

پژوهشکده فناوری اطلاعات گروه سکویهای فناوری اطلاعات و فضای مجازی

گزارش فنی


تحلیل وردنت زبان فارسی در حوزه فاوا

مستخرج از پروژه: توسعه وردنت زبان فارسی در حوزه فاوا

کد پروژه: ۸۹۳۲۴۱۶

محرم منصوری زاده
محرم منصوری زاده، محمد نصیری،
محمد دادرس
ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷.۰۲
۹۱/۰۳/۲۷
۲۰/نهایی


مجری:
تهیه کننده/ تهیه کنندگان:
کد گزارش:
تاریخ ارائه:
نسخه/ وضعیت

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات




در راستای تحقق مأموریت پژوهشگاه ارتباطات و فناوری در فراهم سازی سکویی برای ارتقاء دانش، انتقال فناوری و بومی سازی محصولات و خدمات حوزه فاوا و با هدف جلب مشارکت علاقه مندان در توسعه و بهره مندی از دستاوردهای پژوهشگاه ارتباطات و فناوری اطلاعات، آزاد رسانی این دستاوردها در زمره برنامه های اولویت دار پژوهشگاه به شمار می آید. به همین منظور مستند حاضر تحت مجوز بین المللی **CC-BY-SA-NC** نسخه ۴، در دسترس عموم قرار گرفته است. شایان ذکر است تحت این مجوز، ضمن حفظ مالکیت فکری این مستند برای پژوهشگاه ارتباطات و فناوری اطلاعات، بازانتشار و بکارگیری آن صرفاً برای موارد تحقیقاتی و با ذکر نام پژوهشگاه ارتباطات و فناوری اطلاعات بلامانع است.

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات
شناسنامه گزارش			
شماره نسخه: ۲,۰		عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا	
تاریخ ارائه گزارش: ۹۱/۰۳/۲۷		نوع گزارش: فنی	کد: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲
نوع پروژه: راهبردی-توسعه ای - بنیادی		نام پروژه: توسعه وردنت زبان فارسی در حوزه فاوا	
تاریخ پایان: ۹۱/۱۱/۱۱		تاریخ شروع: ۹۰/۰۴/۲۸	
نام گروه: سکویهای فناوری اطلاعات و فضای مجازی			
شماره و تاریخ قرارداد: ۹۰/۰۴/۲۸-۵۰/۶۶۱۴/ت		کد پروژه: ۸۹۳۲۴۱۶	
ناظر/ ناظرین: مریم محمودی، آنیسا هادیزاده، بهروز مینایی بیدگلی		مجری: محرم منصوری زاده	
تهیه کننده/ تهیه کنندگان: محرم منصوری زاده، محمد نصیری، محمد دادرس			
نام و نشانی مجری: همدان، دانشگاه بوعلی سینا، دانشکده مهندسی، گروه مهندسی کامپیوتر، کد پستی ۶۵۱۴۸۳۳۶۹۵، تلفن: ۰۸-۸۲۹۲۵۰۵-۸۱۱ فکس: ۰۸۱۱-۸۲۹۲۶۳۱			
نام و نشانی حمایت کننده: تهران، انتهای خیابان کارگر شمالی، پژوهشگاه فضای مجازی، کد پستی ۱۴۳۹۹۵۵۴۷۱، تلفن: ۸۴۹۷۷۷۷			
ملاحظات: ندارد			
چکیده:			
<p>در این سند تحلیل نیازها و دامنه پروژه توسعه وردنت زبان فارسی در حوزه فاوا ارائه شده است. تعریف حوزه فاوا و زیردامنه های آن، نحوه جمع آوری واژگان، متدلوژی ساخت وردنت فاوا و همچنین چرخه حیات واژگان به تفصیل در این سند مورد بحث قرار گرفته است. استراتژی اصلی ما ساخت یک وردنت دوزبانه انگلیسی-فارسی فاوا است که طی آن مفاهیم تخصصی حوزه فاوا از منابع متنوعی مانند کتابهای تخصصی، فرهنگ نامه‌ها و سایت‌های وب جمع آوری می‌شوند و سپس ارتباطات بین آنها به کمک روش‌های متن کاوی خودکار و نیمه خودکار استخراج می‌گردد. این مجموعه واژگان و روابط بین آنها وردنت انگلیسی فاوا را تشکیل می‌دهند که در مرحله دوم به فارسی ترجمه شده و روابط ساختاری بین واژگان به آن اضافه می‌شود. روش‌ها و ایده‌های مطرح شده در این سند طی فازهای مختلف پروژه پیاده سازی شده و نتایج آنها در سندهای جداگانه ارائه می‌شود.</p>			
کلمات کلیدی: وردنت فارسی، حوزه فاوا، تحلیل			
وضعیت گزارش: نهایی		زبان گزارش: فارسی	
وضعیت دسترسی: عادی		تعداد صفحات: ۳۲	

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده فناوری اطلاعات
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	

چکیده


در این سند تحلیل نیازها و دامنه پروژه توسعه وردنت زبان فارسی در حوزه فاوا ارائه شده است. تعریف حوزه فاوا و زیردامنه های آن، نحوه جمع آوری واژگان، متدلوژی ساخت وردنت فاوا و همچنین چرخه حیات واژگان به تفصیل در این سند مورد بحث قرار گرفته است. استراتژی اصلی ما ساخت یک وردنت دوزبانه انگلیسی-فارسی فاوا است که طی آن مفاهیم تخصصی حوزه فاوا از منابع متنوعی مانند کتابهای تخصصی، فرهنگ نامه‌ها و سایت‌های وب جمع آوری می‌شوند و سپس ارتباطات بین آنها به کمک روش‌های متن کاوی خودکار و نیمه خودکار استخراج می‌گردد. این مجموعه واژگان و روابط بین آنها وردنت انگلیسی فاوا را تشکیل می‌دهند که در مرحله دوم به فارسی ترجمه شده و روابط ساختاری بین واژگان به آن اضافه می‌شود. روش‌ها و ایده‌های مطرح شده در این سند طی فازهای مختلف پروژه پیاده سازی شده و نتایج آنها در سندهای جداگانه ارائه می‌شود.

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

لیست مستندات مرتبط		
شماره مستند	نوع مستند	نام مستند

لیست تغییرات اعمال شده در نسخه های قبلی گزارش		
شماره نسخه	تاریخ	تغییرات اعمال شده
۱،۰	۱۳۹۰/۰۸/۱۱	نسخه اولیه
۲،۰	۱۳۹۱/۰۳/۲۷	نسخه نهایی


تایید کنندگان				
ملاحظات	امضاء	تاریخ	نام و نام خانوادگی	
			محرم منصوری زاده	مجری پروژه
			محرم منصوری زاده، محمد نصیری، محمد دادرس	تهیه کننده / تهیه کنندگان
			مریم محمودی، آنیته هادیزاده، بهروز مینایی بیدگلی	ناظر پروژه
			علیرضا یاری	مدیر گروه
			زهره ساعی	مسئول مستندات پژوهشکده
			علیرضا یاری	رئیس پژوهشکده / معاون پژوهشی

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

تقدیر و تشکر


بدین وسیله از ناظرین و مشاورین محترم پروژه قدردانی می‌شود. همچنین از سرکار خانم دکتر نعمت زاده و جناب آقای دکتر مینایی به خاطر پیشنهادهای ارزشمندشان سپاسگزاریم. در تدوین این گزارش و همچنین تعریف زیر دامنه های حوزه فاوا از راهنمایی‌ها و مشاوره های بسیار مفید کارشناسان و اساتید فناوری اطلاعات و زبان شناسی در دانشگاهی داخل و خارج کشور بهره برده‌ایم. کمترین کاری که می‌توان کرد، نام بردن از این عزیزان است. امیدواریم در آینده نیز بتوانیم از راهنمایی‌های سودمند این بزرگواران بهره‌مند باشیم.

ردیف	گروه	نام و نام خانوادگی	محل کار
۱	فاوا	دکتر سید وحید ازهری	دانشگاه علم و صنعت ایران
۲	فاوا	مهندس حسن بشیری	دانشگاه صنعتی همدان
۳	فاوا	مهندس نرگس بطحائیان	دانشگاه بوعلی سینا
۴	فاوا	دکتر حسن ختن‌لو	دانشگاه بوعلی سینا
۵	فاوا	دکتر میرحسین دزفولیان	دانشگاه بوعلی سینا
۶	فاوا	دکتر غلامرضا شاه محمدی	دانشگاه علوم انتظامی
۷	فاوا	دکتر محمد قدسی	دانشگاه صنعتی شریف
۸	فاوا	دکتر نصرا...	دانشگاه تربیت مدرس
۹	فاوا	مهندس محسن ملانوری	دانشگاه کالگری، کانادا
۱۰	فاوا	دکتر علی یزدیان	دانشگاه تربیت مدرس
۱۱	زبان‌شناسی	دکتر محمد راسخ مهند	دانشگاه بوعلی سینا
۱۲	زبان‌شناسی	دکتر امید طبیب زاده	دانشگاه بوعلی سینا

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات


فهرست مطالب

شماره صفحه	عنوان
۱	۱- درباره سند تحلیل
۲	۲- تعریف مسئله
۲	۲-۱- لزوم تولید وردنت زبان فارسی در حوزه فاوا
۲	۲-۲- شرح وظایف سیستم
۳	۲-۳- کارفرما، کاربران و همه افراد دخیل در سیستم
۳	۲-۴- بررسی پیشینه کارهای مرتبط
۵	۳- حوزه فناوری اطلاعات و ارتباطات (فاوا)
۵	۳-۱- داده، اطلاعات و دانش
۵	۳-۲- فناوری اطلاعات و ارتباطات چیست؟
۶	۳-۳- زیر دامنه های حوزه فاوا
۷	۳-۴- زبان و واژگان حوزه فاوا
۹	۳-۵- جمع بندی
۱۰	۴- روش شناسی توسعه وردنت فاوا
۱۱	۴-۱- نحوه تعامل تیم زبان شناسی و تیم فناوری اطلاعات
۱۲	۴-۲- توسعه هسته اولیه وردنت فاوا
۱۳	۴-۳- توسعه شمای واژگانی وردنت فاوا
۱۴	۴-۴- توسعه وردنت فاوا به زبان انگلیسی فاوا
۱۴	۴-۴-۱- استخراج واژگان و اصطلاحات تخصصی حوزه فاوا در زبان انگلیسی
۱۵	۴-۴-۲- تعریف مترادفها و استخراج روابط معنایی بین آنها
۱۵	۴-۵- ترجمه وردنت انگلیسی فاوا به فارسی
۱۶	۴-۶- آزمون صحت و کیفیت
۱۷	۵- چرخه حیات یک واژه
۱۷	۵-۱- چرخه حیات واژه در وردنت انگلیسی
۱۸	۵-۲- چرخه حیات واژه در وردنت فارسی
۱۹	۶- جمع بندی

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات


فهرست اشکال

- شکل ۳-۱) نمودار بسامد واژگان مقوله های مختلف دستوری ۹
- شکل ۴-۱) توسعه وردنت دوزبانه انگلیسی-فارسی فاوا ۱۰
- شکل ۴-۲) توسعه حجمی وردنت فاوا ۱۱
- شکل ۴-۳) توسعه هسته اولیه وردنت دوزبانه انگلیسی-فارسی فاوا ۱۳


	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶،۱۱.۷.۰۲	فناوری اطلاعات

فهرست جداول

- جدول ۱-۳) عناوین تعدادی از زیر دامنه های حوزه فاوا ۶
- جدول ۲-۳) بسامد واژگان زبان در مقوله های مختلف ۹

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶،۱۱.۷.۰۲	فناوری اطلاعات

فهرست اختصارات

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات


۱- درباره سند تحلیل

کلیات

هدف اصلی این سند تعریف دقیق نیازمندی‌ها، کاربردها و محدودیت‌های سیستم فوق می‌باشد. به صورت خلاصه، اهداف کلی تنظیم این سند عبارتند از:

- شناسایی مسائل و مشکلاتی که منجر به احساس نیاز برای وردنت فارسی فاوا شده‌اند
- شناسایی و استخراج نیازمندی‌ها و وظایف وردنت فارسی فاوا
- شناسایی کاربران، بهره برداران و دیگر افراد مرتبط با سیستم
- معرفی حوزه فاوا و تعیین دامنه واژگانی آن
- ارائه راه حل پیشنهادی برای توسعه وردنت فاوا

در ادامه و در فصل‌های بعدی جزئیات هر کدام از مباحث فوق به تفصیل توضیح داده می‌شوند.

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده فناوری اطلاعات
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	

۲- تعریف مسئله

۲-۱- لزوم تولید وردنت زبان فارسی در حوزه فاوا


زبان فارسی زبان رسمی کشور است و بر همین اساس زبان اول اسناد و مکاتبات رسمی و علمی است. از گذشته های دور کتابها، مقالات، گفتارها و نشریات متعددی در حوزه های گوناگون به این زبان تولید و منتشر شده است و هر روز هم بر حجم عظیم این محتوای زبانی اضافه می شود. بهره برداری از این همه محتوای تولید شده و پویا با روش های سنتی کتابخانه ای مقدور نیست و نیاز است که برای کاربردهایی نظیر مطالعات زبان شناختی، بازیابی اطلاعات و ترجمه ماشینی؛ ابزارهای نرم افزاری خودکاری تولید شوند تا این منابع ساختار دهی شده و به صورت منسجمی در دسته بندی های دستوری و معنایی قرار گیرند.

تاکنون مطالعات و تحقیقات متعددی در حوزه عمومی زبان انجام یافته و محصولات مفید و کاربردی هم داشته اند. پیکره واژگانی روزنامه همشهری محصول دانشگاه تهران یا طرح واژگان پایه دانش آموزان دوره ابتدایی محصول وزارت آموزش و پرورش نمونه های قابل ذکر در این حیطه هستند. اما در حوزه های تخصصی مانند فنی و مهندسی یا پزشکی فعالیتی جدی انجام نشده است. در این پروژه وردنت فارسی در حوزه تخصصی فناوری اطلاعات و ارتباطات (فاوا) ساخته می شود. این وردنت باید تسهیلات و امکانات نسخه ۳ وردنت زبان انگلیسی را برای زبان فارسی در این حوزه ارائه دهد.

۲-۲- شرح وظایف سیستم

وردنت فارسی فاوا باید مطابق با وردنت زبان انگلیسی نسخه ۳ بوده و امکانات زیر را داشته باشد:

- ۱- ثبت واژگان به زبان و خط فارسی
- ۲- ۳۰۰۰۰ (سی هزار) مدخل از پرکاربردترین واژه ها در حوزه تخصصی فوق
- ۳- پوشش مقوله های دستوری اسم، صفت، فعل و قید


	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

- ۴- درج روابط میان مقوله ای و بین مقوله ای
- ۵- درج ویژگی های نحوی، ساخت واژگی، آوایی، ویژگی های واژه ها و مجموعه ترادف ها
- ۶- امکان جستجوی پیشرفته و دو زبانه
- ۷- قابلیت اتصال به وردنت های دیگر
- ۸- قابلیت گسترش از لحاظ دامنه و پوشش لغوی و قابلیت تغییر، حذف و درج واژگان با استفاده از واسط کاربری واژه های از سوی کاربران مجاز
- ۹- قابلیت دسترسی و بکار گیری به صورت منفک، با واسط و همچنین به صورت API در اختیار سامانه های دیگر
- ۱۰- نرم افزار مدیریت واژگان به منظور اجازه اجرای عملیات حذف، درج و به هنگام سازی واژگان به صورت برخط از روی شبکه وب
- ۱۱- نحوه ساخت به صورت نیمه خودکار
- ۱۲- ارزیابی در چهار مرحله:
 - a. آزمون سازگاری توسط نرم افزار به منظور جلوگیری از ورود اطلاعات ناسازگار
 - b. تایید تیمی تخصصی به سرپرستی زبان شناسان حاذق و متخصص در حوزه فاوا و در حوزه فرهنگ نگاری
 - c. ثبت در حوزه جهانی وردنت
 - d. پذیرش مقالات

۳-۲- کارفرما، کاربران و همه افراد دخیل در سیستم

- کارفرمای این پروژه موسسه تحقیقات ارتباطات و فناوری اطلاعات است و تیمی تخصصی وظیفه نظارت بر پروژه را بر عهده دارند.
- کاربران این پروژه عبارتند از:
- ۱- ویرایشگران: وظیفه ورود و ویرایش اطلاعات را بر عهده دارند.
 - ۲- پژوهشگران: دانشجویان و اساتید حوزه زبان که برای تحقیقات خود از دادگان این مرجع استفاده می کنند
 - ۳- نرم افزارهای دیگر که با استفاده از واسط برنامه نویسی به این نرم افزار متصل می شوند.

۴-۲- بررسی پیشینه کارهای مرتبط

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده فناوری اطلاعات
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	

وردنت‌های موجود در سند مطالعه تطبیقی به تفصیل معرفی و بررسی شده‌اند. برای اطلاع از آخرین وردنت‌های ساخته شده می‌توانید به سایت انجمن جهانی وردنت^۱ مراجعه کنید. البته این سایت هیچ وردنت تخصصی را گزارش نکرده است.^۲


وردنت‌های موجود را می‌توان در قالب دسته‌های زیر مطالعه کرد:

- **وردنت انگلیسی:** وردنت انگلیسی نسخه اصلی وردنت بوده و ایده ساخت چنین شبکه معنایی برای واژگان این وردنت معرفی شد.
- **وردنت‌های تک زبانه غیر انگلیسی:** این وردنت‌ها ایده وردنت انگلیسی را برای زبان‌های مختلف پیاده‌سازی کرده‌اند. بسیاری از آن‌ها مانند وردنت یونانی، وردنت انگلیسی را به زبان دیگر ترجمه کرده‌اند [۱]. برخی از آن‌ها نیز مانند فارس‌نت، قالب وردنت را حفظ کرده و واژگان و روابط معنایی آن‌ها را بر اساس متون زبان مقصد استخراج نموده‌اند.
- **وردنت‌های چند زبانه:** نمونه شاخص این نوع وردنت‌ها، بالکانت است که زبان‌های شبه جزیره بالکان را پوشش می‌دهد.
- **وردنت‌های تخصصی:** وردنت‌های تخصصی موجود عموماً گسترش وردنت عمومی برای واژگان حوزه‌های تخصصی هستند که از میان آنها می‌توان به گسترش وردنت عمومی یونانی به حوزه روانشناسی را نام برد.

مطالعه روش‌شناسی توسعه وردنت‌های عمومی و گسترش آنها به حوزه‌های تخصصی می‌تواند کمک شایانی به توسعه وردنت فارسی فاوا بنماید.

^۱ <http://www.globalwordnet.org/>

^۲ این اطلاعات بر اساس آخرین مراجعه به سایت فوق در تاریخ ۱۳۹۰/۰۸/۰۱ به دست آمده است.

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده فناوری اطلاعات
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	

۳- حوزه فناوری اطلاعات و ارتباطات (فاوا)

برای درک صحیح از حوزه تخصصی فناوری اطلاعات و ارتباطات (فاوا) نیاز است ابتدا مفاهیم و اصطلاحات آن را تعریف کنیم. در اینجا سعی می‌کنیم تعاریف ساده، مختصر و قابل درکی از هر یک از مفاهیم مورد نظر ارائه کنیم. ادعایی نداریم که این تعاریف جامع و کامل می‌باشند اما حداقل برای اهداف ما کافی هستند.

۳-۱- داده، اطلاعات و دانش

داده^۱ محصول اندازه گیری کمیت‌ها یا سیگنال‌هاست. طول و عرض یک جسم، دمای یک اتاق در زمانی معین و نام گل‌های بومی یک منطقه جغرافیایی نمونه‌هایی از کمیت‌ها هستند. مقادیر اندازه گیری این کمیت‌ها اعداد و علایمی هستند که داده‌ها را تشکیل می‌دهند. به عنوان نمونه طول = ۳ متر و عرض = ۴ متر، دما = ۲۲ درجه سانتی‌گراد و گل بومی = زنبق کوهی داده‌های حاصل از کمیت‌های فوق هستند. اطلاعات^۲ به داده معنی‌دار گفته می‌شود. مثلاً اگر دمای اتاق در مقاطع معینی اندازه گیری و ثبت شود، میانگین حسابی این مقادیر بیانگر دمای متوسط اتاق در طول زمان و نوعی **اطلاعات** است. اگر دما از حد معینی بالاتر رود، باید پنجره را باز کرد. این **قانون** در کنار اطلاعات **دانش**^۳ را می‌سازد.


۳-۲- فناوری اطلاعات و ارتباطات چیست؟

فناوری اطلاعات و ارتباطات (فاوا) اصطلاحی است که تعریف واحدی که همه متخصصان و کاربران بر آن اتفاق نظر داشته باشند؛ ندارد. بسته به نیاز، هر شخص، سازمان و کشوری می‌تواند تعریف مورد نظر خود را از

^۱ Data

^۲ Information

^۳ Knowledge

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

آن ارائه کند. بر اساس گزارشی که سازمان جهانی مخابرات (ITU) درباره اندازه گیری فعالیت‌های مرتبط با فاوا در کشورهای مختلف منتشر کرده است؛ هر کشوری تعریف خاص خود را از این مقوله دارد [۲]. البته نکات مشترکی در اغلب این تعاریف وجود دارد که بیان آنها می‌تواند به درک مفهوم این حوزه تخصصی کمک کند. بر اساس تعاریف موجود در فرهنگ لغات، سایت‌های مرجع کتاب‌های تخصصی این حوزه می‌توان چنین استنباط کرد که :

فناوری اطلاعات و ارتباطات به مجموعه دستگاه‌ها، ابزارها و روش‌هایی اطلاق می‌شود که برای تولید، فرآوری، جابجایی و کاربری اطلاعات بکار می‌روند.

بر اساس تعریف فوق، رادیو، تلویزیون، تلفن، دورنگار، کامپیوتر و اینترنت، چاپگر، اسکنر و دوربین‌های دیجیتال را می‌توان از جمله ادوات فاوا نام برد. نرم افزارها و الگوریتم‌های ثبت و ضبط داده‌ها و تحلیل آن‌ها در زمره ابزارهای تولید و فرآوری اطلاعات قرار می‌گیرند. همچنین انواع شبکه‌ها و پروتکل‌های مورد استفاده در آن‌ها نیز وظیفه جابجایی اطلاعات را بر عهده دارند.

۳-۳- زیر دامنه های حوزه فاوا

حوزه فاوا مشتمل بر زیر دامنه های متعدد و متنوعی مانند تئوری اطلاعات، شبکه های کامپیوتری، بازی‌های رایانه ای می‌باشد. مطالعه حوزه فاوا در قالب زیر دامنه‌ها این امکان را ایجاد می‌کند که برای هر مفهوم واژه حوزه کاربردی خاص تعریف و معین گردد.

چندین انجمن و نهاد علمی مانند ^۱ACM، دایرکتوری وب ^۲dmoz و ^۳Wikipedia زیر دامنه های پیشنهادی خود را برای حوزه فاوا معرفی کرده‌اند. در این پروژه از فهرست پیشنهادی ACM تا دو سطح استفاده می‌کنیم. در جدول زیر عناوین تعدادی از این زیر دامنه‌ها را آورده‌ایم. فهرست کامل زیر دامنه‌ها در پیوست ۳ آمده است.


جدول ۳-۱) عناوین تعدادی از زیر دامنه های حوزه فاوا

عنوان فارسی	عنوان انگلیسی
تحلیل الگوریتم‌ها و پیچیدگی مسئله‌ها	ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY
مداری‌های منطقی و محاسباتی	ARITHMETIC AND LOGIC STRUCTURES
هوش مصنوعی	ARTIFICIAL INTELLIGENCE
کدگذاری و تئوری اطلاعات	CODING AND INFORMATION THEORY

^۱ www.acm.org

^۲ www.dmoz.org

^۳ www.wikipedia.com

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات
COMPUTER GRAPHICS	گرافیک کامپیوتری		
COMPUTER-COMMUNICATION NETWORKS	شبکه های ارتباطی کامپیوتری		
DATA ENCRYPTION	رمزنگاری داده ها		
DATA STRUCTURES	ساختار داده ها		
DATABASE MANAGEMENT	مدیریت پایگاه داده ها		
DISCRETE MATHEMATICS	ریاضیات گسسته		
DOCUMENT AND TEXT PROCESSING	پردازش متن و اسناد		
IMAGE PROCESSING AND COMPUTER VISION	پردازش تصویر و بینایی ماشین		
INFORMATION STORAGE AND RETRIEVAL	ذخیره و بازیابی اطلاعات		
MEMORY STRUCTURES	ساختارهای حافظه		
OPERATING SYSTEMS	سیستم های عامل		
PROCESSOR ARCHITECTURES	ساختار پردازنده ها		
PROGRAMMING LANGUAGES	زبان های برنامه سازی		
SOFTWARE ENGINEERING	مهندسی نرم افزار		
GAMING AND GAME THEORY	بازی سازی و تئوری بازی ها		


باید توجه کرد که برخی از این زیر دامنه ها با علوم دیگر نیز خویشاوندی نزدیکی دارند. مثلاً *الگوریتم* و *محاسبات* هم در زمینه فاوا و هم در زمینه ریاضیات قابل طرح و بررسی است.

۳-۴- زبان و واژگان حوزه فاوا

فاوا علمی پویاست و هر روز نوآوری هایی در جنبه های مختلف آن ظاهر می شود. این پویایی ادبیات فاوا را هم دستخوش تغییر می سازد و واژگان جدیدی به آن اضافه می کند. رواج اینترنت سبب ایجاد واژگان جدیدی در حوزه عمومی زبان نیز شده است. حجم این واژگان جدید آنقدر قابل توجه است که به فرهنگ لغات معتبری مانند آکسفورد فصل جدیدی با عنوان لغات حوزه فناوری اطلاعات و اینترنت اضافه شده است [۳]. بر این اساس منابع واژگانی فاوا باید جدید، به روز و غنی باشند. در این راستا می توان از منابع زیر استفاده کرد.

- اصطلاحنامه ها^۱ و فرهنگ لغت های تخصصی فاوا (یک زبانه و دو زبانه)
- کتاب هایی که در حوزه عمومی و کاربردی فاوا منتشر می شوند و معمولاً مطابق با جدیدترین تغییرات فناوری ویرایش و چاپ می شوند.

^۱ Thesaurus

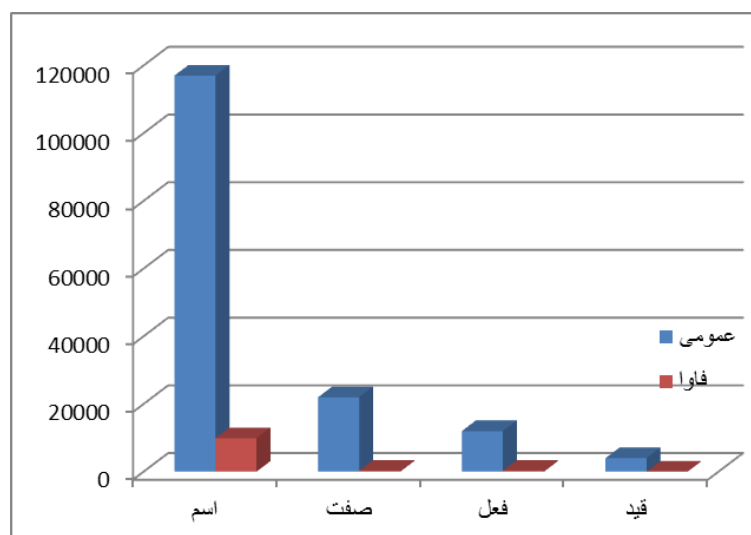
	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

- روزنامه‌ها، مجلات و انواع گاهنامه‌های تخصصی و نیمه تخصصی که عمدتاً تغییرات و اخبار فاوا را منتشر می‌کنند.
- پایگاه‌های وب

بیشتر اصطلاحات فاوا به زبان انگلیسی است و متخصصین این حوزه در محاورات علمی و حتی عمومی از همان صورت انگلیسی استفاده می‌کنند. در مقابل، کاربران عمومی فاوا معمولاً از ترجمه متون به زبان خودشان استفاده می‌کنند. تحول پرشتاب فاوا در این بخش هم اثرگذار است و معادل‌سازی‌ها و ترجمه‌های انجام شده معطل فرایندهای دقیق اما زمان‌بر واژه‌گزینی نمی‌مانند. هر مترجم بنا بر توانایی‌ها و اطلاعات و گاهی سلیقه خودش معادل‌های مناسب را انتخاب می‌کند. به همین دلیل هم هست که در ادبیات فاوا معادل‌های متعددی برای یک اصطلاح یافت می‌شوند. به عنوان نمونه می‌توان به اصطلاح Context Free Grammars اشاره کرد. اگر کتاب‌های طراحی زبان‌های برنامه‌سازی را بررسی کنید به معادل‌هایی مانند *گرامرهای بدون متن، گرامرهای متن باز و گرامرهای مستقل از متن* برخورد خواهید کرد. یک خواننده - به ویژه خواننده ای که از مفهوم این اصطلاح در زبان اصلی آگاهی ندارد- معمولاً در مواجهه با این معادل‌ها دچار ابهام و تردید می‌شود. استفاده از این معادل‌های ناهمگون در پردازش‌های لغوی و زبانی بسیار زیان‌آورتر است. زیرا یک مفهوم در زبان مبدأ با چندین اصطلاح در زبان مقصد معرفی شده که هر کدام از این اصطلاحات حوزه معنایی خاص و متفاوت خود را دارند. مثلاً *گرامر بدون متن* بیانگر این مفهوم است که گرامر مذکور پارامتری با نام متن را ندارد و مثلاً ساده تر و محدود تر از گرامرهای دارای متن است. در حالی که گرامر مستقل از متن نشان دهنده این است که این گرامر نیازی به متن ندارد و دامنه آن وسیع‌تر است. بر اساس این حقایق، پژوهش‌های لغوی و معنایی واژگانی باید بر اساس همان صورت اصلی واژه انجام گیرد و نهایتاً در صورت لزوم یافته‌های آن به زبان‌های مقصد ترجمه گردد.

مجموعه واژگان حوزه فاوا را از منابع فوق‌الذکر می‌توان استخراج کرد. البته تشخیص واژگان فاوا از واژگان غیر فاوا کاملاً خودکار نیست اما می‌توان بخش‌های زیادی از آن را به صورت خودکار انجام داد. مثلاً استخراج واژگان پرکاربرد کتاب‌ها و سپس پالایش خودکار و دستی آن‌ها یکی از این روش‌هاست. در این روش بسامد تک‌تک واژگان کتاب‌ها (یا هر منبع مورد استفاده دیگر) محاسبه شده و واژگانی که بسامد آن‌ها از حد معینی بالاتر باشد؛ استخراج می‌شوند. پر بسامدترین واژگان کتاب‌های فارسی کلمات و حروف ربط مانند که، و، به، می و مانند آن و کتاب‌های انگلیسی ، the, it, is و موارد مشابه هستند. این نوع کلمات، کلمات ایست (توقف) می‌نامند. واژگان پر بسامد بعد از کلمات ایست هستند که امید می‌رود جزو کلمات تخصصی کتاب باشند. بر اساس مطالعاتی که Nation انجام داده است، تنها ۲۰ درصد کلمات پر بسامد کتاب‌های علمی و فنی نیمه تخصصی و حداکثر ۱۰ درصد آن‌ها جزو کلمات تخصصی محسوب می‌شوند [۴]. مجموعه اولیه واژگان ممکن است بیش از صد هزار واژه داشته باشد. انتخاب دستی واژگان تخصصی کاری زمان‌بر بوده و به دقت زیاد نیاز دارد. برای حذف کلمات غیر ضروری می‌توان از فرهنگ لغت‌های عمومی بهره گرفت.

مقوله های دستوری اسم، صفت، فعل و قید جداگانه در وردنت مطالعه می‌شوند. همان‌طوری که در جدول و شکل زیر دیده می‌شود، بیشتر واژگان بررسی شده از مقوله اسم هستند. برای مقایسه حوزه تخصصی فاوا با حوزه عمومی از فرهنگ لغات مایکروسافت استفاده کردیم [۵]. این فرهنگ حدود ده هزار مدخل دارد و مقوله دستوری هر کدام از آن‌ها را نیز مشخص کرده است. در حوزه فاوا نیز مقوله اسم با اختلاف زیاد از سایر مقوله‌ها بیشتر است و مقوله قید تنها با هفت مدخل کمترین بسامد را دارد.




جدول ۳-۲) بسامد واژگان زبان در مقوله های مختلف

فاوا	عمومی	
۹۹۰۰	۱۱۷۰۰۰	اسم
۳۰۲	۲۲۰۰۰	صفت
۳۲۶	۱۲۰۰۰	فعل
۷	۴۰۰۰	قید

شکل ۳-۱) نمودار بسامد واژگان مقوله های مختلف دستوری

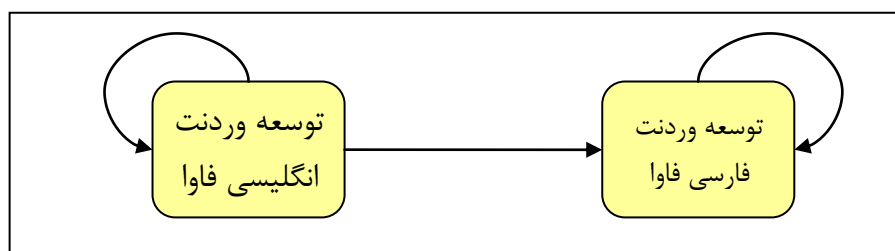
۳-۵- جمع بندی

منابع مهم واژگانی در حوزه فاوا کتاب‌ها و صفحات وب تخصصی هستند و از این منابع می‌توان برای استخراج واژگان استفاده کرد. با توجه به حجم کم واژگان فاوا که معادل‌های فارسی مصوبی دارند، رویکرد اصلی در این پروژه توسعه وردنت فاوا و زبان انگلیسی و سپس ترجمه آن به فارسی است. حوزه تخصصی فاوا دامنه واژگانی محدودتری نسبت به حوزه عمومی زبان دارد اما ترکیب مقوله های مختلف دستوری آن تقریباً شبیه حوزه عمومی است. در این حوزه قید کمترین بسامد را دارد.

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات


۴- روش شناسی توسعه وردنت فاوا

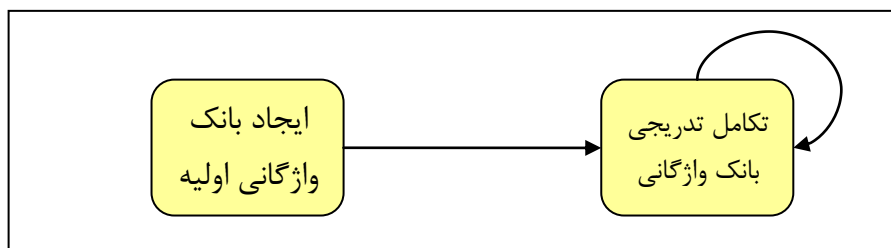
در این بخش متدولوژی ساخت وردنت فاوا را معرفی می‌کنیم. از آنجایی که واژگان فارسی حوزه فاوا از زبان انگلیسی برگرفته شده‌اند، در این متدولوژی ابتدا وردنت فاوا را در زبان انگلیسی توسعه داده و سپس به زبان فارسی ترجمه می‌کنیم. به این ترتیب وردنت توسعه داده شده دارای دو نسخه انگلیسی و فارسی است که در کنار هم قرار دارند و مفاهیم متناظر آنها به همدیگر نگاشت می‌شوند. با توجه به اینکه که در زبان انگلیسی برای واژه‌ها تعاریف مناسبی ارائه شده و همچنین روش‌های کارآمدی برای استخراج خودکار روابط معنایی بین واژگان و مفاهیم ارائه شده است، این روش شناسی فرایند توسعه وردنت فاوا به زبان انگلیسی خیلی سریع اجرا می‌شود. توسعه نسخه فارسی وردنت فاوا مشتمل بر ترجمه مفاهیم از زبان انگلیسی به فارسی و ایجاد روابط ساختاری بین واژگان است.



شکل ۴-۱) توسعه وردنت دوزبانه انگلیسی-فارسی فاوا

اگر از بعد حجم واژگان به توسعه وردنت فاوا نگاه کنیم، این فرایند شامل دو مرحله اصلی است که در مرحله اول آن بانک واژگانی اولیه وردنت فاوا ایجاد می‌شود. این بخش شامل تعریف زیردامنه های فاوا و مجموعه واژگان اولیه آن هاست و فقط یکبار اجرا می‌شود (شکل زیر).

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده فناوری اطلاعات
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	



شکل ۴-۲) توسعه حجمی وردنت فاوا


فرایند تکمیل واژگانی یک فرایند تکراری است که به مرور به حجم واژگان وردنت اضافه می‌شود. این کار از طریق ویرایشگر رومیزی و تحت وب انجام می‌گیرد.

۴-۱- نحوه تعامل تیم زبان شناسی و تیم فناوری اطلاعات

در پروژه توسعه وردنت فاوا با همکاری تیم‌های فناوری اطلاعات و زبان‌شناسی توسعه می‌یابد. وظایف اصلی تیم فناوری اطلاعات عبارتند از:

- ایجاد زیرساخت نرم افزاری برای میزبانی وردنت
 - گزینش و معرفی واژگان تخصصی حوزه فاوا به زبان انگلیسی
 - توسعه روش‌های خودکار استخراج اطلاعات مورد نیاز از متن‌های خام یا ساختاریافته
- تیم زبان‌شناسی بر تولید محتوا نظارت دارد و فرایند های زبانی را برای استخراج انواع اطلاعات پیشنهاد یا طراحی می‌نماید. وظایف اصلی این تیم عبارت است از:
- نظارت بر ترجمه واژگان و مترادفها
 - تعریف ساختارهای واژگانی فارسی و روابط بین آنها
 - طراحی روش‌هایی برای استخراج روابط و اطلاعات ساختاری

برای پرهیز از هرگونه ناسازگاری و ناهماهنگی در بانک اطلاعاتی، ورود و ویرایش دادگان بانک اطلاعاتی تنها به وسیله تیم فناوری اطلاعات انجام می‌گیرد. نحوه تعامل با تیم زبان‌شناسی به این صورت است که داده های مورد نیاز این تیم در قالب‌هایی که استفاده از آنها راحت است، در اختیار این تیم قرار می‌گیرد. پیش بینی شده است که دو قالب اصلی پایگاه داده های Microsoft Access و فایل‌های Microsoft Excel برای ورود یا ویرایش اطلاعات به وسیله تیم زبان شناسی مورد استفاده قرار گیرند. امروزه تقریباً همه دانشگاهیان با این ابزارها آشنا هستند و دوره های معین کسب مهارت‌های پایه فناوری اطلاعات برای آموزش اساتید و دانشجویان همه رشته‌ها برگزار می‌شود. برای آشنایی بیشتر با این نحوه از تعامل، فرایند ترجمه واژگان را به صورت مثال در اینجا ذکر می‌کنیم:

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

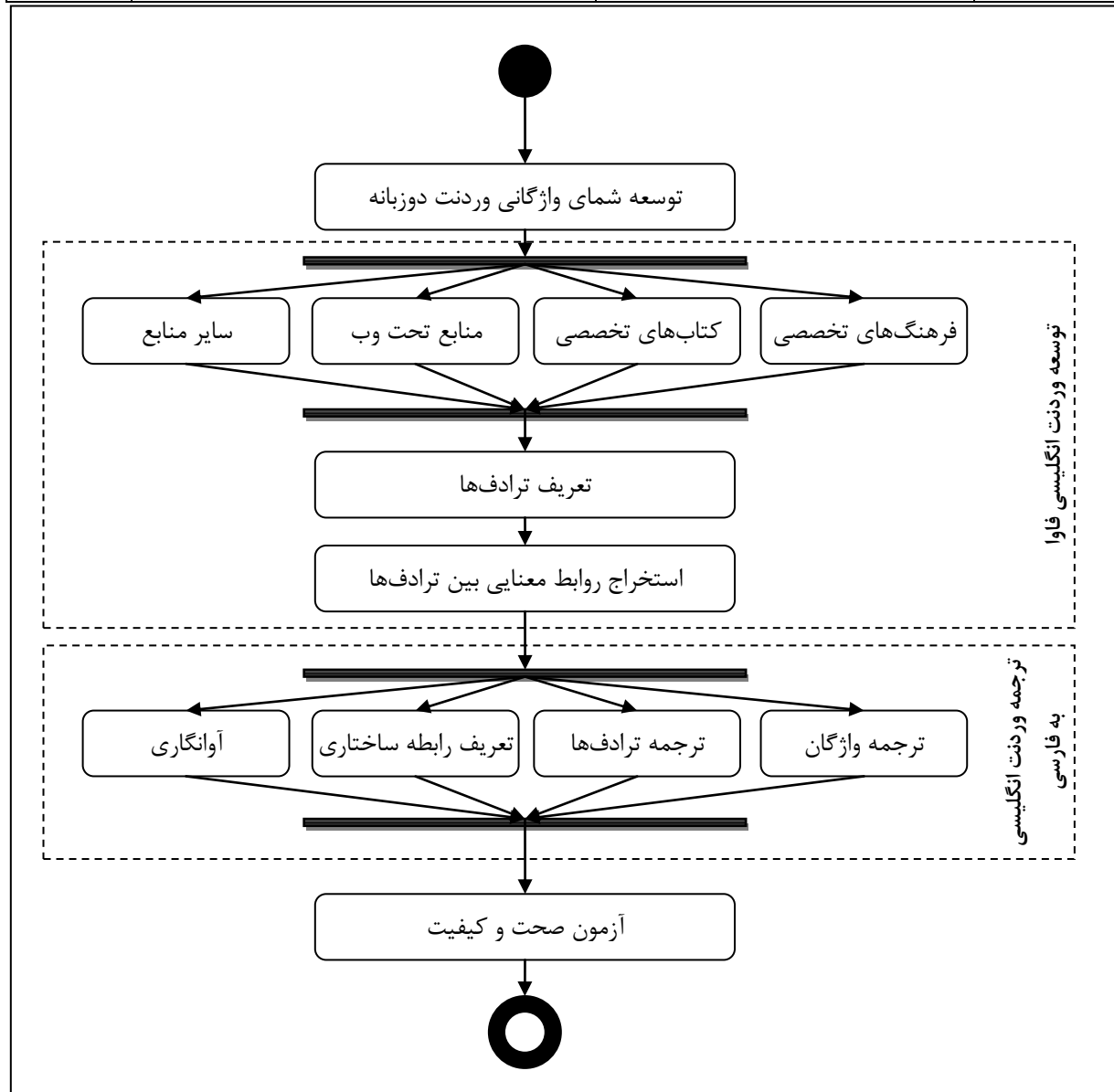
تیم فناوری اطلاعات برای هر کدام از واژگان تخصصی، معادل یا معادل‌های فارسی را پیشنهاد کرده و فهرست این واژگان به همراه معادل‌های فارسی را در قالب فایل Excel در اختیار تیم زبان‌شناسی قرار می‌دهد. تیم زبان‌شناسی این فایل را ویرایش کرده و به تیم فناوری اطلاعات برمی‌گرداند. تیم فناوری اطلاعات دادگان ویرایش شده را به پایگاه اضافه می‌نماید.

پر واضح است که فعالیتی مانند گزینش معادل فارسی برای معمولاً به تعامل و بحث‌های دو سویه بین دو تیم نیاز دارد. به همین سبب جلسات منظمی بین این دو تیم برگزار شده و تبادل نظر انجام می‌گیرد.

۴-۲- توسعه هسته اولیه وردنت فاوا

همان‌طوری که در شکل زیر به صورت نمودار گردش کار نشان داده شده است، فرایند توسعه هسته اولیه وردنت شامل مراحل زیر است. توضیحات مفصل این مراحل در قالب زیربخش‌هایی در ادامه ارائه شده است.


- طراحی شمای واژگانی
- استخراج واژگان تخصصی حوزه فاوا در زبان انگلیسی
- تعریف مترادف‌ها و استخراج روابط معنایی بین آن‌ها در زبان انگلیسی
- ترجمه و ثبت واژگان در وردنت فارسی
- تکمیل اطلاعات ساختاری وردنت فارسی
- آزمون صحت و کیفیت وردنت تولید شده



شکل ۴-۳) توسعه هسته اولیه وردنت دوزبانه انگلیسی-فارسی فاوا

۴-۳- توسعه شمای واژگانی وردنت فاوا

شمای واژگانی وردنت دوزبانه انگلیسی-فارسی فاوا ساختار اطلاعاتی پایگاه داده را تشریح می‌کند که در آن جدول‌ها و نماها برای نگهداری اطلاعات واژگان و روال‌ها برای پرس و جو و ویرایش این اطلاعات تعبیه شده‌اند. این شما نداشت بین وردنت انگلیسی و فارسی فاوا را در سطح واژگان و ترادفها ممکن می‌سازد. نرم افزارهای کاربردی مانند مرورگر و ویرایشگر از این امکانات استفاده کرده و خدمات خاص خود را ارائه می‌دهند. از جمله خدمات این برنامه‌ها می‌توان درج، ویرایش و حذف واژگان و اطلاعات آن‌ها و جستجو در پایگاه و تهیه گزارش‌ها را می‌توان نام برد. برای اطلاع بیشتر درباره شمای واژگانی به سند « طراحی شمای واژگانی » مراجعه نمایید.

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

۴-۴- توسعه وردنت فاوا به زبان انگلیسی فاوا

همان طوری که پیش تر گفتیم، وردنت تخصصی فاوا ابتدا به زبان انگلیسی ساخته می‌شود. در این مرحله واژگان تخصصی استخراج شده و سپس در قالب ترادفها دسته بندی می‌شوند. در انتها روابط معنایی بین ترادفها استخراج می‌شود. در ادامه هر کدام از این کارها را بیشتر توضیح می‌دهیم.

۴-۴-۱- استخراج واژگان و اصطلاحات تخصصی حوزه فاوا در زبان انگلیسی


واژگان حوزه فاوا را از منابع مختلف می‌توان جمع آوری کرد. در زبان فارسی چندین فرهنگ لغت در حوزه فاوا منتشر شده‌اند که تعدادی از آنها ترجمه های متفرقه از فرهنگ لغت‌های دیگر هستند و تعدادی نیز وجود دارند که به صورت تلفیقی از متون فاوا و فرهنگ لغت‌های دیگر ساخته شده‌اند. نمونه‌هایی از فرهنگ لغت‌های قابل استفاده فاوا در این پروژه را می‌توان چنین برشمرد:

- اصطلاح‌نامه‌ها و فرهنگ لغت‌های فارسی
 - اصطلاح‌نامه های علوم، سازمان اسناد و مدارک ملی ایران [۶]
 - فرهنگ معاصر، واژه نامه کامپیوتر [۷]
 - فرهنگ تشریحی کامپیوتر میکروسافت [۸]
- اصطلاح‌نامه‌ها و فرهنگ لغت‌های غیر فارسی
 - فرهنگ علوم و تکنولوژی مک گروهیل [۹]
 - فرهنگ تخصصی میکروسافت [۵]

امروزه تقریباً همه کتاب‌ها به صورت کامپیوتر تایپ و صفحه بندی می‌شوند و نسخه الکترونیکی بسیاری از آنها نیز به همراه نسخه چاپی در قالب محصول جداگانه منتشر می‌شود. به کمک تکنیک‌های آماری و متن کاوی می‌توان واژگان پربسامد این کتاب‌ها را استخراج کرده و پس از پالایش‌های ماشینی انبوه، به صورت دستی نیز آنها را کنترل کرد. روش‌های متعدد آماری و ساختاری برای تشخیص کلمات مرکب وجود دارد که از بین آنها می‌توان تحلیل آماری باهم‌آیی^۱ واژگان را می‌توان نام برد.

در مرحله انتخاب واژگان تخصصی، کتاب‌ها، سایت‌ها، فرهنگ نامه‌ها و مجموعه های لغت که به صورت بالقوه می‌توانند شامل تعداد قابل توجهی از واژگان تخصصی فاوا باشند انتخاب می‌شوند. سپس مجموعه واژگان هر منبع استخراج می‌شود. از میان این منابع کتاب‌های تخصصی و متون مشابهی که ساختار معینی ندارند به پیش پردازش‌هایی مانند محاسبه n-gram ها نیازمند هستند اما منابع ساختاریافته چنین پیش پردازش‌هایی نیاز ندارند. نهایتاً کارشناس متخصص فاوا مجموعه واژگان منتخب را بررسی کرده و واژگان

^۱ Collocation

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

غیر فاوا را حذف می‌نماید. در سند «فرهنگ نگاری و تکمیل بانک واژگانی» نحوه استخراج واژگان تخصصی و منابع آنها به تفصیل معرفی شده است.

۴-۲- تعریف مترادفها و استخراج روابط معنایی بین آنها

واژگان استخراج شده ابتدا در قالب مترادفها دسته بندی می‌شوند و سپس رابطه معنایی بین این مترادفها تعریف می‌شود. برای تعیین مترادفها، به کمک تکنیک‌های مختلف مانند متن کاوی رابطه مترادف بودن بین واژگان کشف می‌شوند.


با نگاهی به هر کدام از وردنت‌های موجود می‌توان دریافت که تعداد مترادف‌های آنها حداکثر حدود ۸۰ درصد تعداد کل واژگان آنهاست. بر این اساس اگر حوزه فاوا دارای حدود ده هزار واژه باشد؛ می‌توان امیدوار بود که حدود هشت هزار مجموعه مترادف داشته باشیم. فرایندهای متن کاوی حداکثر دو برابر این تعداد را به عنوان مجموعه اولیه مترادفها گزارش می‌کنند. پالایش و انتخاب مجموعه نهایی برای این حجم از داده‌ها رami توان به صورت دستی انجام داد. در سند «استخراج روابط معنایی» به صورت مفصل در این باره بحث می‌نماییم.

۴-۵- ترجمه وردنت انگلیسی فاوا به فارسی

ساخت وردنت فارسی فاوا شامل ترجمه واژگان، ترجمه تعریف مترادفها و همچنین تعریف ساختار و روابط ساختاری واژگان است. در ادامه هر کدام از این فعالیتها را بیشتر توضیح می‌دهیم. در فرایند ترجمه سعی می‌شود به کمک متخصصان حوزه فاوا و زبان‌شناسان مجرب در حوزه واژه‌گزینی از بین معادل‌های موجود، مناسب‌ترین معادل فارسی برای واژگان و اصطلاحات فاوا انتخاب گردد. برای واژگانی که معادل فارسی برای آنها هنوز وجود ندارد نیز می‌توان معادل‌های مناسبی با رعایت قواعد واژه‌گزینی مصوب فرهنگستان، پیشنهاد داد.

فرایند ترجمه واژگان و مترادفها برای همه واژگان یکی نیست. ترجمه بسیاری از واژگان در فرهنگ لغت‌ها و واژه نامه های تخصصی معتبر موجود است. منظور از منابع معتبر در اینجا فرهنگ لغت‌های تخصصی مانند فرهنگ کامپیوتر معاصر یا واژه نامه‌ها یا کتاب‌های است که متخصصان برجسته حوزه فاوا آنها را تألیف نموده‌اند. فهرست زیر نمونه‌هایی از این آثار است:


- محمد تقی روحانی رانکوهی، مفاهیم بنیادی پایگاه داده‌ها، نشر جلوه، ۱۳۹۰
- محمد قدسی، داده ساختارها و مبانی الگوریتم‌ها، انتشارات فاطمی، ۱۳۸۸
- محمدرضا محمدی فر، مبانی نمایه سازی، سازمان چاپ و انتشارات وزارت ارشاد، ۱۳۸۱

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده فناوری اطلاعات
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	

اگر ترجمه واژه ای در این قبیل از منابع موجود باشد، همین ترجمه مستقیماً در وردنت فارسی فاوا درج می‌شوند. فرایند کار به این شکل است که تیم ترجمه برای هر دسته معنایی این واژه از بین معادل‌های فارسی موجود مناسب‌ترین گزینه را انتخاب و ثبت می‌نماید. بسیاری دیگر از واژگان تخصصی فاوا اسم خاص هستند و نیازی به ترجمه ندارند. این واژگان به همان صورت انگلیسی وارد پایگاه داده فارسی می‌شوند. برای مجموعه باقیمانده واژه‌ها تیم ترجمه گزینه‌هایی را از کتب تخصصی پر تیراژ انتخاب می‌کند. این کتب ممکن است به اندازه منابع فوق معتبر نباشند اما به هر حال تنها منابع باقیمانده هستند. این ترجمه‌ها در دو مرحله بازبینی و ویرایش می‌شوند. در مرحله اول کارشناسان خبره فاوا و در مرحله دوم کارشناسان زبان شناسی فرایند معادل‌های انتخاب شده را بررسی و در صورت لزوم تصحیح می‌نمایند. ترجمه تعریف مترادف‌ها هم مسیر مشابهی را طی می‌کند. بدین معنی که ابتدا تیم ترجمه برای هر مترادف معنی آن را ترجمه کرده و به تیم کارشناسان فاوا تحویل می‌دهد. سپس این تیم ترجمه‌ها را بازبینی و ویرایش کرده و جهت اعلام نظر به تیم زبان شناسی ارائه می‌دهد. تیم زبان شناسی اصلاحات نهایی را انجام می‌دهد و پس از آن تعریف مترادف در پایگاه وردنت درج می‌گردد. درج اطلاعات ساختاری واژگان مانند ساختار ظرفیتی افعال و همچنین رابطه اشتقاق بین واژگان کلاً به وسیله تیم زبان شناسی انجام می‌پذیرد. بخش‌هایی از این فرایند مانند استخراج ریشه واژگان یا نوع اشتقاق به صورت خودکار صورت می‌گیرد.

۴-۶- آزمون صحت و کیفیت

پس از ساخت وردنت فاوا، محصول نهایی به روش‌های مختلف تحت آزمون قرار می‌گیرد. همان‌طوری که در سند طرح آزمون هم به تفصیل بیان شده است، به صورت عمده دو دسته آزمون نرم افزار و آزمون محتوا در اینجا انجام می‌گیرد. آزمون‌های نرم افزار به منظور اطمینان از صحت عملکرد و ارائه امکانات خواسته شده در تعریف پروژه انجام می‌گیرند. آزمون محتوا نیز به منظور اطمینان از درستی اطلاعات ارائه شده درباره واژگان و روابط بین آن‌ها انجام می‌گیرند. برای اطلاع بیشتر درباره این آزمون‌ها به سند «طرح آزمون وردنت فاوا» مراجعه نمایید.

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

۵- چرخه حیات یک واژه

آنچه که در نسخه های نهایی وردنت وجود دارد، واژگان و اطلاعات آن هاست. برای پرهیز از پیچیدگی های طراحی، اطلاعات جنبی مانند مرجع واژه، وضعیت فعلی واژه یا نام کاربری که آن را وارد پایگاه کرده در به صورت جداگانه در قالب فایل های Excel نگهداری می شود.


۵-۱- چرخه حیات واژه در وردنت انگلیسی

پس از گزینش واژگان تخصصی، برای هر واژه پرونده ای تشکیل می شود. در این پرونده اطلاعاتی مانند مرجع اصلی، تعریف، فونتیک، واژگان مرتبط، ترجمه و همچنین وضعیت فعلی ثبت واژه در پایگاه وردنت ذخیره می شود. وضعیت ثبت واژه پس از انجام هر فعالیتی درباره آن، به روز می شود. هنگام ذخیره واژه قواعد زیر رعایت می شود:

- حداقل اطلاعات لازم برای ذخیره یک واژه در پایگاه صورت املائی آن است. یعنی هیچ واژه بدون صورت املائی در پایگاه وجود ندارد.
- اطلاعات جنبی یک واژه مانند آوانگاری، عضویت در ترادف یا اطلاعات ساختاری آن در هر زمانی پس از ذخیره واژه قابل تعریف یا تغییر است.
- هر واژه تنها در صورتی در پایگاه ذخیره می شود که حداقل یک تعریف برای آن وجود داشته باشد.

شرط لازم و کافی برای اینکه ثبت یک واژه در وردنت فاوا کامل تلقی شود این است که :

- صورت املائی و آوایی آن در پایگاه ثبت شده باشد.
- اطلاعات ساختاری آن مانند ساختار ظرفیتی فعل یا نوع صفت ثبت شده باشد.
- واژه مورد نظر عضو حداقل یک ترادف تعریف شده باشد.

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

بدیهی است که هر واژه ممکن است عضو بیش از یک مترادف باشد یا صورت‌های مختلف املائی و آوایی داشته باشد. وظیفه ثبت نهایی واژگان در وردنت انگلیسی فاوا به عهده ویرایشگران تیم فاوا است. پس از تعریف واژگان، مترادف‌ها را می‌توان تعریف کرد.

شرط لازم و کافی برای ثبت نهایی یک مترادف در وردنت انگلیسی فاوا این است که:


- برای این مترادف تعریف ارائه شود
- مقوله دستوری و مقوله معنایی (زیر دامنه فاوا) آن مشخص شده باشد
- حداقل یک واژه عضو مترادف شود.

در مرحله بعد با استفاده از روش‌هایی که در سند «استخراج روابط معنایی به صورت خودکار» به تفصیل معرفی شده‌اند، روابط معنایی حاکم بین واژگان استخراج می‌شود. از آنجایی که رابطه‌های معنایی بین مترادف‌ها تعریف می‌شوند؛ شرط لازم و کافی برای ایجاد یک رابطه معنایی بین دو مترادف این است که هر دو مترادف و رابطه معنایی مذکور تعریف شده باشند.

۵-۲- چرخه حیات واژه در وردنت فارسی


وردنت فارسی ترجمه واژگان و مترادف‌های وردنت انگلیسی است. بر اساس معانی مختلف، یک واژه در زبان انگلیسی ممکن است به چندین واژه در زبان فارسی ترجمه شود. در حوزه عمومی زبان می‌توان واژه Spring را مثال زد که معانی فارسی آن بهار، چشمه و فنر می‌باشد. در حوزه فاوا نیز می‌توان واژه Index را نام برد که معادل‌های فارسی شاخص، اندیس و شماره ردیف برای آن وجود دارند. برای پرهیز از ابهام، مترادف‌های وردنت انگلیسی و فارسی به همدیگر نگاشت می‌شوند. بر این اساس شرط لازم و کافی برای اینکه یک مترادف در وردنت زبان فارسی وجود داشته باشد این است که مترادف متناظر آن در وردنت زبان انگلیسی وجود داشته باشد.

هنگام ترجمه یک مترادف از زبان انگلیسی به فارسی، برای هر واژه معادل مناسبی گزینش می‌شود. باید توجه داشت که ممکن است دو واژه انگلیسی یک معادل فارسی داشته باشند. به عنوان نمونه در متون فارسی به جای هر دو واژه Efficiency و Performance از کارایی استفاده می‌شود. در اینگونه موارد کافی است واژه کارایی تنها یک‌بار در وردنت فارسی ثبت شود. با ثبت یک واژه در وردنت فارسی، حیات آن آغاز می‌گردد و پرونده‌ای برای آن گشوده می‌شود که بیانگر آخرین اطلاعات درباره ثبت اطلاعات ساختاری و آوایی آن می‌باشد.

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات


۶- جمع بندی

در این سند تحلیل نیازها و دامنه پروژه توسعه وردنت زبان فارسی در حوزه فاوا ارائه شده است. در بخش اول این سند ابعاد پروژه بررسی شده و تعاریف مفاهیم لازم ارائه شده است. در ادامه حوزه فاوا و محدوده واژگانی آن تعیین شده و منابعی برای جمع آوری واژگان این حوزه معرفی شده است. راه حل پیشنهادی ما برای توسعه وردنت فارسی در حوزه فاوا این است که وردنت انگلیسی فاوا را ایجاد نموده و سپس آن را به فارسی ترجمه کنیم. برای آزمون امکان پذیری، یک مطالعه پایلوت انجام داده ایم که خلاصه نتایج آن در این سند و نتایج مبسوط آن در لوح فشرده پیوست ارائه می گردد. این نتایج نشان می دهد که توسعه وردنت در حوزه فاوا امکان پذیر است.

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

مراجع


- [۱] Harry KORNILAKIS, Eleni GALIOTOU \ Maria GRIGORIADOU and Evangelos PPAKITSOS Sofia STAMOU, "The Software Infrastructure for the Development and Validation of the Greek Wordnet," *ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY*, vol. ۷, no. ۱-۲, ۲۰۰۴.
- [۲] ITU, "Measuring ICT: the global status of ICT indicators," Geneva, ۲۰۰۵.
- [۳] Oxford University Press, "Oxford Advanced Learners Dictionary," ۲۰۰۶.
- [۴] I.S.P. Nation, *Learning Vocabulary in another Language*. Cambridge, UK: Cambridge University Press, ۲۰۰۱.
- [۵] Microsoft Corp., *Microsoft Computer Dictionary*, ۵th ed.: Microsoft Press.
- [۶] Iranian Research Institute for Scientific Information and Documentation. (۲۰۱۱) *Irindoc Thesauri*. [Online]. <http://thesauri.irandoc.ac.ir/>
- [۷] تهران: فرهنگ معاصر, ۱۳۸۵. واژه نامه کامپیوتر انگلیسی-فارسی, محمد رضا محمدی فر
- [۸] تهران: فروزش, ۱۳۸۹. فرهنگ تخصصی کامپیوتر, نسرين محمدی
- [۹] McGraw-Hill, *McGraw Dictionary of Scientific and Technical Terms*: McGraw-Hill, ۲۰۰۰.
- [۱۰] A. and Konstantinidi, I. and Papadaki, M. and Keramidas, G. and Grigoriadou, M. Kremizis, "Greek Wordnet Extension in the Domain of Psychology and Computer Science," in *Proceedings of the Ath Hellenic European Research Computer Mathematics and its Applications Conference (HERCMA ۲۰۰۷)*, Athens, ۲۰۰۷.
- [۱۱] J. and Grobelnik, M. and Mladenic, D. Brank, "A survey of ontology evaluation techniques," in *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD ۲۰۰۵)*, ۲۰۰۵.
- [۱۲] P. and Cimiano, P. and Magnini, B. Buitelaar, "Ontology learning from text: methods, evaluation and applications," *Computational Linguistics*, vol. ۳۲, no. ۴, ۲۰۰۵.
- [۱۳] Daniel Jurafsky and James H. Martin, *SPEECH and LANGUAGE PROCESSING*, ۲nd ed.: Pearson Prentice Hall, ۲۰۰۹.


	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

واژه نامه انگلیسی به فارسی

انگلیسی	فارسی
Access	دسترسی
Activity	فعالیت
Application	کاربرد
Applied	کاربردی
Approach	رویکرد
Architecture	معماری
Automatic	خودکار
Automation	خودکارسازی
Backup	پشتیبان
Base	پایگاه
Basic	پایه
Browser	مرورگر
Code	کد
Communication	ارتباط
Communications	ارتباطات
Component	مؤلفه
Connection	اتصال
Contradictor	ناسازگار
Corpus	پیکره
Data	دادگان
Datum	داده
Desktop	رومیزی
Development	توسعه
Dynamic	پویا
Edit	ویرایش
Entry	مدخل
Evaluation	ارزیابی
Expanded	وسیع تر
Experiment	آزمون
Export	صدور
Extraction	استخراج
Feature	ویژگی
Functional	کارکردی
Fuzzy	فازی،
General	عمومی
Import	اکتساب

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات
Inference			استنتاج
Inflections			تصرف
Information and Communication Technology			فناوری اطلاعات و ارتباطات
Information Technology			فناوری اطلاعات
Infrastructure			زیر ساخت
Input			ورودی
Insert			درج
Institute			موسسه
Interface			واسط
Isolated			منفک،
Lexical			لغوی
Linguistic			زبان شناختی
Login			ورود
Mapping			نگاشت
Monitor			ناظر
Network			شبکه
Normal			عادی
Noun			اسم
Online			برخط
On time			به هنگام
Permission			اجازه
Persian			فارسی
Plural			جمع
Primary			ابتدایی
Process			فرایند
Protocol			پروتکل
Reliability			صحت
Retrieval			بازیابی
Risk			ریسک
Scientific			علمی
Secure			امن
Semantic			معنایی
Semantic			نحوی
Sentence			جمله
Specialized			اختصاصی
Spectrum			طیف
Standard			استاندارد
Strategic			راهبردی
Structure			ساختار
Synonym			مترادف
Synset			ترادف


	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ۱۱۰۷۰۲.۸۹۳۲۴۱۶.ITF.ITP.TCH	فناوری اطلاعات
System			سامانه
Thesaurus			تزاروس
Tier			لایه
Tool			وسيله
Tools			ابزارها
Universal			جهانی
Usage			استفاده
User			کاربر
Validity			اعتبار
Verb			فعل
Virtual Space			فضای مجازی
Web			وب
Word			واژه
WorldNet			وردنت
Words			واژگان


	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

واژه نامه فارسی به انگلیسی

انگلیسی	فارسی
Primary	ابتدایی
Tools	ابزارها
Connection	اتصال
Permission	اجازه
Specialized	اختصاصی
Communication	ارتباط
Communications	ارتباطات
Evaluation	ارزیابی
Standard	استاندارد
Extraction	استخراج
Usage	استفاده
Inference	استنتاج
Noun	اسم
Validity	اعتبار
Import	اكتساب
Secure	امن
Experiment	آزمون
Retrieval	بازیابی
Online	برخط
On time	به هنگام
Base	پایگاه
Basic	پایه
Protocol	پروتکل
Backup	پشتیبان
Dynamic	پویا
Corpus	پیکره
Synset	ترادف
Thesaurus	تزاروس
Inflections	تصرف
Development	توسعه
Plural	جمع
Sentence	جمله
Universal	جهانی
Automatic	خودکار
Automation	خودکارسازی
Data	دادگان

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات
Datum			داده
Insert			درج
Access			دسترسی
Strategic			راهبردی
Desktop			رومیزی
Approach			رویکرد
Risk			ریسک
Linguistic			زبان شناختی
Infrastructure			زیر ساخت
Structure			ساختار
System			سامانه
Network			شبکه
Reliability			صحت
Export			صدور
Spectrum			طیف
Normal			عادی
Scientific			علمی
General			عمومی
Persian			فارسی
Fuzzy			فازی،
Process			فرایند
Virtual Space			فضای مجازی
Activity			فعالیت
Verb			فعل
Information Technology			فناوری اطلاعات
Information and Communication Technology			فناوری اطلاعات و ارتباطات
User			کاربر
Application			کاربرد
Applied			کاربردی
Functional			کارکردی
Code			کد
Tier			لایه
Lexical			لغوی
Synonym			مترادف
Entry			مدخل
Browser			مرورگر
Architecture			معماری
Semantic			معنایی
Isolated			منفک،
Institute			موسسه
Component			مؤلفه

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات
Contradictor			ناسازگار
Monitor			ناظر
Semantic			نحوی
Mapping			نگاشت
Words			واژگان
Word			واژه
Interface			واسط
Web			وب
WordNet			وردنت
Login			ورود
Input			ورودی
Expanded			وسیع تر
Tool			وسیله
Edit			ویرایش
Feature			ویژگی


	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

پیوست ۳ سرفصل‌های پیشنهادی حوزه فاوا

سرفصل‌های پیشنهادی ACM برای علوم و مهندسی کامپیوتر

ACM سرفصل‌های حوزه علوم و مهندسی کامپیوتر را به صورت سلسله مراتبی و در سه سطح تنظیم نموده است. در زیر عناوین اصلی این سرفصل‌ها تا سطح دوم آمده است. همچنین فهرست کامل این سرفصل‌ها تا سه سطح در لوح فشرده پیوست این گزارش موجود است.


Code	Title
A.	General Literature
A.1	INTRODUCTORY AND SURVEY
A.2	REFERENCE (e.g., dictionaries, encyclopedias, glossaries)
B.	Hardware
B.1	CONTROL STRUCTURES AND MICROPROGRAMMING (D.3.2)
B.2	ARITHMETIC AND LOGIC STRUCTURES
B.3	MEMORY STRUCTURES
B.4	INPUT/OUTPUT AND DATA COMMUNICATIONS
B.5	REGISTER-TRANSFER-LEVEL IMPLEMENTATION
B.6	LOGIC DESIGN
B.7	INTEGRATED CIRCUITS
B.8	PERFORMANCE AND RELIABILITY (C.4) (NEW)
C.	Computer Systems Organization
C.1	PROCESSOR ARCHITECTURES
C.2	COMPUTER-COMMUNICATION NETWORKS
C.3	SPECIAL-PURPOSE AND APPLICATION-BASED SYSTEMS (J.7)
C.4	PERFORMANCE OF SYSTEMS
C.5	COMPUTER SYSTEM IMPLEMENTATION
D.	Software
D.1	PROGRAMMING TECHNIQUES (E)
D.2	SOFTWARE ENGINEERING (K.6.3)
D.3	PROGRAMMING LANGUAGES
D.4	OPERATING SYSTEMS (C)
E.	Data
E.1	DATA STRUCTURES
E.2	DATA STORAGE REPRESENTATIONS
E.3	DATA ENCRYPTION
E.4	CODING AND INFORMATION THEORY (H.1.1)
E.5	FILES (D.4.3, F.2.2, H.2)
F.	Theory of Computation
F.1	COMPUTATION BY ABSTRACT DEVICES
F.2	ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY (B.6-7, F.1.3)
F.3	LOGICS AND MEANINGS OF PROGRAMS
F.4	MATHEMATICAL LOGIC AND FORMAL LANGUAGES

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	فناوری اطلاعات

- G. Mathematics of Computing
 - G.1 NUMERICAL ANALYSIS
 - G.2 DISCRETE MATHEMATICS
 - G.3 PROBABILITY AND STATISTICS
 - G.4 MATHEMATICAL SOFTWARE
- H. Information Systems
 - H.1 MODELS AND PRINCIPLES
 - H.2 DATABASE MANAGEMENT (E.5)
 - H.3 INFORMATION STORAGE AND RETRIEVAL
 - H.4 INFORMATION SYSTEMS APPLICATIONS
 - H.5 INFORMATION INTERFACES AND PRESENTATION (e.g., HCI) (I.7)
- I. Computing Methodologies
 - I.1 SYMBOLIC AND ALGEBRAIC MANIPULATION (REVISED)
 - I.2 ARTIFICIAL INTELLIGENCE
 - I.3 COMPUTER GRAPHICS
 - I.4 IMAGE PROCESSING AND COMPUTER VISION (REVISED)
 - I.5 PATTERN RECOGNITION
 - I.6 SIMULATION AND MODELING (G.3)
 - I.7 DOCUMENT AND TEXT PROCESSING (H.4-5) (REVISED)
- J. Computer Applications
 - J.1 ADMINISTRATIVE DATA PROCESSING
 - J.2 PHYSICAL SCIENCES AND ENGINEERING
 - J.3 LIFE AND MEDICAL SCIENCES
 - J.4 SOCIAL AND BEHAVIORAL SCIENCES
 - J.5 ARTS AND HUMANITIES
 - J.6 COMPUTER-AIDED ENGINEERING
 - J.7 COMPUTERS IN OTHER SYSTEMS (C.3)
- K. Computing Milieux
 - K.1 THE COMPUTER INDUSTRY
 - K.2 HISTORY OF COMPUTING
 - K.3 COMPUTERS AND EDUCATION
 - K.4 COMPUTERS AND SOCIETY
 - K.5 LEGAL ASPECTS OF COMPUTING
 - K.6 MANAGEMENT OF COMPUTING AND INFORMATION SYSTEMS
 - K.7 THE COMPUTING PROFESSION
 - K.8 PERSONAL COMPUTING

سرفصل‌های دایرکتوری وب^۱ dmoz


^۱ <http://www.dmoz.org>

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده فناوری اطلاعات
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	


دایرکتوری dmoz به صورت دستی مدیریت و بروز رسانی می‌شود. در زیر سرفصل‌های این دایرکتوری زیر عنوان Computers and Internet آمده است.

Title

Algorithms
 Artificial Intelligence
 Artificial Life
 Associations
 Bulletin Board Systems
 CAD and CAM
 Chats and Forums
 Companies
 Computer and Technology Law
 Computer Science
 Conferences
 Consultants
 Consulting
 Customer Relationship Management
 Data Communications
 Data Formats
 Desktop Publishing
 Developers
 Directories
 E-Books
 Education and Training
 Education
 Employment
 Emulators
 Ethics
 FAQs, Help, and Tutorials
 Fonts
 Games
 Graphics
 Hacking
 Hardware
 History
 Home Automation
 Human-Computer Interaction
 Imaging
 Internet Marketing
 Internet
 Intranet
 Legal Information
 Mailing Lists

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶,۱۱.۷.۰۲	فناوری اطلاعات


Marketing and Advertising
 Merchant Account Services
 MIS
 Mobile Computing
 Multimedia
 News and Media
 Newsgroups
 Open Source
 Operating Systems
 Order Fulfillment
 Organizations
 Outsourcing
 Parallel Computing
 Performance and Capacity
 Product Support
 Programming
 Robotics
 Security
 Services
 Shopping
 Software
 Speech Technology
 Strategy
 Supercomputing
 Systems
 Technology Vendors
 Telecommunications
 Usenet
 Virtual Reality
 Website Promotion

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده فناوری اطلاعات
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.۷۰۲	

سرفصل‌های ویکی‌پدیا

دایره‌المعارف عمومی ویکی‌پدیا اسناد خود را به صورت سلسله‌مراتبی دسته‌بندی می‌کند. در این دسته‌بندی کلیه موضوعات اصلی، فرعی و جنبی علوم و مهندسی کامپیوتر و همچنین فاوا به صورت فهرست‌الفبایی آمده است. در زیر عناوین سرفصل‌های اصلی آمده است که سرفصل‌های جزئی‌تر از هرکدام قابل دسترسی است.

- [+] Areas of computer science (16 C)
- [+] Computer science awards (8 C, 55 P)
- [+] Computer science conferences (10 C, 85 P)
- [+] Computer science education (2 C, 34 P)
- [+] Computer science literature (6 C, 5 P)
- [+] Computer science organizations (16 C, 52 P)
- [+] Computer science stubs (12 C, 782 P)
- [+] Computer scientists (10 C, 25 P)
- [+] History of computer science (2 C, 9 P)
- [+] Philosophy of computer science (1 C, 4 P)
- [×] Unsolved problems in computer science (19 P)
- [×] Wikipedia books on computer science (19 P)

	عنوان گزارش: تحلیل وردنت زبان فارسی در حوزه فاوا		پژوهشکده
	وضعیت گزارش: نهایی	کد گزارش: ITF.ITP.TCH.۸۹۳۲۴۱۶,۱۱.۷.۰۲	فناوری اطلاعات

Abstract

This document presents requirement analysis of Persian WordNet in ICT domain. At first, we determine boundary of ICT domain and then identify vocabulary resources. Identification of ICT sub-domains and their related glossaries and life-cycle of words and concepts are described in depth. Our main strategy is to develop an English WordNet for ICT domain and then translate it to Persian. ICT related words and phrases are collected from various resources such as text books and web pages. Using text-mining tools and methods several types of semantic relations are extracted from free text. The set of words and their relations builds up the English ICT WordNet which will be translated to Persian subsequently. Ideas and method introduced in this document are implemented or planned to be implemented and their results will be reported in separate documents.



Information Technology Faculty

Information Technology Platform Group

Technical Report

Analysis Document of Persian WordNet for ICT Domain

Project Name: Persian WordNet for ICT Domain

Project code: ۸۹۳۲۴۱۶

Project Director	Muharram Mansoorizadeh
Author(s)	M. Mansoorizadeh, M. Nassiri, M. Dadrass
Document Code	ITF.ITP.TCH.۸۹۳۲۴۱۶.۱۱.V.۰۲
Preparing Date	۹۱.۰۳.۲۷
Status/Version	Final/۱.۰