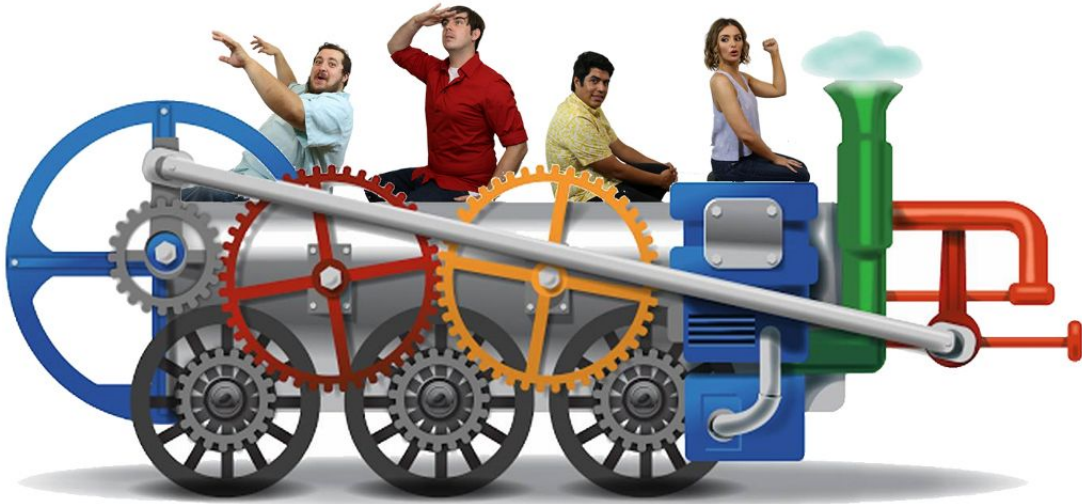


Search Engine Project

Advanced programming techniques



Search Engine



Team no.2 [Semester]

Menna-Tullah Mustafa Fathy - Section 2

Mohamed Atef Mohamed - Section 2

Bassel Hossam Shawkat - Section 1

Needed steps to run the program

- 1- **Install:** python-2.7.amd64.msi
 - 2- Uncompress the “site-packages.rar” in: C:\Python27\Lib\site-packages
 - 3- Run “main.py”
- OR
- 3- If you want to run each module as stand alone you can uncomment programs in “crawler.py” , “indexer.py”

Used Database and how to start service [MongoDB]:

- 1-Install mongodb server
- 2- open cmd
- 3- cd C:\Program Files\MongoDB\Server\3.4\bin
- 4- **Set Data Path to C:\data\db:** md \data\db
- 5- run: mongod
- 6- from a new cmd: cd C:\Program Files\MongoDB\Server\3.4\bin
- 7- run: mongo
- 8- commands:
 - use Indexer
 - db.createCollection("words")

Used Packages:

- 1- **Nodebox English Linguistics library (en):** used for verbs stemming
- 2- **Snowball stemmer from natural language toolkit (nltk):** used for nouns/adj/adv..etc stemming
- 3- **Re:** for regex and get links from html page
- 4- **time:** for sleep and stop multiple visits to same server in less than 5s
- 5- **Htmlparser:** We overwrote the main 5 functions of it
- 6- **Multiprocessing.lock:** to synchronize the indexers
- 7- **Threading:** to make the indexer and crawler multithreaded
- 8- **Os:** To read the html files
- 9- **robotparser:** parsing robots.txt and check if url is allowed or not
- 10- **urllib:** open URL and get html document

11- **urlparse (urlparse, urljoin)**: parsing relative paths to full url and get the base server url