گزارش کلی پروژه مبانی بازیابی اطلاعات

احمد حسینی ۴۰۰۱۴۰۶۹۰۱ - عرفان انصاری ۹۹۱۴۰۶۵۰۴ - امید فرهادی ۹۹۱۴۰۶۵۰۵

خلاصه کلی

این کد یک سیستم بازیابی اطلاعات کامل برای کار با مجموعه داده‌های فارسی همشهری است که شامل پیش‌پردازش متن، ایجاد فهرست (indexing)، فشرده‌سازی، و عملیات جستجو می‌باشد.

منوی اصلی برنامه

برنامه در حالت CLI اجرا می‌شود و شامل گزینه‌های زیر است:

عملکرد	شماره
خواندن سند بعدی از مجموعه	1
نمایش متن و شناسه سند فعلی	2
پردازش سند فعلی	3
نمایش تمام اسناد پردازش‌شده	4
پاک‌سازی ترمینال	5
SPIMI ایندکس‌گذاری با روش	6
BSBI ایندکس‌گذاری با روش	7
JSON ذخیره ایندکس در فایل	8
بارگذاری ایندکس از فایل	9
فشرده‌سازی داده‌ها	10
(یا عبارت‌محور TF-IDF) جستجو	11
خروج از برنامه	0

ویژگی‌های اصلی

1. پیش‌پردازش متن فارسی (persian_text_proccess)

- **نرمال‌سازی**: استفاده از Normalizer کتابخانه hazm
- **توکن‌سازی**: تقسیم متن به کلمات جداگانه
- **حذف اعداد**: فیلتر کردن توکن‌های عددی
- **حذف علائم نگارشی**: پاک کردن نشانه‌های نگارشی فارسی و انگلیسی
- **حذف کلمات ایست**: استفاده از لیست کلمات ایست فارسی
- **ریشه‌یابی (Stemming)**: تبدیل کلمات به ریشه
  - **واژه‌شناسی (Lemmatization)**: تبدیل کلمات به شکل استاندارد

2. الگوریتم‌های ایجاد فهرست

الف) SPIMI (Single-Pass In-Memory Indexing)

- پردازش تمام اسناد در یک بار
- ایجاد فهرست معکوس در حافظه

- مناسب برای مجموعه داده‌های کوچک

ب) BSBI (Blocked Sort-Based Indexing)

- تقسیم اسناد به بلوک‌های کوچک (سایز 3)
- ایجاد فهرست جداگانه برای هر بلوک
- ترکیب نهایی فهرست‌ها
- مناسب برای مجموعه داده‌های بزرگ

3. تکنیک‌های فشرده‌سازی

الف) Dictionary as String

- ذخیره اصطلاحات و شناسه‌های سند به صورت رشته پیوسته
- نگهداری موقعیت‌ها در یک دیکشنری جداگانه
- کاهش overhead ساختار داده‌ای

ب) Blocking Compression

- تقسیم اصطلاحات به بلوک‌های ثابت
- استفاده از اشاره‌گر برای دسترسی سریع
- بهینه‌سازی فضای ذخیره‌سازی

ج) Gamma Code Compression

- تبدیل شناسه‌های سند به اعداد صحیح
- استفاده از Gap Encoding
- فشرده‌سازی با الگوریتم Gamma Code
- بهینه برای توزیع‌های نامتقارن

د) Variable Byte Compression

- مشابه Gamma Code اما با استفاده از Variable Byte
- انعطاف‌پذیری بیشتر در فشرده‌سازی
- کارایی بهتر برای اعداد بزرگ

4. سیستم‌های جستجو

الف) TF-IDF (Term Frequency-Inverse Document Frequency)

- محاسبه وزن اصطلاحات در اسناد
- رتبه‌بندی اسناد بر اساس امتیاز
- نمایش 10 سند برتر
- استفاده از فرمول smoothed IDF

ب) Phrasal Search

- جستجوی عبارات در نقل قول
- ترکیب جستجوی کلمات منفرد و عبارات
- بررسی حضور متوالی کلمات در عبارات
- استفاده از عبارات منظم (Regex)

5. ارزیابی عملکرد

معیارهای ارزیابی:

- **Precision**: دقت نتایج بازیابی شده
- **Recall**: فراخوانی اسناد مرتبط
- **F-measure**: میانگین هارمونیک دقت و فراخوانی
- **MAP (Mean Average Precision)**: میانگین دقت متوسط

نقاط قوت

1. **پشتیبانی کامل از زبان فارسی**: استفاده از کتابخانه hazm
2. **انعطاف‌پذیری**: امکان انتخاب بین روش‌های مختلف
3. **بهینه‌سازی فضا**: تکنیک‌های متنوع فشرده‌سازی
4. **ارزیابی دقیق**: استفاده از معیارهای استاندارد
5. **رابط کاربری ساده**: منوی تعاملی برای کاربر

نقاط ضعف و پیشنهادات بهبود

مسائل فنی:

1. **کیفیت کد**: نیاز به بهبود نام‌گذاری متغیرها
2. **مدیریت حافظه**: عدم بهینه‌سازی برای داده‌های بزرگ
3. **مدیریت خطا**: فقدان کنترل خطاهای احتمالی
4. **مستندات**: نبود توضیحات مفصل

پیشنهادات:

1. **بهبود ساختار کد**: تقسیم به کلاس‌ها و ماژول‌ها
2. **افزودن logging**: ثبت عملیات و خطاها
3. **بهینه‌سازی الگوریتم‌ها**: بهبود کارایی BSBI
4. **رابط کاربری گرافیکی**: جایگزینی منوی متنی
5. **پشتیبانی از فرمت‌های مختلف**: XML، JSON، CSV

عملکرد کلی

کد یک سیستم بازیابی اطلاعات جامع و کاربردی ارائه می‌دهد که قابلیت‌های زیر را دارا است:

- **مقیاس‌پذیری**: قابلیت کار با مجموعه داده‌های مختلف
- **کارایی**: استفاده از الگوریتم‌های بهینه
- **دقت**: ارزیابی دقیق نتایج
- **انعطاف**: امکان تنظیم پارامترهای مختلف

نتیجه‌گیری

این کد یک پیاده‌سازی مناسب برای سیستم بازیابی اطلاعات فارسی محسوب می‌شود که می‌تواند به عنوان پایه‌ای برای توسعه سیستم‌های پیچیده‌تر مورد استفاده قرار گیرد. با اعمال بهبودهای پیشنهادی، می‌توان آن را به یک سیستم حرفه‌ای تبدیل کرد.