برای شروع مواردی که باید انجام شوند را فهرست می کنیم:

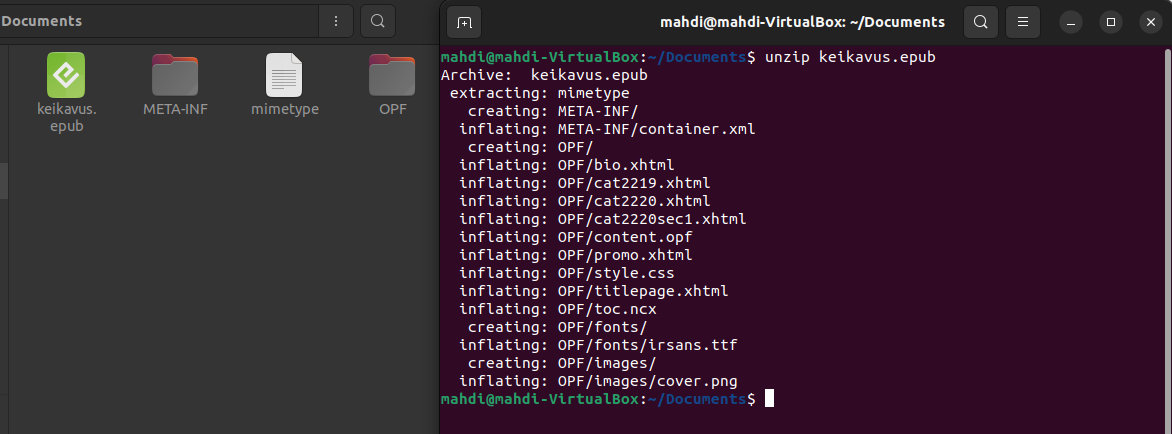
1. پیدا کردن فایل متنی کتب و واژه نامه مغولی
2. تبدیل آنها به فرمت خوانا برای برنامه (مانند txt)
3. نوشتن برنامه
   * باز کردن و دریافت متن از فایل و واژه نامه
   * جدا کردن کلمات از یکدیگر (Word Tokenizing)
   * بررسی ریشه کلمات فایل و مقایسه با واژه نامه
     + تبدیل کلمات به ریشه ها (کتابهایش ← کتاب، می خوردیم ← خور)
     + مقایسه لغات و شمارش آنها
     + خروجی نتیجه خام
4. تحلیل خروجی ها

**پیدا کردن فایل کتب و نوشته های شاعران**

بعد از تحقیقی در اینترنت به سایت متن باز (Open source) [گنجور1](https://ganjoor.net/) می رسیم. در بخش [کتابخانه گنجور](http://epub.ganjoor.net/) می توانیم کتاب ها را با فرمت epub دانلود کنیم.

1.«گنجور» مجموعه‌ای تحت وب از آثار سخنسرایان پارسی‌گو است که با استفاده از آن می‌توان ضمن مرور به تفکیک نام شاعر و نام آثار او، در بین آثار جستجو کرد. این مجموعه در ابتدا به کمک نرم‌افزار مدیریت محتوای وردپرس راه‌اندازی شده بود و از سال ۱۴۰۰ به بعد به پشتوانه نرم‌افزار اختصاصی بازمتن خود کار می‌کند.

سپس فایل را از حالت کتاب الکترونیک (epub) خارج می کنیم.



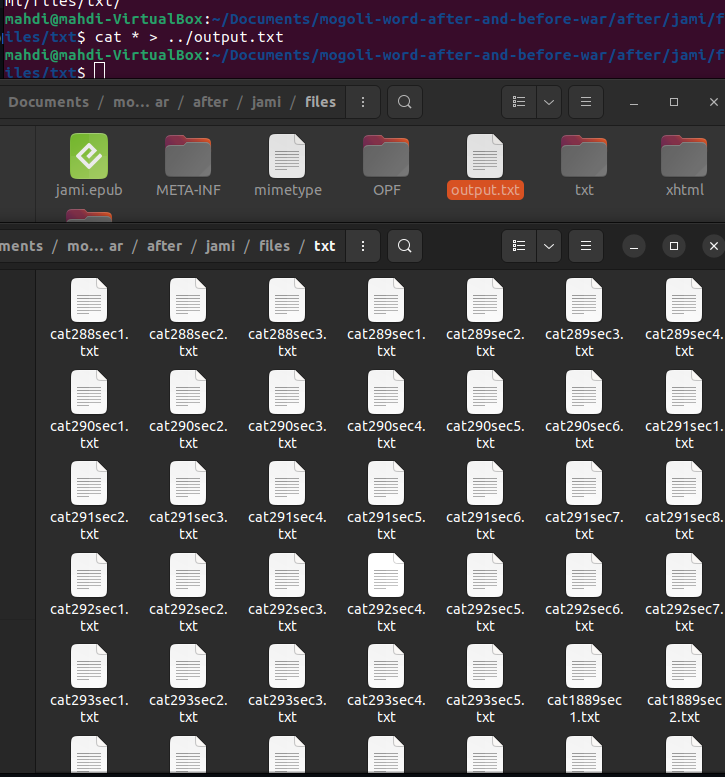
با بررسی فایل ها متوجه می شویم فایل های در پوشه OPF به شکل \*sec\*.xhtml محتوای کتاب ما هستند.

حال باید فایل هایمان را به فرمت بهتری برای بررسی تبدیل کنیم.

سایت ها و برنامه های مختلفی برای تبدیل xhml به txt وجود دارد. ما از [این](https://products.aspose.app/words/conversion/xhtml-to-html) استفاده می کنیم.

حال باید فایل های txt خود را به یک فایل تبدیل کنیم. با دستور زیر می توانیم این کار را انجام دهیم. (فرض شده که تمامی فایل ها در یک پوشه هستند و هیچ فایل دیگری در آن پوشه نیست.)

|  |
| --- |
| cat \* > your\_address/file\_name.txt |



حال فایل نهایی خود را برای دادن به برنامه داریم.

**نوشتن برنامه**