۱- درستی یا نادرستی هر یک از عبارت‌های زیر را با ذکر دلیل مشخص کنید.

الف) در یک سیستم بازیابی بولی (Boolean) استفاده از ریشه‌یابی (stemming) همیشه باعث افزایش دقت (precision) سیستم می‌شود.

ب) در یک سیستم بازیابی بولی، استفاده از ریشه‌یابی، هیچ‌گاه فراخوانی (recall) را کاهش نمی‌دهد.

ج) نرمال‌سازی باعث کاهش اندازه‌ی فرهنگ لغات (vocabulary) می‌شود.

د) در یک سیستم بازیابی اسناد متنی، بهتر است نرمال‌سازی و ریشه‌یابی اسناد در زمان پردازش کوئری انجام شود.

۲- چهار قاعده، برای نرمال‌سازی متون فارسی طراحی کنید.

۳- اندازه‌ی postings listهای تعدادی کلمه در جدول زیر داده شده است:

کلمه	اندازه‌ی postings list
سهام	۲۱۳۳۱۲
بارگزاری	۸۷۰۰۹
سالاد	۱۰۷۹۱۳
جهان	۲۷۱۶۵۸
اسکی	۴۶۶۵۳
طلا	۳۱۶۸۱۲

بهترین ترتیب پردازش (شهودی) برای پرسمان زیر را بیابید.

(اسکی or طلا) and (سالاد or جهان) and (بارگزاری or سهام)

۴- منظور از ماتریس تُنُک (sparse) چیست؟ دلیل تنک بودن ماتریس وقوع (incidence matrix) را بیان کنید.

۵- الف) یک شاخص معکوس مکانی را در نظر بگیرید. نحوه‌ی ذخیره‌سازی این شاخص را به صورت دقیق مشخص کنید. ساختار داده‌ی مورد استفاده را تعیین کرده و اطلاعاتی که در آن ذخیره می‌شود را با ذکر جزییات بیان کنید. جزئیات در حد شبه کد به یک زبان آشنا مانند جاوا یا پایتون بیان شود.

ب) با داشتن این ساختار داده، جستجوی یک پرسمان چگونه انجام می‌شود؟ یک شبه‌کد (pseudocode) برای این عملیات بنویسید.

۶- (اختیاری) الگوریتم ادغام لیست‌های postings را برای هر فرمول ترکیبی از گزاره‌های بولی تعمیم دهید. در مورد پیچیدگی زمانی آن بحث کنید. آیا می‌توان در زمان خطی (بر اساس چه پارامتری) ادغام را انجام داد؟ راه حل بهتری نیز وجود دارد؟ پرسمان زیر را به عنوان نمونه در نظر بگیرید.

(دربی or استقلال) and not (قهرمان or پرسپولیس)

۷- در معماری MapReduce برای ساخت شاخص به صورت توزیع شده، برای هر یک از واحدهای parser و inverter، ورودی، خروجی و عملیاتی که در این واحدها انجام می‌شود را بیان کنید.

۸- برای ذخیره سازی Postings listها دو انتخاب وجود دارد: ۱) ذخیره همه Postings listها در یک فایل، ۲) ذخیره هر یک از Postings listها در یک فایل مجزا. تاثیر این انتخاب در کارآیی الگوریتم شاخص‌گذاری پویایی که از یک شاخص اصلی و یک شاخص کمکی و ادغام این شاخص‌ها استفاده می‌کند چیست؟ توضیح دهید.