راهنمای استفاده از پیکره تک سندی پاسخ

تهیه شده توسط آزمایشگاه فناوری وب دانشگاه فردوسی مشهد

سفارش دهنده : سازمان فناوری اطلاعات و ارتباطات ایران

wtlab.um.ac.ir
ijaz.um.ac.ir

جمهوری اسلامی ایران
وزارت ارتباطات و فناوری اطلاعات

برای تولید پیکره تک سندی در ابتدا باید تعدادی سند از سایتهای خبرگذاری معروف انتخاب شود. در انتخاب خبرگزاری ها سعی شد تا آنهایی انتخاب شوند که بیشترین بازدید کننده را دارند و در ضمن دیدگاههای مختلف را پوشش دهند. بهمین منظور ۱۰۰ خبر با اندازه های مختلف از خبرگزاری‌های جدول۱ دانلود و پالایش گردید:

جدول ۱- لیست خبرگزاری های انتخاب شده برای جمع آوری خبر

ردیف	نام خبرگزاری	آدرس خبرگزاری	شماره کد در نام گذاری
۱	خبرگزاری تابناک	http://www.tabnak.ir	TAB
۲	خبرگزاری پرس تی وی	http://www.presstv.ir	PTV
۳	خبرگزاری فارس	http://www.farsnews.com	FAR
۴	خبرگزاری ایرنا	http://irna.ir	IRN
۵	خبرگزاری همشهری	www.hamshahrionline.ir	HAM
۶	خبرگزاری الف	www.alef.ir	ALF
۷	خبرگزاری جام جم	www.jamejamonline.ir	JAM

این موضوعات در قالب دسته بندی کلی زیر جمع آوری شدند. در جمع آوری این پیکره سعی شده است تا همه انواع خبرها پوشش داده شود.

جدول ۲- دسته بندی موضوعات

ردیف	نام دسته	شماره کد در نام گذاری
۱	اقتصادی	EC
۲	فرهنگی	CU
۳	اجتماعی	SO
۴	سیاسی	PO
۵	ورزشی	SP
۶	علمی	SC

فرمت نام گذاری

جهت دسترسی سریع و مشخص به اسناد، فرمولی برای نام گزاری اسناد انتخاب شده است. اسناد در پیکره با فرمول زیر نام گذاری می شوند:

<شماره سند>.<تاریخ سند>.<کد دسته سند>.<کد خبرگزاری سند>

بنابراین سند با نام PTV.PO.13901228.068.xml یعنی سند با شماره کد ۰۶۸ که از نوع سیاسی بوده و از خبرگزاری پرس تی وی دانلود شده است و تاریخ آن هم مربوط به ۲۸ اسفند سال ۱۳۹۰ می باشد. این اسناد در فرمت xml و با تگ های زیر تولید شدند:

```
<?xml version="1.0" encoding="utf-8" ?>
<DOC>
<DOCNO>010</DOCNO>
<DATE-TIME>1391/02/02</DATE-TIME>
<CATEGORY/>ورزشی<CATEGORY>
<TITLE>
در آکادمی ملی المپیک: دور جدید تمرینات تیم ملی والیبال بانوان از ۵ اردیبهشت می‌شود
</TITLE>
<SUMMARY>
خبرگزاری فارس: مرحله جدید اردوی تیم ملی والیبال بانوان بزرگسال برای حضور
در مسابقات هشت تیم برتر آسیا در آکادمی ملی المپیک برگزار می‌شود.
</SUMMARY>
<TEXT>
به گزارش خبرگزاری فارس، تیم ملی والیبال بانوان بزرگسال ایران که خود را برای حضور در مسابقات هشت تیم برتر آسیا آماده می‌کند به دلیل پر بودن سالن‌های آزادی در باشگاه پیکان تمرین می‌کند.
پس از بازگشایی هتل آکادمی قرار است ملی‌پوشان تمریناتشان را در این مکان پیگیری کنند.
تمرینات تیم ملی دختران نوجوان و جوان نیز بارها به دلیل نبود سالن و خوابگاه تعطیل شده است.
این اردو با حضور ۲۰ بازیکن از ۵ اردیبهشت آغاز می‌شود. پس از بازگشت تیم گیتی‌پسند از مسابقات جام باشگاه‌های آسیا تعدادی
```

فایلهای xml به عنوان یک قالب استاندارد برای انتقال داده در محیطهای مختلف نرم افزاری شناخته می شود. تقریبا در تمام زبانهای برنامه سازی سطح بالا امکانات مناسبی برای کار با فایلهای xml فراهم شده است. استفاده از ساختار سلسله مراتبی و بهره مندی از اصول ساخت یافتگی سبب شده است تا دردسرهای احتمالی استفاده از پایگاههای داده در بسیاری از برنامه‌های کاربردی از بین برود. xml به عنوان یک پایه برای طراحی زبانهای تحت وب نیز تلقی می‌شود. از این رو به اعتقاد بسیاری html در واقع گونه گسترش یافته xml در فضای وب محسوب می‌شود.

همانطور که در تصویر هم مشخص است مشخصات سند در تگهای xml قابل دسترسی است. مشخصات این تگ‌ها در جدول ۳ آمده است:

جدول ۳- لیست تگ های سندهای خبری

تگ	توضیحات
<DOCNO>	حاوی شماره سند می باشد.
<DOC-TIME>	زمان انتشار خبر را مشخص می کند.
<CATEGORY>	دسته خبر را مشخص می کند (سیاسی، اقتصادی، ورزشی و ...)
<TITLE>	عنوان خبر را مشخص می کند.
<SUMMARY>	خلاصه خبر که عموما توسط خبرگزاریها تولید می شود در این تگ قرار می گیرد.
<TEXT>	متن خبر در این تگ قرار داده می شود.

فرآیند تولید پیکره

برای تولید پیکره خلاصه انسانی، ۱۰ نفر از دانشجویان کارشناسی (زن – مرد) بکار گرفته شدند. با توجه به اینکه هرکدام از این افراد سلایق و دانش خاص خود را دارند، به منظور کاهش نظرات شخصی افراد، برای هر متن *۵ نفر خلاصه انسانی* تولید کردند. بدین ترتیب در زمان ارزیابی، نتایج خلاصه های ماشینی با میانگین این ۵ خلاصه انسانی مقایسه می شود و تاثیر نظرات شخصی افراد خلاصه ساز کاهش می یابد.

هر کدام از این افراد، برای موضوعات تعیین شده، یک خلاصه چکیده ای و یک خلاصه ی استخراجی تولید کردند. در خلاصه‌سازی استخراجی، افراد شرکت‌کننده در فرآیند تولید پیکره، برای هر متن تک سندی، بین ۳ تا ۷ عدد از مهمترین جملات متن سند را انتخاب کردند. بنابراین مهمترین مساله در این مدل تعیین مهمترین جملات می‌باشد. در خلاصه‌سازی چکیده‌ای، افراد شرکت‌کننده در فرآیند تولید پیکره، برای هر متن تک سندی، حدود ۳ تا ۷ جمله که بیانگر محتوای اصلی متن می‌باشد، با قلم خود بازنویسی کردند.

بنابراین در مجموع، برای *۱۰۰ متن خبری، ۱۰۰۰ خلاصه* تولید شد، که *۵۰۰ عدد چکیده ای* و *۵۰۰ عدد استخراجی* می‌باشد.

برای تولید پیکره توسط افراد، نرم افزاری به نام خلاصه یار تولید شد که از آن برای تولید خلاصه‌های مختلف می‌توان استفاده نمود. در بخش‌های بعدی به معرفی این سامانه پرداخته شده است.

لیست موضوعات تک سندی

همانطور که پیش تر اشاره شد، ۱۰۰ سند برای خلاصه سازی تک سندی جمع آوری شد که لیست عنوان موضوعات جمع آوری شده در جدول ذیل آورده شده است:

جدول ۴- لیست موضوعات اسناد برای تولید خلاصه های تک سندی

ردیف	موضوع	شماره سند
۱	جدایی نادر از... بالاخره خوب است یا نه	ALF.CU.13910117.019
۲	لزوم گسترش و بکارگیری تجارت الکترونیکی	ALF.EC.13901207.026
۳	قطر بازار جدید صادرات فرش ایرانی	ALF.EC.13910204.022
۴	کالاهایی که ۵ سال در گمرک خاک می‌خورد	ALF.EC.13910204.024
۵	گشایش بزرگترین نمایشگاه تجاری جهان در آلمان	ALF.EC.13910204.025
۶	تازه‌ترین دستاورد دانشمندان برای درمان کچلی	ALF.SC.13910131.018
۷	در گفتگو با یک جامعه شناس مطرح شد؛ کار برای ما ضد ارزش است	ALF.SO.13910126.023
۸	برخورد اماکن با خوابگاههای غیرمجاز	ALF.SO.13910130.017
۹	شمقدری، ایازی، اسلامی‌مهر، آقامحمدیان و...: معرفی اعضای شورای سیاستگذاری نخستین جشنواره فیلم ویدیویی تهران	FAR.CU.13910203.004
۱۰	با همکاری سینمای استرالیا: بازیگر انگلیسی فیلمساز شد	FAR.CU.13910203.005
۱۱	۸ تهدید برای بانک‌های ایرانی در سال ۹۱	FAR.EC.13910127.012
۱۲	نکته ای قابل توجه پس از نشست استانبول: نتایج زود هنگام مذاکرات	FAR.EC.13910129.014

	هسته ای	
۱۳	تأملی درباره گرانی های اخیر	FAR.EC.13910131.011
۱۴	نقش تبلیغات تجاری بر اشاعه‌ی مصرف‌زدگی	FAR.EC.13910203.008
۱۵	بدغذا خوردن افسردگی می‌آورد	FAR.SC.13901223.007
۱۶	۵ عقیده رایج اما اشتباه	FAR.SC.13910122.006
۱۷	در آکادمی ملی المپیک: دور جدید تمرینات تیم ملی والیبال بانوان از ۵ اردیبهشت می‌شود	FAR.SP.13910202.010
۱۸	هابل ۲۲ ساله شد؛ سومین تلسکوپ قدرتمند جهان از دید یک اخترشناس	HAM.SC.13910204.038
۱۹	وزیرخارجه انگلیس نیز بالاخره از خشونت های بحرین ابراز نگرانی کرد	IRN.PO.13910202.036
۲۰	انتخابات ریاست جمهوری فرانسه آغاز شد	IRN.PO.13910203.028
۲۱	نماینده دومای روسیه نتایج نشست استانبول را مثبت ارزیابی کرد	IRN.PO.13910203.037
۲۲	پایگاه خبری مالزیایی: دستیابی ایران به سیستم ساخت پهپاد آمریکایی یک موفقیت بزرگ محسوب می شود	IRN.PO.13910204.031
۲۳	فرمانده نیروی انتظامی: تلفات رانندگی با اجبار کاهش نمی یابد	IRN.SO.13910204.029
۲۴	فرهادی با چه کسانی رقابت کرد؟	JAM.CU.13901101.003
۲۵	کاهش مشارکت در انتخابات فرانسه	JAM.PO.13910203.034
۲۶	انتخابات ریاست جمهوری در فرانسه روز یک شنبه آغاز شد	PTV.PO.13910203.015
۲۷	حمایت از تولید ملی و افزایش صادرات	TAB.EC.13910125.030
۲۸	عباسعلی نورا : کاهش ارزش پول ملی عامل بازدارنده سرمایه‌گذاری	TAB.EC.13910203.035
۲۹	یک روز مانده به انتخابات فرانسه: «سارکوزی» در آستانه شکست!	TAB.PO.13910202.032
۳۰	مورینیو چگونه سد بارسا را شکست؟	TAB.SP.13910203.033
۳۱	"فرمول خون"اوضاع بحرین را نقش برآب کرد	ALF.PO.13910203.021
۳۲	زنی که پایه های دیکتاتوری قذافی را لرزاند	ALF.PO.13910204.020
۳۳	هشدار ارتش کره شمالی به کره جنوبی	ALF.PO.13910204.027
۳۴	برای آموزش قوانین راهنمایی و رانندگی صورت گرفت: ابتکار معلم	FAR.SO.13910202.009

	سیرجانی در یادگیری خوب دانش‌آموزان	
۳۵	از نهم اردیبهشت در مجموعه ورزشی آزادی: بانوان دوچرخه سوار در اولین مرحله لیگ کورسی رقابت می‌کنند	FAR.SP.13910202.013
۳۶	برندگان نوبل فیزیک ۲۰۱۱ معرفی شدند	HAM.SC.13900713.047
۳۷	روایتی از شرکت رهبر انقلاب در سرشماری عمومی نفوس و مسکن	HAM.SO.13900802.050
۳۸	مرحله نیمه نهایی لیگ برتر بسکتبال از امروز کلید می‌خورد: آغاز بازی‌های دشوار برای فینالیست‌شدن	HAM.SP.13910203.046
۳۹	ارکستر آکادمیک تهران با عنوان خلیج فارس به صحنه می رود	IRN.CU.13910203.045
۴۰	مرکز دایره المعارف بزرگ اسلامی با ۴۰ اثر چاپ اولی در نمایشگاه کتاب حضور می یابد	IRN.CU.13910204.055
۴۱	رییس کمیته عالی مستقل انتخابات تونس: در برخی حوزه ها بیش از ۸۰ درصد واجدان شرایط رای دادند	IRN.PO.13900801.051
۴۲	گونه نادر سیاه گوش برای نخستین بار در استان مرکزی مشاهده شد	IRN.SC.13910131.052
۴۳	همزمان با روز زمین، لامپی با طول عمر بیش از ۲۰ سال عرضه می شود	IRN.SC.13910204.053
۴۴	محققان ایرانی نانوکامپوزیت زیستی برای پوشش دهی سیب تولید کردند	IRN.SC.13910204.054
۴۵	دوومیدانی قهرمانی کشور؛ رکورد پرش ارتفاع ایران شکسته شد	IRN.SP.13910201.048
۴۶	اسماعیل پور: رضا یزدانی می تواند در بازی های المپیک لندن جاودانه شود	IRN.SP.13910202.016
۴۷	تیم ملی شنای کشورمان با ۸ ورزشکار عازم مالزی می شود	IRN.SP.13910204.049
۴۸	میراسماعیلی:نتایج تیم ملی جودو در مسابقات آسیایی قابل پیش بینی نیست	IRN.SP.13910204.057
۴۹	هفته نزولی طلا	JAM.EC.13910203.056
۵۰	علم چیست؟	JAM.SC.13910204.002
۵۱	کیپسانگ فاتح ماراتن لندن شد	JAM.SP.13910203.001
۵۲	چند راکت منطقه دیپلماتیک پایتخت افغانستان را لرزاند	PTV.PO.13910127.042
۵۳	"مصر برای اسراییل خطرناک‌تر از ایران است"	PTV.PO.13910203.043

۵۴	اتحادیه اروپا درباره تحریم سوریه به توافق رسیده است	PTV.PO.13910204.040
۵۵	ناظران اعزامی سازمان ملل به سوریه از شهرهای حمص، حما و رستن بازدید کردند	PTV.PO.13910204.044
۵۶	اتحادیه عرب خواهان راه حل سیاسی در سوریه	PTV.PO.13910205.041
۵۷	شکار هواپیمای امریکایی و نگرانی اسرائیل	PTV.PO.13910921.060
۵۸	بر پایه آمار انتخاباتی؛ نمایندگانی که بیشترین درصد رأی را در انتخابات نهم به دست آوردند	TAB.PO.13910202.039
۵۹	قتل تازه در پرونده مردی با سابقه ۳ جنایت	TAB.SO.13910120.059
۶۰	مهمترین‌تفاوت‌های‌اینترنت‌ملی‌بااینترنت‌فعلی	TAB.SO.13910202.058
۶۱	تشدید طرح برخورد با موتورسواران متخلف	ALF.SO.13910203.075
۶۲	آیا مجلس مردم را قانع می‌کند؟جزئیات گزارش اختلاس درکمیسیون اصل۹۰	HAM.EC.13900802.083
۶۳	دانشمندان در جست‌وجوی داوطلب برای شبیه سازی سفر به مریخ	HAM.SC.13901118.085
۶۴	پژوهشگران در مجله لنست اعلام کردند:افزایش شمار خودکشی‌ها در اروپا به دنبال بحران اقتصادی ۲۰۰۸	HAM.SO.13900419.072
۶۵	شمار مظنونان پرونده اسیدپاشی مرگبار به ۵ نفر رسید	HAM.SO.13900524.087
۶۶	ایران قهرمان وزنه برداری جوانان جهان شد	HAM.SP.13900416.076
۶۷	انتظار نیم قرنی والیبال به پایان می‌رسد؟آرزوی رسیدن به المپیک	HAM.SP.13901125.070
۶۸	چشم‌ها به کامپیوتر دروغ نمی‌گویند	JAM.SC.13910204.084
۶۹	روان‌شناسان معتقدند یکی از عوامل اضطراب دانش‌آموزان، فشار روانی والدین بر آنهاست:اضطراب امتحان	JAM.SO.13900307.066
۷۰	اغلب بازنشستگان زیر خط فقر هستند	JAM.SO.13901020.071
۷۱	طی دیماه سال جاری:تهران در صدر آمار طلاق است	JAM.SO.13901203.081
۷۲	بررسی ضرورت ارتباط والدین و مربیان مدرسه:مسیری هموار میان خانه اول و دوم	JAM.SO.13910204.062
۷۳	رییس کمیته فوتسال تشریح کرد:برنامه‌های تیم ملی فوتسال	JAM.SP.13900813.078
۷۴	تظاهرات ضدامریکایی و سعودی در یمن	PTV.PO.13900507.097

۷۵	ادامه سرکوب اعتراضات جنبش ۹۹ درصدی	PTV.PO.13901228.068
۷۶	افزایش تلفات زلزله ترکیه به ۶۰۰ نفر	PTV.SO.13900809.096
۷۷	آیۀالله جوادی: تعیین‌کنندگان زمان ظهورِ دجّالند	TAB.CU.13900209.074
۷۸	اثر جدید استاد فرشچیان رونمایی شد	TAB.CU.13900311.069
۷۹	اقامتگاه امام زمان(عج) کجاست؟	TAB.CU.13900424.073
۸۰	برگزاری هفته هنر یک استان کهن و تاریخی	TAB.CU.13900428.077
۸۱	نگرانی از خاموشی ارکستر سمفونیک تهران	HAM.CU.13900115.094
۸۲	نمایش مستند های جشنواره فیلم شهر در خانه هنرمندان	HAM.CU.13900218.093
۸۳	جذب ۵ هزار مربی قرآن در آموزش و پرورش در ۳ مرحله آزمون انجام شد	HAM.CU.13901206.082
۸۴	مهم‌ترین رویدادهای فرهنگی و ادبی مهر سال ۹۰	HAM.CU.13910107.090
۸۵	۹۰ درصد بودجه قرآنی سال گذشته معادل ۲۰۰ میلیارد تومان به حساب واریز شد	HAM.CU.13910116.061
۸۶	قیمت سکه در بورس، روی مرز ۷۰۰ هزار تومان	HAM.EC.13900619.089
۸۷	کتاب زندگی و آثار خوشنویسی استاد غلامحسین امیرخانی منتشر شد:یک عمر زندگی در خطوط	JAM.CU.13900622.092
۸۸	دری: کتاب «کلنل» دولت‌آبادی را مطالعه می‌کنم، اما ضرورتی برای انتشارش در داخل کشور وجود ندارد:شهرستان‌ها و روستاها؛ کانون توجه هفته کتاب	JAM.CU.13900819.067
۸۹	معاون صدا مطرح کرد:زبان سایبری بهترین روش برای معرفی فرهنگ عاشورا	JAM.CU.13901114.086
۹۰	جشنواره هنرهای تجسمی فجر امروز با معرفی ۱۳۰ برگزیده و اهدای ۹ نشان عالی به کار خود پایان می‌دهد:پایان طولانی‌ترین جشنواره فجر پس از ۳۴ روز	JAM.CU.13901214.063
۹۱	نگاهی به نقش روبات‌ها در پژوهش‌های دیرینه‌زیست‌شناسی:روبات‌های پرنده، کلید حل معمای پرواز می‌شوند	JAM.SC.13900924.065
۹۲	توهم و هذیان در روان‌های پریشان موج می‌زند:روان‌پریشی، فراگیرتر از دیابت	JAM.SC.13901218.064
۹۳	یک اقتصاددان یونانی: یورو فرو خواهد پاشید	PTV.EC.13900917.100

۹۴	عربستان و اردن در پس ناآرامی‌های سوریه	PTV.PO.13900120.099
۹۵	تجزیه لیبی در دستور کار ناتو	PTV.PO.13900202.098
۹۶	۹۱ کشته در بمب‌گذاری و تیراندازی نروژ	PTV.PO.13900501.095
۹۷	هما با اجازه مراجع صبحانه سرو می‌کند	TAB.CU.13900522.091
۹۸	روز شمار بحران ارزی خرداد ماه؛پایان بحران ارزی، در نزدیکی مرز دلار ۱۱۰۰ تومانی	TAB.EC.13900331.079
۹۹	پنجاه درصد از جوانان یونانی بیکار هستند	TAB.EC.13900401.080
۱۰۰	صاحبخانه، مستاجر را مقابل سگ‌های گرسنه قرار داد	TAB.SO.13900412.088