DETECTION OF DUPLICATES AMONG NON-STRUCTURED DATA FROM DIFFERENT DATA SOURCES

Faculté des sciences et de génie
Institut Intelligence et Données (IID)
UNIVERSITÉ LAVAL

SITUATION

When assessing a commercial risk, an insurer needs to gather various information about the risk. This long and complex process implies numerous questions. Thus an insurer is prompt to use an external source to help reduce the number of questions.

Thus, we want to detect duplicate of commercial risk in another data source using as **little** as possible information [**?**].

SITUATION

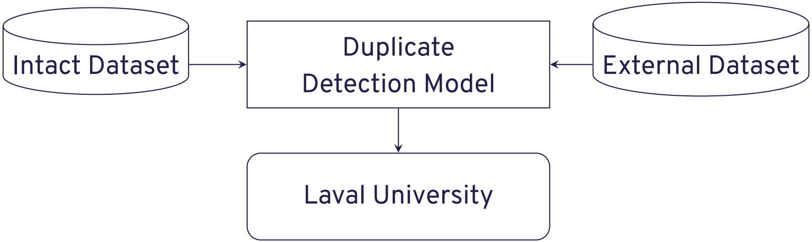
When assessing a commercial risk, an insurer needs to gather various information about the risk. This long and complex process implies numerous questions. Thus an insurer is prompt to use an external source to help reduce the number of questions.

Thus, we want to detect duplicate of commercial risk in another data source using as **little** as possible information **[?]**.

Example

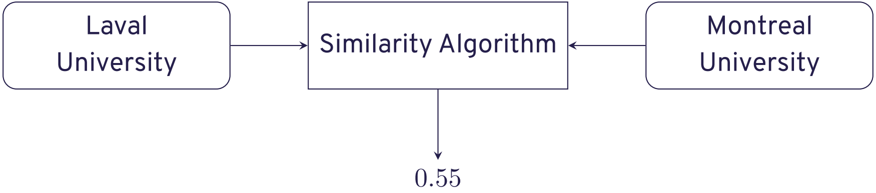
David Beauchemin, the owner of ”Beauchemin inc.”, calls for insurance. Using minimal information, we want to retrieve as much as possible from an external source to ask him as less than the necessary number of questions.

HOW DO WE DETECT DUPLICATE ?



Similarity Between Two Entities

SIMILARITY ALGORITHM



SIMILARITY ALGORITHM

Similarities algorithms are one that measures the resemblance between two strings base on the distance between their tokens.

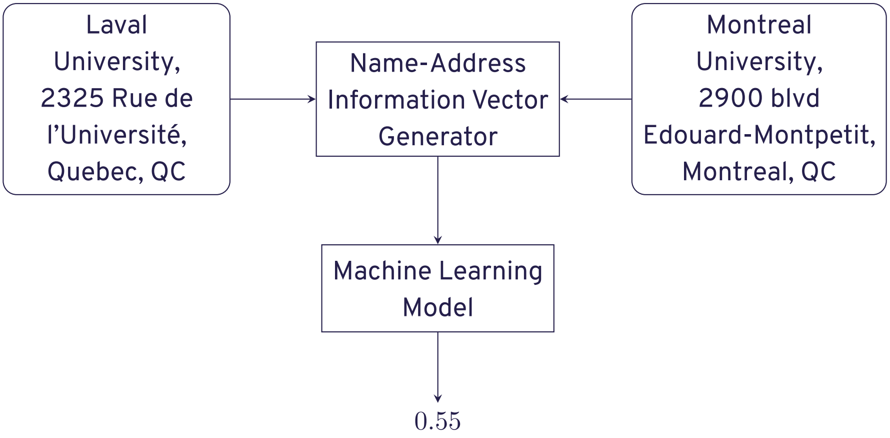
NAME INTERESTING RESULTS

	StoS	Jaccard
Recall (%)	44.15	**65.87**

ADDRESS INTERESTING RESULTS

	StoS	CSS
Recall (%)	13.19	**50.36**

MACHINE LEARNING SIMILARITY ALGORITHM



NAME-ADDRESS INFORMATION VECTOR GENERATOR

We used the previous similarity algorithm to generate an information vector between two entities using the name and address.

Example of an information vector

StoS	Levenshtein	Jaro-Winkler	LCSP	Jaccard	Cosinus	-
0.00	0.15	0.25	0.35	0.15	0.15	-

StoS	Levenshtein	Jaro	LCSP	Jaccard	Cosinus	CSS
0.00	0.16	0.55	0.15	0.45	0.37	0.48

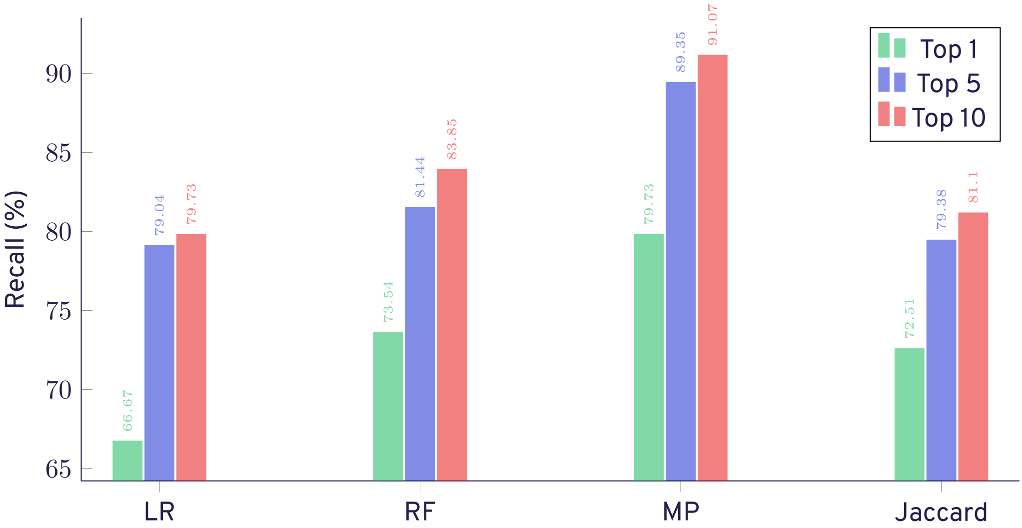
RESULTS

	Logistic Regression	Random Forest	Multilayer perceptron	Jaccard
Recall (%)	66,67	73,54	**79,73**	72,51

IMPROVING THE RESULTS - *N* MOST SIMILAR

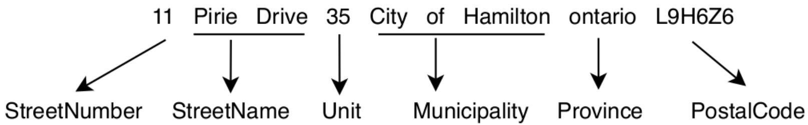
We consider a matching is good when the pair *(commercial risk, REQ entity)* is included in the *N* most similar.

RESULTS



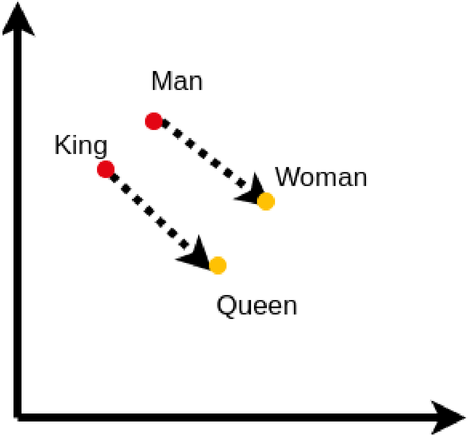
Recall (%)
90
85
80
75
70
65
66.67
79.04
79.73
73.54
81.44
83.85
79.73
89.35
91.07
72.51
79.38
81.1
LR
RF
MP
Jaccard
Top 1
Top 5
Top 10

DEEPPARSE[1]



1. HTTPS://DEEPPARSE.ORG/

SUBWORD EMBEDDINGS



RESULTS

Country	FastText	BPEmb
Canada	98.96	96.98
United States	98.49	96.5

CONCLUSION

- Intact can now detect a duplicate of a commercial risk with the REQ.
- Intact can now use Deepparse to parse multinational addresses.

ACKNOWLEDGMENTS

This research was supported by the Natural Sciences and Engineering Research Council of Canada (IRCPJ 529529-17) and a Canadian insurance company.
Luc Lamontagne for his mentorship and reviewers for their comments.

THANK YOU FOR LISTENING!

REFERENCES i