سوالات استخدامی : بازیابی اطلاعات

۱- عبارت"یک سیستم همیشه باید مشخص ترین قسمت از یک سند را درپاسخ به جستجو برگرداند" به کدام اصل اشاره دارد؟

۱. اصل بازیابی سند عام
۲. اصل بازیابی سند تشخیصی
۳. اصل بازیابی سند ساخت یافته
۴. هر سه مورد

۲- کلمات بسیار عام که ارزش اندکی در کمک به انتخاب سند منطبق با نیاز کاربر دارند را چه می نامند؟

۱. کلمات توقف
۲. نیاز اطلاعاتی
۳. کلمات کم ارزش
۴. هر سه گزینه درست است

۳- "روند استاندارد سازی نشانه ها به طوری که تطبیق علیرغم تفاوت های صوری در دنباله کارکتری نشانه ها رخ دهد" چه نامیده می شود.

۱. تطبیق نشانه ها
۲. قوانین نگاشت
۳. نرمالسازی نشانه ها
۴. نگاشت نشانه ها

۴- برای جستجو در لغت نامه از کدامیک ازموارد زیر می توان بهره برد؟

۱. درختان جستجو
۲. درهم سازی
۳. جستجوهای جایگزین
۴. گزینه ۱و۲

۵- کدام گزینه یک پرس و جوی دنباله دار است؟

۱. S*dney
۲. *idney
۳. Sidne*
۴. هر سه گزینه درست است.

۶- کدام گزینه عبارات را به جای شناسه های آنها در شاخص گزاری استفاده می کند؟

۱. شاخص گزاری برون حافظه ای
۲. شاخص گزاری بلوکی تک گذره
۳. شاخص گزاری بلوکی مبتنی بر مرتب سازی
۴. شاخص گزاری درون حافظه ای تک گذره

۷- کدامیک از گزینه های زیر از اهداف شاخص گزاری لغت نامه نمی باشد؟

۱. گنجاندن لغت نامه در حافظه
۲. سهولت در شاخص گزاری
۳. حفظ حافظه
۴. تسهیل در اشتراک منابع با برنامه های دیگر

۸- به میانگین وزندار هارمونیک از صحت و بازخوانی ..... گفته می شود.

۱. Map-Measure
۲. آماره ی Kappa
۳. F-Measure
۴. نمودار ROC

۹- شباهت بین دو سند در فضای برداری را چگونه ارزیابی می کنند ؟

۱. تفاوت بین دو بردار سند
۲. شباهت کسینوسی
۳. ضرب نقطه ای بردارها
۴. ضرب طول اقلیدسی بردارها

۱۰- ایده ی محاسبه کردن "مجموعه ی r با بالاترین وزن ها برای عبارت t موردنظر" چه نام دارد؟

۱. لیست های قهرمان
۲. اسناد برتر
۳. لیست های مورد علاقه
۴. هر سه گزینه

۱۱- برای بازیابی سند XML ، روش های بازیابی XML متن-محور بر چه اساسی عمل می کنند؟

۱. فیلدهای طولانی متن-تطبیق دقیق-نتایج ربط رتبه بندی شده
۲. فیلدهای طولانی متن-تطبیق نادقیق-نتایج ربط رتبه بندی شده
۳. فیلدهای طولانی متن-تطبیق دقیق-نتایج ربط رتبه بندی نشده
۴. فیلدهای طولانی متن-تطبیق نادقیق-نتایج ربط رتبه بندی نشده

۱۲- "قانون" Zipf " و" قانون Heaps " به ترتیب چه کاربردی دارند؟

۱. مدلسازی توزیع عبارات- تخمین تعداد عبارات
۲. تخمین تعداد عبارات-مدلسازی توزیع عبارات
۳. هر دو مورد برای تخمین تعداد عبارات
۴. هر دو مورد برای مدلسازی توزیع عبارات

۱۳- چالش بزرگ موتورهای جستجوی وب در شاخص گزاری و بازیابی کدام مورد است؟

۱. توزیع نامتوازن کاربران و تنوع سخت افزارها
۲. تنوع جغرافیایی و توزیع نامتوازن کاربران
۳. انتشار متمرکز محتوا با کنترل مرکزی اعتبار
۴. انتشار غیرمتمرکز محتوا بدون کنترل مرکزی اعتبار

۱۴- روش مبارزه با ارسال هرزنامه که متن صفحات وب خود را دستکاری می کند با چه عنوان شناخته می شود؟

۱. تحلیل پیوند
۲. تحلیل متن
۳. تحلیل SEO
۴. تحلیل پوشاندن

۱۵- کدامیک از روش های زیر برای تشخیص صفحات وب دو نسخه ای نزدیک استفاده می شود؟

۱. روش پوشاندن
۲. روش تحلیل پیوند
۳. روش SEO
۴. روش اثرانگشت

۱۶- کدامیک از موارد زیر جزو ویژگی های کدهای گاما V می باشد؟

۱. جهانی بودن
۲. پیشوند آزاد
۳. بدون پارامتر
۴. هر سه مورد صحیح است

۱۷- د رکدگزاری بایت متغییر، هفت بیت آخریک بایت که بخشی از فاصله را کدگزاری می کند،چه می نامند؟

۱. بیت توازن ۲. بیت ادامه ۳. بار ۴. نیبل

۱۸- ایده اصلی الگوریتم های soundex چیست؟

۱. تولید الگوریتم های تبدیل صوت
۲. تولید الگوریتم های تبدیل متن به صوت
۳. تولید یک درهم سازی آوایی
۴. تولید الگوریتم های مقاوم سازی صوت

۱۹- چه زمانی بازخورد ربط به تنهایی کافی نمی باشد؟

۱. املای غلط
۲. بازیابی اطلاعات بین زبانی
۳. پرس و جوهایی که مجموعه پاسخ ذاتا فصلی دارند.
۴. همه موارد

۲۰- مقدار kappa .....خواهد بود اگر دو داور همیشه موافق باشند و ..... خواهد بود اگر آنها در نرخ داده شده توسط شانس به توافق برسند.

۱. یک - منفی ۲. منفی- صفر ۳. یک - صفر ۴. صفر- منفی

سوالات تشریحی

۱- تفاوت مدخل گیری و ریشه گیری را با ذکر مثال توضیح دهید. ۱/۲۰ نمره

۲- گام پیش پردازش هرس خوشه ای را برای خوشه بندی بردارهای سند توضیح دهید. ۱/۲۰ نمره

۳- الگوریتم K گرمی برای پرس و جوهای جایگزین توضیح دهید. ایرادات روش را بیان نمایید. ۱/۲۰ نمره

۴- تفاوت ها و شباهت های شاخص های موقعیتی و غیر موقعیتی را شرح دهید. ۱/۲۰ نمره

۵- بازخورد ربط و شبه بازخورد ربط را به صورت کامل توضیح دهید و تفاوت آن ها را بیان نمایید. ۱/۲۰ نمره

شماره سوال	پاسخ صحیح
1	ج
2	الف
3	ج
4	د
5	ج
6	د
7	ب
8	ج
9	ب
10	د
11	ب
12	الف
13	د
14	الف
15	الف
16	د
17	ج
18	ج
19	د
20	ج

۱- «. . . . . ،. موضوعی است که کاربران علاقه مندند تا درباره آن بیشتر بدانند.»

۱. پرس و جو ۲. سند ۳. نیاز اطلاعاتی ۴. عبارت

۲- توصیف زیر، در مورد نتایج بازگشتی سیستم، به کدام گزینه اشاره دارد؟

«چه کسری از نتایج بازگشتی، به نیاز اطلاعاتی مرتبط است؟»

۱. شاخص ۲. صحّت ۳. بازخوانی ۴. فرهنگ واژگان

۳- . . . . . ، سازماندهی کار پاسخگویی به پرس و جو است به طوریکه حداقل مقدار کار توسط سیستم انجام شود.

۱. پردازش پرس و جو ۲. بهینه‌سازی پرس و جو ۳. اشتراک پرس و جو ۴. شاخص گذاری

۴- . . . . . معمولاً به فرآیند مکاشفه ای خاصی اشاره دارد که اکثر اوقات، انتهای کلمات را به امید دستیابی به یک صورت پایه متعارف، قطع می‌کند و گاهی شامل حذف ضمیمه‌های نحوی است.

۱. مدخل‌گیری ۲. ریشه‌گیری ۳. هم‌ارزی ۴. نرمالسازی

۵- حذف بخش انتهایی کلمه چه نام دارد؟

۱. مدخل‌گیری ۲. حذف ریشه

۳. ریشه‌گیری ۴. همه موارد صحیح است.

۶- در لغت نامه از چه نوع درخت جستجو استفاده می کنیم؟

۱. درخت جستجوی دودویی ۲. درخت جستجوی دودویی بهینه

۳. درخت سیاه قرمز ۴. B-Tree

۷- گاهی اوقات، کلمات بسیار عام که ظاهراً ارزش اندکی در کمک به انتخاب اسناد منطبق با نیاز کاربر دارند، باید از مجموعه واژگان مستثنی شوند. این کلمات . . . . . نامیده می‌شوند.

۱. کلمات زاید ۲. کلمات اصلی ۳. کلمات ناقص ۴. کلمات توقّف

۸- . . . . . به فرآیند مکاشفه‌ای خاصی اشاره دارد که اکثر اوقات، انتهای کلمات را به امید دستیابی به این هدف قطع می‌کند و گاهی شامل حذف ضمیمه‌های نحوی است.

۱. مدخل‌گیری ۲. ریشه‌گیری ۳. حذف ۴. شاخص‌گذاری

۹- هزینه الگوریتم شاخص گذاری بلوکی مبتنی بر مرتب سازی با الگوریتم BSBI چیست؟ (T یک حد بالا برای تعداد عبارتی است که ما باید مرتب کنیم)

۱. $\theta(T)$ ۲. $\theta(T^2)$ ۳. $\theta(T\,Log^T)$ ۴. $\theta(Log^T)$

۱۰- کدام الگوریتم‌ها، مجموعاً به عنوان الگوریتم‌های Soundex معروف هستند.

۱. الگوریتم‌های تشخیص صوت
۲. الگوریتم‌های تبدیل متن به صوت
۳. الگوریتم‌های درهم سازی آوائی
۴. الگوریتم‌های مقاوم‌سازی صوت

۱۱- بخشی از حافظه اصلی را که در بلوک خوانده شده یا نوشته شده در آن ذخیره می‌شود، ..... می‌نامیم.

۱. کَش
۲. پیگرد
۳. میانگیر
۴. گذرگاه

۱۲- در کدگذاری بایت متغیر، ۷ بیت آخر یک بایت، ..... نامیده می‌شود و بخشی از فاصله را کدگذاری می‌کند.

۱. توازن
۲. بار
۳. ادامه
۴. نیبل

۱۳- یک کد با این ویژگی که برای توزیع دلخواه P، فاکتوری از کد بهینه باشد، ..... نامیده می‌شود.

۱. کد بهینه
۲. آنتروپی
۳. کد جهانی
۴. کد پیشوندی

۱۴- برای وزن‌دهی tf-idf، لیست‌های قهرمان، r سندی خواهند بود که بالاترین مقادیر ..... را برای عبارت t دارند.

۱. idf
۲. tf
۳. if
۴. tf + idf

۱۵- جمله "می توان از منابع غیر مستقیم، به جای بازخورد صریح ربط، به عنوان مبنای بازخورد ربط، استفاده کرد." تعریف چیست؟

۱. شبه بازخورد ربط
۲. بازخورد کور
۳. بازخورد ربط روی وب
۴. باز خورد ربط ضمنی

۱۶- توصیف زیر، کدام اصل را توضیح می‌دهد؟

«یک سیستم، همیشه باید مشخّص‌ترین قسمت از یک سند را در پاسخ به پرس و جو برگرداند.»

۱. اصل بازیابی سند ساخت‌یافته
۲. اصل بازیابی تشخیصی
۳. اصل بازیابی سند عام
۴. اصل بازیابی عامل سند

۱۷- کدام گزینه در مورد میانگین متوسط صحّت (MAP) صحیح نیست؟

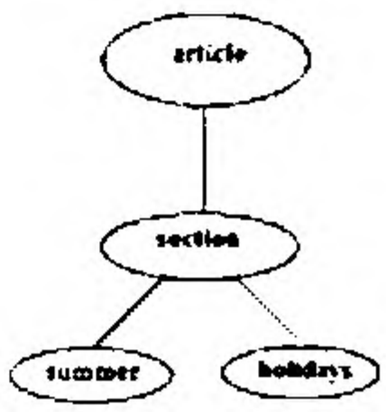
۱. در میان معیارهای ارزیابی، MAP به طور خاص، پایایی و قابلیت جداسازی خوبی دارد.
۲. با استفاده از MAP، سطوح بازخوانی ثابت انتخاب می‌شوند.
۳. با استفاده از MAP، هیچ درونیابی وجود ندارد.
۴. MAP یک معیار تک شکلی از کیفیت را روی سطوح بازخوانی، فراهم می‌آورد.

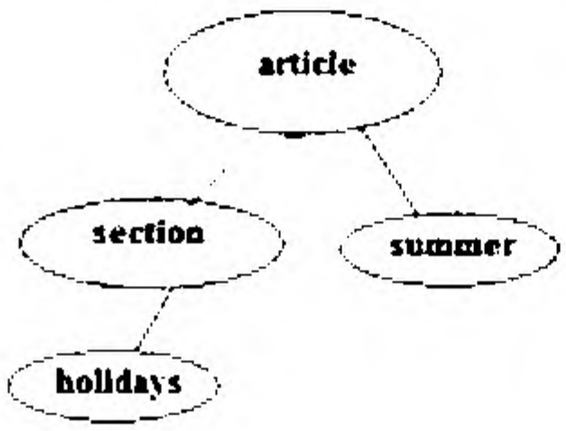
۱۸- در کدام روش، تحلیل محلی خودکار صورت می‌گیرد؟

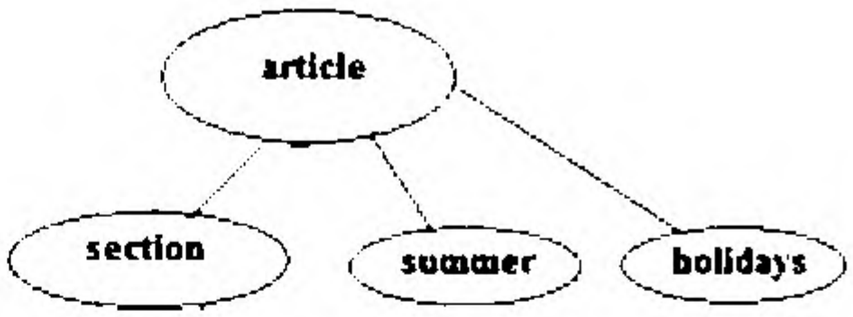
۱. بازخورد ربط کور
۲. شبه بازخورد ربط
۳. بازخورد کاربر
۴. موارد اول و دوم صحیح است.

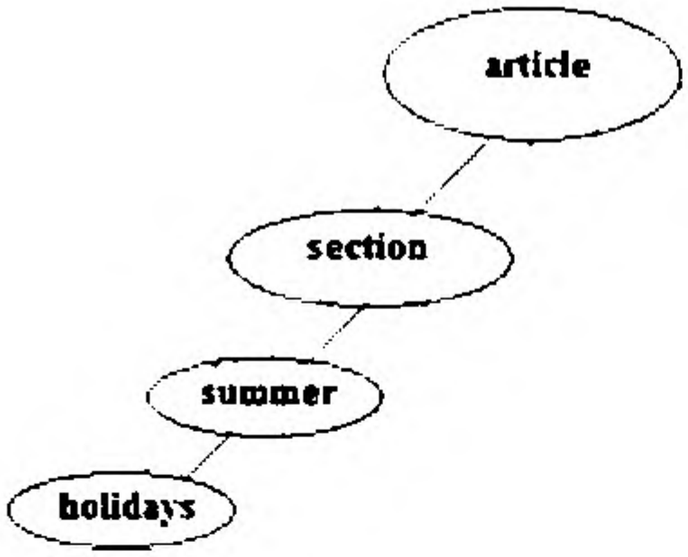
۱۹- پرس و جوی XML در فرمت NEXI در دستور زیر، نمایش جزئی کدام درخت می باشد.

//article
[.//yr = 2001 or .//yr = 2002]
//section
[about(., summer holidays)]

۱.


۲.


۳.


۴.


۲۰- فرمول محاسبه معیار «دقّت» یعنی میزان دسته‌بندی‌های صحیح برابر است با:

توضیحات:

tp = true positive ، fp = false positive

fn = false negative ، tn = true negative

۱. $$\text{accuracy} = \frac{tp + fp}{tp + fp + fn + tn}$$

۲. $$\text{accuracy} = \frac{tp + fn}{tp + fp + fn + tn}$$

۳. $$\text{accuracy} = \frac{tp + fn + tn}{tp + fp + fn + tn}$$

۴. $$\text{accuracy} = \frac{tp + tn}{tp + fp + fn + tn}$$

۲۱- اساس روش بازیابی XML متن-محور برای بازیابی سند XML چیست؟

۱. فیلدهای طولانی متن، تطبیق نادقیق و نتایج ربط رتبه بندی

۲. فیلدهای طولانی متن، تطبیق دقیق و نتایج ربط رتبه بندی

۳. فیلدهای کوتاه متن، تطبیق نادقیق و نتایج ربط رتبه بندی

۴. فیلدهای کوتاه متن، تطبیق دقیق و نتایج ربط رتبه بندی

۲۲- کدام یک از موارد زیر، از خصوصیات و ویژگیهای وب است؟

مورد اوّل: بی‌مانند در مقیاس

مورد دوّم: تقریباً بی‌مانند در نبود هماهنگی کامل در ایجاد

مورد سوم: بی‌مانند در تنوع پس‌زمینه‌ها

مورد چهارم: بی مانند در انگیزه اعضای آن

۱. فقط موارد اول و دوم و سوم

۲. فقط موارد اول و سوم و چهارم

۳. فقط موارد دوم و سوم و چهارم

۴. موارد اول و دوم و سوم و چهارم

۲۳- کدام گزینه، سه دسته عمده از صفحات وب را به خوبی و درستی بیان می‌کند؟

۱. IN ، OUT ، SCC

۲. IN ، PIN ، LINK

۳. OUT ، SCC ، LINK

۴. IN ، OUT ، PIN

۲۴- توصیف زیر، کدام ویژگی یک پیمایشگر مطلوب را توصیف می‌کند.

«با دانستن اینکه کسر قابل توجّهی از تمامی صفحات وب، بهره وری ضعیفی برای ارائه سرویس به نیازهای پرس و جوی کاربر دارند، پیمایشگر باید در ابتدا نسبت به آوردن صفحات مفید بایاس داشته باشد.»

۱. تازگی ۲. کارآیی و بهره‌وری ۳. کیفیت ۴. توسعه‌پذیری

۲۵- کدام گزینه صحیح است.

۱. ترجمه یک URL به شکل متنی، به یک آدرس IP، تحلیل DNS نامیده می‌شود.

۲. سرویس نام دامنه (DNS) ماهیّت توزیع‌شده‌ای دارد.

۳. پیاده‌سازی‌های جستجوی نام دامنه در کتابخانه‌های استاندارد، معمولاً سنکرون هستند.

۴. همه موارد فوق، صحیح می‌باشند.

سوالات تشریحی

۱- الگوریتم ریشه‌گیر Porter برای ریشه‌گیری در زبان انگلیسی را توضیح دهید، مراحل آن را عنوان نموده و شرح دهید. ۱/۲۰ نمره

۲- الگوریتم شاخص‌گذاری بلوکی مبتنی بر مرتب‌سازی را توضیح دهید. ۱/۲۰ نمره

۳- در علوم اجتماعی، یک معیار رایج برای توافق بین داوران آماره (Kappa Statistic) Kappa است. این معیار برای داوریهای قیاسی طراحی شده است و نرخ توافق ساده را برای نرخ توافق تصادفی تصحیح می کند. برای جدول زیر آماره Kappa را بدست آورید. ۱/۲۰ نمره

		ربط داور ۲		
		بله	خیر	کل
ربط داور ۱	بله	۳۰۰	۲۰	۳۲۰
	خیر	۱۰۰	۷۰	۸۰
	کل	۳۱۰	۹۰	۴۰۰

۴- پدیده ناهمگونی شِما در XML را شرح دهید. ۱/۲۰ نمره

۵- پرس و جوهای وب در سه رده وسیع گروه بندی می شوند، این سه رده را نام برده و توضیح مختصری برای هر کدام ارائه دهید. ۱/۲۰ نمره

پاسخ صحیح	شماره سوال
ج	1
ب	2
ب	3
ب	4
ج	5
د	6
د	7
ب	8
ج	9
ج	10
ج	11
ب	12
ج	13
ب	14
د	15
الف	16
ب	17
د	18
الف	19
د	20
الف	21
د	22
الف	23
ج	24
د	25

سوالات تشریحی

۱- پاسخ در صفحه ۵۶ از فصل ۲ منبع درسی.

۲- پاسخ در صفحه ۹۵ از فصل ۴ منبع درسی.

۳-

جواب صفحه ۱۹۳ کتاب مقدمه ای بر بازیابی اطلاعات ترجمه دکتر هدیه ساجدی مهندس زهرا و فرناز سادات تقوی

۴- صفحه ۲۳۱

۵- پاسخ در صفحه ۴۷۱ از فصل ۱۹ منبع درسی.

۱- کدام گزاره‌ها صحیح است؟

الف) یافتن اطلاعات در متن صورت می‌گیرد.

ب) همیشه بازیابی از متن صورت می‌گیرد.

ج) نیاز اطلاعاتی از یک مجموعه بزرگ استخراج می‌شود.

۱. الف . ب ۲. الف و ج ۳. ج ۴. الف، ب وج

۲- عملکرد مجاورت در جستجوی بولی چه چیزی را نشان می‌دهد؟

۱. عملگرهایی که باید نزدیک هم رخ دهند. ۲. اسنادی که باید نزدیک هم رخ دهند.

۳. متن‌هایی که باید نزدیک هم رخ دهند. ۴. عبارت‌هایی که باید نزدیک هم رخ دهند.

۳- کدام گزاره‌ها گام‌های ساخت شاخص وارونه است؟

الف) تجمیع اسنادی که قرار است شاخص‌گذاری شود.

ب) نشانه‌گذاری متن

ج) مشخص کردن کلمات توقف

د) پیش‌پردازش متن از لحاظ کلمات شاخص

۱. الف و ب ۲. الف و ج ۳. ب و ج ۴. الف، ب و د

۴- حذف بخش انتهایی کلمه چه نام دارد؟

۱. مدخل‌گیری ۲. حذف ریشه

۳. ریشه‌گیری ۴. همه موارد صحیح است.

۵- کدام گزاره‌ها در خصوص شاخص موقعیتی صحیح است؟

الف) اگر مقدار /افست موقعیت را فشرده کنیم تعداد پست‌هایی که باید ذخیره کنیم کاهش نمی‌یابد.

ب) استفاده از شاخص موقعیتی پیچیدگی عملیات را کاهش می‌دهد.

ج) تعداد اقلامی که در شاخص موقعیتی باید کنترل شوند، محدود به تعداد اسناد است.

۱. ب و ج ۲. الف و ب ۳. الف ۴. الف، ب و ج

۶- چه نوع داده‌ساختاری برای جستجو استفاده می‌شود؟

۱. AVL ۲. Heap ۳. BST ۴. B-Tree

۷- تصحیح آوایی برای کدام مورد جستجو کاربرد بیشتری دارد؟

۱. جستجوی کلمات هم‌آوا
۲. جستجوی کلمات با ریشه‌ی معنایی یکسان
۳. جستجوی اسامی افراد
۴. همه موارد صحیح است.

۸- بخشی از حافظه که بلوک در آن نوشته و یا از آن خوانده می‌شود، چه نام دارد؟

۱. cash
۲. میانگیر
۳. حافظه مجازی
۴. حافظه بلوک

۹- کدام گزینه یک معماری عمومی برای محاسبات توزیع‌شده است؟

۱. GMP
۲. General Compute
۳. شاخص توزیع شده
۴. MapReduce

۱۰- قانون Zipf چه کاربردی دارد؟

۱. محاسبه میزان فشره‌سازی
۲. محاسبه ضریب اصابت
۳. مدل‌سازی توزیع عبارات
۴. تخمین اندازه مجموعه واژگان

۱۱- کدام گزاره‌ها درخصوص کد گاما صحیح است؟

الف) جهانی بودن
ب) پیشوند آزاد بودن
ج) چندپارامتری بودن

۱. الف و ب
۲. ب و ج
۳. الف و ج
۴. الف، ب و ج

۱۲- صورت خاصی از داده در مورد یک سند مثل مولف سند چه نام دارد؟

۱. فراداده
۲. داده ابتدایی
۳. MFD
۴. خصوصیت سند

۱۳- نرمال‌سازی محوری طول سند چیست؟

۱. تغییر طول سند است.
۲. نوعی روش تصحیح طول سند است.
۳. بهینه‌سازی طول سند است.
۴. همه موارد صحیح است.

۱۴- در هرس خوشه‌ای چه تعداد از اسناد را به عنوان رهبر انتخاب می‌کنیم؟

۱. $\sqrt{N}$
۲. N/4
۳. N/3
۴. N/2

۱۵- کدام گزاره‌ها در خصوص نمایش سند به صورت بردار صحیح است؟

الف) بدون اتلاف است.

ب) ترتیب نسبی عبارت در یک سند کدگذاری سند به عنوان بردار حفظ می‌شود.

ج) شاخص ساخته شده برای بازیابی فضای بردار برای پرس‌وجوی اصطلاحات به طور کلی قابل استفاده نیست.

۱. الف ۲. ب ۳. ج ۴. الف و ب

۱۶- کسری از اسناد بازیابی شده‌ی مرتبط چه نام دارد؟

۱. صحت ۲. دقت ۳. فراخوان ۴. بازیابی

۱۷- کدام آماره نرخ توافق ساده را برای نرخ توافق تصادفی تصحیح می‌کند؟

۱. مد ۲. رگرسیون ۳. واریانس ۴. Kappa

۱۸- دخالت دادن کاربر در روند بازیابی چه نام دارد؟

۱. بازخورد ربط ۲. بازخورد کاربر ۳. کاربر-تعامل ۴. تعامل-بازخورد

۱۹- در کدام روش، تحلیل محلی خودکار صورت می‌گیرد؟

۱. بازخورد ربط کور ۲. شبه بازخورد ربط

۳. بازخورد کاربر ۴. موارد اول و دوم صحیح است.

۲۰- نام دیگر بازیابی XML چیست؟

۱. بازیابی فرامتن ۲. بازیابی نیمه‌ساخت‌یافته

۳. بازیابی فرامدل ۴. بازیابی ساخت‌یافته

۲۱- در XML داده-محور کدام صفات کدگذاری می‌شود؟

۱. داده متنی و عددی ۲. داده عددی و غیرمتنی

۳. داده متنی و غیرعددی ۴. داده غیرمتنی و غیرعددی

۲۲- مشخصه‌ی اساسی وب که منجر به رشد انفجاری آن شد چه بود؟

۱. انتشار متمرکز بدون کنترل مرکزی ۲. انتشار غیرمتمرکز با کنترل مرکزی

۳. انتشار متمرکز با کنترل مرکزی ۴. انتشار غیرمتمرکز بدون کنترل مرکزی

۲۳- نام دیگر جستجوی محض چیست؟

۱. پایه‌ای
۲. الگوریتمی
۳. مبنایی
۴. همه موارد صحیح است.

۲۴- تبدیل URL به IP چه نام دارد؟

۱. جستجوی DNS
۲. تحلیل DNS
۳. NAT
۴. موارد لول و دوم صحیح است.

۲۵- کدام گزینه از راه‌کارهای کاهش حافظه جدول نگاشت بین پرس‌وجوها و صفحات است؟

۱. کدگذاری شکاف‌ها در لیست‌های مرتب شده
۲. استفاده از ویژگی محلیت با استفاده از اعداد بزرگ برای ادرس‌دهی
۳. استفاده از شباهت بین لیست‌ها و ذخیره لیست‌های مشابه در هر درایه
۴. موارد اول و دوم صحیح است.

سوالات تشریحی

۱- Grep کردن چیست؟ به طور کامل شرح دهید.

۲- میزان شباهت کسینوسی این دو سند را محاسبه کنید.

	مهدی زین‌الدین	قاسم سلیمانی	حسین باپا	حسین خرازی	احمد کاظمی	محمدابراهیم همت	احمد متوسلیان
سند ۱	3	4	2	3	1	2	5
سند ۲	1	5	0	2	4	0	4

۳- مطابق جدول زیر میزان بازخوانی و دقت را حساب کنید:

نامرتبط	مرتبط	
۴	۱۰	بازیابی شده
۵۰	۲۵	بازیابی نشده

۴- پدیده ناهمگونی شما در XML را شرح دهید.

پاسخ صحیح	شماره سوال
ب	1
د	2
الف	3
ج	4
ج	5
د	6
ج	7
ب	8
د	9
ج	10
الف	11
الف	12
ب	13
الف	14
ج	15
الف	16
د	17
الف	18
د	19
ب	20
ب	21
د	22
ب	23
د	24
الف	25

سوالات تشریحی

۱- صفحه 23

۲- صفحه 151

۳- صفحه 181

۴- صفحه 231

۱- روش Grep کردن برای کدام نوع از کاربردهای زیر مناسب است؟

۱. بازیابی های رتبه بندی شده
۲. مجموعه اسناد بزرگ با سرعت بالا
۳. پرس و جوهای ساده در مجموعه های نسبتا کوچک
۴. کاربردهای تطبیقی و انعطاف پذیر

۲- کدام گزینه غلط است؟

۱. لیست پستها معمولا در حافظه اصلی و لغت نامه در دیسک نگهداری می شوند تا سرعت بازیابی افزایش یابد.
۲. اگر بروزرسانی اسناد کم باشد، آرایه های با طول متغیر، متراکم تر و سرعت پیمایش سریعتری دارند.
۳. آرایه های با طول متغیر نسبت به لیست های پیوندی، فضای کمتری نیاز دارند و سریعتر هستند
۴. مزیت لیست پیوندی برای نگهداری لیست پستها، هزینه کم درج در آنها است.

۳- در کدام مرحله از تعیین مجموعه واژگان عبارت، امکان حذف کامل جمله "to be or not to be" وجود دارد؟

۱. نشانه گذاری و حذف علائم نگارشی
۲. نرمال سازی
۳. کلمات توقف
۴. ریشه گیری

۴- اگر در پرس و جوهای اصطلاحات، به همراه شاخص دو کلمه ای، شاخص سه کلمه ای نیز داشته باشیم:

۱. مثبت های کاذب (FP) افزایش می یابد
۲. مثبتهای کاذب ارتباطی به تعداد کلمات انتخابی ندارد.
۳. از نظر فضای ذخیره سازی، کاملا عملیاتی است.
۴. مثبتهای کاذب (FP) کاهش می یابد

۵- در تکنیک درهم سازی برای جستجو در لغت نامه:

۱. یافتن انواع فرعی از یک عبارت پرس و جو براحتی امکان پذیر است.
۲. مشکلی برای استفاده در محیطهای پویا و رو به رشد ندارد.
۳. کلیدهای مورد استفاده در مجموعه سند، به یک ترتیب معین سازماندهی می شوند.
۴. یافتن همه اصطلاحاتی که با یک پیشوند خاص آغاز می شوند، ممکن نیست.

۶- اگر در یک فرایند تصحیح املایی به روش شاخص های 2-گرمی، برای کلمه bord عبارت motherboard پیشنهاد شود، ضریب جاکارد برابر کدام گزینه خواهد بود؟

۱. $\frac{3}{11}$
۲. $\frac{2}{11}$
۳. $\frac{3}{13}$
۴. $\frac{2}{13}$

۷- در مورد معماری Map-Reduce برای شاخص گذاری توزیع شده، کدام گزینه نادرست است؟

۱. در مرحله نگاشت، نگاشت داده های ورودی به جفت های کلید-مقدار توسط ماشین تجزیه کننده انجام می شود.

۲. ماشین های تجزیه کننده و وارونه کننده، مجموعه های مجزایی از ماشین ها هستند.

۳. در مرحله کاهش، کلیدها به J افراز تقسیم می شوند.

۴. تعداد افرازهای مرحله کاهش توسط فردی که سیستم شاخص گذاری را اجرا می کند، تعیین می شود.

۸- در شاخص گذاری پویا، اگر n اندازه شاخص کمکی و T تعداد کل پستها باشد، مرتبه زمانی الگوریتم ادغام لگاریتمی برابر کدام است؟

۱. $\theta\left(\log\frac{T}{N}\right)$ ۲. $\theta\left(\frac{T^2}{N}\right)$ ۳. $\theta\left(T\log\frac{T}{N}\right)$ ۴. $\theta\left(\frac{T}{N}\right)$

۹- بر اساس قانون Zipf ، اگر عبارت با بیشترین فراوانی CF بار در سند رخ دهد، دومین عبارت پرتکرار چه مقدار از وقوع های آن را خواهد داشت؟

۱. یک دوم ۲. یک سوم ۳. یک چهارم ۴. یک پنجم

۱۰- کد γ ی 111101001 به کدام مقدار در مبنای ده کدگشایی می شود؟

۱. 9 ۲. 25 ۳. 19 ۴. 24

۱۱- با فرض اینکه مجموعه ای از اسناد داشته باشیم که هر سند دارای چهار ناحیه عنوان، نویسنده، چکیده و بدنه هستند و به ترتیب وزن هر ناحیه برابر 0.2، 0.15، 0.25، و 0.4 باشد، اگر به دنبال عبارت " رده بندی" باشیم و این عبارت در عنوان و بدنه یک سند وجود داشته باشد و در مابقی ناحیه ها نباشد، چه نمره ای به جفت (سند، پرس و جو) داده خواهد شد.

۱. 0.45 ۲. 0.75 ۳. 0.35 ۴. 0.6

۱۲- در وزندهی و فراوانی عبارات.....

۱. tf-idf یک عبارت زمانی کم خواهد بود که تعداد وقوع آن در مجموعه اسناد کم باشد.

۲. با وزندهی tf، عبارات معین و خاص در تعیین مرتبط بودن اسناد به پرس و جو، کم اثر یا بی اثر هستند.

۳. cf به df ترجیح داده می شود.

۴. اندازه idf برای یک کلمه توقف بالاست.

۱۳- در نسخه هرس خوشه ای پارامتری به منظور محاسبه سریع نمره ها، پارامتر b1 به معنای الصاق هر پیرو به b1 رهبر است و پارمتر b2 به معنای در نظرگرفتن تعداد b2 رهبر نزدیک به عبارت پرس و جوی q در هنگام جستجو است. افزایش b1 و b2 چه تاثیری دارد؟

۱. احتمال یافتن k سند با بالاترین نمره علیرقم محاسبات بیشتر، کاهش می دهد.

۲. احتمال اینکه تعداد اسناد یافت شده نهایی از k کمتر باشد، افزایش می دهد.

۳. احتمال یافتن k سند با بالاترین نمره را البته با محاسبات بیشتر افزایش می دهد.

۴. در حالت پایه که b1=b2=1 است، بهترین عملکرد را خواهیم داشت.

۱۴- کدام گزینه صحیح نیست؟

۱. از آنجاییکه داده ها اغلب بسیار نامتوازن هستند، معیار دقت (Accuracy) بسیار خوب عمل می کند.

۲. در معیار F اگر $\beta < 1$ باشد، تاکید بر صحت خواهد بود.

۳. بازخوانی برابر یک، منجر به صحت کم خواهد شد.

۴. در یک سیستم بازیابی، وقتی تعداد اسناد بازیابی زیاد می شود، صحت کاهش می یابد.

۱۵- کدام یک از معیارهای زیر، مساحت زیر منحنی بازخوانی-صحت را برای یک مجموعه از پرس و جوها تخمین می زند؟

۱. MAP ۲. F-Measure ۳. MAE ۴. ROC

۱۶- در رویه اصلی بازخورد ربط، پس از طرح پرس و جو توسط کاربر و نمایش مجموعه ابتدایی از نتایج بازیابی، چه عملی انجام می شود؟

۱. سیستم بر اساس نتایج قبلی، برخی نتایج را بصورت مرتبط یا غیرمرتبط علامتگذاری می کند.

۲. کاربر بعضی از نتایج را بعنوان مرتبط و غیرمرتبط علامتگذاری می کند.

۳. چندین کاربر در یک مکانیزم رای گیری، بعضی از نتایج را بعنوان مرتبط یا غیرمرتبط علامتگذاری می کنند.

۴. چندین سیستم بازیابی اطلاعات بر اساس نتایج قبلی، برخی از نتایج را بصورت مرتبط و غیر مرتبط علامتگذاری می کنند.

۱۷- کدام یک از معیارهای زیر تنها روی k نتیجه بازیابی شده یک سیستم بازیابی اطلاعات ارزیابی را انجام می دهد؟

۱. k precision ۲. NDCG ۳. ROC(k) ۴. گزینه های 1و 2

۱۸- « عناصر، صفات و متن درون عناصر در یک ساختار XML را به شکل گره ها در درخت بیان می کند» تعریف کدام گزینه است؟

۱. XML DTD ۲. Xpath ۳. DOM ۴. NEXI

۱۹- کدام گزینه درست است؟

۱. روشهای XML داده محور، معمولا برای مجموعه های داده با ساختار پیچیده استفاده می شوند که عمدتا دارای داده های غیر متنی هستند.

۲. بازیابی فقط محتوا، افزایش صحت بیشتری نسبت به بازیابی تمام ساختار در ابتدای لیست نتایج فراهم می کند.

۳. XML داده محور، عمدتا مقدار صفت داده غیر عددی و متنی را کدگذاری می کند.

۴. XML متن محور، بر جنبه های ساختاری از اسناد XML و پرس و جوها تاکید دارد.

۲۰- بزرگترین چالش برای موتورهای جستجوی وب در شاخص گذاری و بازیابی چیست؟

۱. انتشار غیرمتمرکز محتوا

۲. تنوع سیستم های سخت افزاری

۳. توزیع نامتوازن کاربران

۴. توزیع نامتوازن داده ها از نظر جغرافیایی

۲۱- کدام گزینه در مورد ساختار وب صحیح نیست؟

۱. گراف جهتدار اتصال دهنده صفحات وب یک شکل پاپیونی دارد.

۲. توزیع درجات ورودی به یک صفحه وب، از توزیع پواسون پیروی می کند.

۳. در ساختار پاپیونی وب، انتقال از صفحات SCC به هر صفحه در IN غیر ممکن است.

۴. توزیع درجات ورودی به یک صفحه وب، از قانون توانی $1/x^{\alpha}$ تبعیت می کند.

۲۲- روش مبارزه با ارسال هرزنامه ای که متن صفحات وب خود را دستکاری می کند، چه نام دارد؟

۱. تحلیل متن

۲. تحلیل SEO

۳. تحلیل پیوند

۴. تحلیل پوشاندن (شینگلینگ)

۲۳- اگر یک کاربر «شرکت هواپیمایی ایران ایر» را در وب جستجو کند، پرس و جوی او جزو کدام نوع پرس و جوها است؟

۱. پرس و جوی اطلاعاتی

۲. پرس و جوی تراکنشی

۳. پرس و جوی تجاری

۴. پرس و جوی راهبردی

۲۴- کدام گزینه چهار ویژگی از ویژگیهایی که یک پیمایشگر وب باید فراهم آورد را به درستی نشان می دهد؟

۱. مقیاس پذیری، توسعه پذیری، سرعت، تازگی

۲. توسعه پذیری، کیفیت، عمومیت، توزیع شده

۳. توزیع شده، مقیاس پذیری، عمومیت، تازگی

۴. مقیاس پذیری، کارایی و بهره وری، توسعه پذیری، تازگی

۲۵- حذف URL های دو نسخه ای با استفاده از کدام روش امکانپذیر است؟

۱. مقایسه ناحیه ای
۲. روش پوشاندن (شینگل)
۳. روش اثرانگشت
۴. گزینه های 2 و 3

سوالات تشریحی

۱- با داشتن جدول احتمال وقوع زیر مقدار معیار های صحت (Precision)، بازخوانی (Recall) و دقت (Accuracy) را محاسبه کنید؟ ۱،۲۰ نمره

	اسناد مرتبط	اسناد غیرمرتبط
اسناد بازیابی شده	43	8
اسناد بازیابی نشده	7	33

۲- الگوریتم اشتراک گیری سریع لیست پستها با استفاده از اشاره گرها را با نمایش دو لیست پست دلخواه توضیح دهید. ۱،۲۰ نمره

۳- توضیح دهید که چگونه کدگذاری بایتهای متغیر می تواند منجر به فشرده سازی فایل پست ها شود؟ ۱،۲۰ نمره

۴- ماتریس عبارت-سند زیر مفروض است. در این ماتریس تعداد وقوع هر عبارت در دو سند آورده شده است. شباهت کسینوسی دو سند را محاسبه نمایید. ۱،۲۰ نمره

	Introduction	to	information	retrieval	in	payame	noor
سند یک	2	3	1	1	1	0	0
سند دو	0	2	1	1	1	1	1

۵- عملیات پایه ی یک پیمایشگر (خزشگر) ابر متن را توصیف نمایید. ۱،۲۰ نمره

پاسخ صحیح	شماره سوال
ج	1
الف	2
ج	3
د	4
د	5
ب	6
ب	7
ج	8
الف	9
ب	10
د	11
ب	12
ج	13
الف	14
الف	15
ب	16
د	17
ج	18
الف	19
الف	20
ب	21
ج	22
د	23
د	24
د	25

سوالات تشریحی

فصل هشتم صفحه 181-1 نمره ۱،۲۰

$P = 43/51 = 0.843$

$R = 43/50 = 0.86$

$A = 76/91 = 0.835$

۲- صفحه 59 و 60 ۱،۲۰ نمره

۳- صفحه 122 پاراگراف وسط صفحه ۱،۲۰ نمره

برای ایجاد نمایش کارآمدتر از فایل پست ها نیازی نیست همه شناسه ها را با کدهای 20 بیتی نمایش دهیم. برخی عبارات پر تکرار که در هر سندی ممکن است بیایند فاصله شناسه اسنادشان برابر یک است. یعنی هر شماره سند تا شماره سند دیگری فقط در یک بیت اختلاف دارند (مثلا the). برخی عبارات مثل computer کمی بیشتر فاصله دارند مثلا در سند شماره 145 آمده و بعدی در 152 و فقط کلمات نادر هستند که به کل فضای 20 بیت ممکن است نیاز داشته باشند. بنابراین بجای 20 بیت برای همه شناسه اسناد، می توان از یک کدگذاری بایتهای متغیر استفاده کرد.

۴- ۱،۲۰ نمره

$$CosineSim(D1.D2) = \frac{D1 \cdot D2}{\|D1\|\|D2\|} = \frac{9}{3 * 4} = 0.75$$

$$\|D1\| = \sqrt{2^2 + 3^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2} = 4$$

$$\|D2\| = \sqrt{0^2 + 2^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2} = 3$$

$$D1 \cdot D2 = 2 * 0 + 3 * 2 + 1 * 1 + 1 * 1 + 1 * 1 + 0 * 1 + 0 * 1 = 9$$

۵- صفحه 485 پاراگراف اول ۱،۲۰ نمره

۱- هر واحدی که ما تصمیم داریم روی آن سیستم بازیابی اطلاعات بسازیم،چه نام دارد؟

۱. مدل ۲. عبارت ۳. اسناد ۴. مجموعه‌ای از نوشته‌ها

۲- توصیف زیر، در مورد نتایج بازگشتی سیستم، به کدام گزینه اشاره دارد؟
«چه کسری از نتایج بازگشتی، به نیاز اطلاعاتی مرتبط است؟»

۱. شاخص ۲. صحّت ۳. بازخوانی ۴. فرهنگ واژگان

۳- یک . . . . . نمونه‌ای از دنباله کاراکترها در اسناد است که باهم به عنوان یک واحد معنایی برای پردازش، در یک گروه قرار گرفته‌اند.

۱. نشانه ۲. نوع ۳. عبارت ۴. سند

۴- گاهی اوقات، کلمات بسیار عام که ظاهراً ارزش اندکی در کمک به انتخاب اسناد منطبق با نیاز کاربر دارند، باید از مجموعه واژگان مستثنی شوند. این کلمات . . . . . نامیده می‌شوند.

۱. کلمات زاید ۲. کلمات اصلی ۳. کلمات ناقص ۴. کلمات توقّف

۵- اگر تمامی عملیات ویرایشی (درج، حذف و جایگزینی) دارای وزن یکسان باشند، آنگاه فاصله ویرایشی cat و dog برابر خواهد بود با:

۱. 3 ۲. 2 ۳. 4 ۴. 1

۶- کدام الگوریتم‌ها، مجموعاً به عنوان الگوریتم‌های Soundex معروف هستند.

۱. الگوریتم‌های تشخیص صوت
۲. الگوریتم‌های تبدیل متن به صوت
۳. الگوریتم‌های درهم سازی آوائی
۴. الگوریتم‌های مقاوم‌سازی صوت

۷- بخشی از حافظه اصلی را که در بلوک خوانده شده یا نوشته شده در آن ذخیره می‌شود، . . . . . می‌نامیم.

۱. کَش ۲. پیگرد ۳. میانگیر ۴. گذرگاه

۸- پیچیدگی زمانی الگوریتم SPIMI (شاخص‌گذاری درون حافظه‌ای تک گذره) برابر است با:

۱. $\theta(T)$ ۲. $\theta(T^2)$ ۳. $\theta(\log T)$ ۴. $\theta(T \log T)$

۹- در شاخص‌گذاری توزیع‌شده به روش MapReduce، تجمیع تمامی مقادیر برای یک کلید معلوم در یک لیست، وظیفه‌ی . . . . . در مرحله کاهش است.

۱. تجزیه‌کننده ۲. توزیع‌کننده ۳. تجمیع‌کننده ۴. وارونه‌کننده

۱۰- غیرحسّاس کردن به حروف کوچک و بزرگ، ریشه‌گیری و حذف کلمات توقّف، صورتهایی از کدام نوع فشرده‌سازی هستند؟

۱. فشرده‌سازی با اتلاف
۲. فشرده‌سازی بدون اتلاف
۳. فشرده‌سازی چرخه‌ای
۴. فشرده‌سازی چندگذره

۱۱- در کدگذاری بایت متغیر، 7 بیت آخر یک بایت، ..... نامیده می‌شود و بخشی از فاصله را کدگذاری می‌کند.

۱. توازن
۲. بار
۳. ادامه
۴. نیبل

۱۲- یک کد با این ویژگی که برای توزیع دلخواه P، فاکتوری از کد بهینه باشد، ..... نامیده می‌شود.

۱. کد بهینه
۲. آنتروپی
۳. کد جهانی
۴. کد پیشوندی

۱۳- برای وزن‌دهی tf-idf، لیست‌های قهرمان، r سندی خواهند بود که بالاترین مقادیر ..... را برای عبارت t دارند.

۱. idf
۲. tf
۳. if
۴. tf + idf

۱۴- کدام یک از موارد زیر صحیح است؟

مورد اول: بازخورد ضمنی از بازخورد صریح اعتبار بیشتری دارد.

مورد دوم: در روش شبه‌بازخورد ربط، کاربر کارایی بازیابی بیشتری را بدون تعامل بیشتر بدست می‌آورد.

مورد سوم: جمع‌آوری بازخورد ضمنی در مقادیر بزرگ برای یک سیستم با حجم بالا، مانند جستجوی وب آسان است.

۱. فقط موارد اول و دوم
۲. فقط موارد دوم و سوم
۳. فقط موارد اول و سوم
۴. موارد اول و دوم و سوم

۱۵- توصیف زیر، کدام اصل را توضیح می‌دهد؟

«یک سیستم، همیشه باید مشخّص‌ترین قسمت از یک سند را در پاسخ به پرس و جو برگرداند.»

۱. اصل بازیابی سند ساخت‌یافته
۲. اصل بازیابی تشخیصی
۳. اصل بازیابی سند عام
۴. اصل بازیابی عامل سند

۱۶- روش گسترش پرس و جو از طریق تولید خودکار فرهنگ لغات جامع و روش بازخورد ربط غیرمستقیم (سراسری) به ترتیب جزو کدام یک از روش‌های حل مشکل هم‌معنائی محسوب می‌شوند؟

۱. سراسری ـ سراسری
۲. سراسری ـ محلّی
۳. محلّی ـ سراسری
۴. محلّی ـ محلّی

۱۷- کدام گزینه صحیح است؟

۱. XML داده - محور عمدتاً مقدار صفت داده عددی و غیر متنی را کدگذاری می‌کند.

۲. XML داده - محور همه انواع مقادیر داده ای شامل داده های عددی و غیرعددی را کدگذاری می‌کند.

۳. اداره محدودیت‌های مرتّب‌سازی و پیوندها در یک مدل بازیابی ساخت‌یافته داده - محور دشوار است.

۴. بسیاری از سیستم‌های بازیابی XML متن - محور توسعه‌هایی از پایگاه داده‌های رابطه‌ای هستند.

۱۸- کدام گزینه در مورد میانگین متوسط صحّت (MAP) صحیح نیست؟

۱. در میان معیارهای ارزیابی، MAP به طور خاص، پایایی و قابلیت جداسازی خوبی دارد.

۲. با استفاده از MAP، سطوح بازخوانی ثابت انتخاب می‌شوند.

۳. با استفاده از MAP، هیچ درونیابی وجود ندارد.

۴. MAP یک معیار تک شکلی از کیفیت را روی سطوح بازخوانی، فراهم می‌آورد.

۱۹- فرمول محاسبه معیار «دقّت» یعنی میزان دسته‌بندی‌های صحیح برابر است با:

توضیحات:

tp = true positive ، fp = false positive

fn = false negative ، tn = true negative

۱. $$\text{accuracy} = \frac{tp + fp}{tp + fp + fn + tn}$$

۲. $$\text{accuracy} = \frac{tp + fn}{tp + fp + fn + tn}$$

۳. $$\text{accuracy} = \frac{tp + fn + tn}{tp + fp + fn + tn}$$

۴. $$\text{accuracy} = \frac{tp + tn}{tp + fp + fn + tn}$$

۲۰- کدام یک از موارد زیر، از خصوصیات و ویژگیهای وب است؟

مورد اوّل: بی‌مانند در مقیاس

مورد دوّم: تقریباً بی‌مانند در نبود هماهنگی کامل در ایجاد

مورد سوم: بی‌مانند در تنوع پس‌زمینه‌ها

مورد چهارم: بی مانند در انگیزه اعضای آن

۱. فقط موارد اول و دوم و سوم

۲. فقط موارد اول و سوم و چهارم

۳. فقط موارد دوم و سوم و چهارم

۴. موارد اول و دوم و سوم و چهارم

۲۱- کدام گزینه، سه دسته عمده از صفحات وب را به خوبی و درستی بیان می‌کند؟

۱. IN ، OUT ، SCC
۲. IN ، PIN ، LINK
۳. OUT ، SCC ، LINK
۴. IN ، OUT ، PIN

۲۲- پروتکل حذف روبات‌ها به چه صورت پیاده سازی می شود؟

۱. با قراردادن فایلی با نام NoRobots.txt در ریشه سلسله مراتب URL در سایت
۲. با قراردادن فایلی با نام Robots.txt در ریشه سلسله مراتب URL در سایت
۳. با قراردادن فایلی با نام NoRobots.txt در نزدیکترین پوشه مربوط به آدرس URL در سایت
۴. با قراردادن فایلی با نام Robots.txt در نزدیکترین پوشه مربوط به آدرس URL در سایت

۲۳- در بسیاری از موتورهای جستجو، برای قرار گرفتن یک صفحه وب در شاخص موتور جستجو، هزینه دریافت می‌شود. این مدل با چه عنوانی شناخته می‌شود؟

۱. جاگذاری لینک
۲. پرداخت الکترونیکی
۳. جاگذاری پرداخت‌شده
۴. تبلیغات اینترنتی

۲۴- در مدل بازیابی اطلاعات مبتنی بر شبکه های بیزی، کدام گزینه صحیح است.

۱. شبکه مجموعه سند، نسبتاً کوچک است و می‌توان از پیش آن را محاسبه کرد.
۲. شبکه پرس و جو، نسبتاً کوچک است امّا هر زمان که یک پرس و جو وارد می‌شود، یک شبکه جدید باید ساخته شود.
۳. شبکه مجموعه سند، بزرگ است و هر زمان که یک پرس و جو وارد می‌شود، یک شبکه جدید باید ساخته شود.
۴. شبکه پرس و جو، نسبتاً بزرگ است ولی می‌توان از پیش آن را محاسبه کرد.

۲۵- کدام گزینه صحیح است.

۱. ترجمه یک URL به شکل متنی، به یک آدرس IP، تحلیل DNS نامیده می‌شود.
۲. سرویس نام دامنه (DNS) ماهیّت توزیع‌شده‌ای دارد.
۳. پیاده‌سازی‌های جستجوی نام دامنه در کتابخانه‌های استاندارد، معمولاً سنکرون هستند.
۴. همه موارد فوق، صحیح می‌باشند.

شماره سوال	پاسخ صحیح
1	ج
2	ب
3	الف
4	د
5	الف
6	ج
7	ج
8	الف
9	د
10	الف
11	ب
12	ج
13	ب
14	ب
15	الف
16	ب
17	الف
18	ب
19	د
20	د
21	الف
22	ب
23	ج
24	ب
25	د

۱- استاندارترین نوع بازیابی اطلاعات که در آن، یک سیستم قسط دارداسناد را از درون مجموعه ای که مرتبط با نیاز اطلاعاتی دالخواه هست، چیست؟

۱. بازیابی بولی ۲. بازیابی داده ۳. بازیابی موردی ۴. بازیابی دانش

۲-

	عبارت	فراوانی سند	اشاره گر به لیست پستها
	a	۶۵۶،۲۶۵	à
	aachen	۶۵	à
	…	…	…
	zulu	۲۲۱	à
: فضای مورد نیاز	۲۰ بایت	۴ بایت	۴ بایت

شکل: ذخیرهی لغت نامه به عنوان آرایهای از مدخل های با عرض ثابت.

ما در لغتنامه بصورت یک رشته ۲۰ بایت را برای خود عبارت، ۴ بایت برای فراوانی سند آن و ۴ بایت برای اشاره گر به لیست پستهای آن تخصیص میدهیم. عبارات در آرایه با جستجوی دودویی جستجو میشود. برای Reuters-RCV1 به $M\times(20+4+4)=400{,}000\times28=11.2$ مگا بایت فضا برای ذخیرهی لغتنامه نیاز داریم. میانگین طول هر عبارت در انگلیسی حدود ۸ کاراکتر است.

به طور متوسط، چند کاراکتر در طرح عرض ثابت هدر می شود؟

۱. ۱۲ ۲. ۸ ۳. ۴ ۴. ۲۰

۳- اولین مفهوم اساسی در بازیابی اطلاعات حائز اهمیت دارد، چیست؟

۱. ذخیره سازی ۲. اندیس گذاری ۳. شاخص گذاری ۴. شاخص وارونه

۴- نشانه گذاری، حذف عبارت متعارف: کلمات متعارف، نرمالسازی و ریشه گیری و مدخل گیری مراحل کدام مفهوم می باشد؟

۱. تعیین مجموعه واژگان عبارت ۲. تعیین مجموعه واژگان جمله

۳. تعیین مجموعه واژگان کلمه ۴. تعیین مجموعه واژگان پاراگراف

۵- هدف ریشه گیری و مدخل گیری هر دو کاهش صورت های صرفی و گاهی صورت های نحوی کلمه است تا به یک صورت پایه ای متعارف برسد. برای مثال:

Am، are ، is àbe

Car، cars، car's، cars'àcar

نتیجه نگاشت the boy's cars are different colors. چیست؟

۱. Boy cars be different color.
۲. The boy car be differ color.
۳. The boy car be different color.
۴. boy car be differ color.

۶- برای کلمه castle شاخص ۳-گرمی برای پرس و جوهای جایگزینی اعمال کنیم کدام گزینه درست می باشد؟

۱. $cas، ast، stl، tle$
۲. $ca،cas، ast، stl، tle، le$
۳. $ca،cas،tle،le$
۴. $cas،tle$

۷- در عملیاتی: ۱. یک کاراکتر را در رشته درج کنید. ۲. یک کاراکتر را از رشته حذف کنید. ۳. یک کاراکتر از یک رشته را با کاراکتر دیگر جایگزین کنید. اسم این عملیات چیست؟

۱. تصحیح املایی
۲. فاصله ویرایشی
۳. تصحیح املایی حساس به متن
۴. تصحیح آوایی Phonetic Correction

۸- پیچیدگی زمانی الگوریتم SPIMI چیست؟ (T یک حد بالا برای تعداد عبارتی است که ما باید مرتب کنیم)

۱. $\theta(T)$
۲. $\theta(T^2)$
۳. $\theta(T\,Log^T)$
۴. $\theta(Log^T)$

۹- برای لغت نامه خصیصه چه نوع ساختمان داده ای استفاده می شود؟

۱. Hash Tree
۲. Binary Tree
۳. درخت سیاه و قرمز
۴. B-Tree

۱۰- پرس و جوی Shakespeare را در مجموعهی که در آن، هر سند سه ناحیه دارد، در نظر بگیرید: title authoc، و body. تابع نمره گذاری بولی برای یک ناحیه مقدار ۱ را اتخاذ می کند اگر عبارت پرس و جوی Shakespeare در آن ناحیه حضور داشته باشد، و در غیر این صورت ۰ خواهد بود. نمره گذاری وزندار ناحیهای د رچندین مجموعهای نیازمند سه وزن g1،g2 و g3 است که به ترتیب متناظر با ناحیههای title، author و body هستند. فرض کنید g1=0.2، g2=0.3 و g3=0.5 را در نظر بگیریم (بطوریکه مجموع سه وزن برابر با ۱ شود)، که این متناظر است با کاربردی که در انطباق در ناحیهی author نسبت به نمرهی کلی کمترین اهمیت، ناحیه title اهمیت بیشتر و ناحیه body بیشترین اهمیت را دارد. بنابراین اگر عبارت Shakespeare در ناحیههای body و title وجود داشته باشد، اما در ناحیهی author سند وجود نداشته باشد، نمره این سند چند خواهد بود؟

۱. ۰،۵
۲. ۰،۸
۳. ۰،۳
۴. ۰،۲

۱۱- پرس و جوی best car insurance را روی مجموعهی فرضی با N=1،000،000 سند در نظر می گیریم که فراوانیهای سند برای car، best، auto و insurance به ترتیب ۱۰۰۰۰، ۵۰۰۰۰، ۵۰۰۰ و ۱۰۰۰ است.

عبارت	پرس و جو				سند				ضرب
	tf	df	idf	Wt،q	tf	wf	Wt،d		
auto	۰	۵۰۰۰	۲،۳	۰	۱	۱	۰،۴۱		۰
best	۱	۵۰۰۰۰	۱،۳	۱،۳	۰	۰	۰		۰
car	۱	۱۰۰۰۰	۲،۰	۲،۰	۱	۱	۰،۴۱		۰،۸۲
insurance	۱	۱۰۰۰	۳،۰	۳،۰	۲	۲	۰،۸۲		۲،۴۶

برای اسناد، وزن دهی tf را بدون استفاده از idf و با استفاده از نرمالسازی اقلیدسی به کار می بریم. اولین مورد، تحت ستونی نمایش داده شده است که به عنوان wf دارد، در حالیکه مورد دوم تحت ستون Wt،d است. نمره وزن دهی tf-idf را بدست بیاورید؟

۱. ۰،۸۲ ۲. ۳،۲۸ ۳. ۱،۲۳ ۴. ۰،۴۱

۱۲- هزینه اجرای الگوریتم مبنا برای محاسبهی نمره های فضای بردار چقدر می باشد.

```
CONSINSCORE(q)
1  float Score[N] = 0
2  Initialize Length[N]
3  For each query term t
4  Do calculate wt،q and fetch postings list for t
   a. For each pair (d، tft،d) in posting list for t
   b. Do Scores[d] += wft،d × wt،q
5  Read the array Length[d]
6  For each d
7  Do Scores[d] = Scores[d]/length[d]
8  []Return Top K components of Scores
```

۱. O(N) ۲. O(N2) ۳. O(NK) ۴. O(logN)

۱۳- مجموعه r سند را لیست قهرمان عبارت t می نامند، مقدار وزن دهی لیست های قهرمان، r سندی که بالاترین مقادیر tf را برای عبارت t دارند، چیست؟

۱. Tf-idf ۲. Idf ۳. Tf ۴. Tf-idft, d

۱۴- در گام پیش پردازش هرس خوشه ای چند سند را بطور تصادفی از مجموعه انتخاب شده، آنها را رهبر (Leader) می نامیم؟

۱. $N + \sqrt{N}$ ۲. $\frac{N}{\sqrt{N}}$ ۳. $\sqrt{N}$ ۴. N

۱۵- مقدار آماری Kappa را برای مثال زیر بدست بیاورید؟

		ربط داور ۲		
		بله	خیر	کل
داور۱	بله	۳۰۰	۲۰	۳۲۰
ربط	خیر	۱۰۰	۷۰	۸۰
	کل	۳۱۰	۹۰	۴۰۰

میزانی که داوران توافق دارند

P(A)=(300+70)/400= 370/400=0.925

حاشیه های ادغام شده

P(nonrelevant)=(80+90)/(400+400)= 170/800=0.2125

P(relevant)=(320+310)/(400+400)=630/800=0.7878

احتمال آنکه دو داور به طور تصادفی توافق کنند

P(E)=P(nonrelevant)2 + P(relevant)2 = 0.21252 +0.78782 = 0.665

۱. ۰،۱۵ ۲. ۰،۷۷۶ ۳. ۱،۱۹ ۴. ۰،۴۵

۱۶- تابع تخمین اندازه مجموعه واژگان در قانون Heaps، مقدار M را با پارامترهای b=0.49، k=44، T = 1000020 تخمین می شود؟

۱. ۲۱،۳۱۲ ۲. ۴۴،۸۸۰ ۳. ۳۸،۳۲۳ ۴. ۳۸،۳۶۵

۱۷- الگوریتم Rocclio چیست؟

۱. برای جداسازی اسناد مرتبط و غیر مرتبط

۲. برای تفاوت برداری میان مراکز اسناد مرتبط و غیر مرتبط

۳. پرس وجوی کاربر و دانش نسبی از اسناد مرتبط و غیر مرتبط

۴. برای پیاده سازی بازخورد ربط

۱۸- در سند XML بین چه برچسبی محصور می شود؟

۱. <title>، </title>

۲. <act>، </act>

۳. <play>، </play>

۴. <scene>، </scene>

۱۹- پرس و جوی XML در فرمت NEXI در دستور زیر، نمایش جزئی کدام درخت می باشد.

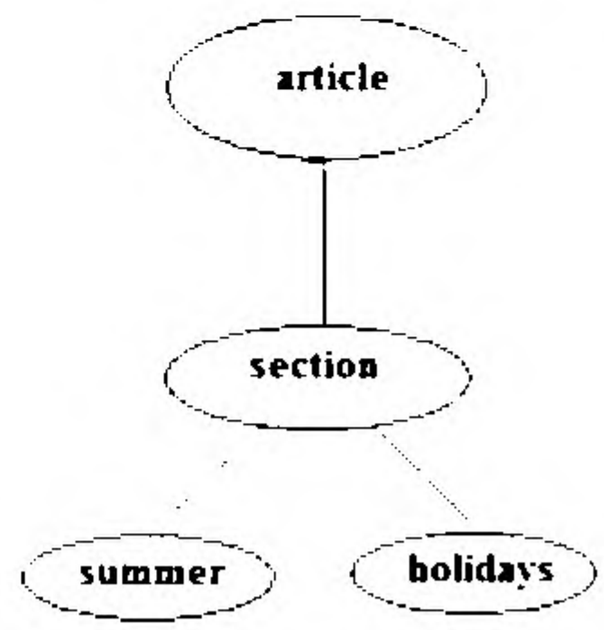
article//

[yr = 2001 or .//yr = 2002//.]

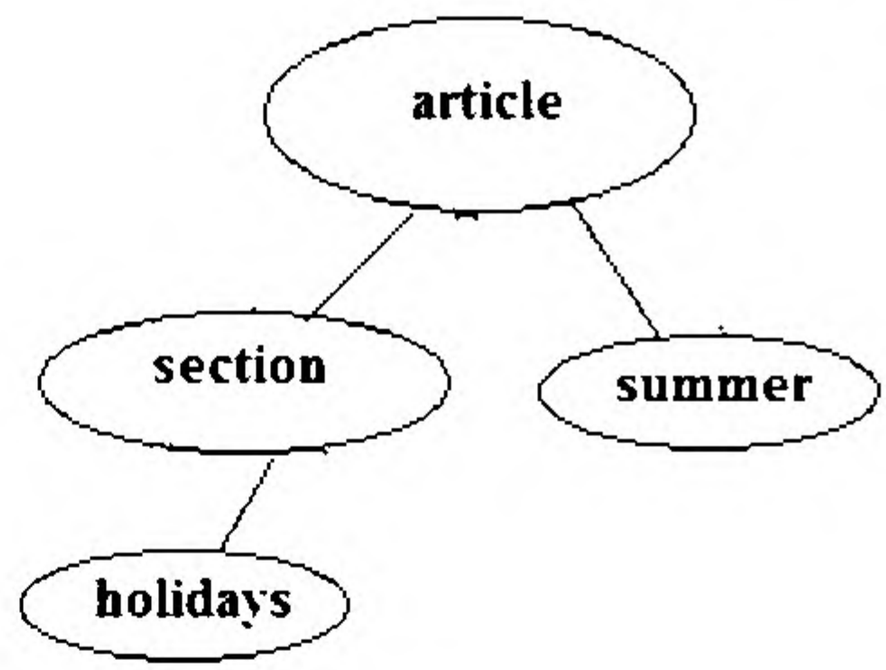
section//

[about(.. summer holidays)]

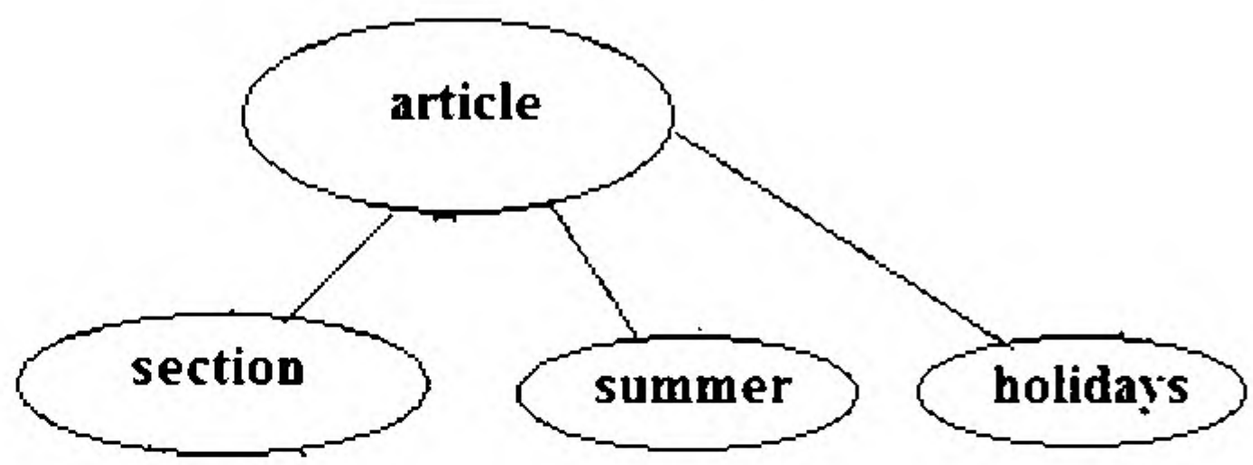
۱.



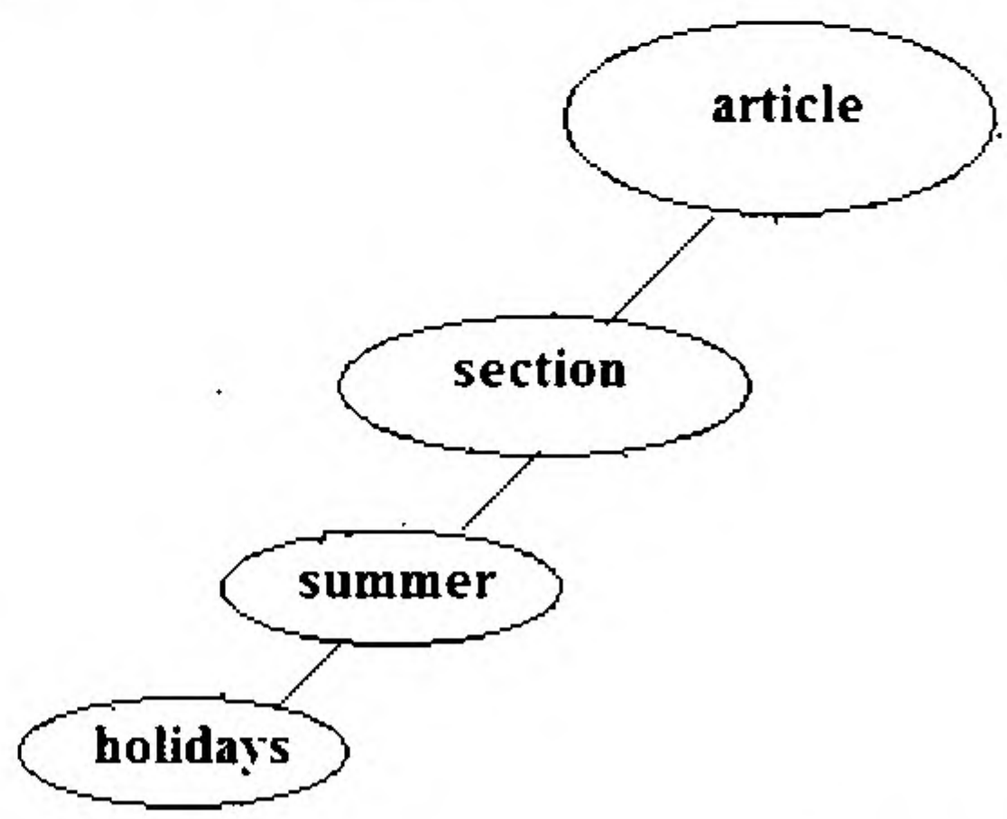
۲.



۳.



۴.



۲۰- اساس روش بازیابی XML متن-محور برای بازیابی سند XML چیست؟

۱. فیلدهای طولانی متن، تطبیق نادقیق و نتایج ربط رتبه بندی

۲. فیلدهای طولانی متن، تطبیق دقیق و نتایج ربط رتبه بندی

۳. فیلدهای کوتاه متن، تطبیق نادقیق و نتایج ربط رتبه بندی

۴. فیلدهای کوتاه متن، تطبیق دقیق و نتایج ربط رتبه بندی

۲۱- مدلهای زبانی برای بازیابی اطلاعات را نام ببرید؟

۱. آتاماتای متناهی و مدلهای زبانی

۲. توزیع چندجمله ای روی کلمات

۳. مدل درستنمایی پرس و جو

۴. برآورد احتمال تولید پرس و جو

۲۲- اصل رتبه بندی احتمالاتی بهینه است، در صورتی که اتلاف مورد انتظار را تحت اتلاف ۰/۱ کمینه کند. تعریف کدام گزینه می باشد؟

۱. استقلال دودویی

۲. Naïve Base

۳. ریسک بیزی Bayes Risk

۴. مورد اتلاف ۰/۱

۲۳- طرح وزن دهی BM25 که اغلب به نام وزن دهی چی شناخته می شود؟

۱. Okapi

۲. Idf

۳. RSVd

۴. NVRt

۲۴- فرض کنید مجموعه سند شامل دو سند باشد: d1: Xyzzy سود را گزارش می کند اما بازده پایین است. d2: Quorus اتلاف را به اندازه ۴/۱ کاهش می دهد، اما بازده بیشتر کاهش می یابد.
این، مدل، مدلهای یک-گرمی برآورد درستنمایی بیشینه اسناد و مجموعه است که با $\lambda=1/2$ ترکیب شده است. فرض کنید پرس و جو "بازده پایین" باشد آنگاه مقدار $p(q|d1)$ و $p(q|d2)$ به ترتیب از چپ به راست چیست؟

۱. $p(q|d1)$ = ۰٫۰۱۱۷ و $p(q|d2)$ = ۰٫۰۰۳۹

۲. $p(q|d1)$ = ۰٫۰۲۳۴ و $p(q|d2)$ = ۰٫۰۰۷۸

۳. $p(q|d1)$ = ۰٫۰۴۶۸ و $p(q|d2)$ = ۰٫۰۰۱۵۶

۴. $p(q|d1)$ = ۰٫۰۹۳۷ و $p(q|d2)$ = ۰٫۰۰۳۱۲

۲۵- برای یافتن احتمال یک دنباله از کلمات، احتمالاتی که مدل به هر یک از کلمات در دنباله می دهد را در احتمال ادامه یا توقف بعد از تولید هر کلمه ضرب می کنیم. برای مثال زیر توزیع احتمال چقدر می باشد؟

P(frog said that toad likes frog)

۱. ۰٫۰۰۰۰۰۰۰۰۰۰۰۳۱۴۶

۲. ۰٫۰۰۰۰۰۰۰۰۰۰۰۶۲۹۲

۳. ۰٫۰۰۰۰۰۰۰۰۰۰۰۱۲۵۸۴

۴. ۰٫۰۰۰۰۰۰۰۰۰۰۰۱۵۷۳

سوالات تشریحی

۱- دو قاعده اصلی در الگوریتم های تصحیح املایی را با مثال نام ببرید؟ ۳/۳۳ نم

۲- در قانون Heaps (تخمین تعداد عبارت) اندازه مجموعه واژگان با مقادیر پارامترهای b=0.49 و k=44 و برای ۱۰۰۰۰۲۰ نشانه را بدست بیاورید؟ ۰/۶۷ نمر

۳- در علوم اجتماعی، یک معیار رایج برای توافق بین داوران آماره (Kappa Statistic) Kappa است. این معیار برای داوریهای قیاسی طراحی شده است و نرخ توافق ساده را برای نرخ توافق تصادفی تصحیح می کند. برای جدول زیر آماره Kappa را بدست آورید. ۰/۶۷ نمر

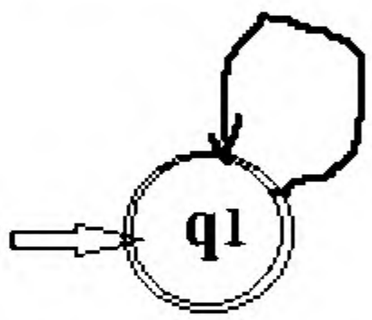
		ربط داور ۲		
		بله	خیر	کل
ربط داور ۱	بله	۳۰۰	۲۰	۳۲۰
	خیر	۱۰۰	۷۰	۸۰
	کل	۳۱۰	۹۰	۴۰۰

۴- فرض ساخت یک مدل زبانی از متن آموزشی زیر را در نظر بگیرید: ۰/۶۷ نمر

The martian has landed on the latin pop sensation ricky martin

- تحت مدل احتمالی یک-گرمی برآورد درستنمایی بیشنه، مقدار احتمال P(the) و P(martian) چقدر است؟
- تحت مدل دو-گرمی برآورد درستنمایی بیشینه، مقادیر احتمال P(sensation | pop) و P(pop | the) چقدر است؟

۵- برای یافتن احتمال یک دنباله از کلمات، احتمالاتی که مدل به هر یک از کلمات در دنباله می دهد را در احتمال ادامه یا توقف بعد از تولید هر کلمه ضرب می کنیم. احتمال یک سند/رشته ویژه را پیدا کنید؟ ۰/۶۶ نمر

 ماشین خودکار متناهی تک حالتی که به عنوان مدل زبان یک-گرمی عمل می کند.	The	۰،۲
	A	۰،۱
	Frog	۰،۰۱
	Toad	۰،۰۱
	Said	۰،۰۳
	Links	۰،۰۲
	That	۰،۰۴
P(stop|q1) = 0.2	...	...

شماره سؤال	پاسخ صحیح
1	3
2	1
3	4
4	1
5	2
6	2
7	2
8	1
9	4
10	2
11	2
12	4
13	1
14	3
15	2
16	3
17	4
18	4
19	1
20	1
21	1
22	3
23	1
24	1
25	4