

# Natural Language Processing (NLP) Capabilities



January 2023

### Guidehouse Artificial Intelligence (AI) and Automation Team

### **Beyond Advanced Analytics**



Combined deep domain and technical knowledge brings a unique value proposition



A specialized team of 250+ Analytics, Automation, and AI technical experts (includes practitioners with TS/SCI+ level clearances)



Access to dedicated resources, advanced technology, and diverse datasets



Recognized as an AI and Automation leader













### **Capabilities**

Outwit Complexity

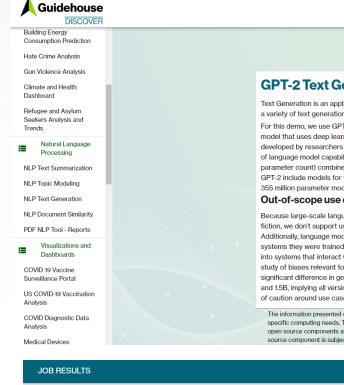
Our advanced capabilities uniquely enable us to develop technical solutions that solve our clients' most complex problems

Combining statistical modeling, data mining, and domain expertise **Data** to help our clients make **Analytics +** data-driven decisions. **Advanced** Automating mundane tasks allows our clients Visualization to use their resources For Machine Learning and Al most effectively. From solutions, such as Natural RPA pilots to Centers of Intelligent Language Processing, we bring a **Machine** Excellence, we use the toolbox of analytic applications, **Automation** leading software Learning + Al including several open source + RPA platforms and tailor them and cloud options. Our to our clients' unique specialists are trained to use needs. TensorFlow, PyTorch, Spark, and all the major cloud providers (AWS, Azure, and GCP). Leveraging big data **Data** Working with organizations technologies, cloud, and AI+ **Engineering** to complete an AI and structured/unstructured **Automation** Automation current state databases to build largeassessment, determine their scale data storage and Strategy **Architecture** target state, and deliver a processing architectures. Enables deep analysis of roadmap to execute the appropriate strategy. large, complex data sets with speed and scale. **Guidehouse** 

### **Discover Platform**

https://discover.guidehouse.com/

- This digital platform showcases our technical solutions – including Advanced Data Analytics and Artificial Intelligence (AI), **Enterprise Digital Modernization and** Systems Integration, and Cloud Adoption and Infrastructure Optimization.
- Discover demonstrates where transformation can happen for agencies and organizations, allowing for clear insights and actionable decisions by leaders. Seeing the future in a test environment provides the inspiration needed to understand where a digital journey could lead.
- Discover features a variety of solutions that demonstrate Guidehouse's breadth of technical experience.



WELCOME WYNN THANE HOME LOGOUT

#### **≛** UPLOAD DATASET FOR NLP TEXT GENERATION

#### **GPT-2 Text Generation**

Text Generation is an applications of NLP that uses deep learning techniques for a variety of text generation tasks.

For this demo, we use GPT-2 (Generative Pre-trained Transformer 2) language model that uses deep learning to produce human-like text. This model was developed by researchers at OpenAI to help us understand how the capabilities of language model capabilities scale as a function of the size of the models (by parameter count) combined with very large internet-scale datasets (WebText). GPT-2 include models for 1.5B, 774M, 355M, and 124M parameters. We use the 355 million parameter model for jobs submitted in this demo.

#### Out-of-scope use cases

Because large-scale language models like GPT-2 do not distinguish fact from fiction, we don't support use-cases that require the generated text to be true. Additionally, language models like GPT-2 reflect the biases inherent to the systems they were trained on, so we do not recommend that they be deployed into systems that interact with humans unless the deployers first carry out a study of biases relevant to the intended use-case. We found no statistically significant difference in gender, race, and religious bias probes between 774M and 1.5B, implying all versions of GPT-2 should be approached with similar levels of caution around use cases that are sensitive to biases around human attributes.

#### **Demo: Generate Text Files**

Using the following form, you can submit a job that generates text using your own text file. You can upload your text file or use one from the dropdown list. This demo only supports .txt, .zip, .tar, and .tar.gz file formats.

The sample text file should contain a paragraph or a leading statement that will be used by the model to generate a large text. For example, "Define the steps to getting started and winning government contracts".

Note: Each run generates different text Input Dataset test-GPT2.txt

Please select from one of the datasets from the dropdown below

RUN LIVE DEMO >

List of recently submitted jobs:

View Demo History

The information presented on the Discover platform is solely for demonstration purposes to illustrate to customers how the algorithms may be customized and implemented to meet specific computing needs. The tools should not be used for any other purpose. Some tools contain third party open source software components, which are owned by third parties. Such open source components are not licensed under Guidehouse's terms, but are instead licensed under the terms of the applicable open source license. Guidehouse's use of each open source component is subject to the terms of each applicable license, which are available to the customer by request. For more information please review our privacy policy

Input File test-GPT2.txt

#### **Text Generation Results**

For many, the benefit cuts are a matter of identity.

"People would rather have a piece of my paycheck, and the state benefits have that piece," said Linda B. Stavivoski, 67, a re The state is losing \$15 for every dollar of added state earnings, the agency projects, and it plans to use every dollar, and

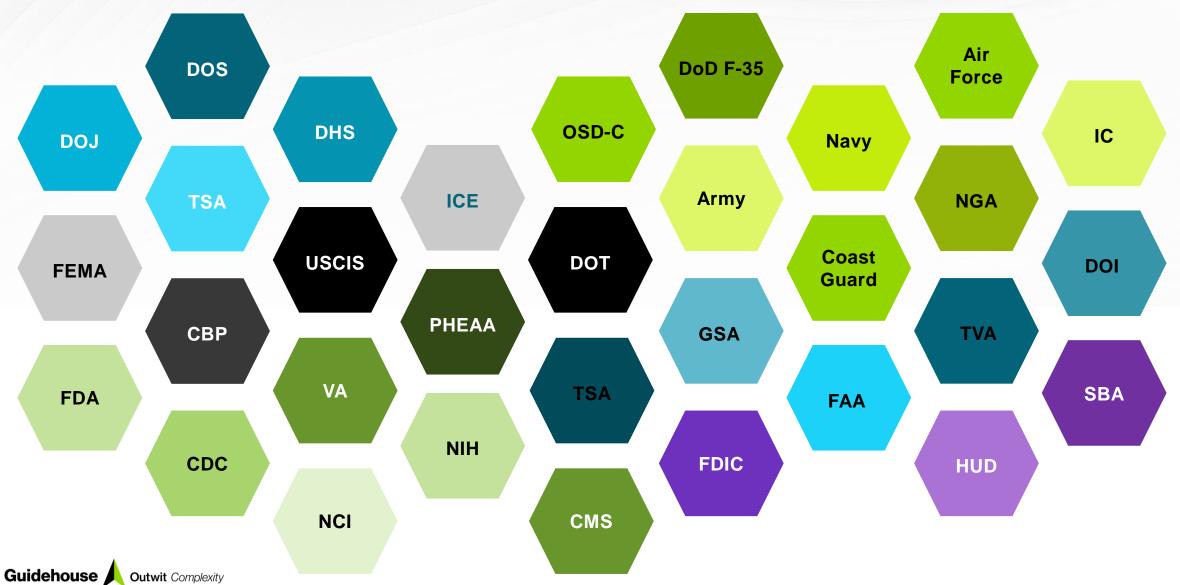
"Some people don't want to get back into paying the bills, and it's going to put pressure on them," said Elizabeth M. Dolan o

We'll tell you what it will take to complete this kickstarter. You'll also receive detailed information so you can make infor

Our goal is to have every backer have access to the KickStarter, with every game you make, from cradle to tinker for yourself



### Serving over 30 federal, state, and local agencies



### **OCR for Faster PDF Processing**

#### **Government Client**

#### **CHALLENGE**

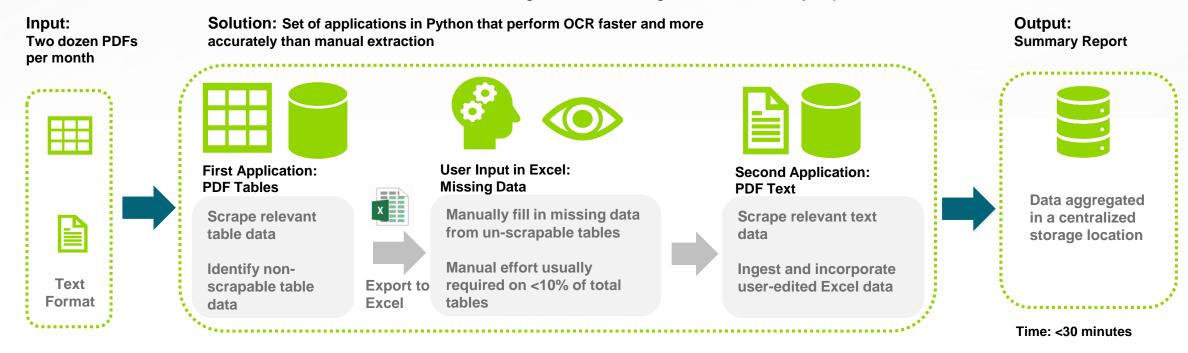
The client regularly needed to extract critical information from PDFs, aggregate the data into a centralized storage location, and combine the data into a summary report. This process was tedious and manual.

#### **SOLUTION**

Guidehouse built a set of applications in Python that use OCR (Optical Character Recognition) to automate PDF data extraction accurately and efficiently.

#### **IMPACT**

The solution minimizes the time needed to aggregate critical data into a central storage location and generate summary reports.





# **Topic Modeling Tool**

#### **CLIENT**

Government Client

#### **CHALLENGE**

The client's program operations require manual review of large amounts of text – e.g. grantee reports, site visit reports, literature, partner studies. Reviewing nonstandardized reports is a cumbersome process that interfered with the client's program and timelines. The client used standardized, quantitative data but treated other qualitative reports as compliance checks rather than data sources. The client's inability to fully leverage qualitative reports hampered its program tracking and monitoring activities, limited the program's ability to quickly reference areas of success, stifling process improvement activities, and prevented the collection of longitudinal and/or empirical evidence.

#### **SOLUTION**

The Guidehouse team developed an NLP prototype tool using Latent Dirichlet Allocation (LDA) to quickly ingest and analyze long text documents. The tool processes the document(s) and creates a topic model, displaying results in two easily digestible formats: a terms table, which categorizes the paragraphs inside document(s) into groups (called "topics") based on their keywords (called "terms") and visualizations with a clickable bar chart detailing the distribution of paragraphs across topics and a word cloud showing the document(s)' vocabulary.

#### **IMPACT**

Guidehouse's tool enables the client to perform the following tasks:

- Regularly review project summaries from the recipients to identify new topics of interest or problems that may require technical assistance
- Review prior media questions and responses to provide consistent responses to media inquiries
- 3. Perform a literature review of emerging threat research and identify a subset of articles most closely related to the client's topic of interest and area of expertise



# **NLP for Public Comment Analysis**

#### **CLIENT**

Government Client

#### **CHALLENGE**

Regulatory bodies across state, local, and non-governmental levels inappropriately interpreted client's guidelines on medication. This misinterpretation led to widespread changes in practices and laws that negatively impacted those living with chronic pain who used medication(s) successfully for decades. The client opened a public comment period to receive feedback from patients, providers, advocacy/industry groups, and caregivers. The client intended to use this feedback to update and clarify its guidelines. The client was quickly inundated with 5,000+comments of varying relevance. Because manual theme identification and tagging was unfeasible given the large number of comments, the client needed to develop automated methods of parsing the comments.

#### **SOLUTION**

Guidehouse developed a RPA bot process, Python scripts, and a R Shiny micro-application to web-scrape the comments, organize them in local file storage, and enable search through and custom tagging of comments to improve thematic analysis. Guidehouse used Natural Language Processing, regular expressions, and Boolean search (e.g. X AND A OR B AND NOT Z) to intelligently filter comments by keywords, sequences of keywords.

#### **IMPACT**

The tool provides functionality to categorize comments into stakeholder groups (patients, providers, advocacy/industry groups, and caregivers), identify content relevance (e.g. which guideline of 12 was being discussed in the comment), and associate keywords (e.g. which medications were being discussed). All comment searches could be saved and reapplied to new comments, ensuring that analysis and data collection could occur in parallel. The solution improved the timeliness of comment analysis and enhanced the client's ability to derive relevant insights from thousands of pages of text.



### **Document Scoring and Topic Modeling Algorithms**

#### **CLIENT**

Government Client

#### **CHALLENGE**

A federal law enforcement agency needed support across all phases of business intelligence and analytics environments to create reports and dashboards. The agency tasked Guidehouse with requirement gathering, design, development, operations and maintenance, and solutions administration to meet its mission, goals, and objectives.

#### **SOLUTION**

As part of the solution, Guidehouse worked with stakeholders spanning across human resources, communications and media relations, and investigations to introduce NLP algorithms (such as topic modeling, relevance scoring, and sentiment analysis) to their workflows. The solutions we developed include a relevance scoring algorithm for the communications and media relations division to quickly find relevant news articles and a topic modeling algorithm to tag and sort the client's troves of unstructured text data and documents.

#### **IMPACT**

Guidehouse's NLP work has reduced the time needed for client analysts to review large corpora of documents and enabled data reporting through Tableau. Guidehouse successfully deployed the first dashboard for the agency and has expanded its reporting capabilities in many areas, such as by improving the hiring pipeline and enhancing COVID-19 preparedness.



# Thank you!

For more information, please contact:

Wynn Thane (202) 222-8607 wthane@guidehousefederal.com

**Bassel Haidar** bhaidar@guidehouse.com

