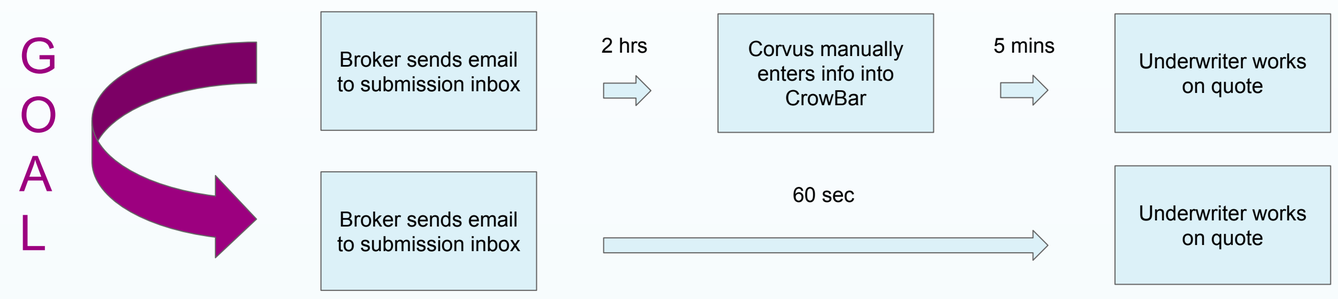
Automated Dataset Creation using PDF Text Extraction

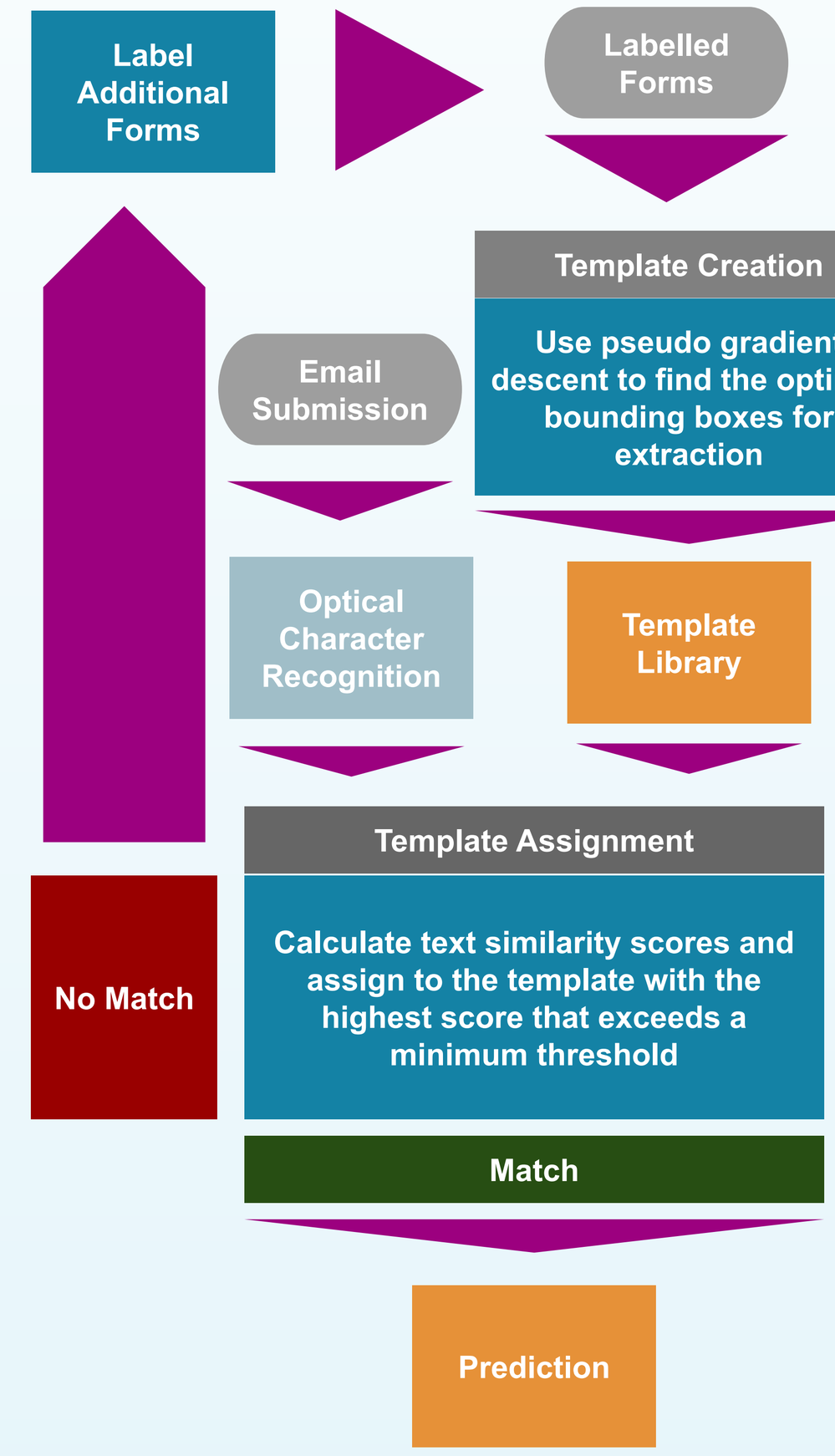
Jack Schooley and John Lazenby

MIT Advisor: Hazhir Rahmandad
Corvus Insurance Advisors: Chris Hedenberg, Michael Simader, Adrien Demelier, and Zak Raicik

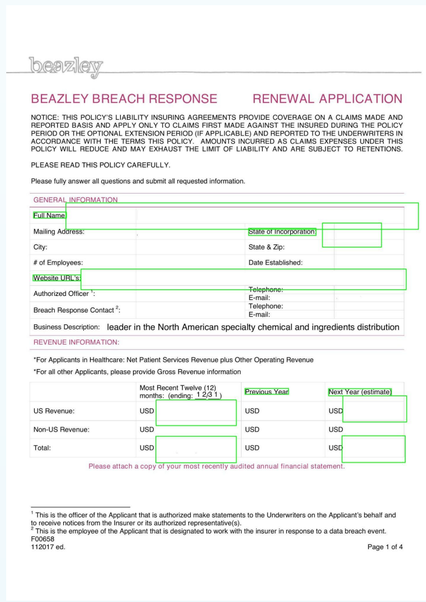
Introduction

Corvus Insurance gets information for quote requests in two ways. The first way is brokers can directly enter the information into Corvus's proprietary software platform, the CrowBar. The second way is brokers send Corvus an email with the information, and this information typically comes in the form of PDF attachments. Currently, the process of taking the information from the PDF attachment and putting it into the CrowBar is done by an administrative support group. This intermediate step means that email quote requests will be slower than broker-entered ones. The goal of our project is remove the intermediate step by automating the data extraction of quote information from the PDFs.





Template Creation


beazley

BEAZLEY BREACH RESPONSE RENEWAL APPLICATION

NOTICE: THIS POLICY'S LIABILITY INSURING AGREEMENTS PROVIDE COVERAGE ON A CLAIMS MADE AND REPORTED BASIS AND APPLY ONLY TO CLAIMS FIRST MADE AGAINST THE INSURED DURING THE POLICY PERIOD OR THE OPTIONAL EXTENSION PERIOD (IF APPLICABLE) AND REPORTED TO THE UNDERWRITERS IN ACCORDANCE WITH THE TERMS THIS POLICY. AMOUNTS INCURRED AS CLAIMS EXPENSES UNDER THIS POLICY WILL REDUCE AND MAY EXHAUST THE LIMIT OF LIABILITY AND ARE SUBJECT TO RETENTIONS.

PLEASE READ THIS POLICY CAREFULLY.

Please fully answer all questions and submit all requested information.

GENERAL INFORMATION

Full Name:
Mailing Address: State of Incorporation:
City: State & Zip:
# of Employees: Date Established:
Website URL's:
Authorized Officer [1]: Telephone: E-mail:
Breach Response Contact [2]: Telephone: E-mail:
Business Description: leader in the North American specialty chemical and ingredients distribution

REVENUE INFORMATION:

*For Applicants in Healthcare: Net Patient Services Revenue plus Other Operating Revenue
*For all other Applicants, please provide Gross Revenue information

	Most Recent Twelve (12) months: (ending 12/31)	Previous Year	Next Year (estimate)
US Revenue:	USD	USD	USD
Non-US Revenue:	USD	USD	USD
Total:	USD	USD	USD

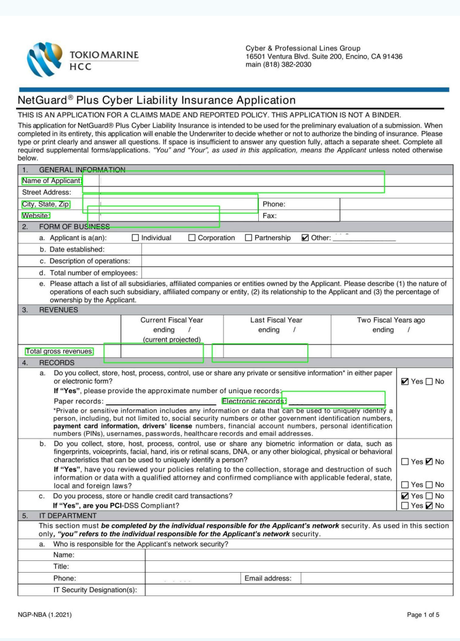
Please attach a copy of your most recently audited annual financial statement.

[1] This is the officer of the Applicant that is authorized make statements to the Underwriters on the Applicant's behalf and to receive notices from the Insurer or its authorized representative(s).
[2] This is the employee of the Applicant that is designated to work with the insurer in response to a data breach event.
F00658
112017 ed.
Page 1 of 4

To extract data, we utilize templates that map the spatial relationship between a keyword phrase and the data to be extracted. We automated the template creation process. For each field we need to extract, we use a pseudo gradient descent algorithm to learn the spatial relationship between the keyword phrase and the associated value.

To do this, we need a set of similar forms that should belong to the same template. We did this by picking one form and querying our PDF database to find similar forms scored by cosine similarity. This is the same process used in the template identification that is elaborated on in the next section.

Optimization occurs by shifting the dimensions of the box, looking at the text within, and comparing it to manually labeled text using string-matching. Intuitively, the algorithm stacks each form on top of each other and selects the box that will capture the correct text for the most forms.


TOKIO MARINE HCC

Cyber & Professional Lines Group
16501 Ventura Blvd. Suite 200, Encino, CA 91436
main (818) 382-2030

NetGuard® Plus Cyber Liability Insurance Application

THIS IS AN APPLICATION FOR A CLAIMS MADE AND REPORTED POLICY. THIS APPLICATION IS NOT A BINDER.

This application for NetGuard® Plus Cyber Liability Insurance is intended to be used for the preliminary evaluation of a submission. When completed in its entirety, this application will enable the Underwriter to decide whether or not to authorize the binding of insurance. Please type or print clearly and answer all questions. If space is insufficient to answer any question fully, attach a separate sheet. Complete all required supplemental forms/applications. *"You" and "Your", as used in this application, means the Applicant* unless noted otherwise below.

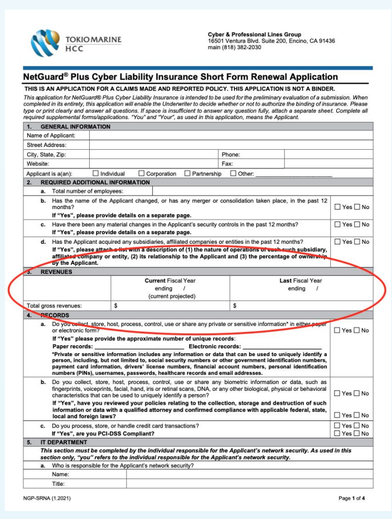
1. GENERAL INFORMATION
Name of Applicant:
Street Address:
City, State, Zip: Phone:
Website: Fax:
2. FORM OF BUSINESS
a. Applicant is a(an): ☐ Individual ☐ Corporation ☐ Partnership ☑ Other: _____
b. Date established:
c. Description of operations:
d. Total number of employees:
e. Please attach a list of all subsidiaries, affiliated companies or entities owned by the Applicant. Please describe (1) the nature of operations of each such subsidiary, affiliated company or entity, (2) its relationship to the Applicant and (3) the percentage of ownership by the Applicant.
3. REVENUES

	Current Fiscal Year ending / (current projected)	Last Fiscal Year ending /	Two Fiscal Years ago ending /
Total gross revenues:			

4. RECORDS
a. Do you collect, store, host, process, control, use or share any private or sensitive information* in either paper or electronic form? ☑ Yes ☐ No
If "Yes", please provide the approximate number of unique records:
Paper records: _____ Electronic records:
*Private or sensitive information includes any information or data that can be used to uniquely identify a person, including, but not limited to, social security numbers or other government identification numbers, **payment card information, drivers' license** numbers, financial account numbers, personal identification numbers (PINs), usernames, passwords, healthcare records and email addresses.
b. Do you collect, store, host, process, control, use or share any biometric information or data, such as fingerprints, voiceprints, facial, hand, iris or retinal scans, DNA, or any other biological, physical or behavioral characteristics that can be used to uniquely identify a person? ☐ Yes ☑ No
If "Yes", have you reviewed your policies relating to the collection, storage and destruction of such information or data with a qualified attorney and confirmed compliance with applicable federal, state, local and foreign laws? ☐ Yes ☐ No
c. Do you process, store or handle credit card transactions? ☑ Yes ☐ No
If "Yes", are you PCI-DSS Compliant? ☐ Yes ☑ No
5. IT DEPARTMENT
This section must ***be completed by the individual responsible for the Applicant's network*** security. As used in this section only, ***"you" refers to the individual responsible for the Applicant's network*** security.
a. Who is responsible for the Applicant's network security?
Name:
Title:
Phone: Email address:
IT Security Designation(s):

NGP-NBA (1.2021) Page 1 of 5

Template Assignment


TOKIO MARINE HCC

Cyber & Professional Lines Group
16501 Ventura Blvd. Suite 200, Encino, CA 91436
main (818) 382-2030

NetGuard® Plus Cyber Liability Insurance Short Form Renewal Application

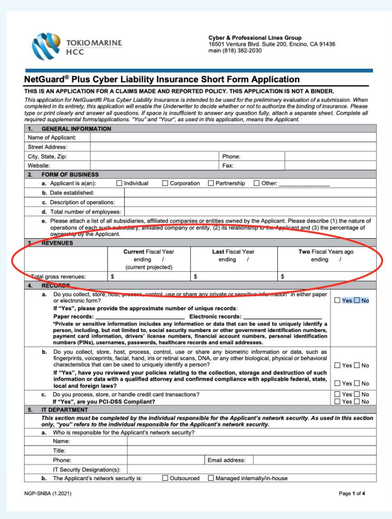
THIS IS AN APPLICATION FOR A CLAIMS MADE AND REPORTED POLICY. THIS APPLICATION IS NOT A BINDER.

This application for NetGuard® Plus Cyber Liability Insurance is intended to be used for the preliminary evaluation of a submission. When completed in its entirety, this application will enable the Underwriter to decide whether or not to authorize the binding of insurance. Please type or print clearly and answer all questions. If space is insufficient to answer any question fully, attach a separate sheet. Complete all required supplemental forms/applications. "You" and "Your", as used in this application, means the Applicant.

1. GENERAL INFORMATION
Name of Applicant:
Street Address:
City, State, Zip: Phone:
Website: Fax:
Applicant is a(an): ☐ Individual ☐ Corporation ☐ Partnership ☐ Other: _____
2. REQUIRED ADDITIONAL INFORMATION
a. Total number of employees:
b. Has the name of the Applicant changed, or has any merger or consolidation taken place, in the past 12 months? ☐ Yes ☐ No
If "Yes", please provide details on a separate page.
c. Have there been any material changes in the Applicant's security controls in the past 12 months? ☐ Yes ☐ No
If "Yes", please provide details on a separate page.
d. Has the Applicant acquired any subsidiaries, affiliated companies or entities in the past 12 months? ☐ Yes ☐ No
If "Yes", please attach, or we will a description of (1) the nature of operations of such each subsidiary, affiliated company or entity, (2) its relationship to the Applicant and (3) the percentage of ownership by the Applicant.
3. REVENUES

	Current Fiscal Year ending / (current projected)	**Last** Fiscal Year ending /
Total gross revenues:	$	$

4. RECORDS
a. Do you collect, store, host, process, control, use or share any private or sensitive information* in either paper or electronic form? ☐ Yes ☐ No
If "Yes", please provide the approximate number of unique records:
Paper records: _____ Electronic records: _____
***Private or sensitive information includes any information or data that can be used to uniquely identify a person, including, but not limited to, social security numbers or other government identification numbers, payment card information, drivers' license numbers, financial account numbers, personal identification numbers (PINs), usernames, passwords, healthcare records and email addresses.**
b. Do you collect, store, host, process, control, use or share any biometric information or data, such as fingerprints, voiceprints, facial, hand, iris or retinal scans, DNA, or any other biological, physical or behavioral characteristics that can be used to uniquely identify a person? ☐ Yes ☐ No
If "Yes", have you reviewed your policies relating to the collection, storage and destruction of such information or data with a qualified attorney and confirmed compliance with applicable federal, state, local and foreign laws? ☐ Yes ☐ No
c. Do you process, store, or handle credit card transactions? ☐ Yes ☐ No
If "Yes", are you PCI-DSS Compliant? ☐ Yes ☐ No
5. IT DEPARTMENT:
This section must be completed by the individual responsible for the Applicant's network security. As used in this section only, "you" refers to the individual responsible for the Applicant's network security.
a. Who is responsible for the Applicant's network security?
Name:
Title:


TOKIO MARINE HCC

Cyber & Professional Lines Group
16501 Ventura Blvd. Suite 200, Encino, CA 91436
main (818) 382-2030

NetGuard® Plus Cyber Liability Insurance Short Form Application

THIS IS AN APPLICATION FOR A CLAIMS MADE AND REPORTED POLICY. THIS APPLICATION IS NOT A BINDER.

This application for NetGuard® Plus Cyber Liability Insurance is intended to be used for the preliminary evaluation of a submission. When completed in its entirety, this application will enable the Underwriter to decide whether or not to authorize the binding of insurance. Please type or print clearly and answer all questions. If space is insufficient to answer any question fully, attach a separate sheet. Complete all required supplemental forms/applications. "You" and "Your", as used in this application, means the Applicant.

1. GENERAL INFORMATION
Name of Applicant:
Street Address:
City, State, Zip: Phone:
Website: Fax:
2. FORM OF BUSINESS
a. Applicant is a(an): ☐ Individual ☐ Corporation ☐ Partnership ☐ Other: _____
b. Date established:
c. Description of operations:
d. Total number of employees:
e. Please attach a list of all subsidiaries, affiliated companies or entities owned by the Applicant. Please describe (1) the nature of operations of each such subsidiary, affiliated company or entity, (2) its relationship to the Applicant and (3) the percentage of ownership by the Applicant.
3. REVENUES

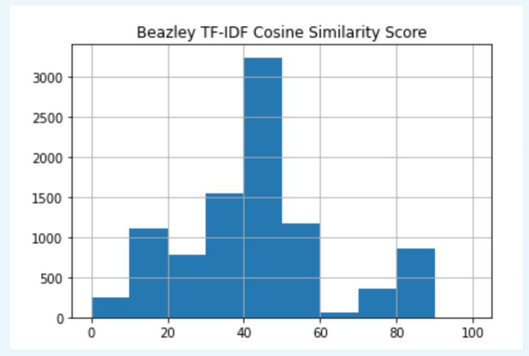
	Current Fiscal Year ending / (current projected)	**Last** Fiscal Year ending /	**Two** Fiscal Years ago ending /
Total gross revenues:	$	$	$

4. RECORDS
a. Do you collect, store, host, process, control, use or share any private or sensitive information* in either paper or electronic form? ☐ Yes ☐ No
If "Yes", please provide the approximate number of unique records:
Paper records: _____ Electronic records: _____
***Private or sensitive information includes any information or data that can be used to uniquely identify a person, including, but not limited to, social security numbers or other government identification numbers, payment card information, drivers' license numbers, financial account numbers, personal identification numbers (PINs), usernames, passwords, healthcare records and email addresses.**
b. Do you collect, store, host, process, control, use or share any biometric information or data, such as fingerprints, voiceprints, facial, hand, iris or retinal scans, DNA, or any other biological, physical or behavioral characteristics that can be used to uniquely identify a person? ☐ Yes ☐ No
If "Yes", have you reviewed your policies relating to the collection, storage and destruction of such information or data with a qualified attorney and confirmed compliance with applicable federal, state, local and foreign laws? ☐ Yes ☐ No
c. Do you process, store, or handle credit card transactions? ☐ Yes ☐ No
If "Yes", are you PCI-DSS Compliant? ☐ Yes ☐ No
5. IT DEPARTMENT
This section must be completed by the individual responsible for the Applicant's network security. As used in this section only, "you" refers to the individual responsible for the Applicant's network security.
a. Who is responsible for the Applicant's network security?
Name:
Title:
Phone: Email address:
IT Security Designation(s):
b. The Applicant's network security is: ☐ Outsourced ☐ Managed internally/in-house

NGP-SMBA (1.2021) Page 1 of 4

The assignment algorithm needs to differentiate between similar looking forms that have material differences in the location of the text to be extracted. For example, there is a key difference in the revenue section of the otherwise similar forms on the left.

We created a vector representation of each form using term frequency - inverse document frequency (TF-IDF) for all of the text recognized by the optical character recognition engine on the first page of the form. We created a similarity score using cosine similarity scores of each vector representation.



Each form was assigned to the template with the highest similarity score so long as that score exceeded a minimum threshold. The Threshold for the Beazley template (with scores shown above) was 75.

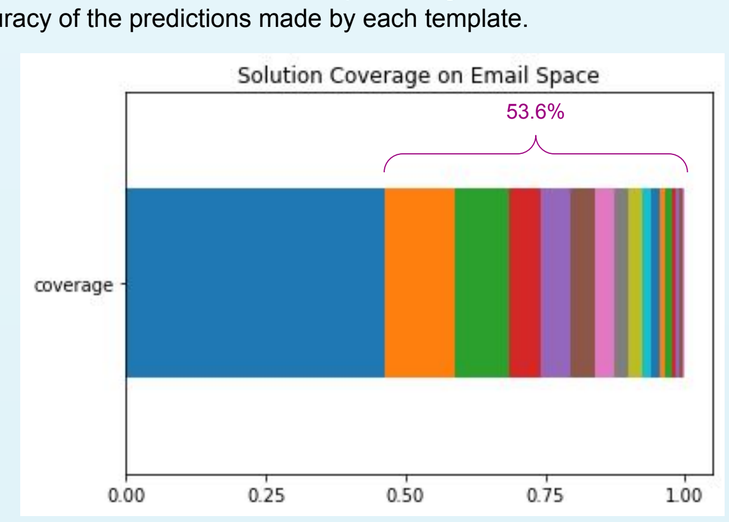
Results

The performance of the extraction process can be measured on two dimensions: the percentage of email submissions that are assigned to a template (coverage), and the accuracy of the predictions made by each template.

Coverage:

We achieve 53.6% coverage on the cyber email submissions using the templates that we created thus far (about 20).

Here we show what each template contributes to entire space. All the space under the bracket represents the amount of emails that are covered by our solution.



Accuracy:

We get correct predictions for all five target fields on over 70% of submissions for two of the most common templates.

Common sources of error are:

- The optical character recognition engine interprets the text incorrectly. This is most common for handwritten forms and forms with poor scan quality
- The extraction bounding box is the incorrect size
- The wrong template is assigned.

The second two errors may be reduced as additional forms are labelled.

Template	Forms Evaluated	All Correct	Accuracy
Beazley	51	74.5%	89.4%
Travelers	50	72%	90.8%
Tokio Marine	52	56%	84.7%
Combined	153	67.8%	88.3%

Impact

Scalability:

Corvus insurance is growing rapidly. A human administrative support team cannot scale at the rate Corvus expects their email submissions to grow. This extraction process by contrast will scale easily.

Time to Quote:

The first company to reply to a quote request is an order of magnitude more likely to win the business. This solution will greatly reduce the time to quote.

Potential to Create Novel Datasets:

The current application of this solution is to collect basic data for quote requests. However, the solution can be easily generalized to generate other data sets from PDF files. Corvus Insurance has expressed interest in using the solution for this purpose.

Cost Savings:

The most immediate benefit to Corvus Insurance comes from time savings for the administrative support team currently performing data entry. Corvus estimates this will save tens of thousands of dollars. This number will increase dramatically as Corvus scales.