Supercharging Expert Networks

Designing a High-Performance Matching Engine to Leverage GLG's Network of Technology and Healthcare Experts to Wide Ranging Customer Requests

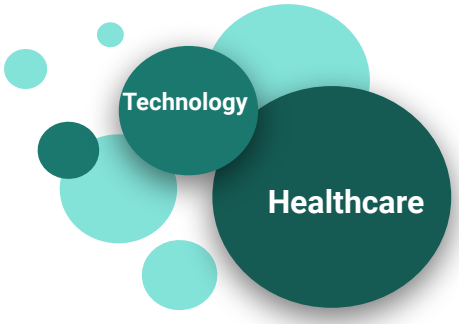
Project GLG | Team MS

Mark
Spencer

Motivation

GLG's Value Proposition

A Diverse Expert Network with
2 Core Competencies



Driving Value to GLG Clients

1 Reduce Client Response Time

Drive **customer satisfaction** by connecting with Experts quickly

2 Increase Matching Accuracy

Improve **customer retention** by getting the match right the first time.

3 Improve Client Retention

Deliver **higher quality matches** to drive customer satisfaction in as little time as possible

4 Increase Customer Lifetime Value

Customers that are more satisfied with their prior experiences with GLG are likely to stay longer and be repeat customers

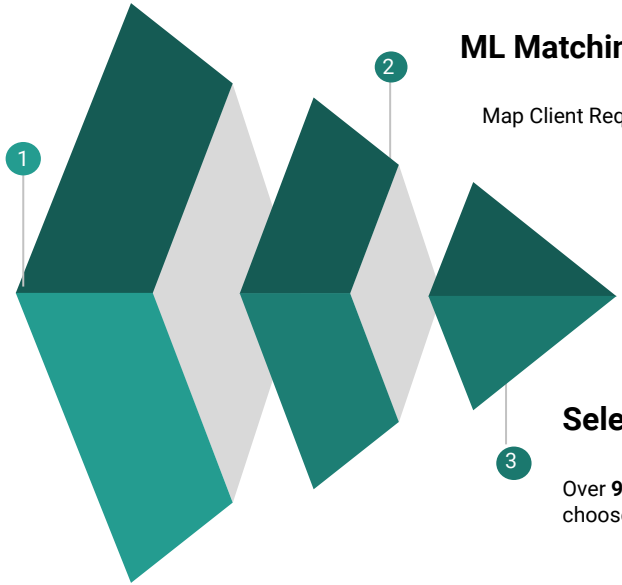
Finding a Needle in a Haystack

Inbound: Client Requests

A *diverse* range of *subjects* to choose from!

"What are the unmet needs of global B2B customers in the organic materials segment?"
Japanese Industrial Conglomerate

"How large is the existing market for robotic process automation (RPA)? Is there room for it to grow? What does the typical RPA customer look like? Which players deliver the most ROI in this space?"
Hedge Fund



ML Matching Engine

Map Client Request -> Expert

Select an Expert

Over **900,000** Experts to choose from!

Visions: Towards A New Engine

Current Engine

Precise Matches

We combine the human insight of our team with an AI-driven matching platform to find the right people.

- Keyword Search
- Human Review
- Contact Client with Chosen Expert

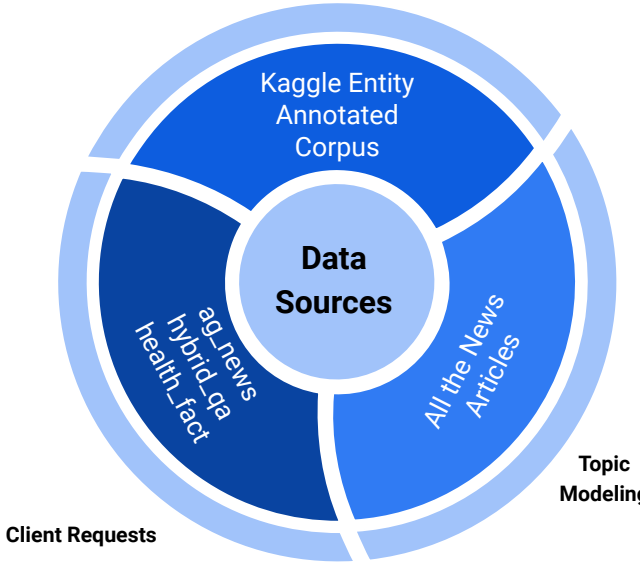
Proposed ML Engine

- Learn **Topics** from text
- Learn **Entity Tags** from text

- Topics + Entities
- Submit a Client Query
- ~~Human Review~~
- Customer Facing Dashboard
- Serve Matched Topics + Entities

Training Data

Named-Entity-Recognition (NER)

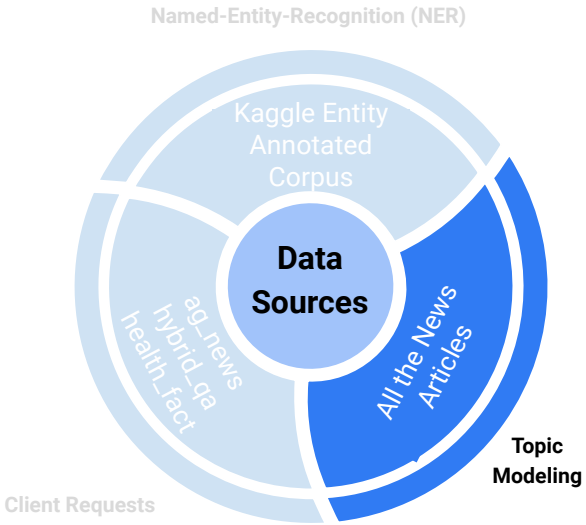


Topic Modeling

Client Requests

Topic Modeling

Training Data



Latent Dirichlet Allocation (LDA)

1 V1.0 Baseline: LDA

- **Training Set:** 250k News Articles
- **Composition**: Many Different News Categories
- **Labels:** 185 Hand Labeled Topics
- **Epochs**: 20X through training set

2 V1.5: LDA

- **Training Set:** 244k News Articles
- **Composition:** Filtered to **Science, Tech** and **Healthcare**
- **Labels:** 265 Hand Labeled Topics
- **Epochs:** 23X through training set

2 V2.0: LDA

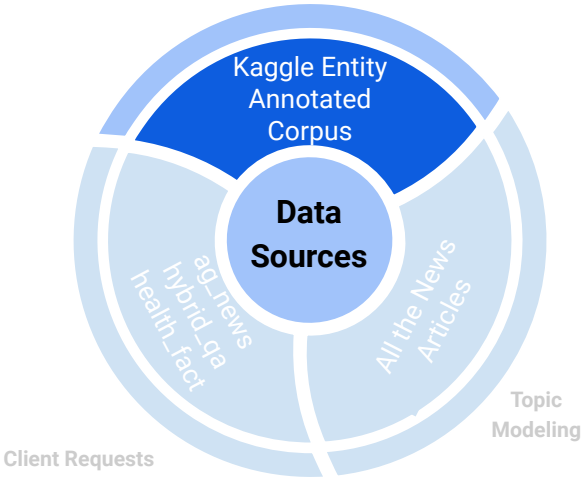
- **Training Set:** 281k News Articles
- **Composition:** Filtered to **Science, Tech** and **Healthcare**
- **Labels:** 265 Hand Labeled Topics
- **Epochs:** 14X through training set

Increasing Complexity

NER Tagging

Training Data

Named-Entity-Recognition (NER)



Increasing Complexity

nltk

Use nltk models with classical NLP features

spaCy

Use spaCy NER models to tag entities

Increasing Complexity

1 **V1.0 Baseline: SpaCy 'en_core_web'**

- **Training Set:** en_core_web
- **Composition**: Generalist
- **Labels:** Annotated Entities

2 **V1.5: CRFSuite**

- **Training Set:** Annotated Entity Corpus (1.35MM Words)
- **Composition:** Generalist
- **Labels:** Annotated Entities

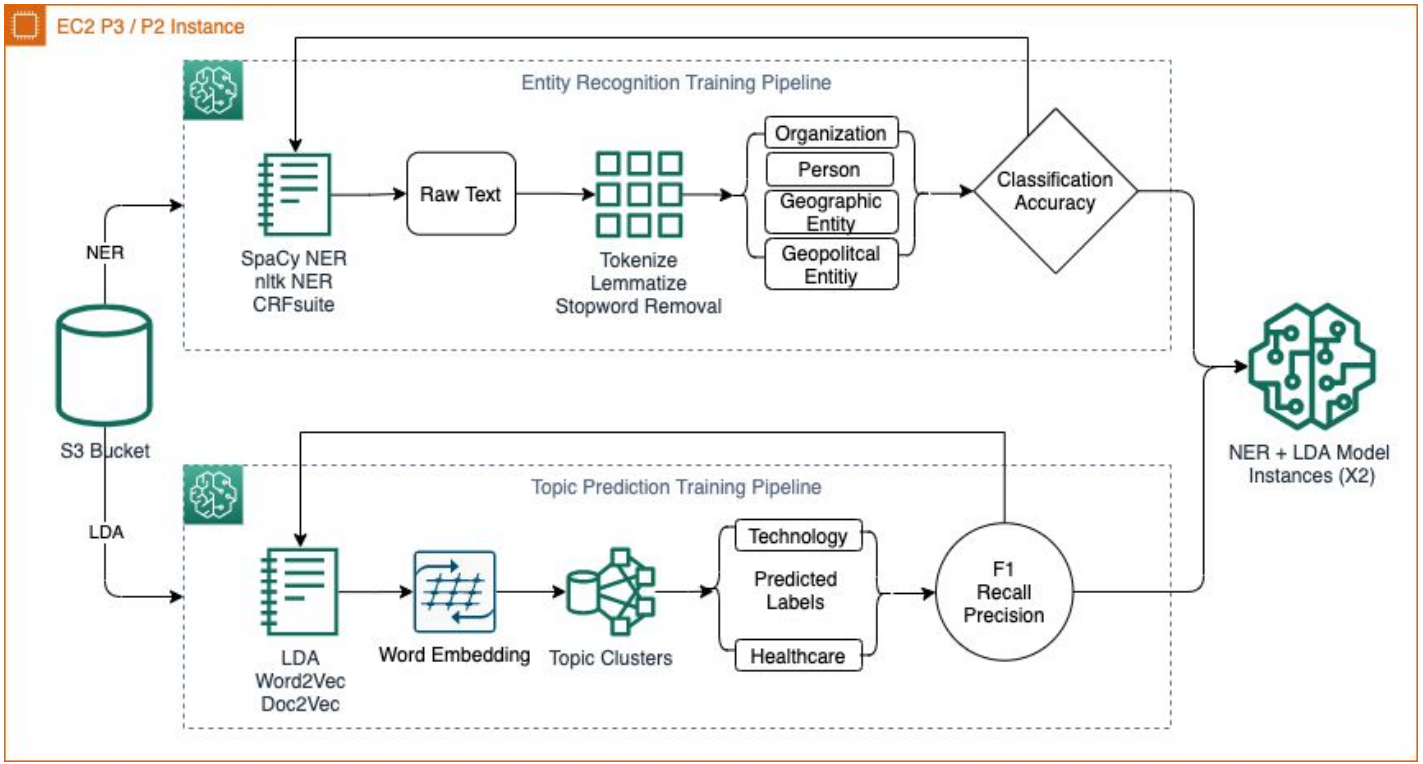
Benchmarking

Test Dataset	Dataset Characteristics	N Examples	Accuracy (%)
ag_news	Short 1-2 line article summaries mapped to 'tech', 'healthcare' or 'other' domains	1900	64.8
GrailQA	Mostly 1-liner documents, mapped to 'tech' or 'other' domains	6763	47.8

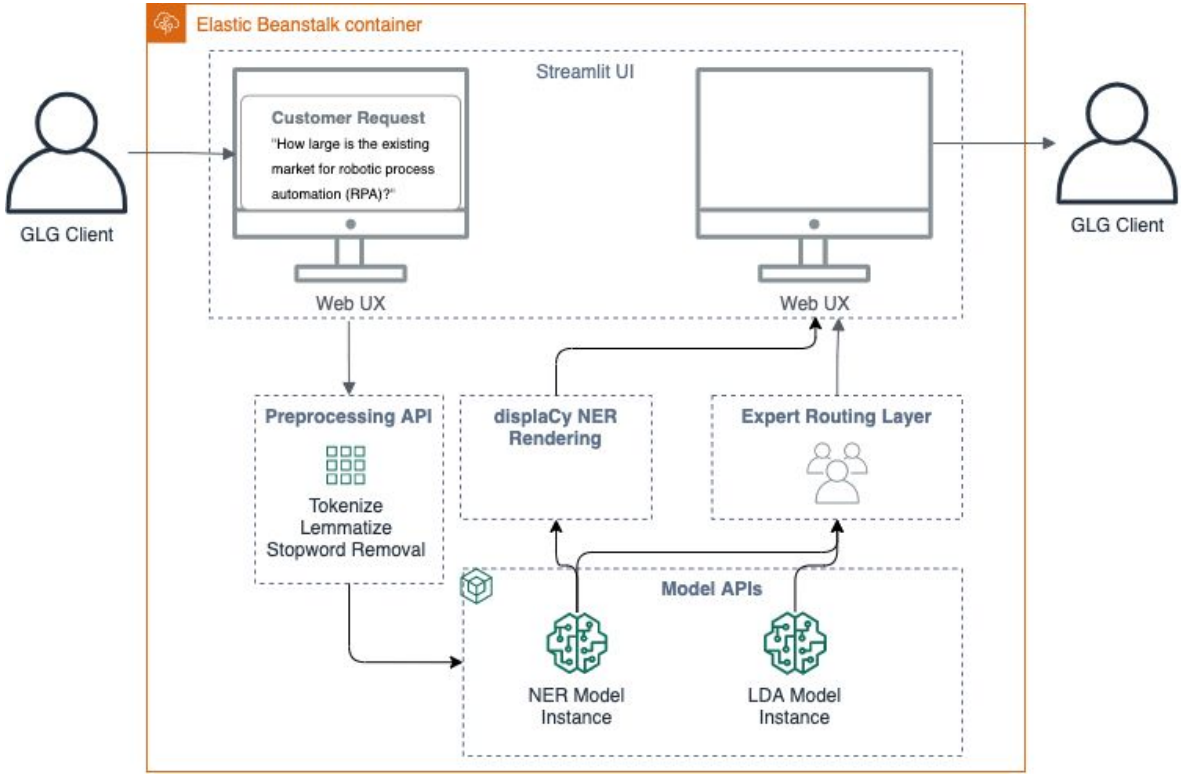
ag_news				
	Precision	**Recall**	**F1-Score**	**Support**
Other	0.88	0.86	0.87	5700
Technology	0.61	0.64	0.63	1900
Accuracy			0.81	7600
Macro Avg	0.74	0.75	0.75	7600
Weighted Avg	0.81	0.81	0.81	7600

grail_qa				
	Precision	**Recall**	**F1-Score**	**Support**
Healthcare	0.26	0.66	0.38	304
Technology	0.17	0.59	0.26	869
Other	0.86	0.45	0.59	5590
Accuracy			0.48	6763
Macro Avg	0.43	0.57	0.41	6763
Weighted Avg	0.75	0.48	0.54	6763

System Design: Training



System Design: Deployment



Ethical Considerations

Data Privacy

- Tool will function using the **minimum amount of customer data** necessary to fulfill the task. No session metadata to be stored.
- Purpose and aggregate-level function of tool not designed to benefit from personal data
- User prompt to state that the request should contain “**No PII** (Personal, Name, Address, SSN or other financial information)”.
- By using this tool, **user consent**s to long-term retention of request for model-training purposes.

Bias

Primary sources of potential bias:

- **Cognitive Bias** in the labeling training dataset
- **Linguistic Bias** inherent in the training data due to the time period from which articles were written, formal styles of language present in news articles, etc.
- **Representation Bias** in the training data where news coverage may lean towards celebrities vs. normal citizens.
- **Gender Imbalance** where men may be featured more prominently over women, etc.

Demo Time!

Q+A