NER in Archival Finding Aids

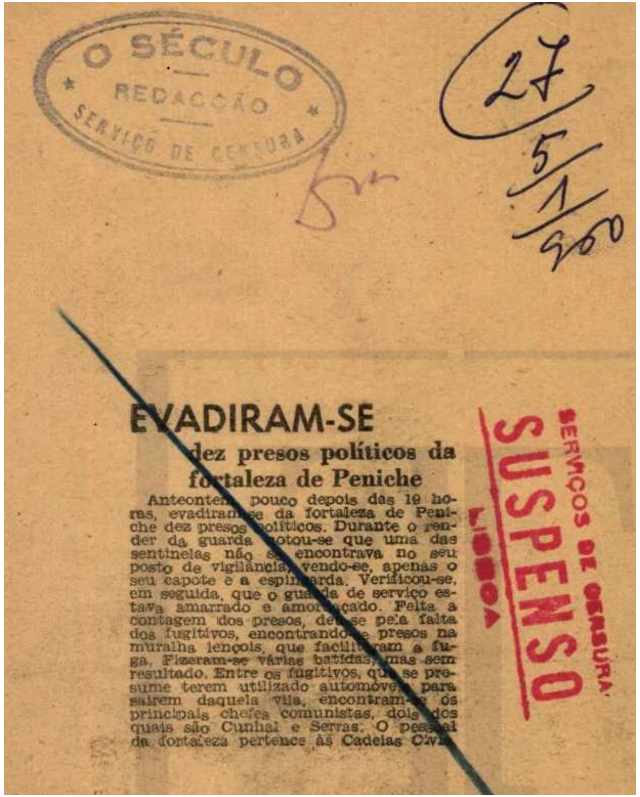
ner.epl.di.uminho.pt

Luís Filipe da Costa Cunha
a83099@alunos.uminho.pt
University of Minho, Portugal

José Carlos Leite Ramalho
jcr@di.uminho.pt
Department of Informatics, University of Minho, Portugal

Introduction

- The first portuguese certificate was issued in 1378 by the TT
- Digital format
- Archives semantic interpretation
- NLP
- Machine Learning algorithms



O SÉCULO
REDACÇÃO
SERVIÇO DE CENSURA

27
5/1/960

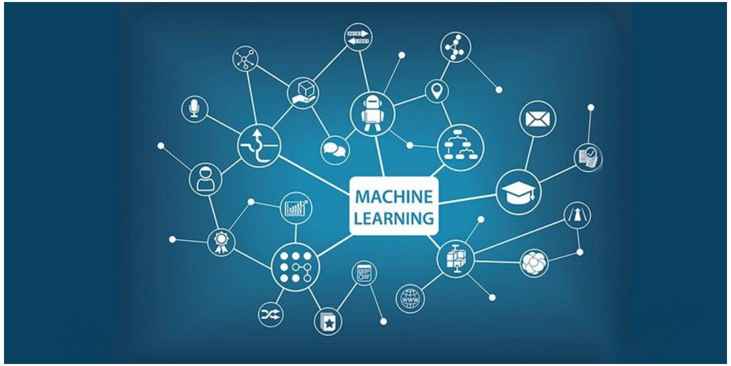
EVADIRAM-SE
dez presos políticos da fortaleza de Peniche

Anteontem, pouco depois das 19 horas, evadiram-se da fortaleza de Peniche dez presos políticos. Durante o render da guarda notou-se que uma das sentinelas não se encontrava no seu posto de vigilância, vendo-se, apenas o seu capote e a espingarda. Verificou-se, em seguida, que o guarda de serviço estava amarrado e amordaçado. Feita a contagem dos presos, deu-se pela falta dos fugitivos, encontrando-se presos na muralha lençois, que facilitaram a fuga. Fizeram-se várias batidas, mas sem resultado. Entre os fugitivos, que se presume terem utilizado automóveis para sairem daquela vila, encontram-se os principais chefes comunistas, dois dos quais são Cunhal e Serras. O pessoal da fortaleza pertence às Cadeias Civis

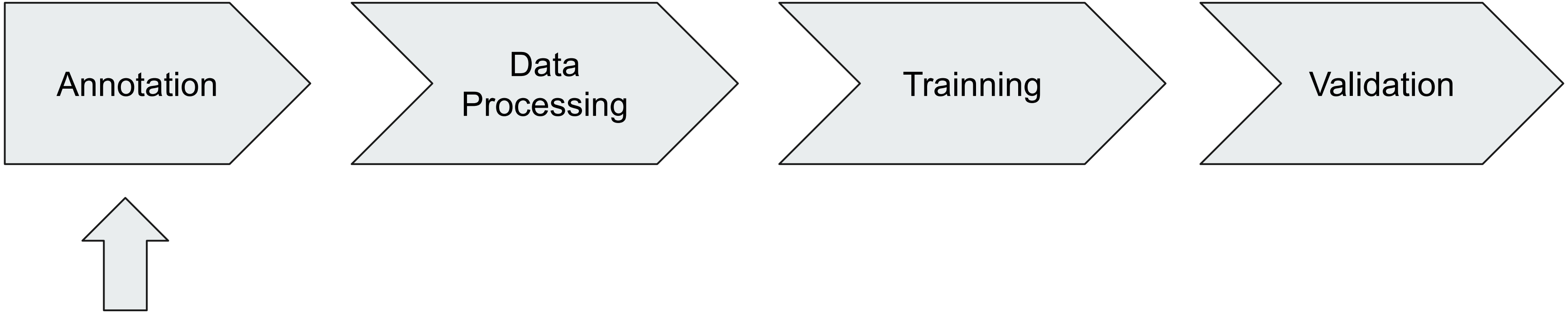
SERVIÇOS DE CENSURA
SUSPENSO
LISBOA

Named Entity Recognition

- Pre-existing Portuguese Models (HAREM and SIGARRA)
- Train new ML models
- OpenNLP
- spaCy
- TensorFlow



Named Entity Recognition Process



Annotation Process

Data Selection

- 8 archival corpus
- Shuffle was performed on each corpus
- Selection of a representative fraction of each corpus

Data Annotation

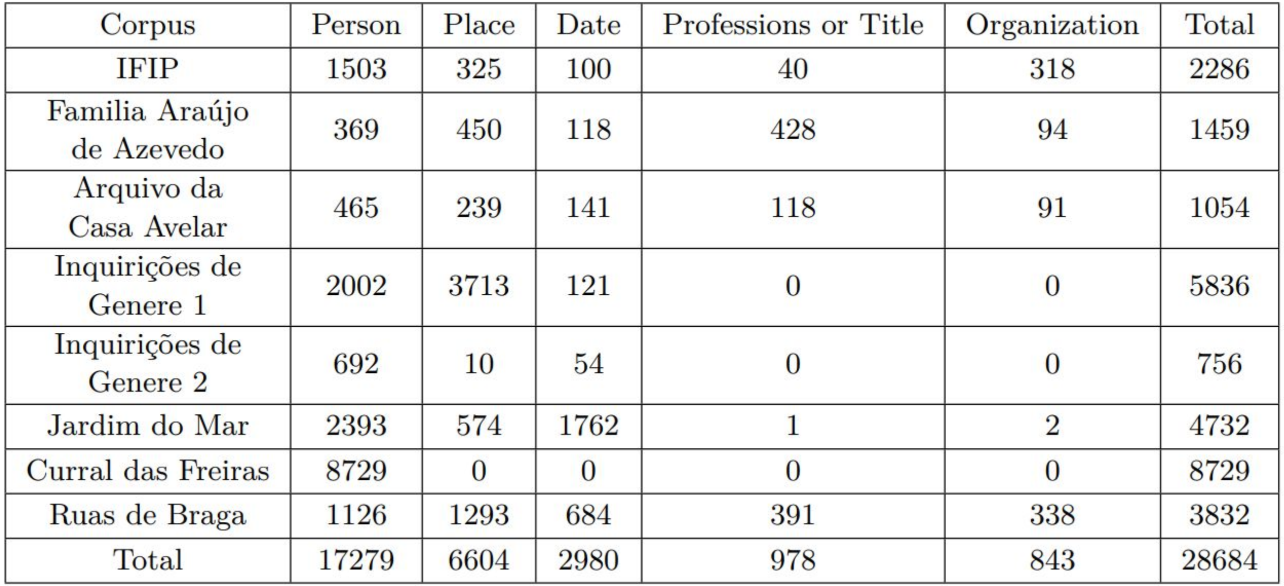
- Manual annotation
- Regular expressions
- Use of a statistical model preceded by correction of the output by the annotator

Annotated Corpora

Nrº of tokens: 164478

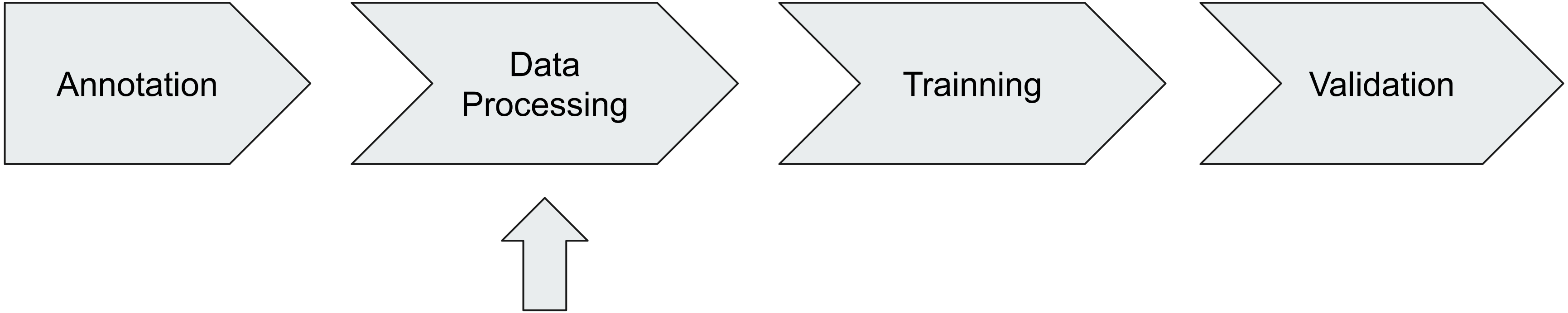
Nrº of phrases: 6302

ner.epl.di.uminho.pt



Corpus	Person	Place	Date	Professions or Title	Organization	Total
IFIP	1503	325	100	40	318	2286
Familia Araújo de Azevedo	369	450	118	428	94	1459
Arquivo da Casa Avelar	465	239	141	118	91	1054
Inquirições de Genere 1	2002	3713	121	0	0	5836
Inquirições de Genere 2	692	10	54	0	0	756
Jardim do Mar	2393	574	1762	1	2	4732
Curral das Freiras	8729	0	0	0	0	8729
Ruas de Braga	1126	1293	684	391	338	3832
Total	17279	6604	2980	978	843	28684

Named Entity Recognition Process



Annotation

Data Processing

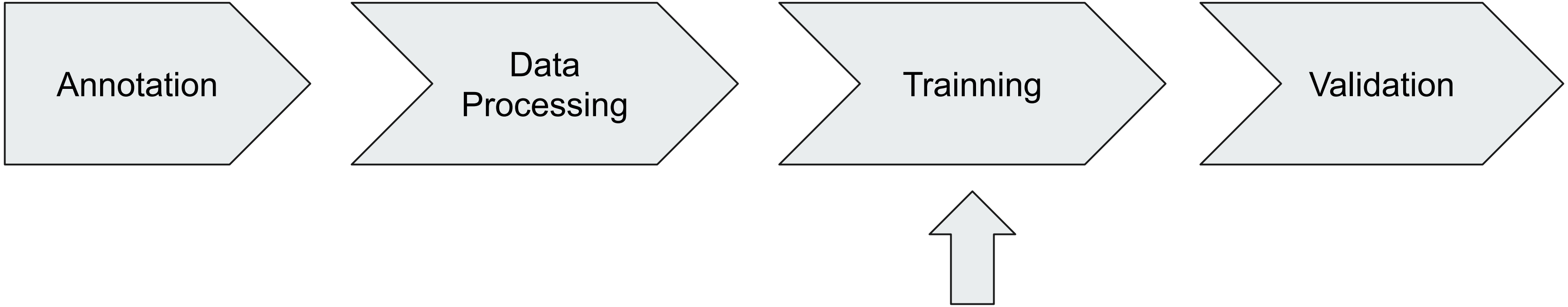
Trainning

Validation

Data processing

- Parse annotated datasets into three different formats
- 70% training, 30% validation
- Tokenization

Named Entity Recognition Process



NLP Tools

OpenNLP

Maximum Entropy (Maxent)

- Entropy
- Features / Restrictions
- Entropy Maximization

Function of Information Entropy:

$$H(X) = -\sum p(a,b) \log p(a,b)$$

spaCy

Convolutional Neural Networks

- “Deep Learning Framework”
- Pre-trained Portuguese word embeddings
- Entity Linking

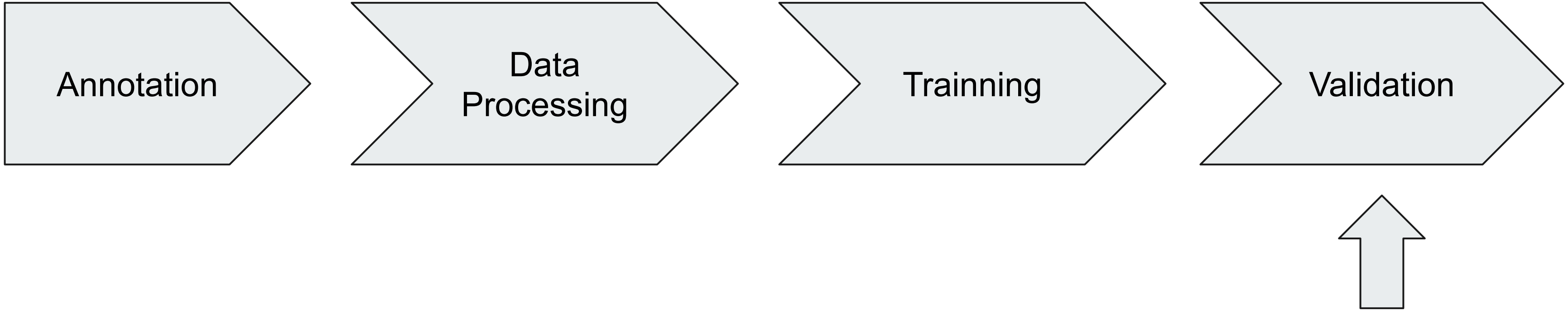
OpenNLP

TensorFlow

Recurrent Neural Networks

- Tokenization
- Vocabulary Creation (limited to training data)
- Word Embeddings
- Bi-LSTM-CRF (Huang, Xu, and Yu, 2015)

Named Entity Recognition Process

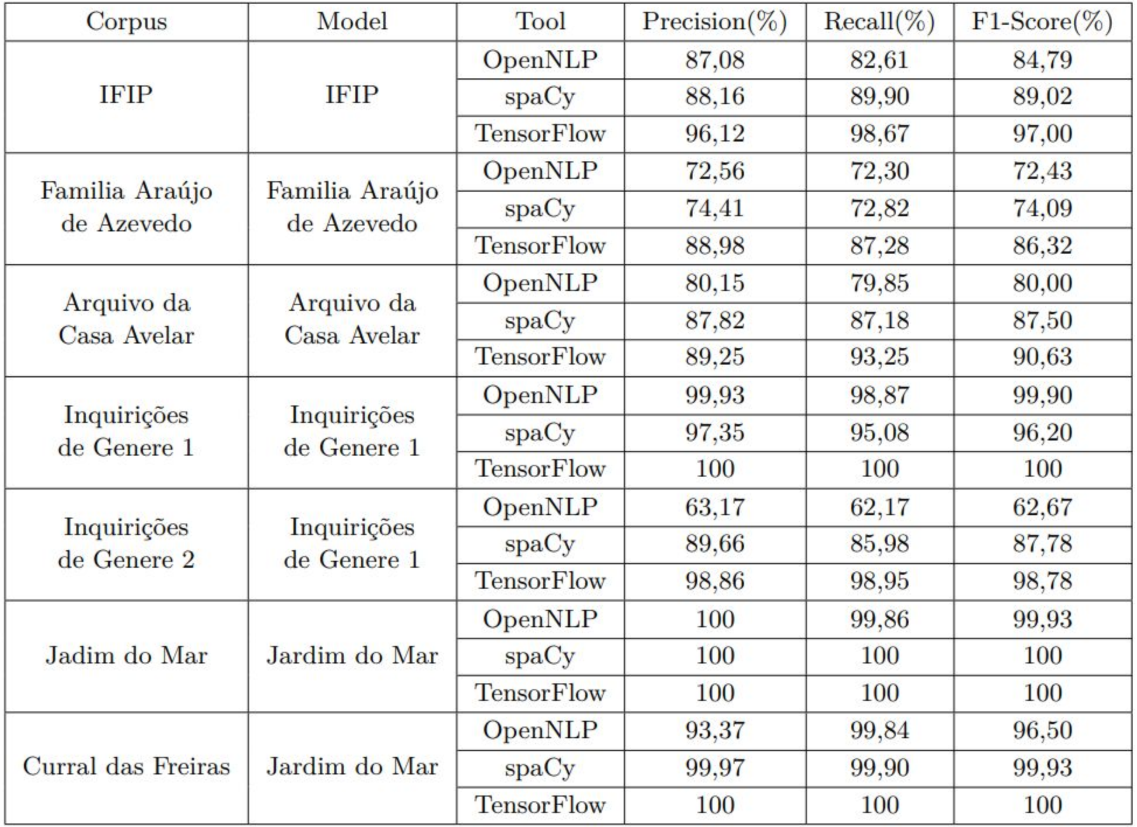


Individual Models

OpenNLP: 62.67% - 100%

spaCy: 70,09% - 100%

TensorFlow: 86,32% - 100%,



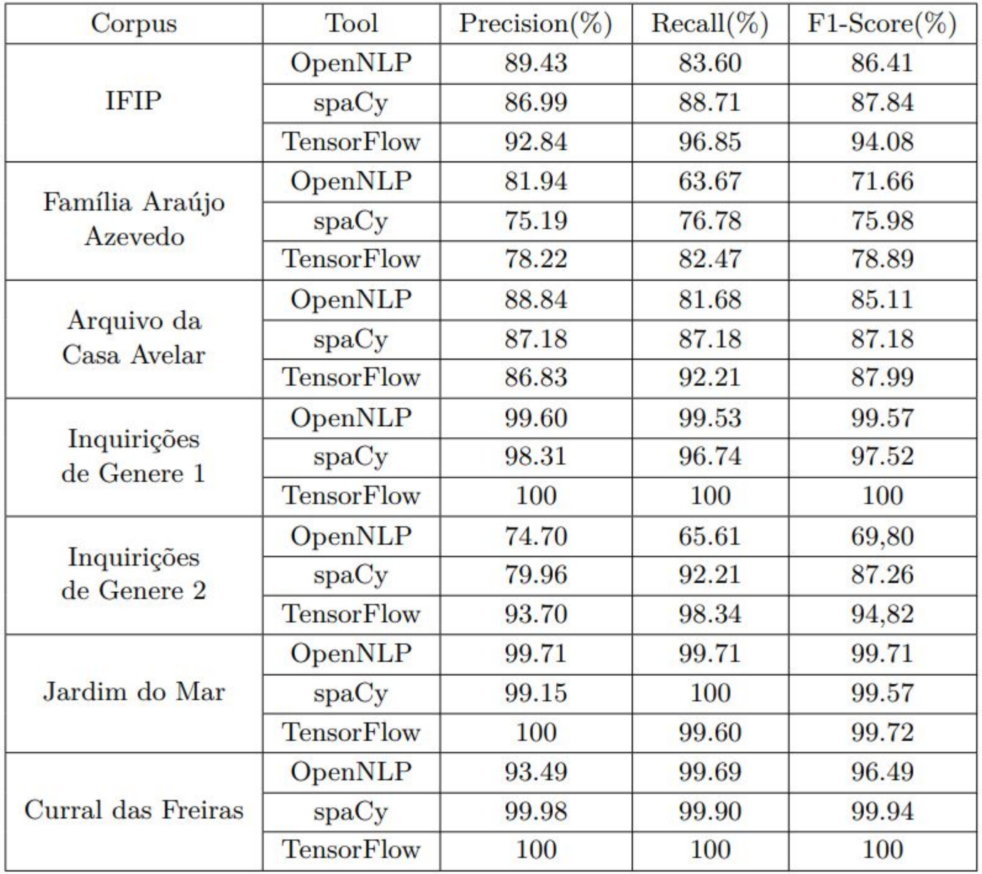
Corpus	Model	Tool	Precision(%)	Recall(%)	F1-Score(%)
IFIP	IFIP	OpenNLP	87,08	82,61	84,79
		spaCy	88,16	89,90	89,02
		TensorFlow	96,12	98,67	97,00
Familia Araújo de Azevedo	Familia Araújo de Azevedo	OpenNLP	72,56	72,30	72,43
		spaCy	74,41	72,82	74,09
		TensorFlow	88,98	87,28	86,32
Arquivo da Casa Avelar	Arquivo da Casa Avelar	OpenNLP	80,15	79,85	80,00
		spaCy	87,82	87,18	87,50
		TensorFlow	89,25	93,25	90,63
Inquirições de Genere 1	Inquirições de Genere 1	OpenNLP	99,93	98,87	99,90
		spaCy	97,35	95,08	96,20
		TensorFlow	100	100	100
Inquirições de Genere 2	Inquirições de Genere 1	OpenNLP	63,17	62,17	62,67
		spaCy	89,66	85,98	87,78
		TensorFlow	98,86	98,95	98,78
Jadim do Mar	Jardim do Mar	OpenNLP	100	99,86	99,93
		spaCy	100	100	100
		TensorFlow	100	100	100
Curral das Freiras	Jardim do Mar	OpenNLP	93,37	99,84	96,50
		spaCy	99,97	99,90	99,93
		TensorFlow	100	100	100

Generalized Model

OpenNLP: 69.80% - 99.71%

spaCy: 75.98% - 99.94%

TensorFlow: 78.89% - 100%



Corpus	Tool	Precision(%)	Recall(%)	F1-Score(%)
IFIP	OpenNLP	89.43	83.60	86.41
	spaCy	86.99	88.71	87.84
	TensorFlow	92.84	96.85	94.08
Família Araújo Azevedo	OpenNLP	81.94	63.67	71.66
	spaCy	75.19	76.78	75.98
	TensorFlow	78.22	82.47	78.89
Arquivo da Casa Avelar	OpenNLP	88.84	81.68	85.11
	spaCy	87.18	87.18	87.18
	TensorFlow	86.83	92.21	87.99
Inquirições de Genere 1	OpenNLP	99.60	99.53	99.57
	spaCy	98.31	96.74	97.52
	TensorFlow	100	100	100
Inquirições de Genere 2	OpenNLP	74.70	65.61	69,80
	spaCy	79.96	92.21	87.26
	TensorFlow	93.70	98.34	94,82
Jardim do Mar	OpenNLP	99.71	99.71	99.71
	spaCy	99.15	100	99.57
	TensorFlow	100	99.60	99.72
Curral das Freiras	OpenNLP	93.49	99.69	96.49
	spaCy	99.98	99.90	99.94
	TensorFlow	100	100	100

Results Comparison

Individual Models Results

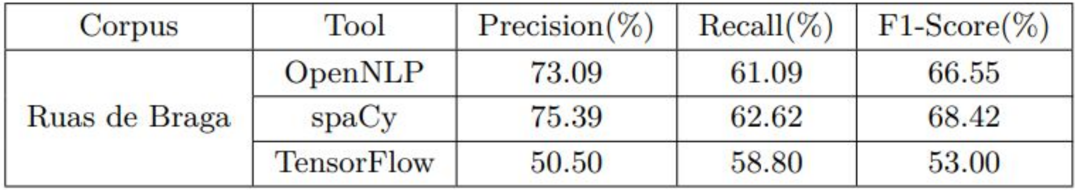
Corpus	OpenNLP	spaCy	TensorFlow
IFIP	F1-score: 84.79%;	F1-score: 89.02%;	F1-score: 97.00%.
Família Araújo Azevedo	F1-score: 72.43%;	F1-score: 74.09%;	F1-score: 86.32%.
Arquivo da Casa Avelar	F1-score: 80.00%;	F1-score: 87.50%;	F1-score: 90.63%.
Inquirições de Genere 1	F1-score: 99.90%;	F1-score: 96.20%;	F1-score: 100%.
Inquirições de Genere 2	F1-score: 62.67%;	F1-score: 87.78%;	F1-score: 98,78%.
Jardim do Mar	F1-score: 99.93%;	F1-score: 100%;	F1-score: 100%.
Curral das Freiras	F1-score: 100%;	F1-score: 100%;	F1-score: 100%.

Generalized Model Results

Corpus	OpenNLP	spaCy	TensorFlow
IFIP	F1-score: 86.41%;	F1-score: 87.84%;	F1-score: 94.08%.
Família Araújo Azevedo	F1-score: 71.66%;	F1-score: 75.98%;	F1-score: 78.89%.
Arquivo da Casa Avelar	F1-score: 85.11%;	F1-score: 87.18%;	F1-score: 87.99%.
Inquirições de Genere 1	F1-score: 99.57%;	F1-score: 97.52%;	F1-score: 100%.
Inquirições de Genere 2	F1-score: 69.80%;	F1-score: 87.26%;	F1-score: 94,82%.
Jardim do Mar	F1-score: 99.71%;	F1-score: 99.57%;	F1-score: 99.72%.
Curral das Freiras	F1-score: 100%;	F1-score: 100%;	F1-score: 100%.

New Corpus

- Ruas de Braga corpus was not used for training
- This corpus has a lot of Profession and Organization entities.
- BI-LSTM-CRF model presented the worst results (vocabulary limited to his training).



Corpus	Tool	Precision(%)	Recall(%)	F1-Score(%)
Ruas de Braga	OpenNLP	73.09	61.09	66.55
	spaCy	75.39	62.62	68.42
	TensorFlow	50.50	58.80	53.00

Conclusion

- By training our own models, it is possible to obtain satisfactory results.
- F1-score values above 86%.
- The use of ML algorithms to perform NER in archival documents is suitable.
- Enables the creation of navigation mechanisms between information records.

NER in Archival Finding Aids

NER (ner.epl.di.uminho.pt)

Luís Filipe da Costa Cunha
a83099@alunos.uminho.pt
University of Minho, Portugal

José Carlos Leite Ramalho
jcr@di.uminho.pt
Department of Informatics, University of Minho, Portugal