E-ISSN: | P-ISSN:

Management Sciences and Decision Analysis

www.msda.reapress.com

Manag. Sci. Decis. Anal. Vol. 3, No. 3 (2025) 332–341.

Paper Type: Original Article

A Data-Driven Decision Model for Identifying Data Scientist Competencies Using Text Mining of Job Advertisements

Abolfazl Nabavi*

Department of Information Technology Management, Allameh Tabatabaei University, Tehran, Iran; a.nabavi124@gmail.com.

Citation:

Nabavi, A. (2025). A data-driven decision model for identifying data scientist competencies using text mining of job advertisements. *Management sciences and decision analysis, 3*(3), 332-341.

Received: 03/11/2024 Reviewed: 12/01/2025 Revised: 12/03/2025 Accepted: 27/04/2025

Abstract

Purpose: Rapid digital transformation and the exponential growth of data have intensified the demand for data scientists across industries. However, the dynamic nature of this profession has created ambiguity regarding the essential competencies required for effective employment. This study aims to develop a data-driven analytical model to systematically identify the key knowledge and skills required for data scientist roles based on real labor market data.

Methodology: The study employs a text mining framework based on the knowledge discovery in databases process. A dataset of 120 job advertisements for data scientist positions posted on LinkedIn in the United States was collected and preprocessed. Text data from job descriptions were cleaned, transformed, and analyzed using TF–IDF weighting to extract significant terms. From a large set of extracted words, the most informative features were selected and categorized into competency groups. Descriptive statistical analysis was also conducted to examine job characteristics, including industry distribution, experience level, and employment type.

Findings: The results reveal that demand for data scientists spans a wide range of industries, with a dominant concentration in IT-related sectors. Most job postings are full-time and target entry-level candidates. Extracted competencies were categorized into two primary groups: hard skills and soft skills, with hard skills accounting for the majority of identified requirements. Technical competencies such as programming, machine learning, data analysis, and statistical modeling were identified as core requirements, while soft skills such as teamwork, problem-solving, and leadership were also found to play a significant supporting role.

Originality/Value: This study contributes to management decision analysis by integrating labor market data and text mining techniques to provide a structured competency identification model. The findings offer practical decision support for educational institutions, human resource managers, and job seekers by aligning training programs and recruitment strategies with real market demands. The proposed framework demonstrates how data-driven analytics can enhance strategic workforce planning in rapidly evolving digital environments.

Keywords: Data scientist, Text mining, Job posting.

Corresponding Author: a.nabavi124@gmail.com https://doi.org/10.22105/msda.v3i3.131

علوم مدیریت و تحلیل تصمیم
دوره ۳، شماره (۳)، (۱۴۰۴)، ۳۴۱-۳۳۲
www.msda.reapress.com

نوع مقاله: پژوهشی

یک مدل تصمیم‌گیری داده‌محور برای شناسایی شایستگی‌های دانشمند داده با استفاده از متن‌کاوی آگهی‌های شغلی

ابوالفضل نبوی*
گروه مدیریت فناوری اطلاعات، دانشگاه علامه طباطبایی، تهران، ایران.

چکیده

هدف: تحول دیجیتال و رشد نمایی داده‌ها موجب افزایش تقاضا برای شغل دانشمند داده در صنایع مختلف شده است. با این حال، ماهیت پویا و در حال تحول این حرفه، ابهاماتی را در خصوص شایستگی‌های کلیدی موردنیاز برای اشتغال موثر ایجاد کرده است. هدف این پژوهش، طراحی یک مدل تحلیلی داده‌محور برای شناسایی نظام‌مند دانش و مهارت‌های موردنیاز شغل دانشمند داده بر اساس داده‌های واقعی بازار کار است.

روش‌شناسی پژوهش: این پژوهش از یک چارچوب متن‌کاو مبتنی بر فرایند کشف دانش در پایگاه داده‌ها استفاده می‌کند. مجموعه‌ای شامل ۱۲۰ آگهی شغلی مرتبط با موقعیت دانشمند داده از شبکه لینکدین برای کشور ایالات متحده گردآوری و پیش‌پردازش شد. داده‌های متنی موجود در شرح شغل‌ها پاک‌سازی، تبدیل و با استفاده از معیار *TF–IDF* تحلیل شدند تا واژگان کلیدی استخراج گردد. سپس مهم‌ترین ویژگی‌های استخراج‌شده انتخاب و در قالب گروه‌های شایستگی طبقه‌بندی شدند. همچنین تحلیل‌های توصیفی برای بررسی ویژگی‌های بازار کار شامل صنعت، سطح تجربه و نوع استخدام انجام شد.

یافته‌ها: نتایج نشان داد تقاضا برای دانشمندان داده در طیف وسیعی از صنایع وجود دارد که بیشترین تمرکز در حوزه فناوری اطلاعات است. اغلب آگهی‌های شغلی به صورت تمام‌وقت و در سطح ورود به بازار کار منتشر شده‌اند. شایستگی‌های استخراج‌شده در دو دسته اصلی مهارت‌های سخت و نرم طبقه‌بندی شدند که سهم مهارت‌های سخت بیشتر است. مهارت‌های فنی نظیر برنامه‌نویسی، یادگیری ماشین، تحلیل داده و مدل‌سازی آماری به‌عنوان الزامات اصلی شناسایی شدند، در حالی که مهارت‌های نرم مانند کار تیمی، حل مسئله و رهبری نیز نقش مکمل مهمی دارند.

اصالت/ارزش افزوده علمی: این پژوهش با تلفیق داده‌های واقعی بازار کار و تکنیک‌های متن‌کاوی، یک مدل ساختاریافته برای شناسایی شایستگی‌ها ارائه می‌دهد که به ادبیات تحلیل تصمیم مدیریتی افزوده می‌شود. نتایج، ابزار تصمیم‌یار کاربردی برای موسسات آموزشی، مدیران منابع انسانی و متقاضیان شغلی فراهم می‌آورد و نشان می‌دهد چگونه تحلیل داده‌محور می‌تواند برنامه‌ریزی نیروی انسانی و طراحی آموزش‌ها را در محیط‌های پویا بهبود دهد.

کلیدواژه‌ها: دانشمند داده، متن‌کاوی، آگهی شغلی.

۱- مقدمه

در [1] به نقل از جفری استنتون، علم داده[1] را رشته در حال ظهوری که به جمع‌آوری، آماده‌سازی، تحلیل، بصری‌سازی، مدیریت و نگهداشت اطلاعات در حجم بالا می‌پردازد و به نقل از جان فورمن، علم داده را تبدیل داده‌ها به بینش‌ها، تصمیم‌ها و محصولات ارزشمند با استفاده از ریاضیات و آمار تعریف شده است.

[1] Data science

دون پورت و پاتیل [2] دانشمند داده[1] را جذاب‌ترین شغل قرن ۲۱ معرفی کرده است. در [3] دانشمند داده به عنوان یک شغل با آینده روشن این‌گونه تعریف شده است: مجموعه‌ای از تکنیک‌ها و برنامه‌های کاربردی تحلیلی برای تبدیل داده‌های خام به اطلاعات مفید با استفاده از زبان‌های برنامه‌نویسی داده گرا و نرم‌افزارهای مصورسازی. به‌کار گیری داده‌کاوی، مدل‌سازی داده، پردازش زبان طبیعی و یادگیری ماشین برای استخراج و تجزیه‌وتحلیل اطلاعات از مجموعه داده‌های بزرگ ساختار یافته و غیرساختار یافته. مصورسازی، تفسیر و گزارش دهی یافته‌های داده‌ای. شاید ایجاد گزارش‌های داده‌ای پویا. کاری که بیش از هر چیز دانشمندان داده انجام می‌دهند این است که در حین شنا در داده‌ها، کشفیاتی انجام می‌دهند. آن‌ها در قلمرو دیجیتالی می‌توانند به مقادیر زیادی از داده‌های بدون شکل، ساختار بدهند و تجزیه‌وتحلیل‌شان را ممکن سازند. آن‌ها در یک محیط رقابتی که چالش‌ها مدام در حال تغییر هستند و جریان داده‌ها هرگز قطع نمی‌شود به تصمیم گیرندگان کمک می‌کنند تا از تجزیه‌وتحلیل موردی داده‌ها به تعامل پیوسته با داده‌ها رو بیاورند [2].

آگهی‌های شغلی[2] مهم‌ترین کانال برای جذب کارمندان جدید هستند [4]. آگهی‌های شغلی منبع داده مهمی هستند زیرا حاوی اطلاعاتی در مورد مهارت‌ها و دانش موردنیاز برای مشاغل خاص هستند [5]. آگهی‌های شغلی منبعی مرتبط از اطلاعات درباره مهارت‌ها و دانش موردنیاز هستند که می‌توانند برای ایجاد بینش سریع درباره تغییرات در پروفایل‌های شغلی مورد استفاده قرار بگیرند [4].

آگهی‌های شغلی به صورت سنتی در روزنامه‌های چاپی منتشر می‌شوند، اما در دو دهه گذشته، غالبا به صورت آنلاین یا در سایت‌های مشخص یا در رسانه‌های اجتماعی منتشر می‌شوند [4]. به دلیل تغییرات سریع ویژگی‌های شغلی در حوزه‌هایی مانند صنعت ٤ به ابزاری سریع برای تحلیل آگهی‌های شغلی نیاز است. به عبارتی بروشی داینامیک‌تر نیاز است [4]. با توجه به هدف همیشه مرتبط قابلیت استخدام و تغییر سریع بازار کار، داشتن بینش عمیق در موردنیازهای مهارتی و دانشی کنونی اهمیت فزآینده‌ای پیدا کرده است [5]. دو روش اصلی برای تحلیل آگهی‌های شغلی تحلیل محتوای دستی و تحلیل متن اتوماتیک (اغلب با عنوان متن‌کاوی شناخته می‌شود) است. متن‌کاوی در مقایسه با تحلیل محتوای دستی زمان و هزینه کمتری نیاز دارد [4]. استفاده از متن‌کاوی در تجزیه‌وتحلیل آگهی‌های شغلی در حال افزایش است. سخت‌افزار قدرتمندتر امکان نمونه‌برداری از مقادیر بیشتری از داده‌ها را در بازه‌های زمانی کوچک‌تر و امکان بررسی آگهی‌های شغلی بیشتری را در مدت زمان کوتاه‌تری فراهم می‌کند [5].

متن‌کاوی فیلدی بین رشته‌ای است که از بازیابی اطلاعات، داده‌کاوی، یادگیری ماشینی، آمار و زبان‌شناسی محاسباتی استفاده می‌کند و هدف مهمش استخراج اطلاعات با کیفیت بالا از متن است [6]. با توجه به سه دسته ارایه شده در [7] برای وب کاوی[3]، متن‌کاوی آگهی‌های شغلی منتشر شده بر روی وب در دسته کاوش محتوای وب[4] قرار می‌گیرد.

از آن‌جا که شغل دانشمند داده هم‌زمان با توسعه تکنولوژی در حال تکامل است، این پژوهش با هدف کشف دانش و مهارت‌های موردنیاز کنونی شغل دانشمند داده با متن‌کاوی آگهی‌های شغلی انجام شده و نتایج در دو بخش تحلیل توصیفی و متن‌کاوی ارایه شده است. نتایج این پژوهش برای چه کسانی مفید است و سود می‌برند [4]؟

۱. موسسات آموزش عالی: بر اساس نتایج این پژوهش می‌توانند برنامه‌ها و دوره‌های آموزشی خود را توسعه دهند.
۲. متخصصین منابع انسانی: بر اساس نتایج این پژوهش می‌توانند تصمیمات آگاهانه‌تری برای استخدام بگیرند.
۳. کارشناسانی که قبلا استخدام شده‌اند یا می‌خواهند استخدام شوند: بر اساس نتایج این پژوهش می‌توانند سطح مهارت‌های خود را برای شغل فعلی و مشاغل موردنظرشان بسنجند.

مقاله علاوه بر این بخش "مقدمه" به این ترتیب تنظیم شده است: بخش "پیشینه پژوهش" که در آن تعدادی از پژوهش‌های پیشین که متن‌کاوی آگهی‌های شغلی را برای شغل‌های مختلف انجام داده‌اند مرور شده است. بخش "روش پژوهش" که در آن روش انجام کار از جمع‌آوری داده تا

[1] Data scientist
[2] Job advertisements
[3] Web mining
[4] Web content mining

تحلیل نتایج تشریح شده است. بخش "نتایج" که در آن نتایج در دو بخش تحلیل توصیفی و متن‌کاوی بیان شده است. بخش "بحث و نتیجه‌گیری" که در آن ضمن بیان کلیات نتایج، محدودیت‌های پژوهش و پیشنهادهای برای پژوهش‌های آتی ذکر شده است.

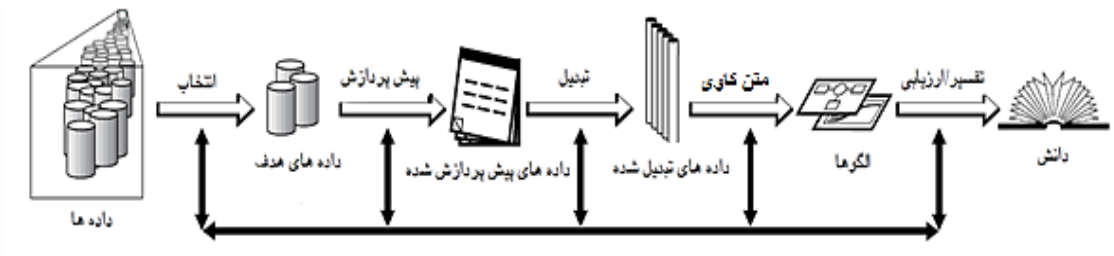
۲- پیشینه پژوهش

در این بخش تعدادی از پژوهش‌های پیشین که متن‌کاوی آگهی‌های شغلی را برای شغل‌های مختلف انجام داده‌اند مرور شده است.

۱. هدف [4] توسعه پروفایلی از آگهی‌های شغلی صنعت ۴ که اغلب به عنوان ابزاری برای جمع‌آوری اطلاعات مرتبط درباره مهارت‌ها و دانش موردنیاز در صنایع با تغییرات سریع استفاده می‌شود، با استفاده از متن‌کاوی آگهی‌های شغلی بوده است.
۲. با توجه به ادعای نویسندگان [5] این اولین مطالعه‌ای است که از متن‌کاوی برای تجزیه‌وتحلیل آگهی‌های شغلی برای انفورماتیک پزشکی استفاده می‌کند؛ زیرا مطالعات قبلی این نوع داده‌ها را به صورت دستی بررسی کرده‌اند. این پژوهش یک پروفایل شایستگی ایجاد کرده است که می‌تواند به عنوان مبنایی برای تجزیه‌وتحلیل و بهبود برنامه درسی انفورماتیک سلامت عمل می‌کند.
۳. در [8] تلاش شده است با استفاده از روش‌های متن‌کاوی مهارت‌های موردنیاز بازار کار در حوزه علم داده شناسایی شود. به این منظور، متن آگهی‌های شغلی برترین شرکت‌ها در حوزه علم داده از سطح وب‌سایت‌ها جمع‌آوری شده و مورد متن‌کاوی قرار گرفته است.
۴. اسمالدونه و همکاران [9] با هدف به‌دست آوردن بینشی در مورد انتظارات فعلی کارفرمایانی که به دنبال استخدام دانشمندان داده هستند انجام شده است. این پژوهش به‌طور سیستماتیک با جمع‌آوری و تجزیه‌وتحلیل داده‌های مرتبط از آگهی‌های شغلی منتشر شده در وب‌سایت‌های کاریابی ایالات‌متحده انجام شده است.
۵. در [10] سعی شده است تا مهارت‌ها و شایستگی‌های موردنیاز در محل کار برای انواع مختلف موقعیت‌های شغلی تجزیه‌وتحلیل کسب‌وکار (شامل چهار نوع شغل اصلی یعنی تحلیلگر کسب‌وکار، تحلیلگر داده، دانشمند داده و مدیر تجزیه‌وتحلیل داده‌ها) بر اساس موقعیت آگهی شده در دو محبوب‌ترین و بزرگترین وب‌سایت‌های جستجوی کار در ایالات‌متحده با کمک متن‌کاوی شناسایی و مشخص شود.
۶. جیانگ و چن [11] به بررسی مهارت‌های یافت شده در شرح وظایف هفت عنوان شغلی مرتبط با علم داده، مهارت‌های شناسایی شده در محتوای دوره‌های موردنیاز برای گواهینامه‌های کارشناسی علوم داده در دانشگاه‌ها و مقایسه مهارت‌های موردنیاز در مشاغل و مهارت‌های آموزش داده شده در گواهینامه‌ها پرداخته است.

۳- روش پژوهش

اصطلاح کشف دانش در پایگاه داده[1] در اولین کارگاه آموزشی *KDD* در سال ۱۹۸۹ ابداع شد تا تاکید کند دانش محصول نهایی کشف داده محور است. فرایند *KDD* فرایند غیرمستقیم شناسایی الگوهای باارزش، جدید، بالقوه موثر و نهایتا قابل‌فهم در داده‌ها است. فرایند *KDD* تعاملی و تکراری (با بسیاری تصمیم گرفته‌شده کاربر) است [12]. این پژوهش بر اساس فرایند *KDD* (با تغییر نام مرحله داده‌کاوی به متن‌کاوی) با زبان *R* طی مراحل زیر انجام شده است (**شکل ۱**):



شکل ۱- فرایند متن‌کاوی بر اساس فرایند KDD.

Figure 1- Text mining process based on the KDD process.

[1] Knowledge Discovery in Databases (KDD)

۳-۱- انتخاب

منبع داده مورد استفاده در این پژوهش، آگهی‌های شغلی منتشر شده بر روی سایت لینکدین است. با فیلتر بخش *Job* سایت لینکدین تعداد ۱۲۰ آگهی شغلی که در اواخر ژانویه ۲۰۲۲ برای کشور آمریکا منتشر شده بودند و عنوان شغلی آن‌ها شامل عبارت کلیدی "دانشمند داده[1]" بودند، به عنوان منبع داده انتخاب شدند. آگهی‌های شغلی دارای فیلدهای **جدول ۱** بوده‌اند.

به سه دلیل از آگهی‌های شغلی سایت لینکدین استفاده شده است [4]: اولا، لینکدین یکی از رهبران مهم در انتشار آگهی‌های شغلی که دامنه وسیعی از سازمان‌ها، کشورها و نوع شغل‌ها را پوشش می‌دهد شده است. دوما، آگهی‌های شغلی در لینکدین شکل نیمه ساختار یافته دارند که برای متن‌کاوی مناسب هستند. سوما، لینکدین را می‌توان به عنوان یک منبع باز مناسب برای نمونه‌گیری درنظر گرفت.

جدول ۱- فیلدهای یک آگهی شغلی منتشر شده در لینکدین.
Table 1- Fields of a job posting published on LinkedIn.

فیلد	نوع فیلد		توضیحات
	ساختاریافته	غیرساختار یافته	
Company		✓	
Company industry		✓	
Job location		✓	
Job title		✓	
Workplace type	✓		دسته‌بندی از پیش تعریف شده: On-site, Remote, Hybrid
Employment type	✓		دسته‌بندی از پیش تعریف شده: Full-time, Part-time, Contract, Temporary, Volunteer, Internship
Experience level	✓		دسته‌بندی از پیش تعریف شده: Internship, Entry level, Associate, Mid-Senior level, Director, Executive
Job description		✓	

۳-۲- پیش پردازش

در منبع داده انتخاب شده فیلدهایی *Workplace type*، *Employment type*، *Experience level* و *Company industry* دارای داده گمشده[2] بودند که با مقدار مناسب پر شدند.

۳-۳- تبدیل

مشکلات زمان خواندن منبع داده در نرم‌افزار هم با انجام پیش‌پردازشی‌هایی بر روی فیلد *Job description* حل شدند. بعد از ورود منبع داده به نرم‌افزار با انجام تبدیلات موردنیاز در فیلد *Job description* شامل حذف علامات نگارشی، حذف اعداد، تبدیل تمام حروف به حروف کوچک و... این فیلد برای انجام متن‌کاوی آماده شد.

[1] Data scientist

[2] Missing data

۴-۳- متن‌کاوی

بر روی منبع داده آماده شده، تحلیل‌های توصیفی و متن‌کاوی انجام شد که نتایج آن‌ها در بخش نتایج آورده شده است.

تحلیل‌های توصیفی بر اساس فیلدهای زیر انجام شد:

۱. *Job location*.
۲. *Company industry*.
۳. *Experience level*.
۴. *Employment type*.
۵. *Workplace type*.

متن‌کاوی بر اساس فیلد *Job description*، طی مراحل زیر انجام شد:

۱. تعداد ۳۹۱۰ کلمه شناسایی شد.
۲. وزن ۳۹۱۰ کلمه شناسایی شده بر اساس معیار *TF-IDF*[1] تعیین شد.

معیار *TF* برای اندازه‌گیری تعداد دفعات وقوع یک کلمه یا عبارت در یک سند استفاده می‌شود. فرض کنید سند "*T1*" حاوی ۵۰۰۰ کلمه داریم و کلمه *Alpha* دقیقا ۱۰ بار در سند وجود دارد. با توجه به این واقعیت کاملا مشخص که طول اسناد می‌تواند از بسیار کوچک تا بزرگ متفاوت باشد، این احتمال وجود دارد که تعداد دفعات وقوع هر کلمه‌ای در اسناد بزرگ در مقایسه با اسناد کوچک بیشتر باشد. بنابراین، برای رفع این مشکل، تعداد دفعات وقوع هر کلمه در یک سند را بر مجموع کلمات موجود در آن سند تقسیم می‌کنیم تا فراوانی کلمه پیدا شود. بنابراین، در این مورد، معیار *TF* برای کلمه "*Alpha*" در سند "*T1*" برابر است با [13]:

$$TF = 10.500 = 0.002.$$

معیار *IDF* یک معیار وابسته به مجموعه اسناد است که به نفع اصطلاحات متمرکز در چند سند یک مجموعه است. فرض کنید ما ۱۰ سند داریم و کلمه *Alpha* در ۵ مورد از آن اسناد وجود دارد، بنابراین، معیار *IDF* برای کلمه *technology* در مجموعه اسناد برابر است با [13]:

$$IDF = log(10.5) = 0.3010.$$

معیار *TF-IDF* از ضرب معیار *TF* در معیار *IDF* به‌دست می‌آید. بنابراین، معیار *TF-IDF* برای کلمه "*Alpha*" برابر است با [13]:

$$TF - IDF = 0.002 \times 0.3010 = 0.000602.$$

از بین ۲۰۰ کلمه با بالاترین معیار *TF-IDF* تعداد ۱۰۰ کلمه که بهتر بیانگر دانش موجود در آگهی‌های شغلی یعنی بیانگر نیازمندی‌های دانشی و مهارتی بازار کار برای دانشمند داده بودند، انتخاب شدند. سعی شد ۱۰۰ کلمه منتخب در سه دسته مهم زیر که هر یک بر مولفه‌های متفاوتی از اشتغال‌پذیری فارغ‌التحصیلان تمرکز دارند [14]، دسته‌بندی و ارایه شوند که البته تمام ۱۰۰ کلمه منتخب در دو دسته مهارت‌های سخت و نرم قرار گرفتند.

۱. مسایل خاص کسب‌وکار (دانش و مهارت‌های سخت مرتبط با کسب‌وکار).
۲. شایستگی‌های بین فردی (مهارت‌های نرم مرتبط با کسب‌وکار).
۳. تجربه کاری و یادگیری مبتنی بر کار.

[1] Term Frequency-Inverse Document Frequency (TF-IDF)

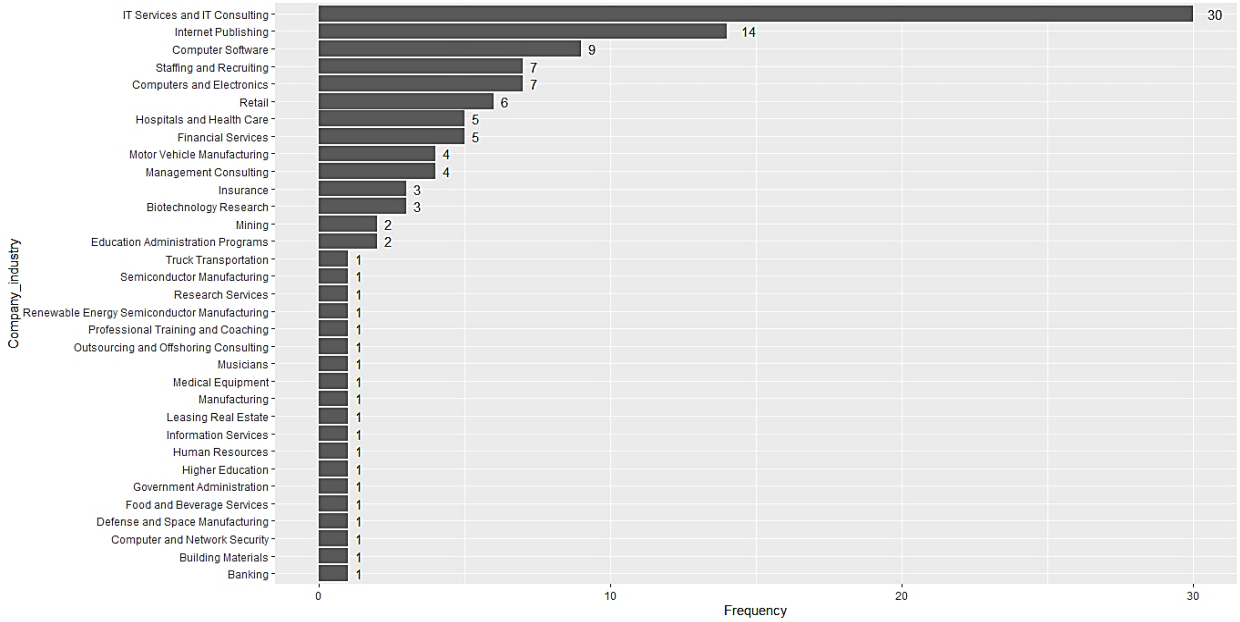
۵-۳- تفسیر / ارزیابی

با هدف کشف اطلاعات و دانش مفید، نتایج تحلیل توصیفی و متن‌کاوی تفسیر و ارزیابی شد که در بخش نتایج تشریح شده است.

۴- نتایج

۱-۴- تحلیل توصیفی

در **شکل ۲** دسته‌بندی فیلد *Company industry* به همراه فراوانی نمایش داده شده است.



شکل ۲- دسته‌بندی فیلد Company industry.

Figure 2- Company industry field classification.

با توجه به **شکل ۲** اطلاعات زیر قابل نتیجه‌گیری است:

۱. ۲۳ صنعت مختلف درخواست برای دانشمند داده داشته‌اند. این تنوع نشانه نیاز صنایع مختلف به دانشمند داده است.
۲. بیشترین درصد (۲۵%) از فیلد *Company industry* صنعت *IT Services and IT Consulting* بوده است.

در **جدول ۲** فراوانی متقابل دو فیلد *Experience level* و *Employment type* نمایش داده شده است.

جدول ۲- فراوانی متقابل دو فیلد Experience level و Employment type.

Table 2- Mutual frequency of the two fields Experience level and Employment type.

		Employment Type	
		Full-Time	**Total**
Experience level	Associate	5	5
	Director	2	2
	Entry level	93	93
	Mid-Senior level	20	20
	Total	120	120

با توجه به **جدول ۲** اطلاعات زیر قابل نتیجه‌گیری است:

۱. از فیلد *Experience level*، ۸۷% مقدار *Entry level* بوده است.
۲. کمترین درصد (۲۵%)، از فیلد *Experience level* مقدار *Director* بوده است.

۳. بیشترین درصد (۱۰۰%)، از فیلد *Employment type* مقدار *Full-time* بوده است.
٤. بیشترین درصد (۷۸%)، از آگهی‌های شغلی در سطح *Entry level* و نوع *Full-time* بوده‌اند.
٥. کمترین درصد (۲%)، از آگهی‌های شغلی در سطح *Director* و نوع *Full-time* بوده‌اند.

در **جدول ۳** فراوانی متقابل دو فیلد *Experience level* و *Workplace type* نمایش داده شده است.

جدول ۳- فراوانی متقابل دو فیلد Experience level وWorkplace type .

Table 3- Mutual frequency of the two fields Experience level and Workplace type.

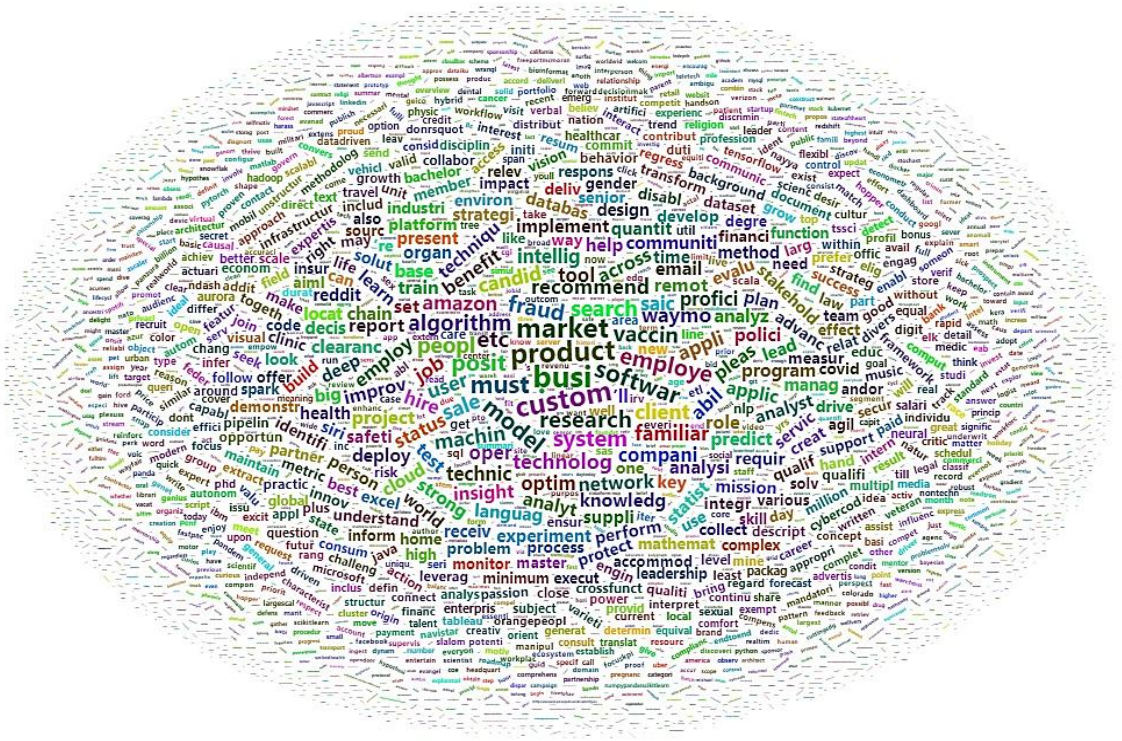
		Workplace type			
		On-site	**Remote**	**Hybrid**	**Total**
Experience level	Associate	2	0	3	5
	Director	2	0	0	2
	Entry level	72	21	0	93
	Mid-Senior level	14	5	1	20
	Total	90	26	4	120

با توجه به **جدول ۳** اطلاعات زیر قابل نتیجه‌گیری است:

۱. بیشترین درصد (۷۸%)، از فیلد *Experience level* مقدار *Entry level* بوده است.
۲. کمترین درصد (۲%)، از فیلد *Experience level* مقدار *Director* بوده است.
۳. بیشترین درصد (۷۵%)، از فیلد *Workplace type* مقدار *On-site* بوده است.
٤. کمترین درصد (۳%)، از فیلد *Workplace type* مقدار *Hybrid* بوده است.
٥. بیشترین درصد (۶۰%)، از آگهی‌های شغلی در سطح *Entry level* و نوع *On-site* بوده‌اند.

۲-۴- متن‌کاوی

در **شکل ۳** ابر کلمات[1] با ۳۹۱۰ کلمه شناسایی شده نمایش داده شده است.

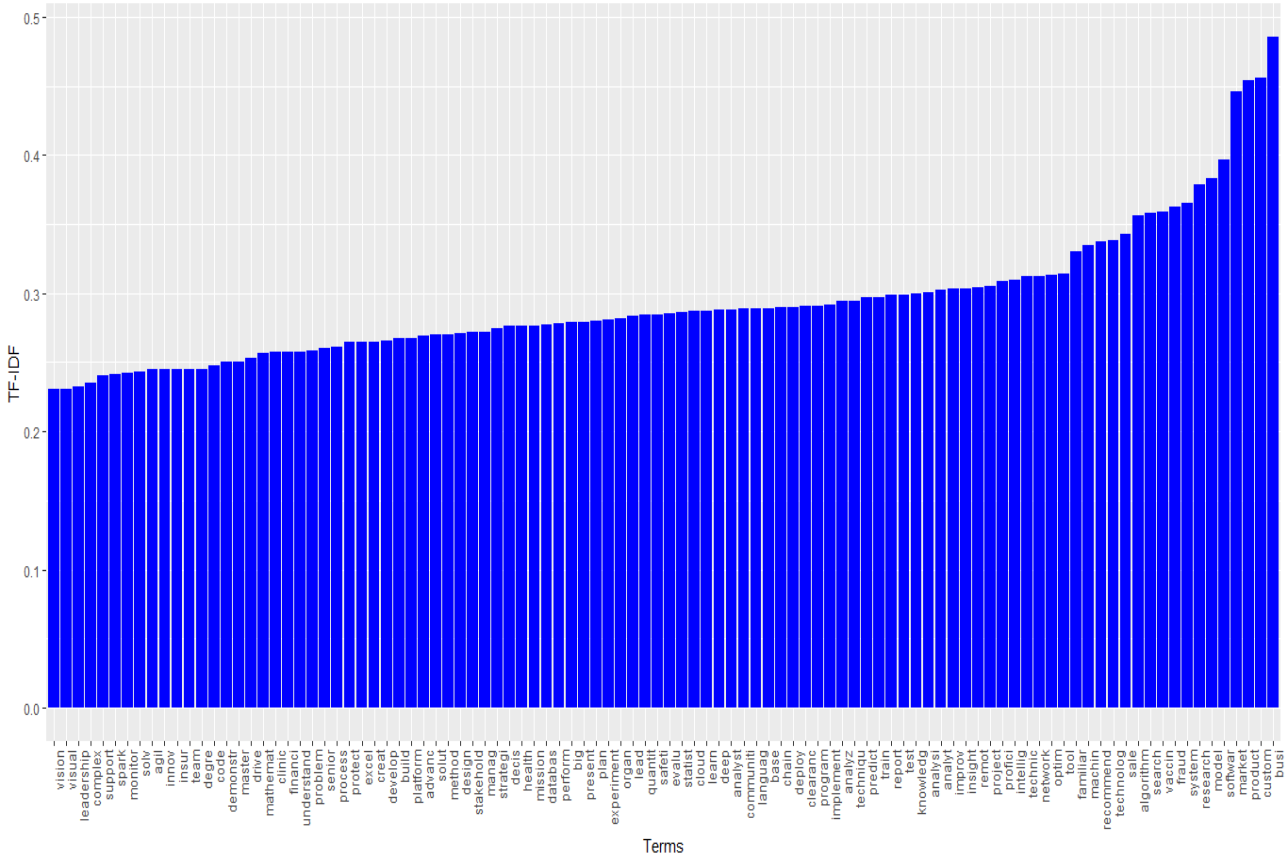


شکل ۳- ابر کلمات با ۳۹۱۰ کلمه.

Figure 3- Word cloud with 3910 words.

[1] Wordcloud

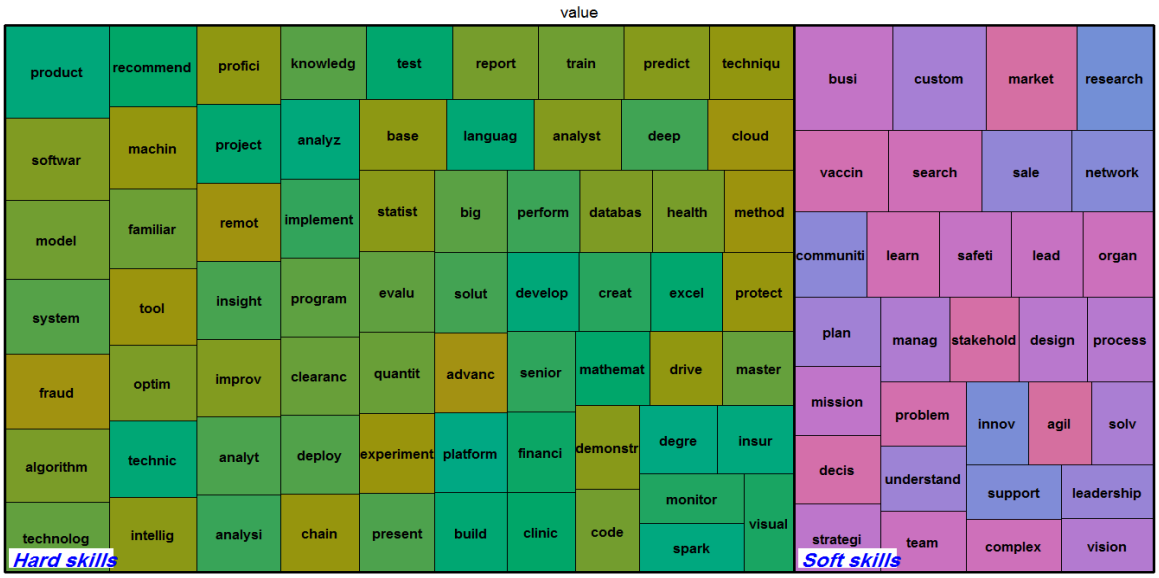
در **شکل ۴** تعداد ۱۰۰ کلمه منتخب از ۲۰۰ کلمه با بالاترین معیار *TF-IDF*، نمایش داده شده است.



شکل ۴- ۱۰۰ کلمه منتخب از ۲۰۰ کلمه با بالاترین معیار TF-IDF.

Figure 4- 100 words selected from 200 words with the highest TF-IDF score.

در **شکل ۵** تصویر *treemap* با دو دسته مهارت‌های نرم و سخت شامل ۱۰۰ کلمه منتخب نمایش داده شده است.



شکل ۵- تصویر treemap با دو دسته مهارت‌های نرم و سخت شامل ۱۰۰ کلمه منتخب.

Figure 5- Treemap image with two categories of soft and hard skills including 100 selected words.

با توجه به **شکل ۵** اطلاعات زیر قابل نتیجه‌گیری است:

۱. از ۱۰۰ کلمه منتخب ۶۹%، در دسته مهارت‌های سخت قرار گرفته‌اند.
۲. از ۱۰۰ کلمه منتخب ۳۱%، در دسته مهارت‌های نرم قرار گرفته‌اند.
۳. مجموع معیار *TF-IDF* برای دسته مهارت‌های سخت ۲۰/۱۸ است.
٤. مجموع معیار *TF-IDF* برای دسته مهارت‌های نرم ۹/۱۸ است.
۵. برای کسب جایگاه شغلی دانشمند داده، باید علاوه بر مهارت‌های سخت مهارت‌های نرم را هم آموزش دید.
٦. کلمات زیر در دسته مهارت‌های سخت، مشخص‌تر به حوزه کاربرد اشاره دارند:

Fraud health– Finance- Clinic

۷. کلمات زیر در دسته مهارت‌های سخت، مشخص‌تر به دانش مورد انتظار اشاره دارند.

System– Intellig- Knowledge– Analyt– Analysi– Analyz– Analyst– Statist mathemat- Model– Machine– Predict- Algorithm– Train– Test– Deep

۸. کلمات زیر در دسته مهارت‌های سخت، مشخص‌تر به مهارت مورد انتظار اشاره دارند.

Program– Language- Software– Tool– Develop– Deploy– Build– Implement– Code– Base– Cloud– Big– Spark– Excel – Report– Visual– Present

۹. کلمات زیر در دسته مهارت‌های نرم، مشخص‌تر به مهارت مورد انتظار اشاره دارند.

Leadership- Lead– Manag- Plan– Mission– Vision– Strategi– Team– Network- Research– Search– Busi– Market– Sale- Problem- Solv– Agil- Innov

۵- نتیجه‌گیری

مقدار زیاد داده‌های ایجادشده و انباشته شده در عصر دیجیتال، نیاز به شغل دانشمند داده را برای تبدیل این داده‌های خام به اطلاعات دانش و بینش‌های مفید ایجاد کرده است. کاری که بیش از هر چیز دانشمندان داده انجام می‌دهند این است که در حین شنا در داده‌ها، کشفیاتی انجام می‌دهند. از آن‌جا که این شغل هم‌زمان با توسعه تکنولوزی در حال تکامل است این پژوهش با هدف کشف دانش و مهارت‌های موردنیاز شغل دانشمند داده با متن‌کاوی آگهی‌های شغلی که بر روی شبکه اجتماعی لینکدین برای کشور آمریکا منتشر شده بودند انجام شده و نتایج در دو بخش تحلیل توصیفی و متن‌کاوی ارایه شده است.

نتایج این پژوهش با آشکارسازی نیاز صنایع مختلف به دانشمند داده دانش و مهارت‌های موردنیاز این شغل را در دو دسته مهارت‌های سخت و نرم ارایه کرده که می‌تواند برای موسسات آموزشی متخصصین منابع انسانی کارشناسان متقاضی این شغل مفید باشد. این پژوهش دارای این محدودیت بوده است که فقط از آگهی‌های شغلی که بر روی شبکه اجتماعی لینکدین برای کشور آمریکا به انگلیسی و در یک محدوده زمانی محدود منتشر شده بودند، استفاده کرده است. برای پژوهش‌های آتی متن‌کاوی آگهی‌های شغلی که در سایر منابع مانند سایت‌های کاریابی، برای سایر کشورها، به زبان‌هایی غیر انگلیسی و در مدت زمان طولانی‌تر منتشر می‌شوند، پیشنهاد می‌شود.

منابع

[1] Sohrabi, B., & Iraj, H. (2015). Data science: Concepts and skills. *Tehran University Jihad Organization*. (**In Persian**). https://B2n.ir/wy2944

[2] Davenport, T. H., & Patil, D. J. (2012). Data scientist. *Harvard business review*, 90(5), 70–76. https://blogs.sun.ac.za/open-day/files/2022/03/Data-Scientist-Harvard-review.pdf

[3] Scientists, D. (2022). *Develop and implement a set of techniques or analytics applications to transform raw data into meaningful information using data-oriented programming languages and visualization software*. https://www.onetonline.org/link/summary/15-2051.00

[4] Pejic-Bach, M., Bertoncel, T., Meško, M., & Krstić, Ž. (2020). Text mining of industry 4.0 job advertisements. *International journal of information management,* 50, 416–431. https://doi.org/10.1016/j.ijinfomgt.2019.07.014

[5] Schedlbauer, J., Raptis, G., & Ludwig, B. (2021). Medical informatics labor market analysis using web crawling, web scraping, and text mining. *International journal of medical informatics,* 150, 104453. https://doi.org/10.1016/j.ijmedinf.2021.104453

[6] Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and. *Techniques, morgan kauffman,* 68. https://homes.di.unimi.it/ceselli/IM/slides/16Mining_Freq_Patterns-part2.pdf

[7] Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *ACM sigkdd explorations newsletter*, 2(1), 1–15. https://doi.org/10.1145/360402.360406

[8] Mozaffari, M. (2019). Identifying the skills required for the data science job market in the european union using text mining methods [Thesis]. (**In Persian**). https://B2n.ir/mf6737

[9] Smaldone, F., Ippolito, A., Lagger, J., & Pellicano, M. (2022). Employability skills: Profiling data scientists in the digital labour market. *European management journal*, 40(5), 671–684. https://doi.org/10.1016/j.emj.2022.05.005

[10] Leon, L. A., Seal, K. C., Przasnyski, Z. H., & Wiedenman, I. (2017). Skills and competencies required for jobs in business analytics: A content analysis of job advertisements using text mining. *International journal of business intelligence research (IJBIR)*, 8(1), 1–25. 10.4018/IJBIR.2017010101

[11] Jiang, H., & Chen, C. (2022). Data science skills and graduate certificates: A quantitative text analysis. *Journal of computer information systems*, 62(3), 463–479. https://doi.org/10.1080/08874417.2020.1852628

[12] Chaharsouki, S. K., Nabavi, A., & Teimourpour, B. (2019). A model for predicting the remaining operating time until the critical state based on engine oil analysis records with a data mining approach. *Journal of logistics thought scientific publication,* 18(70), 77-96. (**In Persian**). https://www.magiran.com/p2087205

[13] Qaiser, S., & Ali, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. *International journal of computer applications*, 181(1), 25–29. http://dx.doi.org/10.5120/ijca2018917395

[14] Andrews, J., & Higson, H. (2008). Graduate employability, ‘soft skills’ versus ‘hard’ business knowledge: A European study. *Higher education in europe*, *33*(4), 411–422. https://doi.org/10.1080/03797720802522627