مؤسسه فرهنگی هنری
دیباگران تهران

به نام خدا

علم داده در پایتون

مفاهیم و کاربردها

مؤلفان:

محسن یزدی نژاد(کاندیدای دکتری هوش مصنوعی)

محمدرضا پور (استادیار،عضو هیئت علمی وزارت علوم، تحقیقات و فناوری)

معصومه جوادی نیا(کارشناسی مهندسی نرم افزار)

فرهاد استوارنژاد(کارشناسی ارشد حسابداری)

- عنوان کتاب: **علم داده در پایتون،مفاهیم و کاربردها**
- مولفان : **محسن یزدی نژاد- محمد رضاپور- معصومه جوادی نیا- فرهاد استوارنژاد**
- **ناشر: موسسه فرهنگی هنری دیباگران تهران**
- **ویراستار:** نرگس مهربد
- **صفحه آرایی:**نازنین نصیری
- **نوبت چاپ:** اول
- **تاریخ نشر:** ۱۴۰۱
- **چاپ و صحافی:**صدف
- **تیراژ:**۱۰۰ جلد
- **قیمت:** ۱۱۶۰۰۰۰ ریال
- **شابک:** ۹-۵۲۹-۲۱۸-۶۲۲-۹۷۸

نشانی واحد فروش: تهران، میدان انقلاب، خ کارگر جنوبی، روبروی پاساژ مهستان، پلاک ۱۲۵۱-تلفن: ۶۶۴۱۰۰۴۶-۲۲۰۸۵۱۱۱

فروشگاههای اینترنتی دیباگران تهران :
WWW.MFTBOOK.IR
www.dibagarantehran.com

عنوان و نام پدیدآور:علم داده در پایتون:مفاهیم و کاربردها/
مولفان:محسن یزدی نژاد و...دیگران؛
ویراستار:نرگس مهربد.
مشخصات نشر: تهران : دیباگران تهران :۱۴۰۱
مشخصات ظاهری:۲۰۰ ص:مصور،جدول
شابک: ۹-۵۲۹-۲۱۸-۶۲۲-۹۷۸
وضعیت فهرست نویسی: فیپا
یادداشت:مولفان:محسن یزدی نژاد،محمد رضاپور،معصومه جوادی نیا،فرهاد استوارنژاد.
موضوع:پایتون(زبان برنامه نویسی کامپیوتر)
موضوع:Python(computer program language)
موضوع: زبان شناسی کامپیوتری computational linguistics
موضوع: داده کاوی Data mining
شناسه افزوده:یزدی نژاد،محسن،۱۳۶۲-
رده بندی کنگره: ۷۶/۷۳ QA
رده بندی دیویی:۰۰۵/۱۳۳
شماره کتابشناسی ملی: ۸۸۰۵۲۴۶

فهرست مطالب

مقدمه ناشر

خط مشی انتشارات مؤسسه فرهنگی هنری دیباگران تهران در عرصه کتاب هایی با کیفیت عالی است که بتواند خواسته های به روز جامعه فرهنگی و علمی کشور را تا حد امکان پوشش دهد.

هر کتاب دیباگران تهران، یک فرصت جدید شغلی و علمی

حمد و سپاس ایزد منان را که با الطاف بیکران خود این توفیق را به ما ارزانی داشت تا بتوانیم در راه ارتقای دانش عمومی و فرهنگی این مرز و بوم در زمینه چاپ و نشر کتب علمی و آموزشی گام‌هایی هرچند کوچک برداشته و در انجام رسالتی که بر عهده داریم، مؤثر واقع شویم.

گستردگی علوم وسرعت توسعه روزافزون آن، شرایطی را به وجود آورده که هر روز شاهد تحولات اساسی چشمگیری در سطح جهان هستیم. این گسترش و توسعه، نیاز به منابع مختلف از جمله کتاب را به عنوان قدیمی‌ترین و راحت‌ترین راه دستیابی به اطلاعات و اطلاع‌رسانی، بیش از پیش برجسته نموده است.

در این راستا، واحد انتشارات مؤسسه فرهنگی هنری دیباگران تهران با همکاری اساتید، مؤلفان، مترجمان، متخصصان، پژوهشگران و محققان در زمینه های گوناگون و مورد نیاز جامعه تلاش نموده برای رفع کمبودها و نیازهای موجود، منابعی پُربار، معتبر و با کیفیت مناسب در اختیار علاقمندان قرار دهد.

کتابی که در دست دارید تألیف " **دکتر محسن یزدی نژاد،دکتر محمدرضاپور،مهندس معصومه جوادی نیا،مهندس فرهاد استوارنژاد**" است که با تلاش همکاران ما در نشر دیباگران تهران منتشرگشته و شایسته است از یکایک این گرامیان تشکر و قدردانی کنیم.

با نظرات خود مشوق و راهنمای ما باشید

با ارائه نظرات و پیشنهادات وخواسته های خود،به ما کمک کنید تا بهتر و دقیق تر در جهت رفع نیازهای علمی و آموزشی کشورمان قدم برداریم.برای رساندن پیام هایتان به ما از رسانه های دیباگران تهران شامل سایتهای فروشگاهی و صفحه اینستاگرام و شماره های تماس که در صفحه شناسنامه کتاب آمده استفاده نمایید.

مدیر انتشارات

مؤسسه فرهنگی هنری دیباگران تهران
dibagaran@mftplus.com

مقدمه نویسندگان

جایگاه «هوش مصنوعی» در دنیا از سطح متخصصین و مهندسان فراتر رفته و مورد توجّه رهبران برجسته‌ی جهان قرار گرفته است. این‌که در کشور ما در چه وضعیتی قرار داریم مشخص است که از نظر «نیروی انسانیِ» توانمند چیزی کم از کشورهای پیشرفته در این عرصه کم نداریم؛ همچنین علاوه بر جوانانِ پای کار و پرتلاش، اکنون می‌توان در اغلب فراخوان‌های مجلات علمی و همایش‌ها دید که در هر رشته‌ای که باشند جویای کاربرد هوش مصنوعی در آن حوزه بوده و خیلی مشتاق‌اند که بدانند در زمینه‌ی تخصصی‌شان چه توصیفی وجود دارد و اگر هم بشود از وضعیت آینده‌شان باخبر شوند.

در طرح‌های پژوهشی که از سوی سازمان‌ها مورد حمایت قرار می‌گیرند نیز محورهایی از این زیرشاخه‌های هوش مصنوعی برجسته می‌شوند و بنابراین استقبال دانشگاه و صنعت از هوش مصنوعی قابل ملاحظه می‌باشد. هر چند سهم کمتر از یک درصدی شرکت‌های دانش‌بنیان از تولید ناخالص داخلی، مطلوب نبوده و رهبر عزیز(حفظه الله تعالی) نیز تأکید کرده‌اند: «این سهم باید حتماً افزایش یابد و در طول ۳ تا ۴ سال، حداقل به ۵ درصد برسد.»[۱]

تا اینجای بحث، به دو عامل «حمایت حاکمیتی» و «نیروی انسانی» اشاره کردیم اما محقّق شدنِ این چشم‌انداز آیا تا همین جای کار مسلّم خواهد بود و به عامل زمینه‌ساز دیگری نیاز ندارد؟ باید خیلی صریح پاسخ بگوئیم که: خیر! اجازه دهید علت را با مثالی به عرض شما خوانندگان عزیز برسانیم:

در سال ۱۹۹۸ میلادی دو دانشجوی دکترای دانشگاه استنفورد، ابزاری برای وبگردی ساختند که کم کم نام این ابزار شد Google! مایکروسافت هم در فضای رقابتی موجود، بدش نمی‌آمد گوی و میدان را از گوگل برُباید و مانند سیستم‌عامل کدبسته‌ی ویندوز، محصول خودش در این عرصه را نیز به جهانیان دیکته کند! بنابراین، پس از یک سرمایه‌گذاری عظیم و البته کارهای زرنگ‌مآبانه و عادت غربی‌ها در جذب نخبگان، و این بار با جذب نخبگانی از تیم فنّی گوگل توسط مایکروسافت، بالاخره موتور جستجوی مایکروسافتی با نام Bing در سال ۲۰۰۹، یعنی چیزی در حدود ۱۳ سال بعد از گوگل، رونمایی شد.

امّا این بار دیگر عوامل غیرمعمول – شاید شانس و شاید حمایتهای ویژه‌ی دولتی- منجر به پیشتازی بیل‌گیتس نشد و موتور بینگ نتوانست در زمره‌ی برگ زرّین دیگری در خاطرات این مرد پولدار نقش ببندد و با حضور در بین مهندسین، خیلی سربلند بگوید که من اخراجی دانشگاهم ولی از همه شما موفق‌تر و در واقع همان پولدارترم!

«گیتس» هر چه کرد نشد که نشد و طبق آخرین برآوردها، بیش از ۸۰٪ بازارِ موتورهای جستجو همچنان به گوگل اختصاص یافت و سهم بینگ در خوشبینانه‌ترین حالت به ۵٪ هم نمی‌رسد! این بار نه سرمایه‌گذاری

---

[۱] دیدار نخبگان و استعدادهای برتر علمی با رهبر انقلاب، ۲۶ آذر ۱۴۰۰

مایکروسافت از گوگل کمتر بود و نه تیم فنّی ضعیف‌تری داشت زیرا اگر هر کدام از این علّت‌ها باعث این ناکامی شده بود، قطعا «بیل» وارد چنین رقابتی نمی‌شد.

البته در این زمینه چندان نمی‌توان بیل‌گیتس را سرزنش کرد زیرا علت این اتفاق در آن زمان قابل پیش‌بینی نبود، و شاید تصور چنین بوده که ۱۳ سال تأخیر از گوگل را با سرمایه‌گذاری بیشتر بتوان جبران نمود. اما در این حوزه، یعنی «هوش مصنوعی» و هسته‌ی آن «یادگیری ماشین»، داده‌ها اهمیت بالایی دارند و به‌ویژه، جمع‌آوری رفتارهای کاربران توسط الگوریتم‌های هوشمند جستجو، خزیدن (کروالینگ) در بین وبلاگ‌ها، اطلاع از نیازهای مشتریان و نهایتاً برچسب زنیِ دادگان و وجود «داده‌های برچسب‌خورده»، نقش تعیین‌کننده‌ای در «پایدار ماندنِ مزیت رقابتی» یک فناوری دارند که همه‌ی اینها را مدت‌ها قبل گوگل شروع کرده بود و آن‌قدر فاصله گرفته بود که مایکروسافت هم نتوانست از وی سبقت بگیرد.

البته فناوری‌های از این نوع، یک ویژگی غیرملموس هم دارند و آن اینکه علاوه بر «خود ساختن» که در پاراگراف قبل اشاره شد، از «تکمیل‌شدنی توسط کاربران» نیز برخوردارند؛ یعنی شما وقتی از یک محصول خودرویی استفاده می‌کنید، به‌عنوان یک مصرف‌کننده نقشی در آموزش دادنِ خودرو از رفتارتان و از انتظاری که از خودرو دارید نخواهید داشت؛ به بیان ملموس‌تر، انسان‌ها در تعامل با همدیگر و در زندگی، همچنان که از خدمات هم بهره‌مند می‌شوند، به مرور باعث تکامل طرف مقابل نیز بایستی بشوند (البته وقتی با دوست خوب مواجهیم!!) و سامانه‌های مبتنی بر «هوش مصنوعی» در این ویژگی‌هاست که به انسان‌ها شبیه می‌شوند، زیرا این‌گونه سامانه‌ها اساساً حیات‌شان وابسته به «آموزش دیدن» و «تکمیل شدنِ مداوم» است؛ آری، گوگل این ثانیه، با گوگل ثانیه‌ی قبل فرق دارد، و هر ثانیه‌اش آموزش‌دیده‌تر از ثانیه‌ی قبلش شده و این روند ۱۳ سال قبل از بینگ شروع شده بود و کاربران هم در آن ۱۳ سال به گوگل خودشان کمک کرده بودند تا بفهمد از وی چه می‌خواهند و در ادامه نیز وقتی به نیاز همدیگر پی بُردند، طبیعی بود که از وی به ابزار دیگری که تازه راه رفتن را داشت یاد می‌گرفت، روی نیاورند. در واقع، کودک نحیفِ بینگ، حتما نمی‌توانست جای نوجوان فربه‌ی گوگل را در این کارزار بستاند هر چند که مایکروسافت، لباس‌های قشنگ و گران بر تنش بکشاند!

از مثال فوق – بدون آنکه خواسته باشیم روی لوگوی خاصی تأکید کرده باشیم- بگذریم و به عامل سوم موفقیت هوش مصنوعی بپردازیم و آن هم وابستگی شدیدش به «داده‌های تمیز و برچسب‌خورده‌ای» است که کاربران نیز نقش مؤثری در این آماده‌سازی داده‌ها ایفا می‌کنند.

وضعیت فعلی در کشور، که نمی‌توان اهمیت یافتن «هوش مصنوعی» را در آن کتمان کرد، مرهون تأکید رهبر عزیز می‌باشد که در فراراهبرد مد نظر ایشان «این مسئله در اداره آینده دنیا نقش دارد و باید به‌گونه‌ای عمل کنیم که ایران جزو ۱۰ کشور برتر هوش مصنوعی در دنیا قرار بگیرد.»[۱]

---

[۱] همان مأخذ قبل

بنابراین کتاب حاضر در راستای چند کتابی که در عرصه‌ی هوش مصنوعی با عناوین «**هم‌اکنون، داده‌کاوی**»، «**ارائه راهکار برای تمام مسائل یادگیری ماشین**» و «**آمار کاربردی در علم داده**» ارائه کرده‌ایم، به مبحث جدّی و سازنده‌ی «علم داده‌ها» می‌پردازد و ابزار انتخاب شده هم پایتون است زیرا هم کدباز است هم یادگیری آن برای کاربران با سطوح مختلف قابل شروع کردن است. کدباز بودن برنامه فقط منجر به راحتی و آماده‌بودنِ قطعه‌کدها توسط دیگر برنامه‌نویسان نبوده، بلکه علاوه بر این هم‌افزایی، قابلیت این نرم‌افزار و چند هزار کتابخانه‌ی آن در پردازش انواع مختلف داده‌ها هم مطلب دیگر برای انگیزه بهره‌برداری از آن بوده و به‌علاوه، رعایت حق کپی مسأله‌ای است که شما اگر مقاله‌ای چاپ کنید و به نرم‌افزاری تجاری برای تحلیل داده‌هایتان اشاره کنید که کدباز نباشد، حتما لازم است اجازه‌ی شرکت مربوطه و مسائل مالی آنرا اثبات نمائید که در پایتون با چنین مشکلی هم مواجه نخواهید بود. نکته‌ی پایانی اینکه کدهای کتاب در آدرس مندرج در آخرین منبع کتاب نیز در دسترس عزیزان علاقمند قرار داده شده است.

متن کتاب حاضر در ایام بزرگداشت حکیم نظامی تفرشی به اتمام رسید و دیباچه را نیز با بیتی از آن حکیم پایان می‌بریم:

خداوندا در توفیق بگشای نظامی را ره تحقیق بنمای

و در این حد اشاره نماییم که «تحقیق» به‌دنبال حقیقت بودن است، و علم داده‌ها نیز به دنبال چنین معرفتی است...

فصل اول:

مقدمه‌ای بر داده‌کاوی

python