به نام خدا

پروژه سوم درس مبانی بازیابی اطلاعات و جستجوی وب

سعید عطایی – 9212430100

در این پروژه قرار است با استفاده از ابزار lucene بر روی تعدادی داکیومنت و کوئری آماده، یک سیستم جستجو پیاده سازی شود.

جستجو باید با استفاده از 3 الگوریتم مختلف انجام شود؛ الگوریتم پیش فرض lucene و دو الگوریتم بهینه شده tf.idf پیشنهادی.

مراحل کار به صورت ذیل خواهد بود:

1. دانلود کتابخانه مناسب از lucene
2. نمایه سازی داکیومنت ها
3. پیاده سازی الگوریتم جستجو
4. انجام جستجو با استفاده از کوئری های آماده
5. ارزیابی الگوریتم ها با استفاده از ابزار trec_eval

1- دانلود کتابخانه مناسب از lucene

با مراجعه به سایت https://lucene.apache.org فایل های این ابزار را با ورژن مناسب کار دانلود می کنیم. آخرین نسخه بارگذاری شده در این سایت ورژن 6.6.0 است. اما از آنجایی که هر ورژن و نسخه نسبت به نسخه های قبلی دارای ناسازگاری و تفاوت هایی است و با توجه به توضیحات ارائه شده برای انجام پروژه، ورژن 5.1.0 را انتخاب و دانلود می کنیم.

فایل های زیر را به IDE مورد استفاده اضافه می کنیم:

- lucene-analyzers-common-5.1.0.jar
- lucene-core-5.1.0.jar
- lucene-queryparser-5.1.0.jar

2- نمایه سازی داکیومنت ها

تمامی اسناد داده شده در پروژه در فایل cran.all.1400 قرار دارند. نمایه سازی اسناد باید به صورت جدا انجام پذیرد، لذا محتوای این فایل را خوانده و اسناد را به صورت جداگانه در پوشه docs ذخیره می کنیم.

نمایه سازی این اسناد با استفاده از توابع ابزار lucene به صورت زیر انجام می پذیرد:

```
public int index(String dataDir, FileFilter filter) throws Exception {
    File[] files = new File(dataDir).listFiles();
    for (File f : files) {
        if (filter == null || filter.accept(f)) {
            indexFile(f);
        }
    }
    return writer.numDocs();
}
```

نکته حائز اهمیت این است که هر سند برای نمایه سازی دارای خصیصه های زیر می باشد:

- I : شماره
- T: عنوان
- A: نویسنده
- B: توضیح اضافی
- W: متن

3- پیاده سازی الگوریتم جستجو

برای جستجو ابتدا از روش پیش فرض ابزار lucene استفاده می کنیم:

```
is.setSimilarity(new DefaultSimilarity());
```

اولین تغییری که در الگوریتم tf.idf ایجاد می کنیم، حذف تعداد کل اسناد از قسمت idf است. چرا که عددی است که تغییر نمی کند و برای همه عبارات و کوئری ها ثابت است. لذا اعمال آن در فرمول تاثیری ندارد.

شکل جدید فرمول به این صورت خواهد بود:

[1 + log(tf)] * [log(1/df)]

برای تغییر دوم محاسبه idf را به جای log گرفتن با sqrt حساب کردن انجام می دهیم. یعنی داریم:

[1 + log(tf)] * [sqrt(N/df)]

4- انجام جستجو با استفاده از کوئری های آماده

تمامی کوئری ها در فایل cran.qry وجود دارند. لذا محتوای فایل را خوانده و کوئری ها را به ترتیب در نمایه مورد جستجو قرار می دهیم و 100 سند برتر را برای هر کوئری به کمک ابزار lucene پیدا می کنیم:

```
TopDocs hits = is.search(query, 100);
```

خروجی را که شامل لیست مرتب اسناد به همراه امتیازشان است را برای هر الگوریتم با فرمت زیر در پوشه output می ریزیم:

```
topic_id  \t Q0 \t document_id \t rank \t score \t your_login
queryNumber + " \tQ0\t" + doc.get("I") + "\t" + (i ++) + "\t" +
scoreDoc.score + "\ttest1\n"
```

5- ارزیابی الگوریتم ها با استفاده از ابزار trec_eval

Trec_eval یک ابزار استاندارد است که توسط جامعه trec برای ارزیابی یک اجرای ad hoc بازیابی مورد استفاده قرار می گیرد. این ابزار را از سایت http://trec.nist.gov/trec_eval دانلود می کنیم.

برای استفاده و اجرای این نرم افزار باید از سیستم عامل لینوکس استفاده کنیم.

اجرای آن در ویندوز نیز امکان پذیر است و باید ابزار Cygwin را دانلود و استفاده کنیم که لزومی برای این کار نیست و طبق توضیحات ارائه شده برای انجام پروژه از سیستم عامل لینوکس استفاده می کنیم – مرجع:
https://github.com/usnistgov/trec_eval/blob/master/README.windows.md

پس از گرفتن خروجی با فرمت داده شده از الگوریتم ها، با استفاده از این ابزار نتایج را ارزیابی می کنیم.

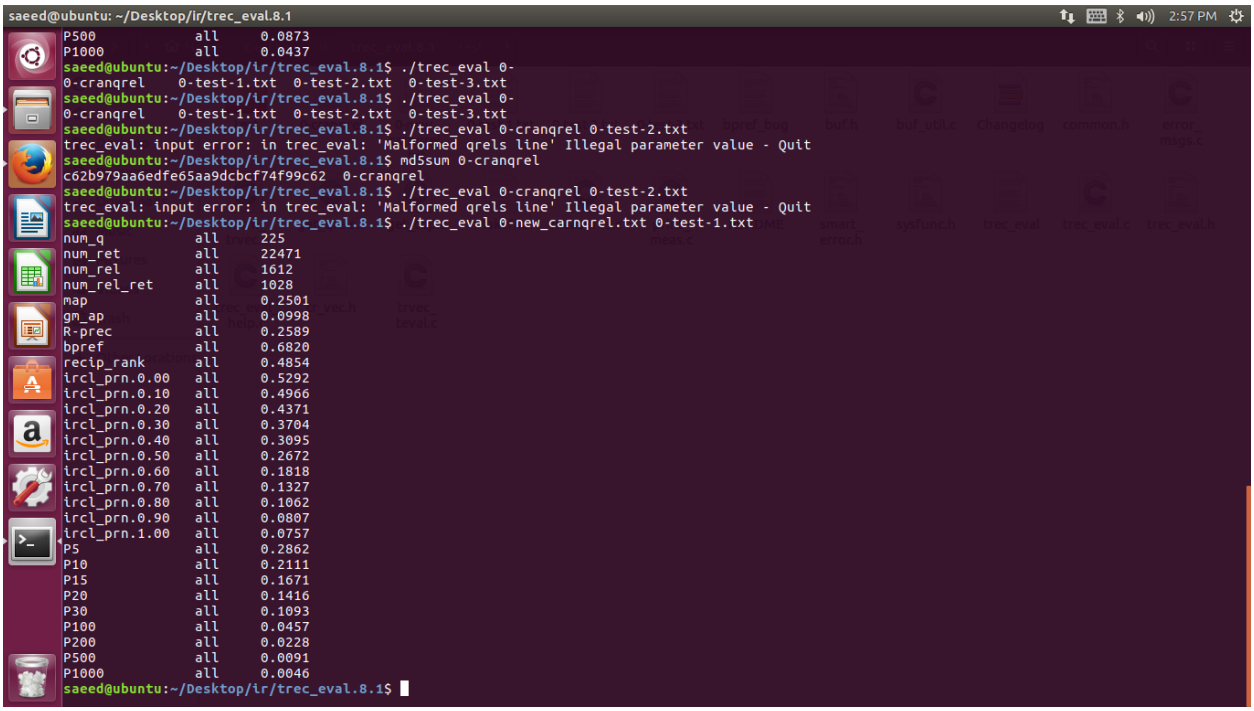
فایل خروجی داده شده اولیه خطای زیر را ایجاد می کرد :

trec_eval: input error: in trec_eval: 'Malformed qrels line' Illegal parameter value – Quit

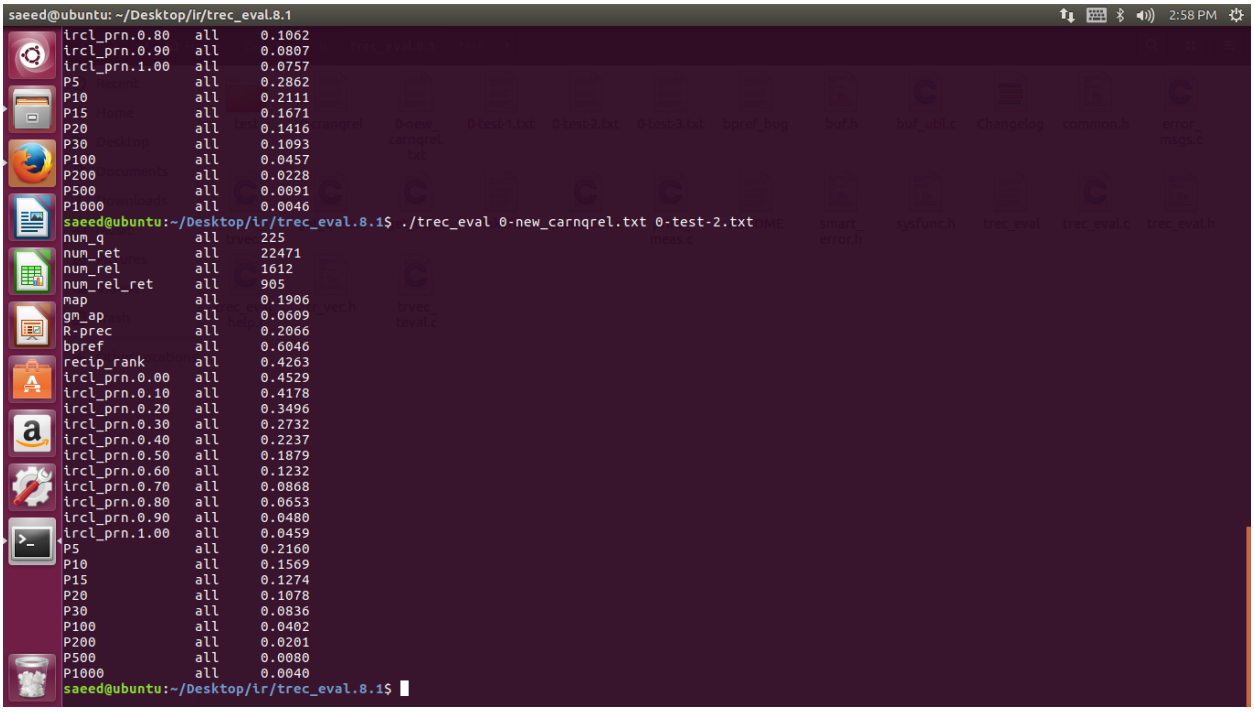
که علت آن فرمت اشتباه فایل بود. لذا فایل جایگزین درست تهیه و استفاده شد.

دستور ارزیابی با توجه به منابع موجود :

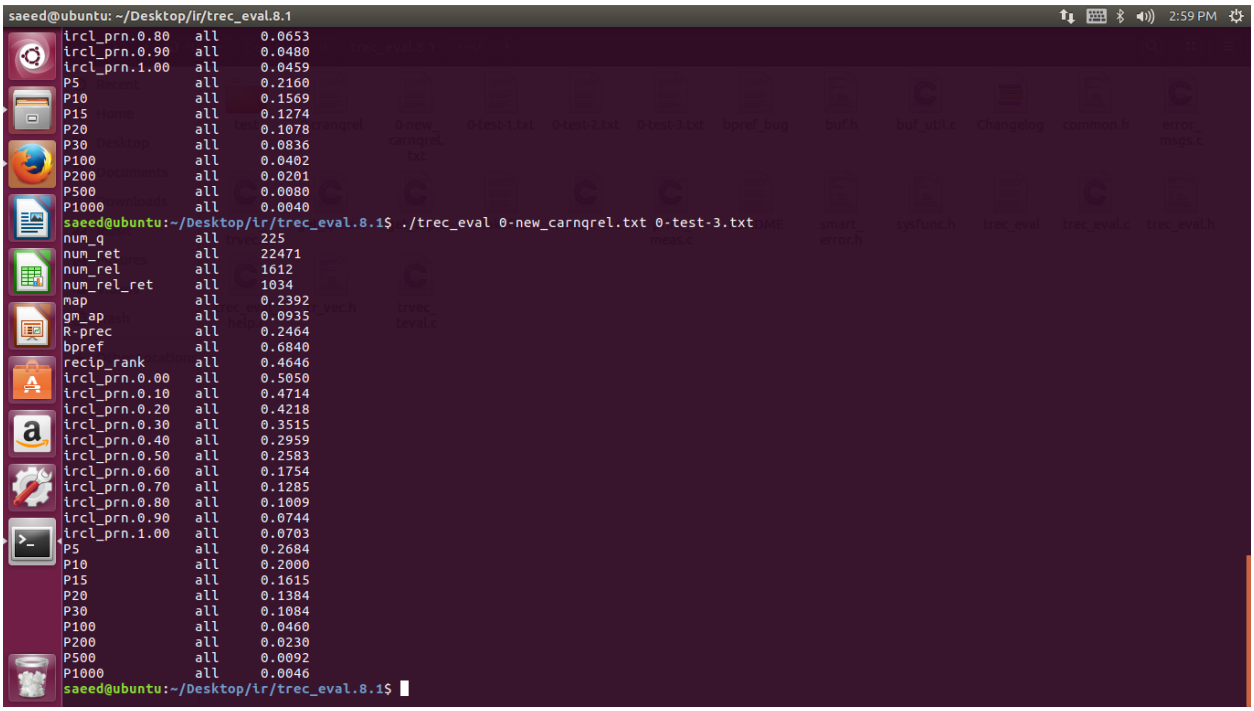
./trec_eval 0-new_carnqrel.txt 0-test-1.txt



(ارزیابی الگوریتم پیش فرض)



(ارزیابی الگوریتم 1)



(ارزیابی الگوریتم 2)

منبع مورد استفاده برای ابزار:

https://infoscience.epfl.ch/record/115460/files/Free_sofware_for_IR.pdf

http://www.inf.ed.ac.uk/teaching/courses/tts/assessed/assessment1.html