تفکیک متون پذیرفت.

• پاسخ پرسش ج:

هدف از طرح این پرسش، ارائه الگویی عملیاتی برای دسته‌بندی متونی است که در هویت نویسنده‌ی آن تردید داریم (در پژوهش‌های شناسایی سبک نویسنده، گزینه‌هایی از نویسنده‌های احتمالی باید در دست باشد تا متون مورد مناقشه با آثار آنها مورد مقایسه قرار گیرند). تعریف مسئله‌ی این پژوهش و پیاده‌سازی آزمایش‌های آن در قالب مسئله‌ی خوشه‌بندی^{۴۲} (و نه در قالب طبقه‌بندی^{۴۳}) ممکن است موجب شود که پیاده‌سازی الگوی پیشنهادی برای حل مسائل دنیای واقعی (که بیشتر از نوع طبقه‌بندی هستند) نیاز به جرح و تعدیل‌هایی داشته باشد. قبل از پرداختن به الگوی پیشنهادی، توضیحی ساده درباره فرق خوشه‌بندی و طبقه‌بندی خالی از فایده نخواهد بود. در مسائل خوشه‌بندی، داده‌ها فارغ از اینکه چه نوع دسته‌بندی‌ای از آنها صحیح است، توسط الگوریتمی خوشه‌بندی می‌شوند؛ مسائل خوشه‌بندی بیشتر ماهیت کاوشی دارند و پاسخ درست یا غلط در آنها مدنظر نیست. اما در مسائل طبقه‌بندی، الگوریتم بر اساس داده‌هایی که مقوله‌ی آنها مشخص است آموزش داده می‌شود و سپس بر روی داده‌هایی که مقوله‌شان نامشخص است اعمال می‌شود تا طبقه‌ی آن معلوم شود. در این حالت، پاسخ درست و غلط از یکدیگر مشخص هستند. آنچه در این پژوهش انجام شد حالتی بینایی داشت، به این ترتیب که مسئله در اصل به صورت مسئله‌ی خوشه‌بندی تعریف شد اما معیار خوشه‌بندی صحیح متون نیز وجود داشت که عبارت بود از تشکیل ۵ خوشه (بر اساس ۵ نویسنده) و قرارگیری تمام متون یک نویسنده در داخل خوشه‌ی مربوط به همان نویسنده.

بر اساس آنچه گفته شد، الگوی عملیاتی برای تشخیص هویت نویسنده‌ی یک متن بی‌هویت شامل مراحل زیر است:

۱. متون همه‌ی نویسندگان مظنون و متن بی‌هویتی که گمان می‌رود توسط یکی از این نویسندگان نوشته شده، در فایل‌های جداگانه به صورت فایل متنی (txt) ذخیره می‌شوند.
۲. فایل‌ها با استفاده از نرم‌افزار خطایابی، استانداردسازی می‌شوند تا شیوه‌ی نگارش آنها همگون شده و خطاهای املایی‌شان تصحیح شود.
۳. بسته‌ی تحلیل سبکی stylo در نرم‌افزار R نصب شده و آماده استفاده می‌شود.

⁴² clustering

⁴³ classification

۴. بر اساس دستورات بسته‌ی stylo، فایل‌های متنی نویسنده‌ها قطعه‌بندی (tokenization) می‌شوند تا متن به کلمات قابل پردازش تبدیل شود.

۵. سپس از فایل‌های متنی نویسنده‌ها و متن بی‌هویت نمونه‌متن‌هایی (حداقل ۴۰۰۰ کلمه‌ای) نمونه‌گیری می‌شود.

۶. پربسامدترین واژه‌های نمونه‌های متنی استخراج می‌شود و تعداد مورد نظر از آنها در جدولی ذخیره می‌شوند.

۷. سپس گزینش واژه‌های دستوری (که تعیین می‌کند حداکثر واژه‌های با بسامد صفر در متون چقدر باشد) به صورت درصدی اعمال می‌شود.

۸. پس از گزینش درصدی واژه‌ها، واژه‌های حاصل‌شده در جدولی ذخیره می‌شوند تا به صورت دستی واژه‌های غیردستوری از آنها حذف شود.

۹. پس از حذف واژه‌های غیردستوری از جدول مربوطه، جدول (که شامل واژه‌های دستوری پربسامد به همراه متون مربوطه است) مجدداً توسط بسته‌ی stylo بارگذاری می‌شود و متون توسط فرامین بسته‌ی stylo و با استفاده از الگوریتم دلتا خوشه‌بندی می‌شوند.

۳-۵. پیشنهاداتی برای پژوهش‌های آتی

در این پژوهش، کارایی واژه‌های دستوری در شناسایی نویسنده‌ی یک متن فارسی بررسی شد. استفاده از روش‌های چندمتغیری تحلیل مولفه‌های اصلی و نیز تحلیل خوشه‌ای نشان داد که ۲۰ واژه‌ی پربسامد دستوری قادرند در سطح کمینه‌ی ۴۰۰۰ کلمه‌ی متنی میان نوشته‌های نویسندگان مختلف تفکیکی با دقت ۱۰۰٪ قائل شوند. با این حال، عوامل دیگری ممکن است در کارایی واژه‌های دستوری در شناسایی نویسنده تاثیرگذار باشند که می‌توان در پژوهش‌های آتی آنها را مد نظر قرار داد:

- تعداد واژه‌های دستوری. در این پژوهش برای راحتی کار و پیچیده‌نشدن فرآیند تحلیل پیکره‌ای و نیز مدنظر قراردادن کمبود واژه در تحلیل‌های چندنگاشتی و متون کم‌حجم، تعداد ۲۰ عدد از واژه‌های دستوری پربسامد مورد تحلیل قرار گرفتند. این تعداد واژه دستوری به نظر می‌رسد که در کد کمینه‌ی تحلیل‌های سبک‌شناختی قرار دارد و بسیاری از پژوهش‌ها واژه‌هایی با تعداد بیش از ۱۰۰ عدد را به کار می‌گیرند. بنابراین، انجام پژوهش‌های مشابه با تعداد واژه‌های دستوری متغیر می‌تواند متغیر دیگری را که همان تعداد واژه‌های

دستوری است در جریان تحلیل وارد و نتایج را در سایه‌ی این متغیر جدید مجدداً ارزیابی کرد.

- حجم کلمات متون. در این پژوهش حجم‌های متعددی از متون مورد بررسی قرار گرفتند و در نهایت نتیجه گرفته شد که واژه‌های دستوری (تک‌نگاشتی) در سطح ۴۰۰۰ واژه دقتی ۱۰۰٪ در تفکیک متون نویسنده‌های مختلف دارند. این نتیجه‌گیری نیاز به بررسی‌های بیشتر دارد. همانطور که پیشتر نیز به آن اشاره شد، حجم کمینه‌ی موردتوافقی برای متون وجود ندارد که بتوان عنوان حجم کمینه‌ی قابل اعتماد برای تحلیل‌های سبک‌سنجی را بر آن نهاد.

منابع

- Binongo, J. 2003. Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, 16, 9-17.
- Barlow, M. 2010. Individual usage: a corpus-based study of idiolects. In *LAUD Symposium*. Landau, Germany.
- Burrows, J. F. 2002. Delta: A measure of stylistic difference and a guide to likely authorship, *Literary and Linguistic Computing*, 17, 267-287.
- Carroll, D. 2008. Psychology of Language (5rd ed.). Wadsworth.
- Coulthard, R. M. 1995. Questioning statements: forensic applications of linguistics. Text of inaugural lecture, Birmingham, English Language Research.
- Coulthard, R. M. 1998. Identifying the author. *Cahiers de Linguistique Francaise*, 20, 139-161.
- Eder, M. 2013. Does size matter? Authorship attribution, small samples, big problem. *Literary and Linguistic Computing*. DOI: <http://dx.doi.org/10.1093/llc/fqt066>
- Eder, M., Kestemont, M., and Rybicki, J. 2013. Stylometry with R: a suite of tools. In *Digital Humanities 2013: Conference Abstracts*. University of Nebraska-Lincoln, NE, pp. 487-89.
- 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.
- Farahmandpour, Z. and Nikmehr, H. 2015. A study on intelligent authorship methods in Persian language. *Journal of Computing and Security*, 2(1), 63-76.
- Kestemont, M. 2014. Function words in authorship attribution: from black magic to theory. *Proceedings of the 3rd Workshop on Computational Linguistics for Literature*, Gothenburg, Sweden, pp. 59-66.
- Koppel, M., & Schler, J. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on*

- Computational Approaches to Style Analysis and Synthesis*, pp. 69-72.
- Gamon, M. 2004. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 611-617.
- Gries, S. 2006. Corpus-based methods and cognitive semantics: The many senses of to run. In Gries, S. and Stefanowitsch, A. (eds.), *Corpora in Cognitive Linguistics. Corpus-Based Approaches to Syntax and Lexis*, 57-100. Berlin: Mouton de Gruyter.
- Gries, S. 2009. *Statistics for Linguistics with R: A Practical Introduction* (2nd ed.). Berlin: de Gruyter.
- Hedegaard, S. & Simonsen, J. G. 2011. Lost in translation: Authorship attribution using frame semantics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. June 19-24, Portland, Oregon, pp. 65-70.
- 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.
- Hoover, D. L. 2004. Testing Burrows's Delta. *Literary and Linguistic Computing*, (4) 19, 475-453.
- Hubert, L. and Arabie, P. 1985. Comparing partitions. *Journal of Classification*, 2(1), 193-218.
- Jannidis, F., Pielstrom, S., Schoch, C., and Vitt, T. 2015. Improving Burrows' Delta – An empirical evaluation of text distance measures. In Digital Humanities Conference 2015, Sydney.
- Johansson, V. 2008. Lexical diversity and lexical density in speech and writing: A developmental perspective. Lund University, Department of Linguistics and Phonetics: Working Papers, 53, 61-79.

- Johnson, A. and Wright, D. 2014. Identifying idiolect in forensic authorship attribution. *Language and Law/Linguagem e Direito*, Vol. 1(1), 37-69.
- Juola, P. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3), 233-334.
- 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.
- Mollin, S. 2009. "I entirely understand" is a Blairism: the methodology of identifying idiolectal collocations. *International Journal of Corpus Linguistics*, 14 (3), 367-392.
- Mosteller, F., & Wallace, D.L. 1964. Inference and disputed authorship: The Feralist. Reading, MA: Addison-Wesley.
- Panicheva, P., Cardiffm, J., and Rosso, P. 2010. Personal sense and idiolect: Combining authorship attribution and opinion analysis. *Seventh International Conference on Language Resources and Evaluation*, Malta.
- Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, 846-850.
- R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://R-project.org/>.
- Sandra, D. & Rice, S. 1995. Network analyses of prepositional meaning: mirroring whose mind- the linguist's or the language user's? *Cognitive Linguistics*, 6(1), 89-130.
- Zhao, Y. and Zobel, J. 2005. Effective and scalable authorship attribution using function words. In *Proceedings of 2nd Asian Information Retrieval Symposium, volume 3689 of LNCS*, Jejulsland, Korea, pp. 174-189.

- Stamatatos, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538-556.
- Tweedie, F. J., Singh, S., and Holmes, D. I. 1996. Neural network applications in stylometry: The Federalist Papers, *Computers and the Humanities*, 30(1), 1-10.
- Wardhaugh, R. & Fuller, J. M. 2015. An Introduction to Sociolinguistics (7th ed.). Wiley-Blackwell.
- Whitelaw, C. & Argamon, S. 2004. Systemic functional features in stylistic text classification. In *Proceedings of AAAI Fall Symposium on Style and Meaning in Language, Art, and Music*.
- Whitelaw, C. & Patrick, J. 2004. Selecting systemic features for text classification. In *Proceedings of Australian Language Technology Workshop*, Sydney, Australia, pp. 93-100.

ABSTRACT

Idiolect is a person's unique way of using linguistic knowledge encompassing words, grammatical structures and pronunciations, that has attracted the attention of researchers in the field of computational stylistics and authorship attribution. One of the linguistic units that have been shown to be the seat of idiolect is function words. Function words are attractive to stylometry research because they are frequent, topic-independent, and unconsciously used. In this project, the possibility of distinguishing texts of various contemporary authors in Persian language has been investigated by analyzing their individual use of function words. First written works of five Iranian writers and academics were gathered in separate corpora and normalized using Vafa spellchecker. Then the Stylo package was used to carry out stylometric analyses in R. Analyses included principal component analysis and cluster analysis using delta distance measure. The results showed that function words in Persian could be considered reliable style markers in distinguishing an author's writings from those of others. In the attribution process, uni-grams performed far better than bi- and tri-grams when the size of texts dropped to 1000 words and their performance picked at 4000 words based on 20 most frequent function words.

Keywords: Author attribution; corpus analysis, cluster analysis, delta method, principal component analysis, stylometry



**Iranian Research Institute
for Information Science and Technology**

Research Project

**Authorship identification based on
idiolect:
A corpus-based study in Persian**

By:

Ramin Golshaie

Summer 2016