PURDUE UNIVERSITY | The Data Mine

# Where Is iT

LOCKHEED MARTIN

## About

Lockheed Martin is an aerospace corporation with operations worldwide, though concentrated in the United States. The company operate in arms, defense, security and technology industries

### CURRENT LOCKHEED MARTIN PROJECTS

ORION SPACECRAFT

F-16 FIGHTER JET

## Project Where Is iT

Project WIT is a Data Mine project that aims to address the searching a directory problem

Lockheed Martin a large amount of information and documents within their company and employees often need to parse these documents for a specific term or phrase but have no reliable way to do so.


**Our Team:**

TAs: Bryan Jacobs, Shashank Namboodiri, Ben Moorman
Students: John Motzel, Mason Wilcox, Brian Abramov, Austin Zapata, Alex Ip, Aidan Redmond

## Objectives

### Primary Goals:

- •Search
- Parse
- Write to output file
- User-friendly
- Filter

### Secondary Goals:

- Identify and index all files
- Search for the input query
- Utilize Apache Tika and pySolr
- Implement pySimpleGui

## Research Methodology

### Assumptions:

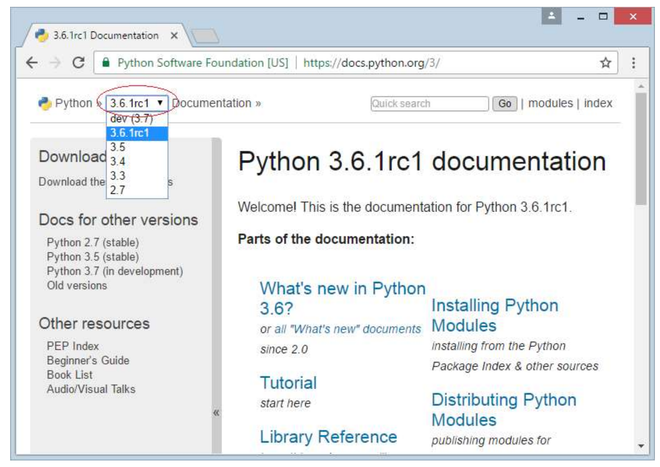
- Research improves likelihood of developing a qualitative product
- Most valuable coding library information found in the original documentation

### Results:

- Utilized Apache suite to convert documents to a readable format for parsing
- Incorporated a highly user-friendly GUI package from python
- Os library interacts with directories and files
  - Can identify file type, location, and some preliminary metadata

### Process:

- Evaluated stakeholders' needs during the first semester
- Continued research and began developing software in 2nd Semester
- Testing scripts frequently updated to provide best user experience




## What is Apache?

### Apache is an open-source software license that provides us with two useful functions:

Tika:

- Apache Tika is our primary parser
- Converts specific file types into metadata
- Each document type has its own specific metadata

PySimpleGUI

Though not a part of the Apache suite, the GUI we intend to use has important functionality in creating our product:

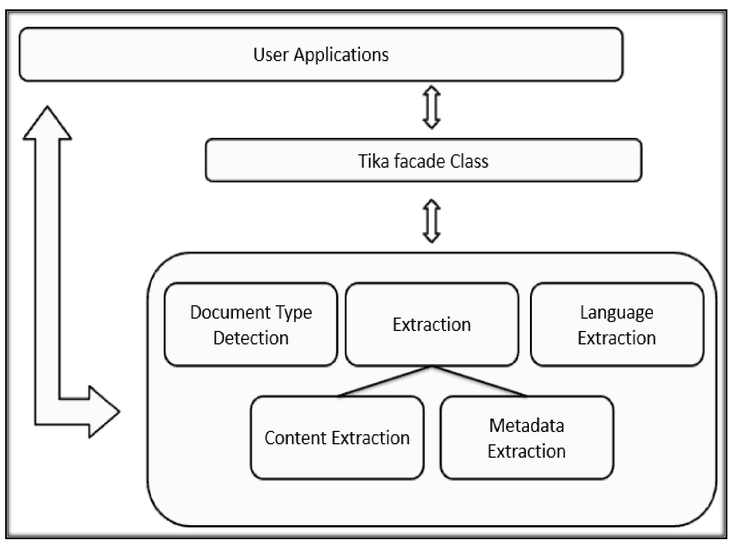
- Enables users to navigate with ease
- Some of the filters will be:
  - File Creation
  - File Size

Apache Tika

WIT


### PySolr:

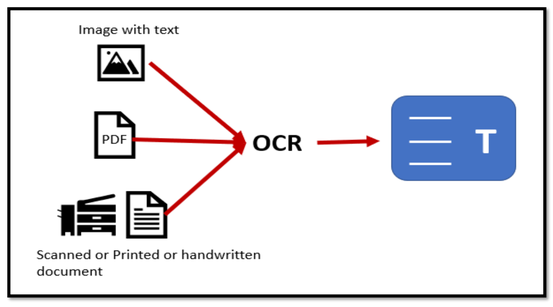
- Apache Solr as a python library
- Searches with Apache Lucene
- Locates specific keywords or phrases
- Functions in tandem with Tika
- Searches a directory in near-real time




## Future Goals

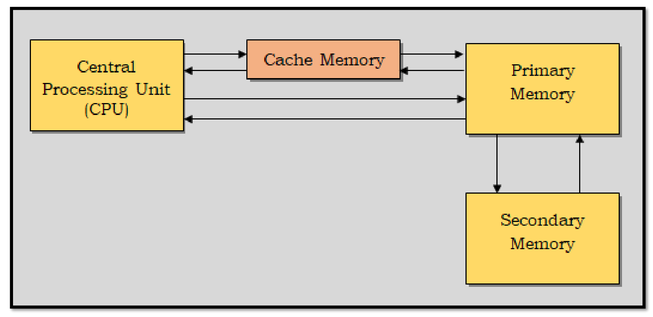
### OPTICAL CHARACTER RECOGNITION (OCR)

- **WHAT?** Extracting text from images and scanned documents
- **WHY?** It would allow us to expand the types of documents we can search
- **HOW?** Our parsing program may have the capabilities built in
- **WHEN?** This will not be completed this semester, so it would be completed in future semester




### IMPLEMENTING A CACHE SYSTEM

- **WHAT?** A component that would store frequently used data to make is run faster
- **WHY?** It would improve the speed of our program
- **HOW?** Currently we haven't done enough research to answer this question
- **WHEN?** This will not be completed this semester, likely in the future semesters




## References & Acknowledgements

### Conclusion

Throughout these two semesters, Purdue's Tech-Strategy team has worked diligently to create a tool that can eliminate the need for Lockheed's Engineers to manually review document artifacts. We are proud to have met the expectations of our corporate partner: Lockheed Martin and to have established a connection that will provide our students with new and exciting opportunities in the many years to come.

**Special Thanks** to **Jennifer Bierbauer, Kelsey Cannon** and **Alan Moshinsky** for supporting us throughout the project!

### References

*OS - miscellaneous operating system interfaces*¶. os - Miscellaneous operating system interfaces - Python 3.10.4 documentation. (n.d.). Retrieved April 3, 2022, from https://docs.python.org/3/library/os.html

*How do I extract data from a DOC/DOCX file using python*. NewbeDEV. (n.d.). Retrieved April 3, 2022, from https://newbedev.com/how-do-i-extract-data-from-a-doc-docx-file-using-python

Ekanayake, V. (n.d.). *Reading and writing Microsoft word docx files with python*. Virantha Namal Ekanayake Full Atom. Retrieved April 3, 2022, from https://virantha.com/2013/08/16/reading-and-writing-microsoft-word-docx-files-with-python/

Quang, T. N. (n.d.). *Description*. DocFetcher. Retrieved April 3, 2022, from http://docfetcher.sourceforge.net/en/index.html