



**Calhoun: The NPS Institutional Archive**  
**DSpace Repository**

---

Acquisition Research Program

Acquisition Research Symposium

---

2022-05-06

# Recommending Recommendations to Support the Defense Acquisition Workforce

Lipizzi, Carlo

Monterey, California. Naval Postgraduate School

---

<http://hdl.handle.net/10945/70249>

---

Copyright is reserved by the copyright owner

*Downloaded from NPS Archive: Calhoun*



Calhoun is the Naval Postgraduate School's public access digital repository for research materials and institutional publications created by the NPS community. Calhoun is named for Professor of Mathematics Guy K. Calhoun, NPS's first appointed -- and published -- scholarly author.

**Dudley Knox Library / Naval Postgraduate School**  
**411 Dyer Road / 1 University Circle**  
**Monterey, California USA 93943**

<http://www.nps.edu/library>



**STEVENS**  
INSTITUTE *of* TECHNOLOGY  
THE INNOVATION UNIVERSITY®



# Recommending recommendations to support the Defense Acquisition Workforce

## Natural Language Processing at work

by

**Dr. Carlo Lipizzi**

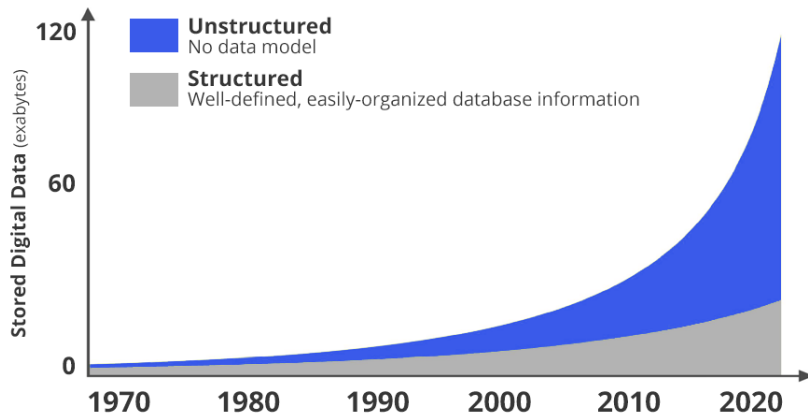
*clipizzi@stevens.edu*

### **Project Team:**

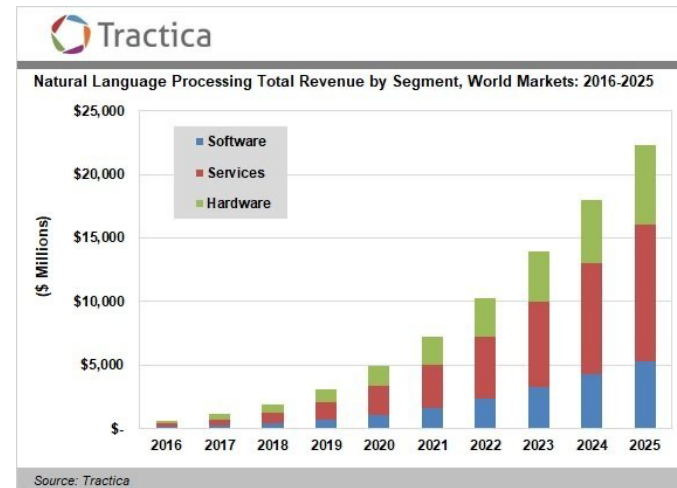
Dr. Carlo Lipizzi; Mr. Hojat Behrooz; Mr. Michael Dressman; Mr. Arya Guddemane Vishwakumar; Mr. Kunal Batra; Dr. Philip Anton; Mr. Irv Blickstein

- The focus is on recommendations contained in relevant Department of Defense (DoD) and private sector studies on acquisition policies and practices, including—
  - the extent to which recommendations have been enacted into law by Congress;
  - extent to which the recommendations have been adopted through the issuance or revision of regulations;
  - the extent to which the recommendations have been adopted through issuance of an appropriate implementing directive or other form of guidance
- Recommendations can be hundreds, with lengths from few pages to hundreds of pages
- Some recommendations or some parts of them may be more relevant to the Defense Acquisition Workforce

- 85-90 percent of all corporate data is in some kind of unstructured form, such as text and multimedia [Gartner, 2019]
- Tapping into these information sources is a need to stay competitive



Source: m-files.com



- Examples of application of **Natural Language Processing**: insurance (claim processing); law (court orders); academic research (research articles); finance (reports analysis); medicine (discharge summaries); technology (patent files); marketing (customer comments)

- Semantic ambiguity and context sensitivity
  - automobile = car = vehicle = Toyota
  - Apple (the company) or apple (the fruit)
- Syntactic/formal ambiguity
  - Misspelling
  - Different words for the same concept (e.g.: street; st.)
- Implicit knowledge
  - We talk about things giving for granted common or specific knowledge

- Understanding Language is not “just” processing. Understanding is a human characteristic, analyzed by philosophers as part of Epistemology
- An accurate (by human standard) “understanding” can come only from a model of human mind
- The current leading models in NLP/”NLU” are focused on the algorithmic part, missing a real model representing how the knowledge is created and used. It is basically representing the brain, not the mind. The leading model for NLP (GPT-3 by Open-AI) has 175 billion parameters, feeding a neural network providing results as a black box

- Language is changing constantly, and NLP is following the changes, going from processing based on predefined structures (taxonomies/ontologies, syntax) to structures deduced from the text itself

## **Limitations of the traditional-deductive- "symbolic" approach**

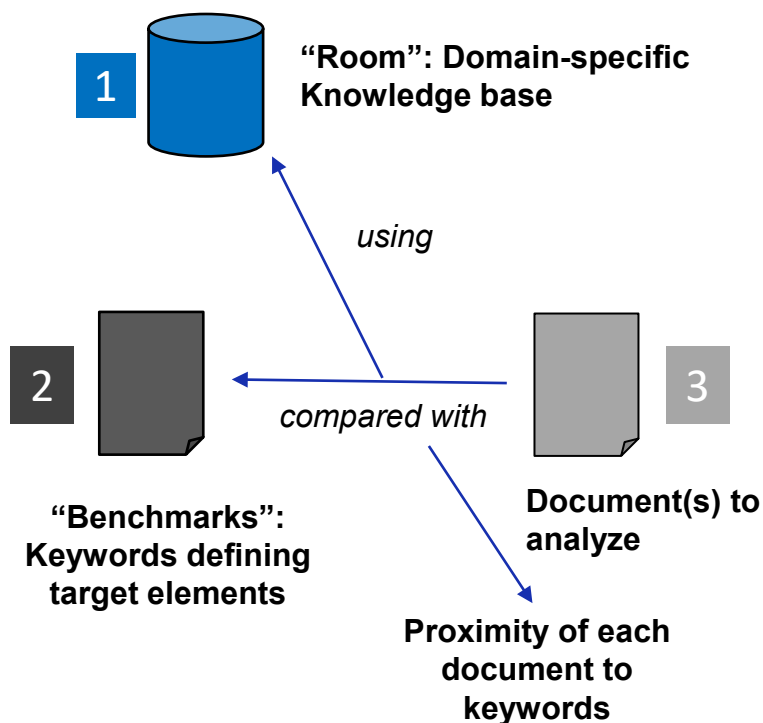
- Predefined structures (ontologies and taxonomies) are used to extract semantic elements
- Today language is more fragmented, has less structure, has more jargons
- Different points of view may provide different interpretations

## **Machine Learning/inductive approach**

- Employing complex "deep learning" systems inspired by the human brain structure
- They do not consider how humans represent their knowledge and how we achieve the understanding of a problem
- They model the brain, not the mind/the way knowledge is created and used

- Our approach is a combination of Symbolic and Machine Learning, with an additional layer of user interface and visualization, to make the findings more usable by the Defense Acquisition Workforce
- For the development of the prototype, we focused on 1. creating a symbolic model for the text understanding and 2. design and implement the process to apply it
- The prototype is based on previous projects we developed for the DoD over the last few years, employing a team of 25 researchers and relying on theories and components we developed. The algorithm/method we used is named “the room theory”, that is a combination of symbolic and machine learning

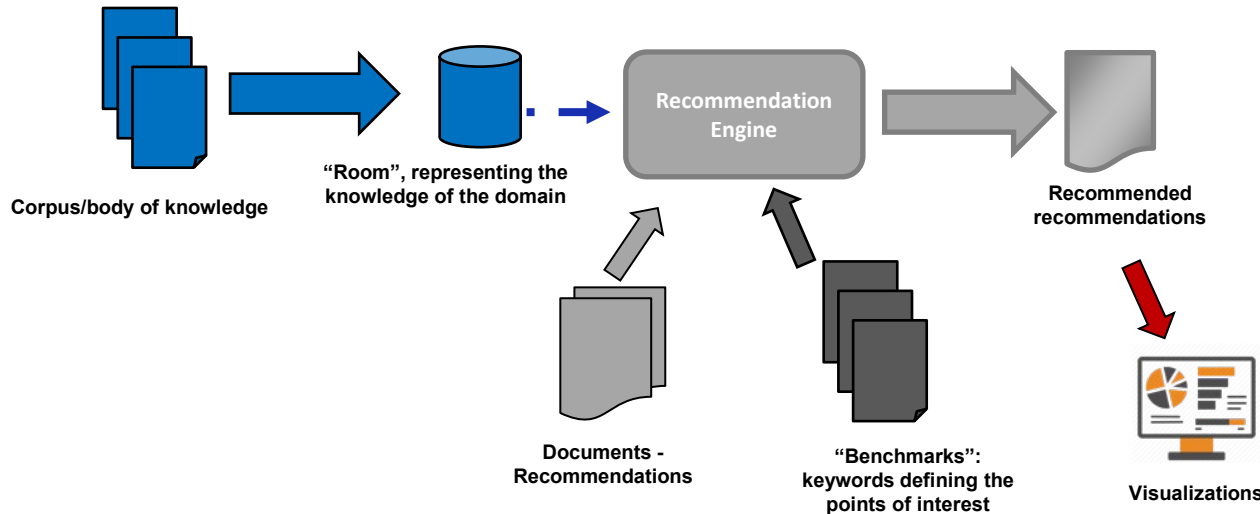




- “Room theory” enables the use of context-subjectivity in the analysis of the incoming documents
- Context-subjectivity can be the point of view of a subject matter expert
- The context-subjectivity in the analysis is represented by a domain specific numerical knowledge base, created from a large domain specific & representative corpus that is then transformed into a numerical dataset (“embeddings table”)

- The key components are:

1. A point of view for the comparison (the “room”). This is represented by a table of vectors extracted from a large/representative corpus from the specific domain
2. A list of “extended” keywords (using synonyms and misspellings) to be used for the analysis (the “benchmark”)

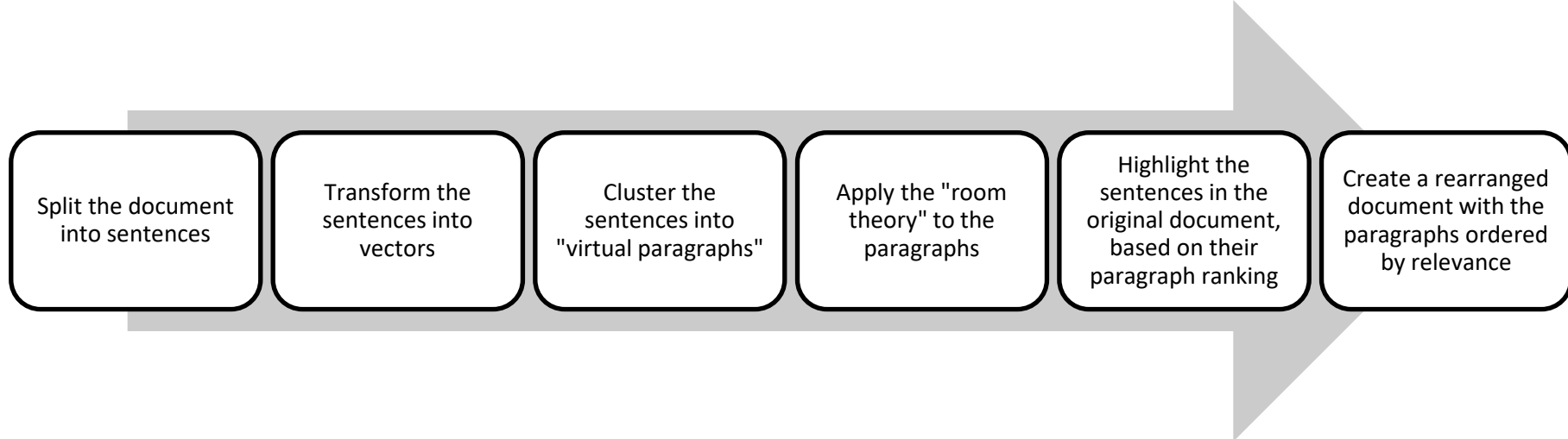


- The “room”/knowledge base has been generated from a corpus collected for a previous DAU project
- The corpus representing a contracting officer’s knowledge base is composed by 537 documents, for a total of 119,941 unique words

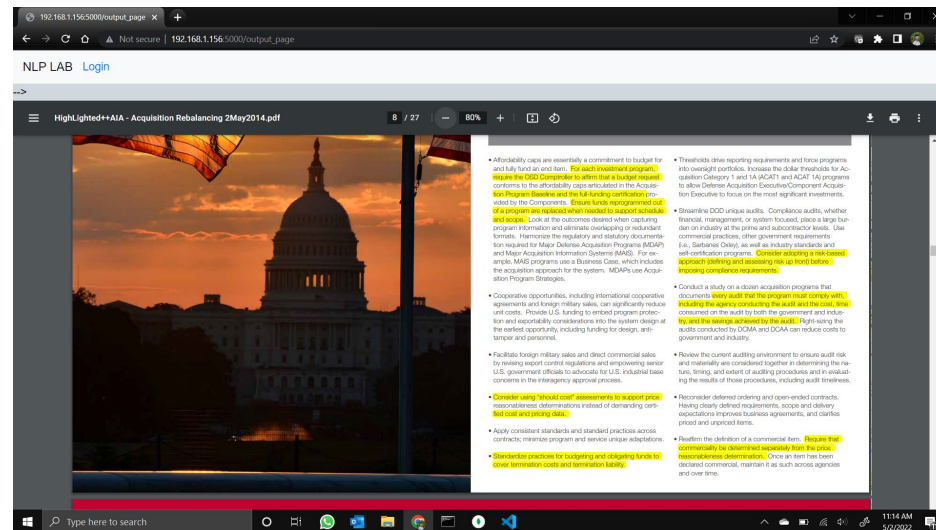
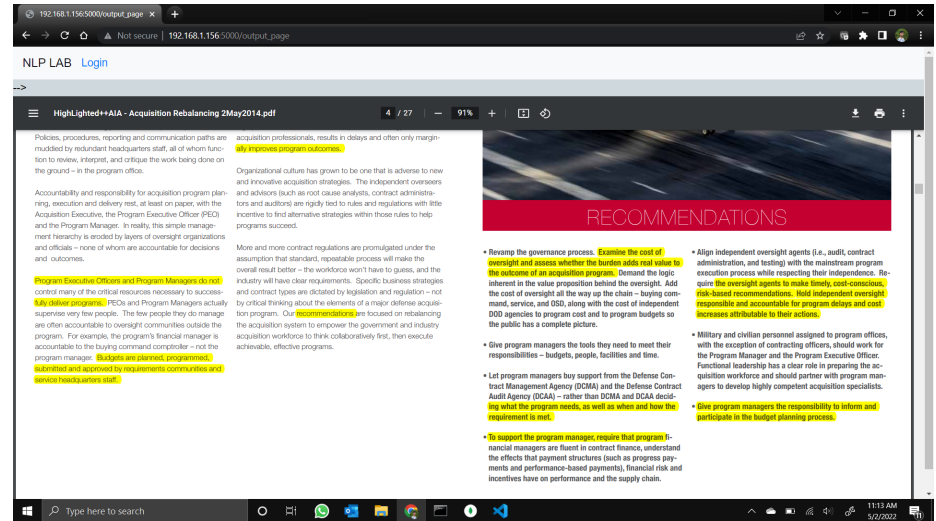
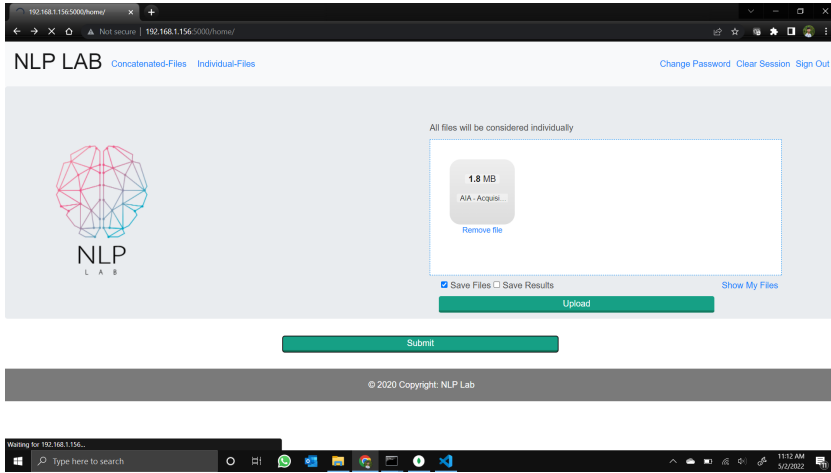
- The Benchmarks is a list of keywords and related weights put together with the SMEs in our team (*175 benchmark words/phrases*)
- We used a total of about 30 pdf and word documents, ranging from 1 to 500+ pages
- We rank the document using our algorithms via the available Room

- We provide graphic visualizations to help user get insights from the results
- A graphical user interface has been created to get data and to deliver the results

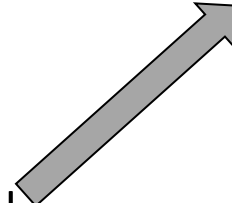
- Large documents cannot be considered either “recommended” or “not – recommended”:
  - In 500 pages there could be some sentences that are relevant, (many) other that may not be
  - The same logical concept can be in multiple pages
- We developed a method for “re-paragraphing” documents

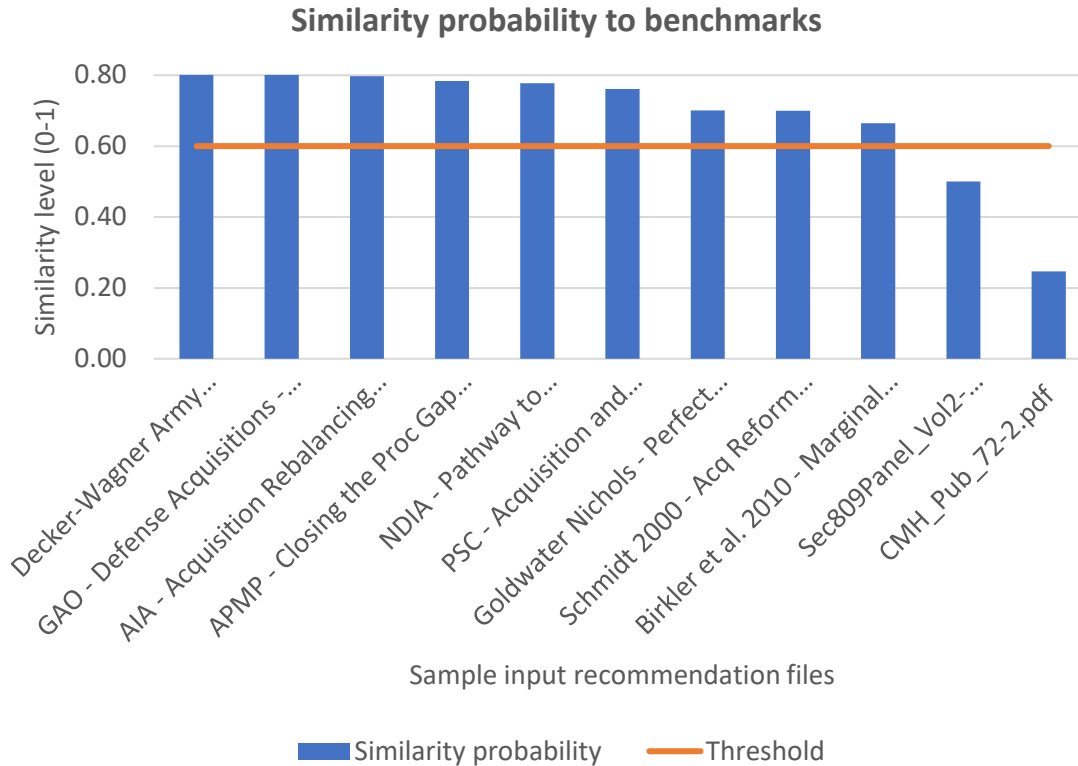


## Input the document



Visualize the original document with highlighted recommended parts



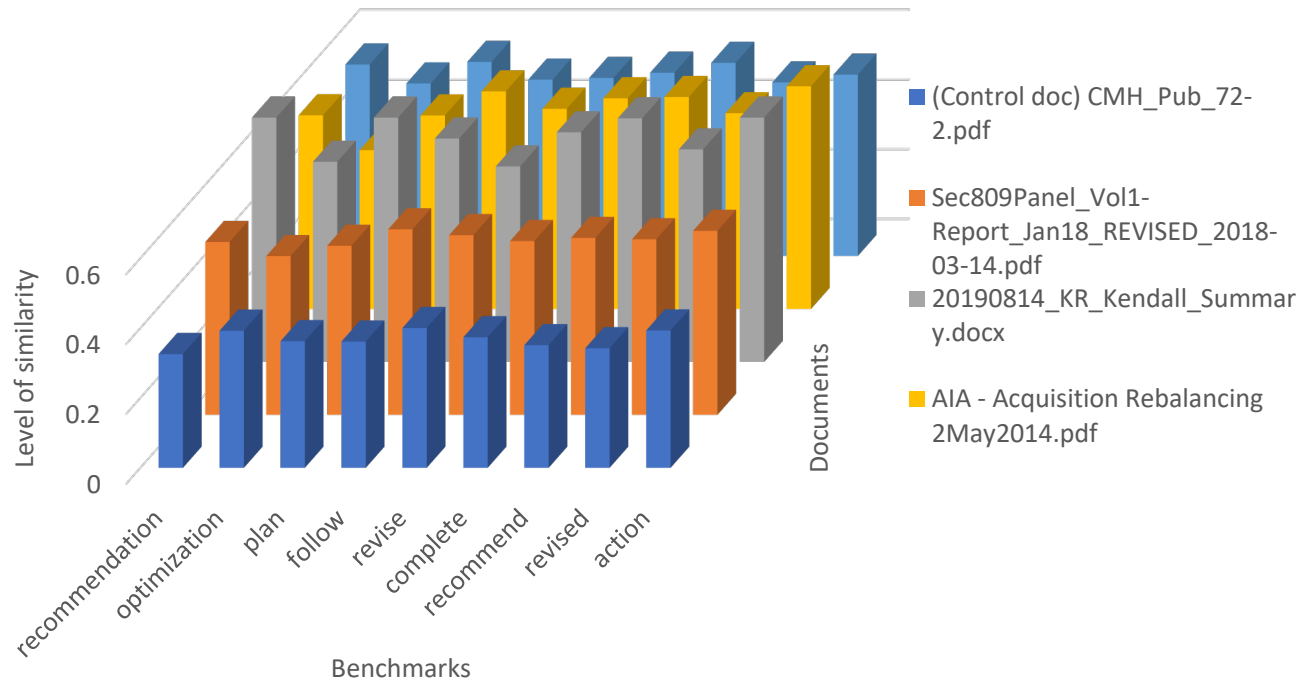


- This is a representation of the potential interest of 10 recommendation files + 1 control file (that is not related to recommendation)
- Results are not yet weighted by a normalized percentage of interest by paragraphs

# The output – comparing multiple files



Level of similarity of each document to each benchmark



- This is a representation of how individual benchmarks match individual documents. There are 3 recommendation files + 1 control file (that is not related to recommendation)
- Results are not yet weighted by a normalized percentage of interest by paragraphs



- Improve/expand the “room”/knowledge base with more problem-specific corpora
- Expand the benchmarks with synonyms and misspellings
- Revise the “paragraphing” subsystem with better clustering and better trace back to the original document
- Reevaluate the document recommendation level using the relevance of its paragraphs
- Integrate the “paragraphing” with the graphs
- Improve the user interface
- Integrate the graphs in the user interface
- Optimize the system for larger scale of operation (more/larger documents)
- Continue the debugging and the testing on more documents



**STEVENS**  
INSTITUTE *of* TECHNOLOGY  

---

THE INNOVATION UNIVERSITY®

*Thank you!*

**Dr. Carlo Lipizzi**  
[clipizzi@stevens.edu](mailto:clipizzi@stevens.edu)