# Digital Assistance for System Requirement Discovery and Analysis using Machine Learning Natural Language Processing Algorithm

Nipa Phojanamongkolkij, Braxton VanGundy, Ian Levitt, and Heidi Glaudel

**NASA Langley Research Center** 



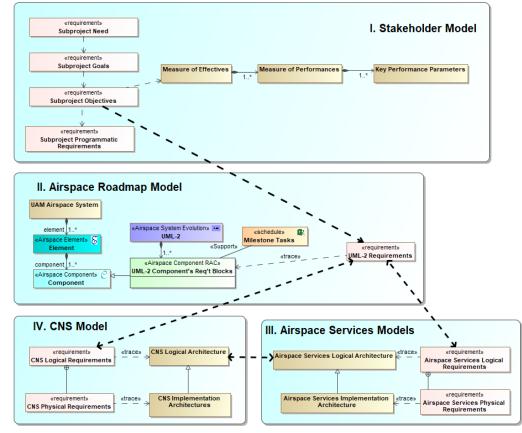
- Introduction
- Machine Learning Natural Language Processing Application
  - The Application Look and Feel
  - What happen behind the scenes?
- Benefits
- Lessons Learned and Next Steps



- NASA's Air Traffic Management-Exploration (ATM-X) Urban Air Mobility (UAM) Subproject is conducting research that evolves UAM airspace towards a highly automated and operationally flexible system of the future.
  - See <u>https://www.nasa.gov/uam-overview/</u> for more information
- The complexity of UAM airspace evolution requires a **plan** to effectively organize, integrate, and communicate NASA's research and development.
  - This planning tool is called **the UAM airspace research roadmap**.
  - Implemented through Model-Based Systems Engineering (MBSE) methodology
  - Leveraging machine learning natural language processing (ML NLP, or just **NLP**).



## Systems Models for Roadmap Development



- Meta-model definition showing structure and high-level dependencies within and between models.
- Dependencies between models are through requirement traceability.
- High cognitive demand to perform Roadmap's requirement discovery and analysis to identify its rationale, reference sources, and dependencies among requirements.
- This challenge can be overcome by the application of **ML NLP methodology.**



# Machine Learning Natural Language Processing Application

- NASA has been transforming our R&D, system engineering, and project management processes to fully leverage evolving digital technologies
  - to view digitized records virtually across diverse organizations, thus improving design efficiency, collaboration, and safety
  - to streamline our processes to take advantage of new digital tools to optimize process flows.
- Perfect opportunity to infuse a new way in the requirement discovery and analysis.
- Developed an application which is an intelligent data interface that rigorously searches a corpus of UAM technical manuscripts, FAA regulations, etc. in near realtime.
  - 20 pdf documents, each ranging from 20 to a few hundred pages long.





# The Application – Look and Feel

AutoSave 💽 🕅 🖫 🌱 v 🖓 v 🤍 🗸 TM_v1.2_requirements.xkx v	P Search	VanGundy, Braxton (LARC-8702)[EAST2] 🌇 🗖 — 🗗 🗙
File Home Insert Page Layout Formulas Data Review View Help Lessons Learned Bot Lessons Learned Bot Team		
	om Data-set urrently Selected Data-set Training Mode: Training not in progress Training Mode: Training not in progress Stop Training Training Status Training Status	↓ Import Data-set ⑦ Help Menu ↑ Export Currently Selected Data-set ⑦ Help Menu mport/Export/Removal - Status: ① About Import / Export Data-sets Help ^
C3 Services (AIS) is the FAA's civil aviation authority providing the foundations for flight in the national airspace system. This AIS includes charted routes, active Special Activity Airspaces (SAAs), NOTAMs, etc. UAM Operators will *		
I. UAM Model is selected by the user ReqText	c	D E F C Lessons Learned Bot Results V X
	Rationale (draft)     The FAA's primary role is to provide a regulatory and operational	aircraft operations in all airspace, and the regulator and oversight authority for civil aircraft operations in
2 275 UML-2.AM ATC will manage the configuration of the airspace for all airspace users.	framework for operations and to provide FAA-originated airspace constraint data to airspace users (e.g., airspace restrictions, facility maps, active Special Activity Airspace [SAA]). (UTM NextGen ConOos v20, Section 2.3.1)	the NAS. The FAA maintains an operating environment that ensures airspace users have access to the
	Aeronautical Information Services (AIS) is the FAA's civil aviation authority providing the foundations for flight in the national airspace system. This AIS includes charted routes, active Special Activity Airspaces (SAAS), NOTAMs, etc. UAM Operators will use	2. UAM ConOps v1.0.4.3.1.3 FAA NAS Data Exchange FAA NAS data sources are available to UAM operations via FAA-industry exchange protocols. This allows for authorized data flow between the UAM
3 273 UML-2.AM UAM Operators will use available Aeronautical Information Services.	this AIS for planning and executing UAM operations. According to UAM ConOps v1.0 ction 5.1, the new AIS that will	
	be required to be used by UAN regulation, for a direct connect to FAA systems, or for tion must be qualified and collection in preparation of	between the FAA and UAM stakeholders is a gateway such that external entities do not have direct 3. UAM ConOps v1.0.4.3.1.1 FAA Regulatory
4 272 UML-2.AM Aeronautical Information Services that the certified or qualified for use.	this qualification is necessary in the UML-2 phase.	The FAA is the federal authority over aircraft
2. User selects Excel	The ConOps, and other UAM	operations in all airspace and the regulatory and oversight authority for civil operations in the NAS. The
5 274 UML-2.AM UAM Ope cells with UAM	the documents are displayed to	FAA maintains an operating environment that ensures airspace users have access to the resources needed to
requirement data	UAN the user	meet specific operational objecti 4. 20210016168
6 19 UML-2.AM The UAM Operator, Fleet Manager, and PIC will receive services from ATC where applicable.	necessary to mitigate the risk. (UAM ConOps v1.0, Section 4.3.1.2 and Section 6.2.2)	MJohnson VertiportAtmtnConOpsRprt final correcte
The UAM Operators, and PSU Operators where applicable, should collect data that supports anticipated 7 269 UML-2.AM changes to how the UAM traffic may be managed.	To facilitate the applicable FAA organization's Safety Assurance processes (Order_8000.369C, section 3.4 Safety Assurance) of how the UAM traffic may be managed, the UAM Operators, and PSU Operators where applicable will collect, manage, and monitor operational data and knowledge of the aerospace system during the UML-2 phase.	d 4.2.4 Vertipo Figure 14 provides a geographic view of the vertiport
RequirementList		Copy Results to Clipboard

6



- PyInstaller package
  - No need for the user to install a Python environment in order to run our application.
  - Allows for the seamless integration of Python Machine Learning capabilities directly into Microsoft Excel.
- Textract, PdfMiner, XIrd packages
  - Supports the extraction of the following data types for training:
    - .docx | .pptx | .txt
    - .pdf
    - .xlsx and .xls
- PyDoE package
  - Python experimental design package.
  - Used to generate a matrix for the Application (called **Lessons Learned Bot or LLB**) auto-train functionality.



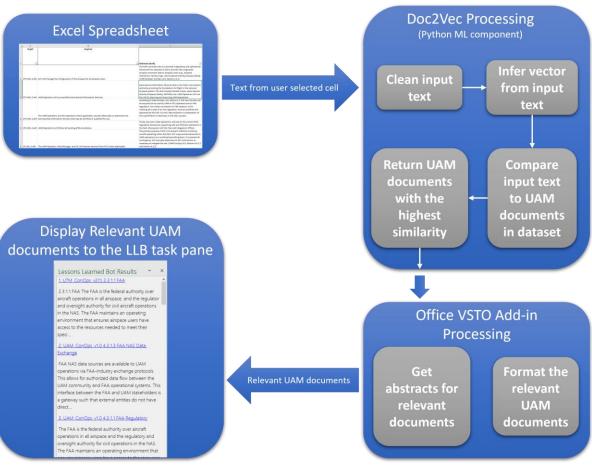
- Doc2Vec package from GenSim
  - A NLP Machine Learning approach for finding related documents using their corresponding vector representations.
  - Based on a single-layer neural network model
    - Input -> Hidden Layer -> Output with a Softmax Classifier.
  - A total of 19 hyper-parameters for training. Examples are:
    - Vector Size Size of Neural Network Input / Output Nodes, or the dimensionality of the feature vectors
    - Window Sliding window size of words in a sentence to be considered as a context
    - Min. Count Minimum threshold frequency (count) of words to be discarded during the preprocessing of the training dataset. (Less frequent words)
    - Sample Threshold value to compute the probability of words with high frequency (with less added information) that will be discarded. (Too frequent words)



- Auto Training vs. Set Default Parameters (or Out-of-the-Box)
  - Two user training modes built into the LLB.
  - Auto Train -> Model generator runs through various hyper-parameter setting scenarios, evaluates the generated model, and selects the best one. Produces best model but is computationally intensive.
  - Out-of-the-Box -> A set of static default parameters is used to train the models. These parameters were verified in the literature to work adequately over a wide range of models but may not perform as good as Auto Train does. Only requires 1 training run vs. multiple runs with Auto Train, therefore it is much faster.



### **Document Recommendation Process**





# Benefits

### **Today's practice**

- Systems engineers (SE) performing reviews of this corpus.
- SEs identify requirement's rationale and references.

### New practice

- The team, consisting of SEs and the NLP expert, uses this corpus to train the Doc2Vec model.
- After validation of the newly trained model, the team can package this add-in application and deploy it to users.
- SEs and/or general users can then install and use the application to search through the UAM body of knowledge efficiently for rationale and references

### **Benefits:**

- Proven a significant time saving (~12x faster for one pdf) over today's practice.
- Reusable trained model of the UAM body of knowledge beyond the current study to survey literature relevant to UAM research questions (as opposed to just requirements).
- Extendable to incorporate the Advanced Air Mobility (AAM), NAS, and National Transportation System (NTS) knowledge corpus.



- Lessons Learned:
  - Pre-processing pdf before training
    - One record per page: does not produce a good result because:
      - a page could have different research concepts especially at the end of one section and the beginning of the next
      - a page could show an incomplete concept when it takes several pages to articulate the concept.
    - One record per section: produces a better result and is used in our application.
    - Manual process in dataset preparation because each pdf comes in various section layouts. Scripting will need to be customized for each pdf.
  - Grouping the corpus into different themes and building a model for each theme.
- Next Steps:
  - More documents to be included in this corpus.
  - A new use case for this NLP tool to find "similar" requirements.
    - A trained model for a corpus of "requirements" for identifying requirement's dependencies.