سؤال 1
درست
نمره 1.00 از 1.00
علامت زدن سؤال

در صورتی که اندازه شاخص معکوس غیرمکانی یک مجموعه سند انگلیسی یک گیگابایت باشد، حد بالای تقریبی اندازه خود مجموعه داده بر حسب گیگابایت چقدر است؟

a. 4
b. 8
c. 12 ✔
d. 16

پاسخ درست «12» است.

سؤال 2
درست
نمره 1.00 از 1.00
علامت زدن سؤال

کدام یک از موارد زیر در مورد بهینه سازی پرسمان (query optimization) درست نیست؟

a. پرسمان را بر اساس اندازه بخش های مختلف آن بازآرائی میکند.
b. می تواند منجر به افزایش سرعت شود
c. می تواند منجر به افزایش دقت شود. ✔
d. تغییری در خود الگوریتم جستجو ایجاد نمی کند.

پاسخ درست «می تواند منجر به افزایش دقت شود.» است.

سؤال 3
درست
نمره 1.00 از 1.00
علامت زدن سؤال

تبدیل درختان به درختها در پردازش متن ورودی جزء کدام یک از پیش پردازش های زیر است؟

a. حذف stop wordها
b. ریشه یابی
c. نرمالسازی ✔
d. توکن سازی

پاسخ درست «نرمالسازی» است.

سؤال 4
نادرست
نمره -0.33 از 1.00
علامت زدن سؤال

کدام یک از موارد زیر از ویژگی های شاخص دو کلمه ای (biword index) برای پاسخگویی به پرسمان های عبارتی نیست؟

a. اندازه شاخص آن در بدترین حالت می تواند تا مرتبه T*T افزایش پیدا کند.
b. در بازیابی عبارات دو کلمه ای سرعت بالایی دارد.
c. در بازیابی عبارات دو کلمه ای هیچ خطای مثبت اشتباهی ندارد. ✖
d. در صورت استفاده از شاخص مکانی دیگر از این روش استفاده نمی شود.

پاسخ درست «
در صورت استفاده از شاخص مکانی دیگر از این روش استفاده نمی شود.» است.

سؤال 5
درست
نمره 1.00 از 1.00
علامت زدن سؤال

چرا تصحیح خطاهای تایپی در حالتی که در شاخص سازی پویا از چند شاخص استفاده می شود، مشکل خواهد بود؟

a. وقوع خطا در شاخص های جانبی بیشتر از شاخص اصلی است.
b. جمع آوری آماره های سطح مجموعه اسناد در این حالت دشوارتر است. ✔
c. نرخ وقوع خطا در میان کلمات مختلف متفاوت است.
d. به دلیل استفاده از چند شاخص ناگزیر به استفاده از بیت های invalidation هستیم.

پاسخ درست «
جمع آوری آماره های سطح مجموعه اسناد در این حالت دشوارتر است.» است.

سؤال 6
درست
نمره 1.00 از 1.00
علامت زدن سؤال

در صورتی که اندازه تعداد کلمات در ادغام لگاریتمی برابر با T=11n باشد، کدام یک از شاخص های زیر در حافظه جانبی در اتمام کار ساخت شاخص خالی خواهد بود؟

a. I0
b. I1
c. I2 ✔
d. I3

پاسخ درست «
I2» است.

سؤال 7
درست
نمره 1.00 از 1.00
علامت زدن سؤال

حداقل تعداد دفعاتی که یک posting ممکن است در روش ادغام لگاریتمی در عملیات ادغام شرکت کند چند بار است؟

a. صفر ✔
b. یک
c. log T
d. log T/n

پاسخ درست «
صفر» است.

سؤال 8
درست
نمره 1.00 از 1.00
علامت زدن سؤال

اگر در ذخیره سازی postings Listها برای هر postings list یک فایل جدا در نظر گرفته شود کدام یک از کارهای زیر با دشواری روبرو می شود؟

a. اضافه کردن یک سند جدید به یک postings list
b. ادغام postings listهای دو کلمه
c. مدیریت فایل ها ✔
d. ادغام شاخص های معکوس اصلی و موقتی

پاسخ درست «
مدیریت فایل ها» است.

سؤال 9
درست
نمره 1.00 از 1.00
علامت زدن سؤال

زمانبرترین عمل در در دسترسی به حافظه جانبی کدام یک از اعمال زیر است؟

a. جستجوی بلاک ✔
b. خواندن بلاک
c. نوشتن بلاک
d. انتقال بلاک

پاسخ درست «
جستجوی بلاک» است.

کوییز ۲

سوال 1
درست
نمره 1.00 از 1.00
علامت زدن سوال

در بحث فشرده سازی دیکشنری، در روش بلوکی (Blocking) ساخت دیکشنری هر چه اندازه بلوک ها (k) بیشتر شود، کدام موارد درست است؟

- a. زمان جستجوی کلمات کندتر می شود. ✔
- b. اندازه کل دیکشنری کمتر میشود. ✔
- c. طول آرایه مورد نیاز برای ذخیره سازی کلمات کمتر می شود.
- d. اندازه (طول) اشاره گرهای مورد نیاز افزایش می یابد.

سوال 2
درست
نمره 1.00 از 1.00
علامت زدن سوال

دیکشنری با 200000 کلمه انگلیسی داریم. حداقل طول کلمات 1، حداکثر آنها 50 و متوسط طول کلمات 10 کاراکتر است. اندازه حافظه مورد نیاز برای ذخیره سازی کلمات دیکشنری و اشاره گرهای آنها (بدون در نظر گرفتن حافظه لازم برای ذخیره سازی فرکانس کلمات و اشاره گرهای به لیست های پست ها) در روش ذخیره‌سازی دیکشنری به عنوان یک رشته (Dictionary-as-a-String) را بر حسب مگابایت کدام است؟

- a. 2.6 ✔
- b. 2
- c. 2.4
- d. 2.2

سوال 3
درست
نمره 1.00 از 1.00
علامت زدن سوال

کدام یک از موارد زیر از مزایای فشرده سازی دیکشنری است؟

- a. ایجاد مزیت رقابتی نسبت به محصولات مشابه ✔
- b. جا شدن بخشی از لیست‌های پست‌ها (postings lists) در حافظه اصلی. ✔
- c. جا شدن دیکشنری در حافظه اصلی ✔
- d. افزایش سرعت بالا آمدن سیستم ✔
- e. افزایش سرعت جستجوی کلمات در دیکشنری

سوال 4
درست
نمره 1.00 از 1.00
علامت زدن سوال

در عمل از قانون هیپس (Heaps' law) برای چه کاری می توان استفاده کرد؟

- a. ایجاد یک رابطه خطی میان تعداد کلمات دیکشنری و تعداد اسناد
- b. تخمین اندازه دیکشنری برای مجموعه اسناد بزرگ ✔
- c. بررسی و استفاده از فرکانس تکرار کلمات مختلف در مجموعه اسناد

بررسی و استفاده از فرکانس تکرار کلمات مختلف در مجموعه اسناد

- d. کاهش اندازه دیکشنری

سؤال 5
درست
نمره 1.00 از 1.00
علامت زدن سؤال

کدام یک از موارد زیر در مورد مدل زبانی تک کلمه ای (unigram) درست است؟

a. با شرط استخراج دیکشنری از اسناد، نیاز به هموارسازی ندارد. ✔

b. نسبت به مدل زبانی دو کلمه ای (bigram) تعداد پارامترهای کمتری دارد. ✔
تعداد پارامترهای کمتری دارد.

c. احتمال وقوع هر جمله برابر با حاصلضرب احتمالات کلمات آن جمله است. ✔

d. کمکی به در نظر گرفتن زمینه (context) در اصلاح خطا نمی کند. ✔

سؤال 6
درست
نمره 1.00 از 1.00
علامت زدن سؤال

در مدل ساده بیز برای تصمیم گیری برای محاسبه امتیاز کلمه درست X برای کلمه اشتباه Y در عمل کدام احتمال را می توان **محاسبه نکرد؟**

a. احتمال وقوع کلمه X

b. احتمال X به شرط Y

c. احتمال Y به شرط X

d. احتمال وقوع کلمه Y ✔

سؤال 7
درست
نمره 1.00 از 1.00
علامت زدن سؤال

در یک مجموعه سند فرضی به زبان انگلیسی، حرف A 200 بار در کلمات مختلف ظاهر شده است که در 10 مورد به اشتباه S نوشته شده است. در صورت استفاده از هموارسازی با اضافه کردن یک (Add-1 Smoothing) احتمال وقوع S به شرط A کدام است؟

a. 0.04867 ✔

b. 0.05500

c. 0.04889

d. 0.04444

سؤال 8
درست
نمره 1.00 از 1.00
علامت زدن سؤال

در تصحیح خطای وابسته به متن اگر X جمله مشاهده شده و W جمله صحیح مد نظر کاربر باشد

a. احتمال وقوع X به شرط W توسط مدل کانال مشخص می شود. ✔

b. احتمال وقوع W توسط مدل زبانی مشخص می شود. ✔

c. احتمال وقوع X توسط مدل زبانی مشخص می شود.

d. احتمال وقوع W توسط توسط مدل کانال مشخص می شود.

سؤال 9
درست
نمره 1.00 از 1.00
علامت زدن سؤال

در اصلاح خطا با کمک مدل کانال نویزی، در صورتی که X کلمه مد نظر کاربر و کلمه Y کلمه تایپ شده دارای اشتباه باشد، کدام یک از احتمال های زیر بیانگر مدل کانال است؟

a. احتمال X به شرط Y

b. آحتمال وقوع کلمه Y

c. احتمال وقوع کلمه X

d. احتمال Y به شرط X ✔

سؤال 10
درست
نمره 1.00 از 1.00
علامت زدن سؤال

در روش شاخص جایگشتی (permuterm index) برای یافتن کلماتی که حاوی عبارت ba (در هر جایی از کلمه) باشند، چه عبارتی را می بایست جستجو کرد؟

a. ba* ✔

b. $ba*

c. ba$*

d. $*ba*

سؤال 11
درست
نمره 1.00 از 1.00
علامت زدن سؤال

منظور از فاصله ویرایشی لونشتاین دو کلمه چیست؟

a. حداکثر احتمال تبدیل یک کلمه به کلمه دیگر

b. حداقل تعداد ویرایش های مورد نیاز برای تبدیل یک کلمه به کلمه دیگر ✔

c. حداقل احتمال تبدیل یک کلمه به کلمه دیگر

d. حداکثر تعداد ویرایش‌های مورد نیاز برای تبدیل یک کلمه به کلمه دیگر

سؤال 12
پاسخ نیمه درست
نمره 0.50 از 1.00
علامت زدن سؤال

کدام یک از موارد زیر در مورد کد گاما درست است؟

a. به دلیل نیاز به عملیات در سطح بیت در عمل زیاد از آن استفاده نمی شود. ✔

b. طول کد تولیدی آن برای هر عدد در بدترین حالت دو برابر طول کد بهینه است.

c. در حالت توزیع یکنواخت اعداد بهینه است.

d. کوتاهترین کد گاما متعلق به عدد صفر است.

سؤال 13
درست
نمره 1.00 از 1.00
علامت زدن سؤال

دنباله باینری زیر حاصل کدگذاری لیست پست‌های یک کلمه با استفاده از روش بایت های متغیر (VB) است. کدام **شناسه سند (DocID)** جزء شناسه های موجود در لیست پست‌های این کلمه است؟
برای سادگی خواندن، بایت های رشته نهایی به ترتیب از بالا به پایین هر کدام در یک خط قرار داده شده اند. خط اول اولین بایت رشته است.

00000000
10001000
10000001
10000101

a. 0 (صفر)

b. 1030 ✔

c. 257

d. 1026

کوییز ۳

سوال 1
درست
نمره 1.00 از 1.00
علامت زدن سوال

سند A و کوئری Q موجود است. در صورتی که محتوای سند A را عینا در آن تکرار کنیم (مثلا اگر سند A قبلا "موفق باشید." بوده، حالا "موفق باشید. موفق باشید." می شود)، کدام یک از موارد زیر درست است؟

a. فرکانس کلمات تغییری نمی کند.
b. فاصله اقلیدسی بردارهای A و Q ثابت باقی می‌ماند.
c. شباهت جکارد A و Q دقیقا یکسان باقی می‌ماند. ✔
d. شباهت کسینوسی A و Q تقریبا بدون تغییر باقی می ماند. ✔

The correct answers are:
شباهت کسینوسی A و Q تقریبا بدون تغییر باقی می ماند.,
شباهت جکارد A و Q دقیقا یکسان باقی می‌ماند.

سوال 2
درست
نمره 1.00 از 1.00
علامت زدن سوال

چرا فاصله اقلیدسی در مقایسه با شباهت کسینوسی معیار مناسبی برای اندازه‌گیری میزان شباهت بین دو سند نیست؟

a. وابستگی آن به فرکانس کلمه در سند
b. وابستگی آن به تنوع کلمات موجود در سند
c. وابستگی آن به طول سند ✔
d. وابستگی آن به عکس فرکانس سند (idf)

پاسخ درست «
وابستگی آن به طول سند» است.

سوال 3
درست
نمره 1.00 از 1.00
علامت زدن سوال

در صورتی که کوئری q تنها از یک کلمه w تشکیل شده باشد و اسناد بر اساس شباهت کسینوسی خود با کوئری مرتب شوند، رتبه بندی حاصل برای اسناد مساوی با کدام یک از حالات زیر **نیست**؟ (اسنادی که حاوی کلمه w نیستند در نتایج در نظر گرفته نمی شوند)

a. حالتی که اسناد بر اساس log(tf+1) کلمه مرتب شوند.
b. حالتی که اسناد بر اساس تعداد تکرار کلمه w در هر سند مرتب شوند.
c. حالتی که اسناد بر اساس tf-idf مرتب شده باشند.
d. حالتی که اسناد بر اساس idf کلمه w مرتب شوند. ✔

پاسخ درست «
حالتی که اسناد بر اساس idf کلمه w مرتب شوند.» است.

سوال 4
درست
نمره 1.00 از 1.00
علامت زدن سوال

در محاسبه شباهت کسینوسی بردار کوئری q با بردار سند d اگر بردار d نرمال شده باشد (دارای طول یک باشد) اما بردار q نرمال نشده باشد، و نتیجه جستجو با مرتب سازی مجموعه اسناد بر اساس مقدار q.d (حاصلضرب داخلی بردارهای سند و کوئری) به دست آید، کدام مورد درست است؟

a. نتایج جستجو با حالتی که بردار q نرمال شود تفاوتی ندارد. ✔
b. عدد حاصل از q.d در این حالت برابر با شباهت کسینوسی است.
c. فرکانس کلمات در این حالت در نظر گرفته نمی شود.
d. نتیجه جستجو غلط خواهد بود.

پاسخ درست «
نتایج جستجو با حالتی که بردار q نرمال شود تفاوتی ندارد.» است.

سوال 5
نادرست
نمره -0.33 از 1.00
علامت زدن سوال

کدام یک از معیارهای ارزیابی زیر برای ارزیابی یک موتور جستجوی وب مناسب تر است؟

a. MRR ✘
b. NDCG
c. MAP
d. Precision@k

پاسخ درست «
NDCG» است.

سوال 6

در دو جستجوی مستقل انجام شده توسط یک کاربر، جواب مورد نظر برای اولین بار در اولین و دومین نتیجه ارائه شده توسط موتور جستجو وجود داشته است. مقدار MRR این موتور جستجو چقدر است؟

a. 0.75

b. 1 ✖

c. 1.5

d. 0.5

علامت زدن سوال

پاسخ درست «0.75» است.

سوال 7

درست

نمره 1.00 از 1.00

علامت زدن سوال

در بازیابی سه سند با میزان ارتباط صفر، یک و دو با کوئری کاربر توسط یک موتور جستجوی اسناد ابتدا سند با ارتباط 1، سپس سند با ارتباط 2 و در آخر سند با ارتباط صفر برگردانده شده اند. مقدار NDCG (بهره تجمعی کاهشی نرمال شده) چقدر است؟ به جای لگاریتم رتبه (log i) در بخش کاهشی، از خود رتبه (i) استفاده کنید.

Answer: 0.8 ✔

پاسخ درست: 0.8

سوال 8

درست

نمره 1.00 از 1.00

علامت زدن سوال

منظور از Presentation bias در روش رتبه‌بندی جایگذاری (interleaved ranking) چیست؟

a. بایاس موتورهای جستجوی به واسطه نحوه بازنمائی اسناد

b. بایاس کردن ذهن کاربر به دلیل نحوه نمایش نتایج ✔

c. بایاس طراح واسط گرافیکی سامانه

d. بایاس الگوریتم های جستجوی استفاده شده

پاسخ درست «بایاس کردن ذهن کاربر به دلیل نحوه نمایش نتایج» است.

سوال 9

درست

نمره 1.00 از 1.00

علامت زدن سوال

کدام یک از موارد زیر د رمورد آزمون A/B درست است؟

a. در زمان آزمون، سامانه قدیمی نیز در حال اجرا است. ✔

b. با بخش کوچکی از کاربران انجام می شود. ✔

c. نتایج موتور جدید به صورت کامل یا در ترکیب با مدل قبلی به کاربر ارائه می شود. ✔

d. برای ارزیابی تغییرات اعمال شده در موتور جستجو استفاده می شود. ✔

کوییز ۴

سؤال 1
درست
نمره 1.00 از 1.00
علامت زدن سؤال

جستجویی با یک کوئری چهار کلمه ای متشکل از کلمات w1، w2، w3 و w4 در حال انجام است. از روش WAND برای محاسبه کارای امتیاز اسناد استفاده می شود.
برای هر یک از چهار کلمه به ترتیب، محل قرارگیری نشانگر آن کلمه، حد بالای امتیاز (UB) آن کلمه و همچنین حد آستانه فعلی مورد استفاده در الگوریتم بیان شده است. کدام یک از اسناد زیر ممکن است در ادامه مورد ارزیابی قرار گیرد و شباهت آن با کوئری ورودی محاسبه خواهد شد؟ گزینه های پاسخ شماره اسناد هستند.

W1, Pointer=230 , UB=2.3
W1, Pointer=263 , UB=1.1
W1, Pointer=275 , UB=2.0
W1, Pointer=302 , UB=2.3
Threshold= 5.1

a. 330 ✔
b. 245
c. 270
d. 280 ✔
e. هیچ سند دیگری ارزیابی نخواهد شد

سؤال 2
درست
نمره 1.00 از 1.00
علامت زدن سؤال

کدام یک از ترتیب های زیر برای لیست پست ها (postings list) یک ترتیب مشترک (common order) محسوب می شود؟

a. امتیاز کل (net-score=g(d)+cos(w,d))
b. تکرار کلمه در سند
c. عکس فرکانس سند (idf)
d. امتیاز کیفیت استاتیک g(d) ✔
e. شماره سند (doc-id) ✔

سؤال 3
درست
نمره 1.00 از 1.00
علامت زدن سؤال

کدام یک از روش های بهبود سرعت بازیابی زیر یک روش امن (safe) است؟

a. روش لیست قهرمانان (champion list)
b. روش WAND ✔
c. استفاده از اسنادی که حداقل چند کلمه از کلمات ورودی را داشته باشتند
d. استفاده از اسناد با idf بالا
e. استفاده از هیپ برای انتخاب به جای مرتب سازی ✔

سؤال 4
درست
نمره 1.00 از 1.00
علامت زدن سؤال

در روش خوشه بندی اگر پارامترهای b1 و b2 را زیاد کنیم

a. دقت بازیابی بیشتر می شود. ✔

b. تعداد خوشه ها بیشتر می شود.

c. سرعت بازیابی کمتر می شود. ✔

d. سرعت بازیابی بیشتر می شود.

سؤال 5
درست
نمره 1.00 از 1.00
علامت زدن سؤال

در روش دسته بندی k نزدیک ترین همسایه (KNN) هر چه اندازه k بیشتر شود.

a. مرز بین کلاس ها هموارتر می شود. ✔

b. واریانس مدل کاهش می یابد. ✔

c. بایاس بیشتر می شود. ✔

d. حساسیت به نویز کمتر می شود. ✔

سؤال 6
درست
نمره 1.00 از 1.00
علامت زدن سؤال

علت مشکل روش دسته بندی Rocchio در کلاس های چندریختی (polymorphic) چیست؟

a. دور بودن میانگین داده های یک کلاس از داده های آن کلاس ✔

b. استفاده از فاصله کسینوسی برا اندازه گیری میزان شباهت

c. استفاده از ضرائب منفی در محاسبه وزن کلمات

d. مبتنی بر نمونه بودن الگوریتم.

سؤال 7
درست
نمره 1.00 از 1.00
علامت زدن سؤال

چرا به الگوریتم دسته بندی نزدیک ترین همسایه الگوریتم تنبل گفته می شود؟

a. مبتنی بر نمونه بودن الگوریتم

b. ساده بودن زیاد فاز یادگیری الگوریتم ✔

c. استفاده زیاد از حافظه

d. در نظر گرفتن فرض پیوستگی (contiguity) فضای برداری

سؤال 8
درست
نمره 1.00 از 1.00
علامت زدن سؤال

برای دسته بندی یک مجموعه از اسناد از یک مدل ساده بیز استفاده می کنیم. در صورتی که تعداد ویژگی های مورد استفاده m، تعداد دسته ها (کلاس ها) c، و تعداد کلمات موجود در فرهنگ کلمات v باشد، تعداد پارامترهای مدل بیز که می بایست محاسبه شوند حدودا چند تا است؟

a. m*c

b. v*(c+1)

c. v*c

d. c*(m+1) ✔

سؤال 9
درست
نمره 1.00 از 1.00
علامت زدن سؤال

در روش بیز برای محاسبه احتمال تعلق داده ورودی x به کلاس c از چه احتمالی استفاده **نمی شود؟**

a. احتمال وقوع داده x به شرط کلاس c

b. احتمال وقوع کلاس c

c. احتمال وقوع داده x ✔

d. احتمال وقوع کلاس c به شرط داده x

سؤال 10
درست
نمره 1.00 از 1.00
علامت زدن سؤال

در روش بازنمایئ کیسه کلمات (Bag of words)

a. ترتیب کلمات حفظ نمی شود. ✔

b. تعداد تکرار کلمات حفظ نمی شود.

c. برای دسته بندی با استفاده از روش بیز قابل استفاده نیست.