



# THE DEVELOPER'S CONFERENCE

## Experimento de Extração de Relações com algoritmo Conditional Random Field

Trilha Inteligência Artificial

**Patricia Nunes Gonçalves**



PHD Engenharia da Computação – Universidade de Lisboa

Arquitetura Corporativa  **Sicredi**

# Agenda



- Introdução PLN
- Introdução Inteligência Competitiva
- Framework Extração Relações
- Experimento
- Resultados
- Conclusão



## A primeira instituição financeira cooperativa do Brasil

Em um mundo cada vez mais colaborativo, queremos unir pessoas com os mesmos propósitos. Por isso, há mais de 116 anos, acreditamos que é possível lidar com o dinheiro de uma maneira diferente.

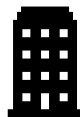
Saber mais:  
[www.sicredi.com.br](http://www.sicredi.com.br)



4 milhões de Clientes



R\$ 2,7 bilhões em resultados



R\$ 15 bilhões patrimônio líquido

# Inteligência Competitiva



**Modelo estruturado** de levantamento e análise de dados e fatos não estruturados com o objetivo de **apoiar a tomada de decisão**.

*Estruturar olhar **para o mercado**, avaliando não só concorrentes, como o setor, macro tendências, insights, sinais fracos, para assim **gerar análises e criar hipóteses e ações**.*

## Suporte para decisão

Processo sistemático, estruturado e ético

Caráter antecipativo  
Macrotendências

Insumos de mercado para direcionamentos e priorizações



Informação é matéria-prima

Escopo amplo, não apenas concorrentes

Processo recorrente de revisão da estratégia

# Processamento de linguagem natural



THE  
DEVELOPER'S  
CONFERENCE

Processamento de linguagem natural (PLN) é uma subárea da ciência da computação, inteligência artificial e da linguística que estuda os problemas da geração e compreensão automática de línguas humanas naturais.

## Exemplos de aplicações que envolvem PLN:

- ❖ Sumarização Automática
- ❖ Chatbots
- ❖ Tradução Automática
- ❖ Speech-to-Text
- ❖ Resolução de Correferência
- ❖ Extração Entidades Nomeadas – NER
- ❖ Extração de Relações

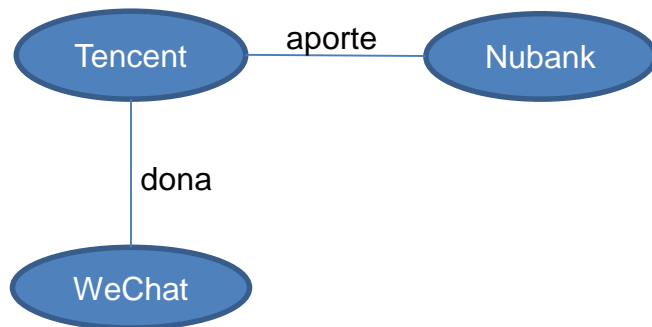
# O que é Extração de Relações?

Um exemplo de extração de relações



THE  
DEVELOPER'S  
CONFERENCE

- “A empresa **Tencent**, dona da **WeChat**, fez um aporte de 200 milhões de dólares no **Nubank**”



# Trabalho construído em parceria



Patricia Nunes Gonçalves  
Gustavo Rossi  
Matheus Quevedo



Sandra Collovini  
Joaquim Francisco dos Santos Neto  
Guilherme Bonfada Cavalheiro

Prof. Dra. Renata Vieira  
<https://www.inf.pucrs.br/~rvieira/>



Grupo de Processamento da Linguagem Natural – PLN PUCRS  
<http://www.inf.pucrs.br/linatural/wordpress/>

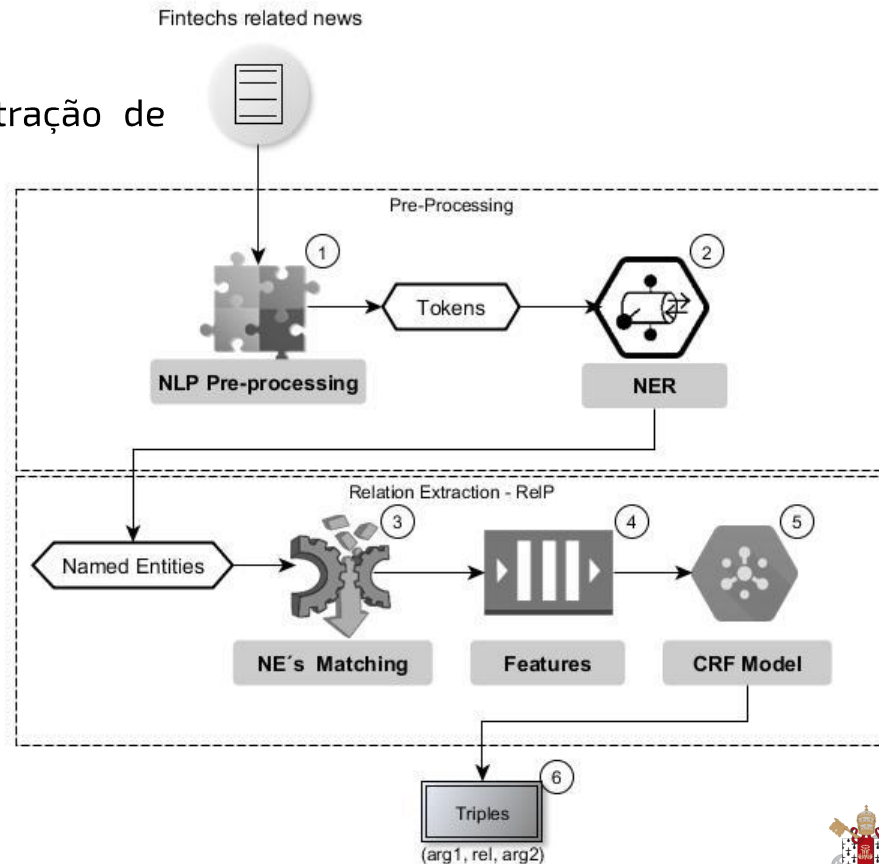
# Framework

Pipeline de processamento end-to-end para Extração de Relações

- Desenvolvido em Open-source:
  - NLTK - <https://www.nltk.org>
  - Cogroo - <http://ccsl.ime.usp.br/cogroo/>
  - NER<sup>1</sup> – Redes Neurais (BiLSTM-CRF Neural Network) implementadas em Python
  - CRF Model<sup>2</sup> – Baseado no sistema RelP desenvolvido em python



<https://github.com/relplusplus>



<sup>1</sup> - Santos, J., Consoli, B., dos Santos, C., Terra, J., Collonini, S., Vieira, R.: Assessing the impact of contextual embeddings for portuguese named entity recognition. In: 8th Brazilian Conference on Intelligent Systems, BRACIS 2019, Bahia, Brazil, October 15-18, 2019. pp. 437–442 (2019)

<sup>2</sup> -Abreu, S.C., Bonamigo, T.L., Vieira, R.: A review on relation extraction with an eye on portuguese. Journal of the Brazilian Computer Society pp. 1–19 (2013)

O Framework foi desenvolvido pela equipe de PLN da PUCRS





# Exemplo

Exemplo do processamento passo-a-passo no Framework



THE  
DEVELOPER'S  
CONFERENCE

## Notícia

Novo aporte na Warren. O aporte foi de R\$ 25 milhões. A rodada foi liderada pela Kaszek Ventures, investidora do Nubank, e Ribbit Capital, que investiu no Guiabolso. Criada em 2017, a startup conta com 60 mil clientes.

## NER + NE Matching

- ✗ Novo aporte em a <ORG>Warren<ORG>.
- ✗ O aporte foi de R\$ 25 milhões.
- ✓ A rodada foi liderada por a <ORG>Kaszek\_Ventures<ORG>, a investidora de o <ORG>Nubank<ORG>, e <ORG>Ribbit\_Capital<ORG>, que investiu em o <ORG>Guiabolso<ORG>.
- ✗ Criada em 2017, a startup conta com 60 mil clientes.

## NE Matching – Seleção de pares de entidades:

- ✓ A rodada foi liderada por a <ORG>Kaszek\_Ventures<ORG>, a investidora de o <ORG>Nubank<ORG>, e <ORG>Ribbit\_Capital<ORG>, que investiu em o <ORG>Guiabolso<ORG>.
- Pares Candidatos:
- Kaszek\_Ventures<ORG> - Nubank<ORG>
  - Kaszek\_Ventures<ORG> - Ribbit\_Capital<ORG>
  - Kaszek\_Ventures<ORG> - Guiabolso<ORG>
  - Nubank<ORG> - Ribbit\_Capital<ORG>
  - Nubank<ORG> - Guiabolso<ORG>
  - Ribbit\_Capital<ORG>- Guiabolso<ORG>

## Extração de Relações

Kaszek\_Ventures<ORG> a investidora de Nubank<ORG>  
Kaszek\_Ventures<ORG> e Ribbit\_Capital<ORG>  
Kaszek\_Ventures<ORG> que investiu em Guiabolso<ORG>  
Nubank<ORG> NULL Ribbit\_Capital<ORG>  
Nubank<ORG> NULL Guiabolso<ORG>  
Ribbit\_Capital<ORG> NULL Guiabolso<ORG>

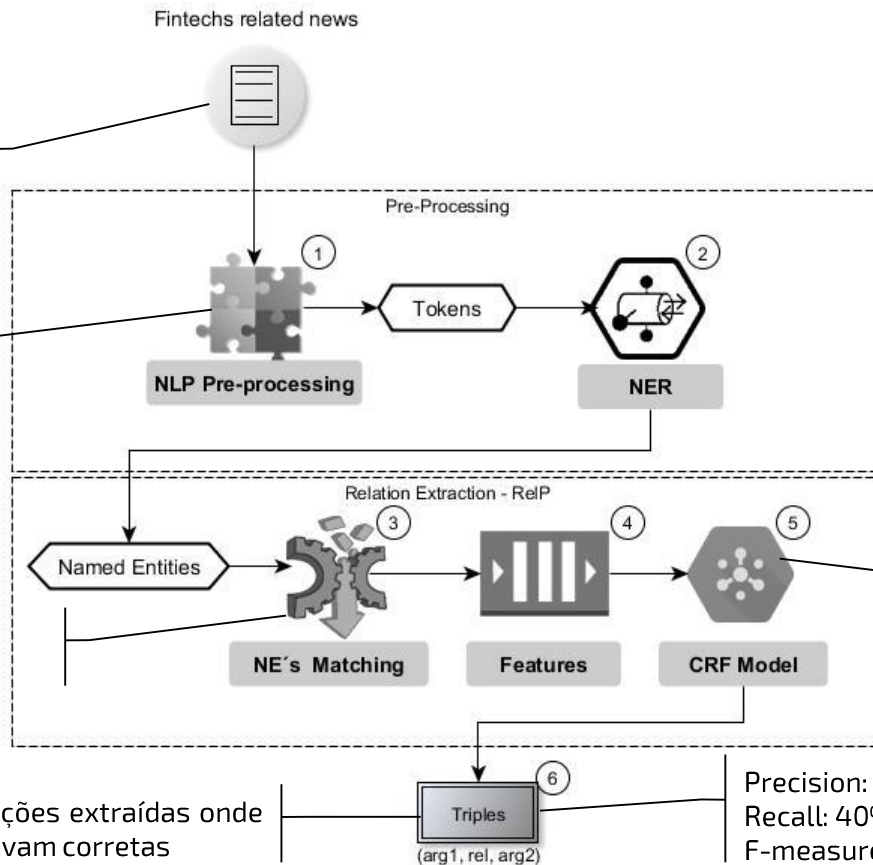
# Experimento

200 notícias selecionadas para o Radar de Inteligência Competitiva sob o tema Fintechs

~1.200 sentenças

142 sentenças (com 2 ou mais NE)

68 relações extraídas onde 47 estavam corretas



403 pares candidatos para Extração de Relações

Precision: 72%  
Recall: 40%  
F-measure: 51%

## Análise de Erro



- Notícias sobre Startups eram muito complexas
- Muitas vezes a relação estavam na frase seguinte (problema de correferência)
- Problema de delimitação do NER:

O programa de capacitação Startup SC, iniciativa do Sebrae_de_Santa_Catarina, (...)	(SC, iniciativa de, Sebrae_de_Santa_Catarina)
--	--

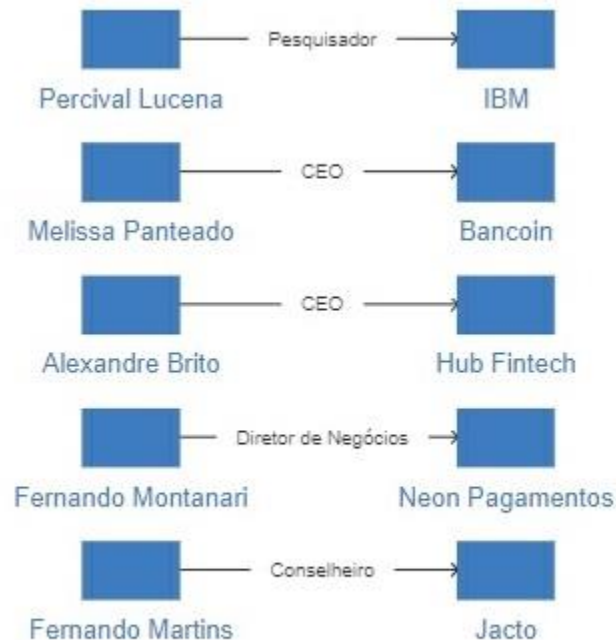
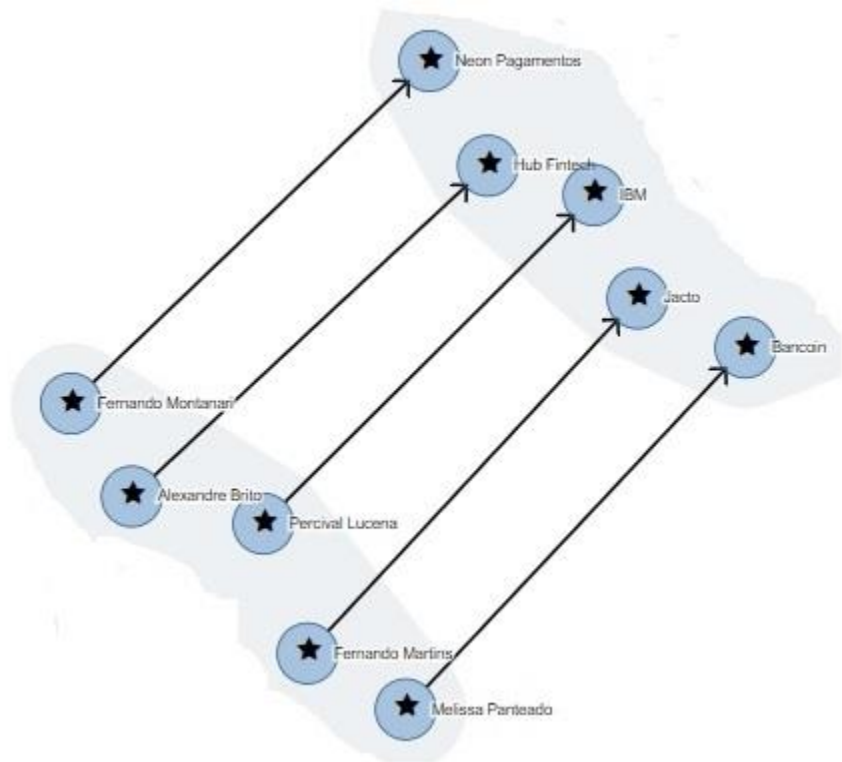
- O nome das empresas eram bem complexos
- Muitas entidades que não eram Organizações:
  - Programa de Inovação
  - Laboratório
  - Iniciativa
- Nosso conjunto de treino precisa ser maior

# Visualização dos Dados

Usamos o ArDoq para disponibilizar as relações entre entidades para análise e consumo do Planejamento Estratégico do Sicredi



THE  
DEVELOPER'S  
CONFERENCE

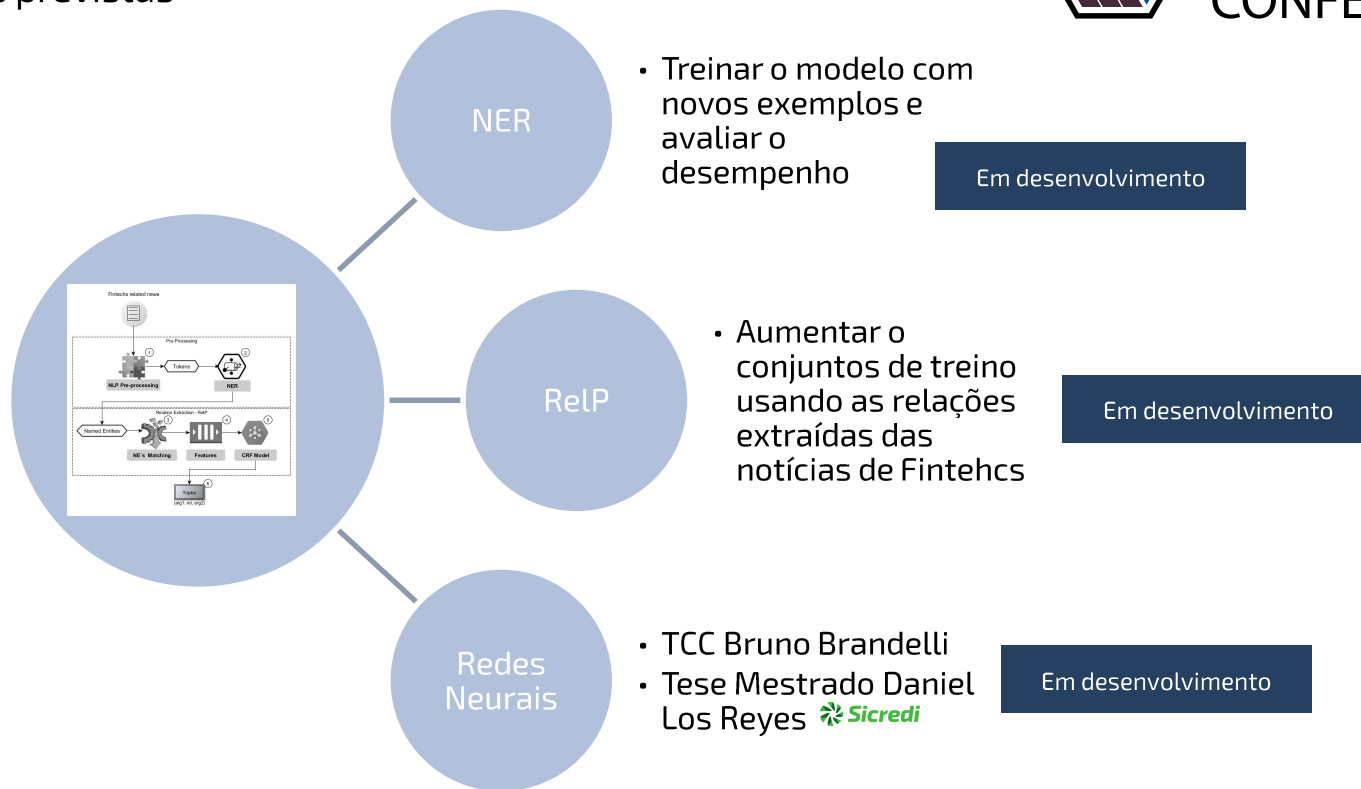


# Próximos passos

Modificações previstas



THE  
DEVELOPER'S  
CONFERENCE



## Conclusão



THE  
DEVELOPER'S  
CONFERENCE



Vantagens para Sicredi: trabalho desenvolvido sem custo  
Vantagens para os pesquisadores: evolução dos sistemas desenvolvidos




Artigo: International Conference on the  
Computational Processing of Portuguese



THE  
DEVELOPER'S  
CONFERENCE

Apresentação na trilha de  
IA e Machine Learning no TDC Porto Alegre



Novos trabalhos sendo desenvolvidos: TCC Bruno Brandelli,  
Aprimoramento Tese Mestrado Joaquim Santos, Nova Tese  
Mestrado Daniel Los Reys 

# Agradecimentos



**Guilherme Bonfada Cavalheiro**  
**Gustavo Rossi**  
**Joaquim Francisco dos Santos Neto**  
**Matheus Quevedo**  
**Renata Vieira**  
**Sandra Collovini**





Patricia Nunes Gonçalves

Sicredi

QUARTA | 27/11

## Trilha Inteligência Artificial

O mundo está mudando e a inteligência artificial e a grande responsável por essa revolução, que vai mudar praticamente tudo como conhecemos.

QUINTA | 28/11

## Trilha Machine Learning

Machine Learning e Deep Learning além do buzzword

Muito se fala de Machine Learning, Deep Learning e seu poder supostamente revolucionário, mas poucas vezes se mostra o que há por trás desses



THE DEVELOPER'S CONFERENCE

QUINTA | 28/11

## Trilha Arquitetura de Dados

Discutir e compartilhar experiências sobre arquitetura de dados RDBMS e NoSQL

Nos dias de hoje com a infinidade de tecnologias de



12:10  
às  
13:00

**Experimento de Extração de Relações com algoritmo Conditional Random Field** ▾

*Patricia Nunes Gonçalves*

**Machine Learning Canvas: Da coleta de dados à geração de valor** ▾

*Alexandre Ray Silva*



11:10  
às  
12:00

**Usando LDA para seleção de tópicos para Recuperação de Informação aplicado a Inteligência de Mercado** ▾

*Patricia Nunes Gonçalves*

**OCR: o que é e como usar?** ▾

*Guilherme Malta*



10:10  
às  
11:00

**Um estudo sobre implementação de ILM (Information Lifecycle Management) no Sicredi** ▾

*Patricia Nunes Gonçalves*

**Kafka e Dataflow: uma história de amor** ▾

*Leonardo Alves Miguel*



# Perguntas?



 [patricia\\_ngoncalves@sicredi.com.br](mailto:patricia_ngoncalves@sicredi.com.br)

 <https://dblp.org/pid/14/6534>

 <https://www.linkedin.com/in/patricianunesgonc/>

