**بسمه تعالی**

**تکلیف 4**

ساخت خزنده

* نسخه منتشر شده Apache Nutch 1.17 را از آدرس <https://nutch.apache.org/downloads.html> دانلود کنید و آن را به روی سیستم خود نصب کنید.
* با اجرای "bin/nutch" صحت نصب خود را بررسی کنید. خروجی صحیح به صورت زیر می باشد:

Usage: nutch COMMAND where command is one of:

readdb read / dump crawl db

mergedb merge crawldb-s, with optional filtering

readlinkdb read / dump link db

inject inject new urls into the database

generate generate new segments to fetch from crawl db

freegen generate new segments to fetch from text files

fetch fetch a segment's pages

* یک Seed list برای خزنده خود ایجاد کنید. (باتوجه به شماره گروه خود از ابتدای لیست سایت ها، یک سایت را برداشته و به همین ترتیب برای گروه های بعدی- مثال: گروه 1 سایت 1، گروه2سایت 2، گروه3 سایت 3، گروه4 سایت 1، گروه5 سایت2 و همین طور تا انتها)
* فایل conf/regex-urlfilter.txt را ویرایش کنید به نحوی که صفحات خارج از دامنه سایت های مدنظر خزش نشود.
* یک Fetch list مناسب ایجاد نموده و فرآیند واکشی صفحات را انجام دهید.
* پارسر مناسب برای سایت های مدنظر ایجاد کرده و فقط اطلاعات زیر را از آنان استخراج کرده:
  + عنوان خبر
  + آدرس سایت
  + متن خبر
  + تاریخ انتشار خبر
  + برچسب خبر( مثال:TRUE-NOT TRUE-FALSE-…)
* نمره اضافی:
* نام نویسنده خبر
* موضوع خبر
* لیست سایت های مدنظر برای خزش:

1. <https://www.politifact.com/>
2. <https://www.truthorfiction.com/>
3. <https://www.snopes.com/fact-check/>