الفبای فارسی در یونیکد

فهرست مطالب
۱ مقدمه
۲ تاریخچه
۳ ساختار
۴ حروف
۵ اعداد
۶ علائم نگارشی
۷ کاراکترهای کنترلی
۸ ترکیب حروف
۹ شکل‌های مختلف
۱۰ حروف اضافی
۱۱ کدگذاری
۱۲ UTF-8
۱۳ UTF-16
۱۴ UTF-32
۱۵ BOM
۱۶ نرمال‌سازی
۱۷ مشکلات رایج
۱۸ فونت‌ها
۱۹ پشتیبانی سیستم‌عامل‌ها
۲۰ ویندوز
۲۱ لینوکس
۲۲ مک‌او‌اس
۲۳ اندروید
۲۴ iOS
۲۵ مرورگرها
۲۶ ورد‌پردازها
۲۷ ویرایشگرها
۲۸ صفحه‌کلید
۲۹ ورودی
۳۰ خروجی
۳۱ چاپ
۳۲ وب
۳۳ HTML
۳۴ CSS
۳۵ جاوااسکریپت
۳۶ پایگاه داده
۳۷ جستجو
۳۸ مرتب‌سازی
۳۹ تبدیل
۴۰ ابزارها
۴۱ منابع
۴۲ مراجع
۴۳ پیوندها

الفبای فارسی در یونیکد مجموعه‌ای از نویسه‌ها است که برای نمایش متون فارسی در سامانه‌های رایانه‌ای و شبکه‌های ارتباطی تعریف شده است. این مجموعه بخشی از استاندارد یونیکد است و هدف آن ایجاد امکان تبادل، ذخیره و پردازش یکسان متون فارسی در محیط‌های گوناگون است. پیش از پیدایش یونیکد، کدگذاری‌های متعددی برای فارسی وجود داشت که با یکدیگر سازگار نبودند و همین امر مشکلات فراوانی در تبادل اطلاعات ایجاد می‌کرد.

تاریخچه این کار به دهه ۱۹۹۰ میلادی بازمی‌گردد. در آن زمان شرکت‌ها و سازمان‌های ایرانی هر کدام از کدگذاری خاص خود استفاده می‌کردند و متن نوشته‌شده در یک سیستم در سیستم دیگر به‌درستی نمایش داده نمی‌شد. با گسترش اینترنت این مشکل بیشتر آشکار شد و ضرورت یک استاندارد واحد احساس شد. نخستین نسخه‌ای که حروف فارسی را در یونیکد معرفی کرد در سال ۱۹۹۲ منتشر شد و از آن پس در نسخه‌های بعدی تکمیل و اصلاح شد.

ساختار الفبای فارسی در یونیکد بر پایه جدول کاراکترها است. هر نویسه یک شماره یکتا دارد که به آن نقطه کد می‌گویند. این شماره‌ها در محدوده‌ای از جدول قرار گرفته‌اند که به زبان‌های خاورمیانه اختصاص دارد. حروف اصلی الفبا، ارقام، علائم نگارشی و نویسه‌های کنترلی هر کدام در بخش مشخصی از این محدوده جای گرفته‌اند تا پیدا کردن و استفاده از آن‌ها آسان باشد.

حروف الفبای فارسی شامل ۳۲ حرف اصلی است. علاوه بر این، حروف اضافی مانند پ، چ، ژ و گ نیز در یونیکد وجود دارند. هر حرف در حالت‌های مختلف شکل‌گیری ممکن است شکل متفاوتی داشته باشد؛ برای مثال حرف ب در ابتدای کلمه، وسط کلمه، پایان کلمه و به‌صورت مستقل شکل‌های گوناگونی دارد. یونیکد برای هر یک از این شکل‌ها نقطه کد جداگانه تعریف کرده است.

ارقام فارسی نیز در یونیکد جداگانه تعریف شده‌اند. ارقام فارسی با ارقام عربی که در انگلیسی استفاده می‌شوند متفاوت‌اند و هر کدام نقطه کد خاص خود را دارند. علائم نگارشی مانند نقطه، ویرگول، علامت سؤال و پرانتز نیز در بخش مشترک جدول قرار گرفته‌اند تا با سایر زبان‌ها سازگار باشند.

کدگذاری یونیکد روش‌های مختلفی دارد که رایج‌ترین آن‌ها UTF-8 است. در این روش هر نویسه با یک تا چهار بایت نمایش داده می‌شود. UTF-16 و UTF-32 نیز وجود دارند اما کمتر استفاده می‌شوند. BOM یا نشانه ترتیب بایت در ابتدای فایل قرار می‌گیرد و مشخص می‌کند که بایت‌ها به چه ترتیبی خوانده شوند.

نرمال‌سازی یکی از مهم‌ترین مباحث در پردازش متون فارسی است. چون یک کلمه ممکن است به چند شکل مختلف در یونیکد نوشته شود، برنامه‌ها باید آن‌ها را به یک شکل استاندارد تبدیل کنند تا جستجو و مقایسه درست انجام شود. فرم‌های نرمال‌سازی مختلفی تعریف شده‌اند که هر کدام قواعد خاص خود را دارند.

مشکلات رایج در استفاده از الفبای فارسی در یونیکد شامل نمایش نادرست حروف، جهت نوشتار و ترکیب حروف است. این مشکلات معمولاً به دلیل پشتیبانی ناقص فونت‌ها یا نرم‌افزارها رخ می‌دهند. فونت‌هایی که برای فارسی طراحی شده‌اند باید تمام شکل‌های حروف را داشته باشند تا متن به‌درستی نمایش داده شود.

سیستم‌عامل‌های مختلف پشتیبانی متفاوتی از یونیکد دارند. ویندوز، لینوکس، مک‌او‌اس، اندروید و iOS هر کدام روش خاص خود را برای ورودی و نمایش متون فارسی دارند. مرورگرهای وب نیز در نحوه رندر کردن متون فارسی تفاوت‌هایی دارند که گاهی باعث بروز مشکل می‌شود.

در برنامه‌نویسی وب، HTML و CSS ابزارهایی برای مدیریت متون فارسی فراهم می‌کنند. ویژگی‌هایی مانند dir برای جهت نوشتار و lang برای مشخص کردن زبان متن به موتورهای جستجو کمک می‌کنند تا محتوا را بهتر ایندکس کنند. جاوااسکریپت نیز توابعی برای کار با رشته‌های یونیکد دارد.

پایگاه‌های داده رابطه‌ای و غیررابطه‌ای باید برای ذخیره متون فارسی تنظیم شوند. انتخاب نادرست مجموعه کاراکتر می‌تواند باعث از دست رفتن اطلاعات شود. جستجو و مرتب‌سازی متون فارسی نیز نیاز به الگوریتم‌های خاص دارد که با ترتیب الفبایی فارسی سازگار باشند.

ابزارهای متعددی برای تبدیل متون از کدگذاری‌های قدیمی به یونیکد وجود دارند. این ابزارها معمولاً بر پایه جدول‌های تبدیل کار می‌کنند و در مواردی که متن ترکیبی از چند کدگذاری باشد ممکن است نتیجه دقیق نباشد. بهتر است از ابتدا همه سامانه‌ها را روی یونیکد تنظیم کرد.

منابع بیشتر درباره الفبای فارسی در یونیکد شامل مستندات رسمی یونیکد، راهنماهای توسعه‌دهندگان و مقالات تخصصی است. مراجع این مقاله نیز فهرستی از این منابع را ارائه می‌دهد. پیوندهای مرتبط به صفحه‌هایی اشاره دارند که اطلاعات تکمیلی درباره جدول کاراکترها و ابزارها ارائه می‌کنند.