به نام خدا

نیمسال اول 1400
مدرس : فاطمه لشکری
شماره درس: 40324

پایان ترم **زمان: 175 دقیقه**

لطفا نام و نام خانوادگی و شماره دانشجویی را بروی تمام برگه های جواب بنویسید.

سوالات کوتاه (مختصر توضیح دهید) 40 نمره

1. آیا LSA مشکل چند معنایی را حل می کند؟ از چه روشی برای dimension reduction استفاده می کند؟ (4 نمره)
2. آیا ضریب جاکارد مجموعه های شینگل دو سند یکی از روش های کاربردی برای یافتن شباهت دو سند است؟ (2 نمره)
3. چه طور تعداد مناسب K را در روش خوشه بندی انتخاب می کنیم؟ آیا این روش برای خوشه بندی مسطح و خوشه بندی سلسه مراتبی مناسب است؟ (6 نمره)
4. ایا مدل Bag-of-Word را می توان نمونه خاصی از مدل Vector space در نظر گرفت؟ (2 نمره)
5. کدامیک از روشهای Add-1 smoothing و Dirichlet Prior smoothing را ترجیح می دهید؟ حداقل دو دلیل ذکر کنید. (4 نمره)
6. آیا purity به خودی خود معیار ارزیابی خوبی است؟ (2 نمره)
7. آیا در روش خوشه بندی K-means می توانیم تضمین کنیم که نقطات نزدیک در خوشهای مشابهی قرار می گیرد؟(اگر درست است توضیح دهید و اگر نادرست است شما چه راه حلی پیشنهاد می کنید) (4 نمره)
8. چه طور می توان سبب کاهش overfitting و افزایش درستی یک classifier شد؟ (درستی روش خود را توضیح دهید) (4 نمره)
9. آیا می توان شباهت معنایی بین صفحات وب را با استفاده ازروش LSH تشخیص داد؟ (2 نمره)
10. آیا فرض "مستقل بودن و توزیع یکسان کلمات " در اسناد، پایه و اساس مدل های زبانی است؟ (2 نمره)
11. آیا تکنیک هایی مانند stemming و lower case به دسته بندی متن کمک می کند؟ (2 نمره)
12. کیفیت مرتبط بودن یک سند بر اساس تمام کلمات کوری داده شده قضاوت می شود.؟ (2 نمره)
13. تفاوت بین بازخورد مرتبط (relevance feedback) و گسترش کوری(query expansion) را از نظر تعامل با کاربر شرح دهید. (4 نمره)

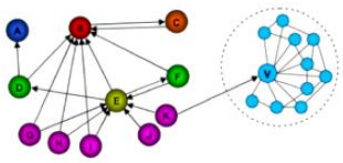
سوالات توضیحی (60 نمره)

14. **تعریف** سیستم بازیابی اطلاعات---------- (12 نمره)

الف) سیستم بازیابی اطلاعات را تعریف کرد وهدف اصلی یک سیستم بازیابی اطلاعات چیست؟ (4 نمره)
ب)بخش های اصلی یک سیستم بازیابی اطلاعات را نام ببرید و اهداف آنها را توضیح دهید. (8 نمره)
برای راحتی می توانید معماری سیستم بازیابی اطلاعات را به کمک شکل نشان دهید و نحوه جریان انتقال ازاطلاعات را نشان دهید

15. **Page Rank** ---------- (7 نمره)

الف) آیا page rank یک صفحه جدید وب نمایش درستی از ارزش آن صفحه را نمایش می دهد؟ مختصر توضیح دهید (2 نمره)
ب) شکل زیر نمایشگر چه مشکلی در وب است. الگوریتم page rankچه طور می تواند این مشکل را بر طرف کرد؟ (5 نمره)



به نام خدا

نیمسال اول 1400
مدرس : فاطمه لشکری
شماره درس: 40324

پایان ترم **زمان: 175 دقیقه**

16. خزشگرCrawler ---------- (18 نمره)

الف) دو نکته حیاتی در پیمایش صفحات وب توسط یک خزشگروجود دارد آنها را نام برده و توضیح دهید؟ (4 نمره)

ب) در کدام بخش از یک خزشگر می توان این دو نکته را مدیریت و پیاده کرد؟ در مورد مکانیزم آن دقیق توضیح دهید.(8نمره)

ج) نقش Host splitter , Duplicate eliminator در معماری یک خزشگر را توضیح دهید (6نمره)

17. Support vector machine---------- (7 نمره)

توضیح دهید چگونه حالت بهینه SVM را به صورت زیر تعریف می شود؟

(راهنمایی) برای فرمول جبری، محدودیت استاندارد را اعمال می کنیم که$\vec{w}^T \vec{x}i + b = +1$ اگر $y_i = 1$ و $\vec{w}^T \vec{x}i + b = -1$ اگر $y_i = -1$ است، و سپس به کمینه کردن $||w||$ ادامه می دهیم. و ....

Optimization problem solved by SVMs

Find $\vec{w}$ and b such that:

- $\frac{1}{2}\vec{w}^T\vec{w}$ is minimized (because $|\vec{w}| = \sqrt{\vec{w}^T\vec{w}}$), and
- for all $\{(\vec{x}_i, y_i)\}$, $y_i(\vec{w}^T\vec{x}_i + b) \geq 1$

18. شما یک روش جدید برای تجزیه اسناد ایجاد کرده اید که از اطلاعات معنایی برای تصمیم گیری در مورد اینکه کدام جملات index شوند و از کدام رد شوند استفاده می کنید. چگونه روش پیشنهادی را با روش index کردن تمام جملات با هدف بازیابی جوابهای مرتبط بیشترمقایسه می کنید؟ (6 نمره)

19. دانشکده کامپیوتر دانشگاه صنعتی شریف قصد دارد به طور خودکار یک نمایه متنی(text profile) (به عنوان مثال، فهرستی از کلمات یا عبارات کلیدی) برای تخصص تحقیقاتی هر عضو هیئت علمی با توجه به مقالات آنها بسازد. از شما به عنوان مشاور این پروژه دعوت شده است. آیا می توانید حداقل دو راه حل مختلف برای این کار پیشنهاد دهید؟ لطفا مختصر توضیح دهید. (4 نمره)

20. در نمودار زیر نقاط مشخص شده حاوی چه اطلاعاتی در مورد اسناد بازیابی شده بر اساس ارتباط بین precision و recall هستند؟ (6 نمره)