MCZA017-13
Processamento de Linguagem Natural

Sobre as avaliações: Prova e projetos

Prof. Jesús P. Mena-Chalco
jesus.mena@ufabc.edu.br

2Q-2019

Sobre a Prova

Sobre a prova – única

A prova consta de duas partes (5 perguntas):

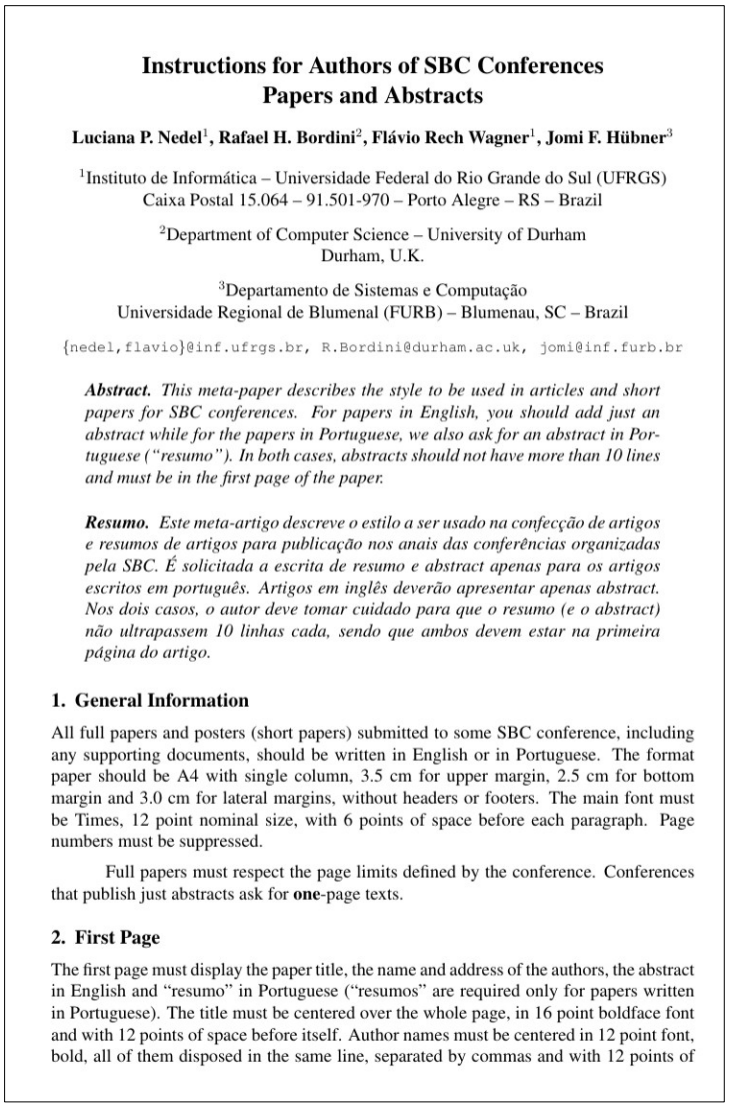
- **Parte 1 (2 perguntas, 40%):** Realizada em sala de aula.
  21h: 15/agosto
- **Parte 2 (3 perguntas, 60%):** Entregue através do Tidia.
  Deadline: 18/agosto

Cuidado com plágio:

Será tomado especial atenção para respostas. Código de ética.

Sobre a entrega Nro 3 do projeto: Relatório + Código + Dados

Sobre o modelo de relatório



Instructions for Authors of SBC Conferences Papers and Abstracts

Luciana P. Nedel[1], Rafael H. Bordini[2], Flávio Rech Wagner[1], Jomi F. Hübner[3]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

[2]Department of Computer Science – University of Durham
Durham, U.K.

[3]Departamento de Sistemas e Computação
Universidade Regional de Blumenal (FURB) – Blumenau, SC – Brazil

{nedel,flavio}@inf.ufrgs.br, R.Bordini@durham.ac.uk, jomi@inf.furb.br

Abstract. *This meta-paper describes the style to be used in articles and short papers for SBC conferences. For papers in English, you should add just an abstract while for the papers in Portuguese, we also ask for an abstract in Portuguese ("resumo"). In both cases, abstracts should not have more than 10 lines and must be in the first page of the paper.*

Resumo. *Este meta-artigo descreve o estilo a ser usado na confecção de artigos e resumos de artigos para publicação nos anais das conferências organizadas pela SBC. É solicitada a escrita de resumo e abstract apenas para os artigos escritos em português. Artigos em inglês deverão apresentar apenas abstract. Nos dois casos, o autor deve tomar cuidado para que o resumo (e o abstract) não ultrapassem 10 linhas cada, sendo que ambos devem estar na primeira página do artigo.*

1. General Information

All full papers and posters (short papers) submitted to some SBC conference, including any supporting documents, should be written in English or in Portuguese. The format paper should be A4 with single column, 3.5 cm for upper margin, 2.5 cm for bottom margin and 3.0 cm for lateral margins, without headers or footers. The main font must be Times, 12 point nominal size, with 6 points of space before each paragraph. Page numbers must be suppressed.

Full papers must respect the page limits defined by the conference. Conferences that publish just abstracts ask for **one**-page texts.

2. First Page

The first page must display the paper title, the name and address of the authors, the abstract in English and "resumo" in Portuguese ("resumos" are required only for papers written in Portuguese). The title must be centered over the whole page, in 16 point boldface font and with 12 points of space before itself. Author names must be centered in 12 point font, bold, all of them disposed in the same line, separated by commas and with 12 points of

Pode usar **qualquer modelo de relatório**, entretanto é fortemente sugerido o formato da Sociedade Brasileira de Computação (SBC).

No relatório evidencia o trabalho que vocês realizaram ao implementar a(s) proposta(s).

- Deadline:
  ~~18/agosto~~ -> 22/agosto (19h)

https://www.overleaf.com/latex/templates/sbc-conferences-template/blbxwjwzdngr

Sobre o código fonte de dados

Enviar conjuntamente com o relatório um arquivo **zip** contendo todo o material utilizados assim como o código fonte.

Sobre as apresentações dos projetos

As aulas dedicadas para apresentação também são candidatas para resumo. Isto é, deverá enviar resumo através do Tidia as aulas dos dias 22, 26, 28 e 30.

Sobre as apresentações

- **Dias de agosto:**

  Rep. do 20/jun

  Rep. do 8/jul

  22(21h), 26(19h), 27(21h) e 29(19h).

- **Ordem de apresentação:**

  Ordem de entrega do Relatório 2.

- A apresentação será em **15min**

  Sendo de **5min** para perguntas.

- Todos os membros do grupo devem apresentar oralmente (tempo homogeneo).

- Recomendável: **uso de slides**.

REPOSIÇÕES DOS FERIADOS				
Quadr	Feriado		Reposição	
2019.1	02 de março	sábado	06 de maio	segunda-feira
	04 de março	segunda-feira	07 de maio	terça-feira
	05 de março	terça-feira	08 de maio	quarta-feira
	06 de março	quarta-feira	09 de maio	quinta-feira
	08 de abril	segunda-feira	10 de maio	sexta-feira
	19 de abril	sexta-feira	13 de maio	segunda-feira
	20 de abril	sábado	11 de maio	sábado
	01 de maio	quarta-feira	14 de maio	terça-feira
	11 de março	segunda-feira	15 de maio	quarta-feira
2019.2	20 de junho	quinta-feira	27 de agosto	terça-feira
	21 de junho	sexta-feira	28 de agosto	quarta-feira
	22 de junho	sábado	31 de agosto	sábado
	08 de julho	segunda-feira	29 de agosto	quinta-feira
	09 de julho	terça-feira	30 de agosto	sexta-feira
	19 de agosto	segunda-feira	02 de setembro	segunda-feira
	20 de agosto	terça-feira	03 de setembro	terça-feira

Sobre os projetos e algumas considerações

A aula do segunda-feira (12/agosto) será de plantão de dúvidas.
Nesse dia não teremos envio de resumos

Artigo

Ordem dos projetos: data submissão

International Journal of Web & Semantic Technology (IJWesT) Vol.4, No.4, October 2013

Architecture of an Ontology-Based Domain-Specific Natural Language Question Answering System

Athira P. M., Sreeja M. and P. C. Reghuraj

Department of Computer Science and Engineering, Government Engineering College, Sreekrishnapuram, Palakkad Kerala, India, 678633

ABSTRACT

Question answering (QA) system aims at retrieving precise information from a large collection of documents against a query. This paper describes the architecture of a Natural Language Question Answering (NLQA) system for a specific domain based on the ontological information, a step towards semantic web question answering. The proposed architecture defines four basic modules suitable for enhancing current QA capabilities with the ability of processing complex questions. The first module was the question processing, which analyses and classifies the question and also reformulates the user query. The second module allows the process of retrieving the relevant documents. The next module processes the retrieved documents, and the last module performs the extraction and generation of a response. Natural language processing techniques are used for processing the question and documents and also for answer extraction. Ontology and domain knowledge are used for reformulating queries and identifying the relations. The aim of the system is to generate short and specific answer to the question that is asked in the natural language in a specific domain. We have achieved 94 % accuracy of natural language question answering in our implementation.

KEYWORDS

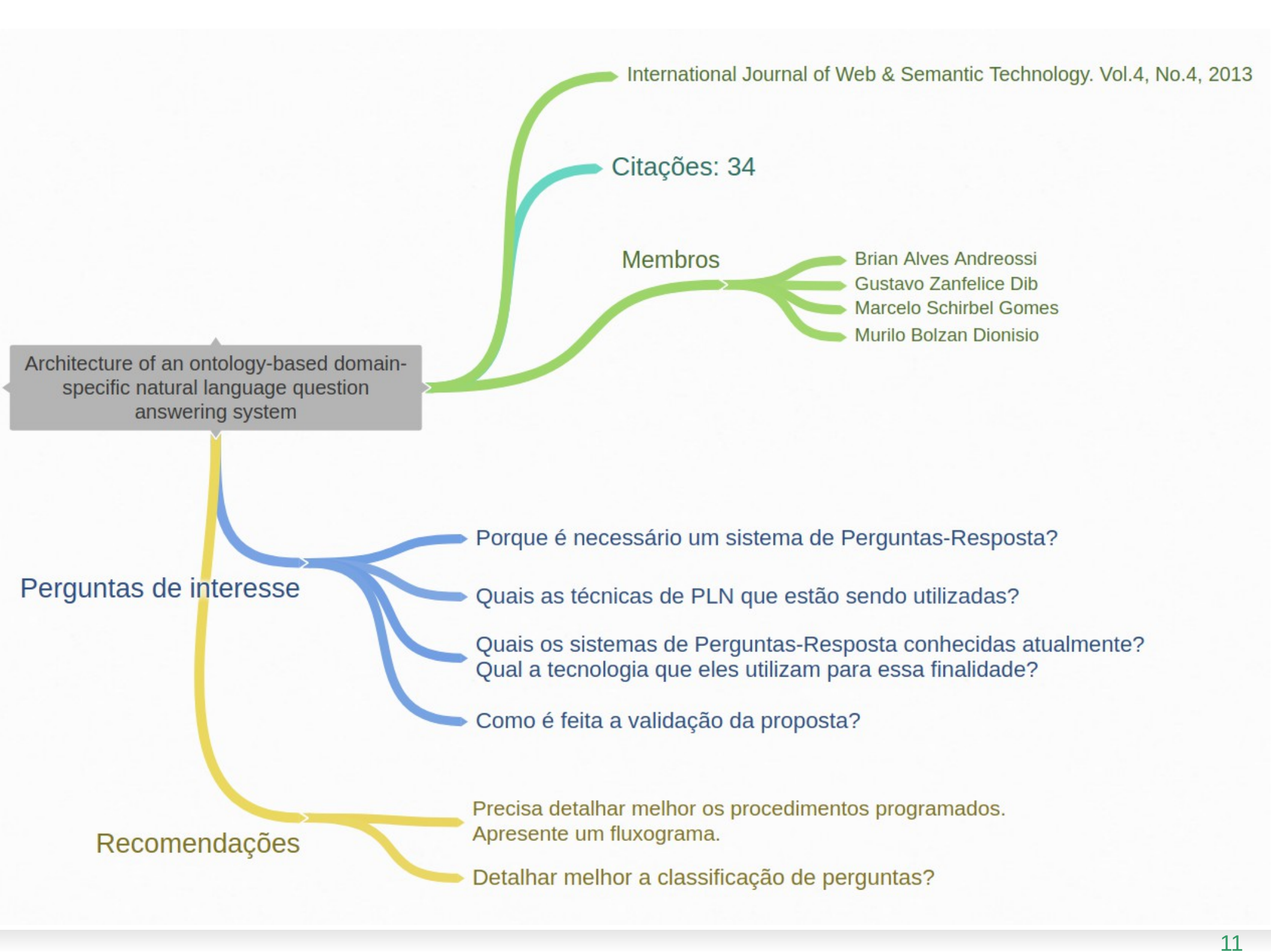
Natural Language Processing, Question Answering, Ontology, Semantic Role Labeling

1. INTRODUCTION

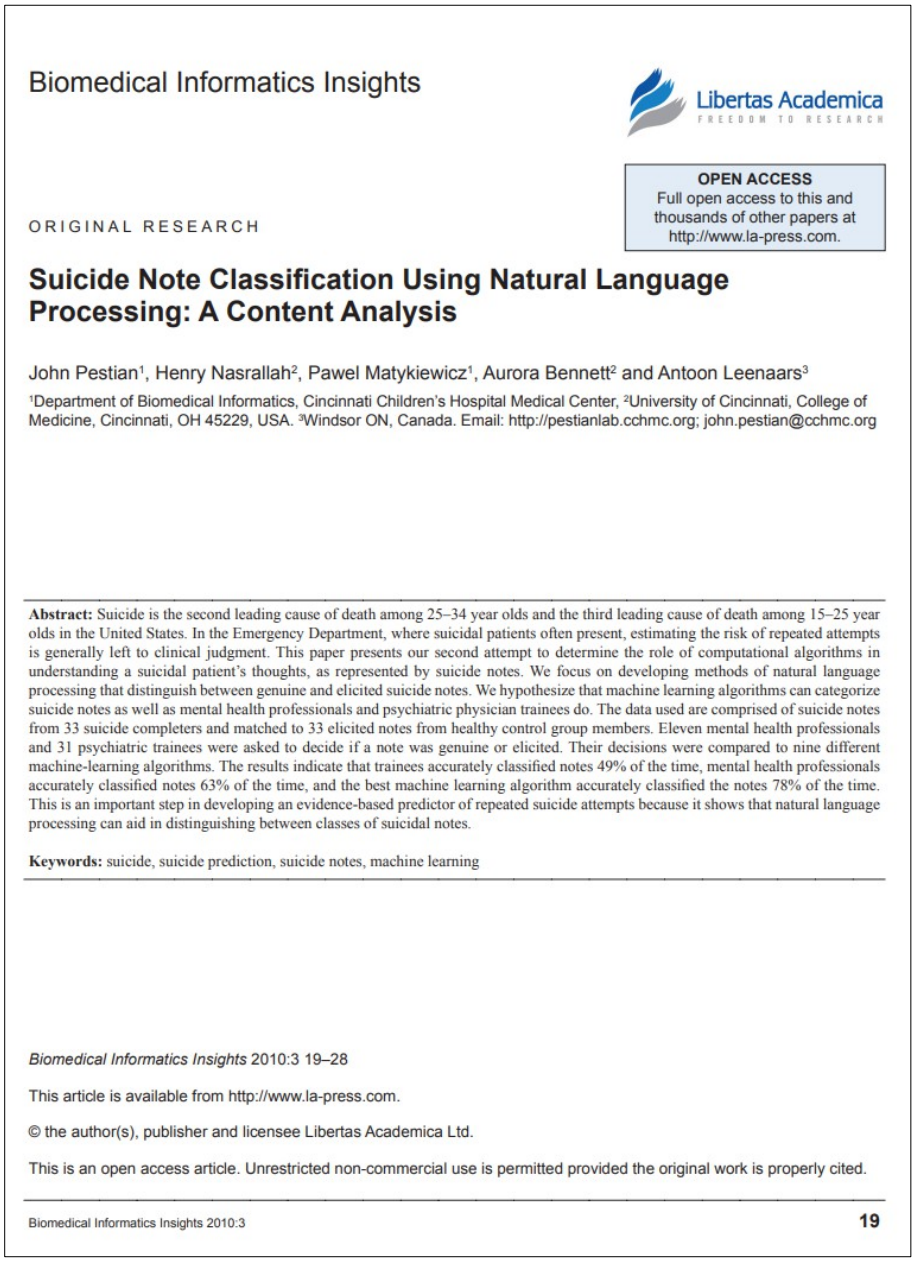
Question Answering is the process of extracting answers to natural language questions. A QA system takes questions in natural language as input, searches for answers in a set of documents, and extracts and frames concise answers. QA systems provide answers to the natural language questions by considering an archive of documents. Instead of providing the precise answers, in most of the current information retrieval systems the users have to select the required information from a ranked list of documents. Information Extraction (IE) is the name given to any process which selectively structures and combines data which is found, explicitly stated or implied, in one or more texts [5]. After finding the significant documents, the IR system submits those to the user. The scope of the QA has been constrained to domain specific systems, due to the complications in natural language processing (NLP) techniques [4]. Current search engines can return ranked lists of documents, but not the answers to the user queries.

DOI : 10.5121/ijwest.2013.4403

O objetivo do artigo é apresentar uma arquitetura para um sistema de pergunta-resposta usando ontologias.


International Journal of Web & Semantic Technology. Vol.4, No.4, 2013
Citações: 34
Membros
Brian Alves Andreossi
Gustavo Zanfelice Dib
Marcelo Schirbel Gomes
Murilo Bolzan Dionisio
Architecture of an ontology-based domain-specific natural language question answering system
Perguntas de interesse
Porque é necessário um sistema de Perguntas-Resposta?
Quais as técnicas de PLN que estão sendo utilizadas?
Quais os sistemas de Perguntas-Resposta conhecidas atualmente?
Qual a tecnologia que eles utilizam para essa finalidade?
Como é feita a validação da proposta?
Recomendações
Precisa detalhar melhor os procedimentos programados.
Apresente um fluxograma.
Detalhar melhor a classificação de perguntas?

Artigo



Biomedical Informatics Insights

Libertas Academica
FREEDOM TO RESEARCH

OPEN ACCESS
Full open access to this and thousands of other papers at http://www.la-press.com.

ORIGINAL RESEARCH

Suicide Note Classification Using Natural Language Processing: A Content Analysis

John Pestian[1], Henry Nasrallah[2], Pawel Matykiewicz[1], Aurora Bennett[2] and Antoon Leenaars[3]

[1]Department of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, [2]University of Cincinnati, College of Medicine, Cincinnati, OH 45229, USA. [3]Windsor ON, Canada. Email: http://pestianlab.cchmc.org; john.pestian@cchmc.org

Abstract: Suicide is the second leading cause of death among 25–34 year olds and the third leading cause of death among 15–25 year olds in the United States. In the Emergency Department, where suicidal patients often present, estimating the risk of repeated attempts is generally left to clinical judgment. This paper presents our second attempt to determine the role of computational algorithms in understanding a suicidal patient's thoughts, as represented by suicide notes. We focus on developing methods of natural language processing that distinguish between genuine and elicited suicide notes. We hypothesize that machine learning algorithms can categorize suicide notes as well as mental health professionals and psychiatric physician trainees do. The data used are comprised of suicide notes from 33 suicide completers and matched to 33 elicited notes from healthy control group members. Eleven mental health professionals and 31 psychiatric trainees were asked to decide if a note was genuine or elicited. Their decisions were compared to nine different machine-learning algorithms. The results indicate that trainees accurately classified notes 49% of the time, mental health professionals accurately classified notes 63% of the time, and the best machine learning algorithm accurately classified the notes 78% of the time. This is an important step in developing an evidence-based predictor of repeated suicide attempts because it shows that natural language processing can aid in distinguishing between classes of suicidal notes.

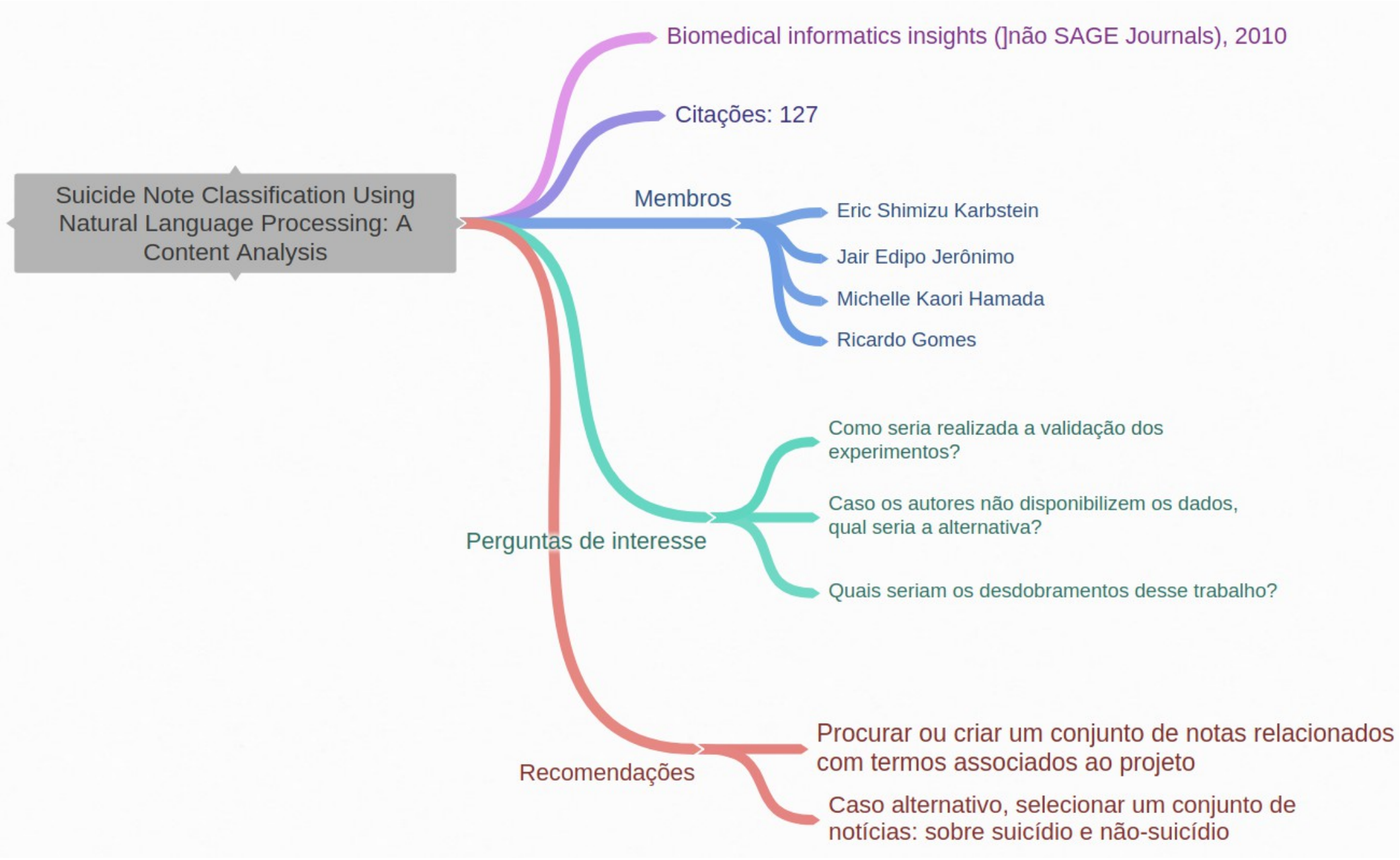
Keywords: suicide, suicide prediction, suicide notes, machine learning

Biomedical Informatics Insights 2010:3 19–28

This article is available from http://www.la-press.com.

O objetivo do artigo é apresentar um método para classificação de texto relacionado a notas de suicídio: genuínas e não-genuínas.

Foram considerados 9 algoritmos de classificação, entre eles, árvores de decisão, SVM, AdaBoost

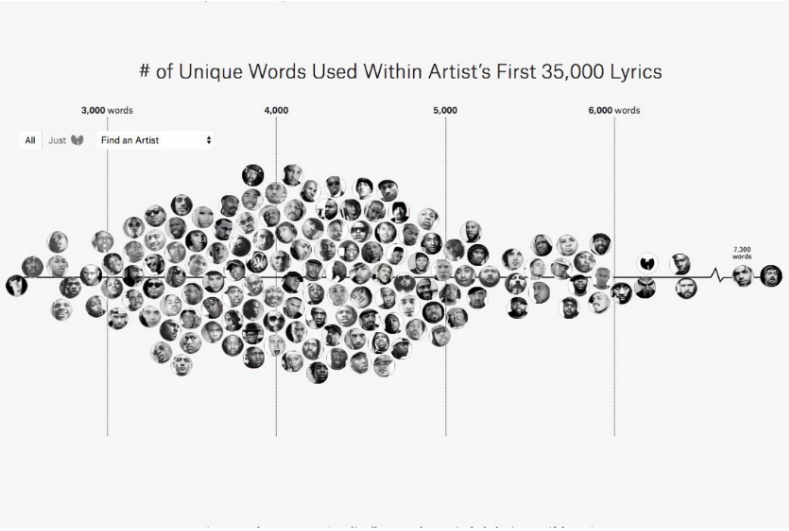

Suicide Note Classification Using Natural Language Processing: A Content Analysis
Biomedical informatics insights (]não SAGE Journals), 2010
Citações: 127
Membros
Eric Shimizu Karbstein
Jair Edipo Jerônimo
Michelle Kaori Hamada
Ricardo Gomes
Perguntas de interesse
Como seria realizada a validação dos experimentos?
Caso os autores não disponibilizem os dados, qual seria a alternativa?
Quais seriam os desdobramentos desse trabalho?
Recomendações
Procurar ou criar um conjunto de notas relacionados com termos associados ao projeto
Caso alternativo, selecionar um conjunto de notícias: sobre suicídio e não-suicídio

Artigo

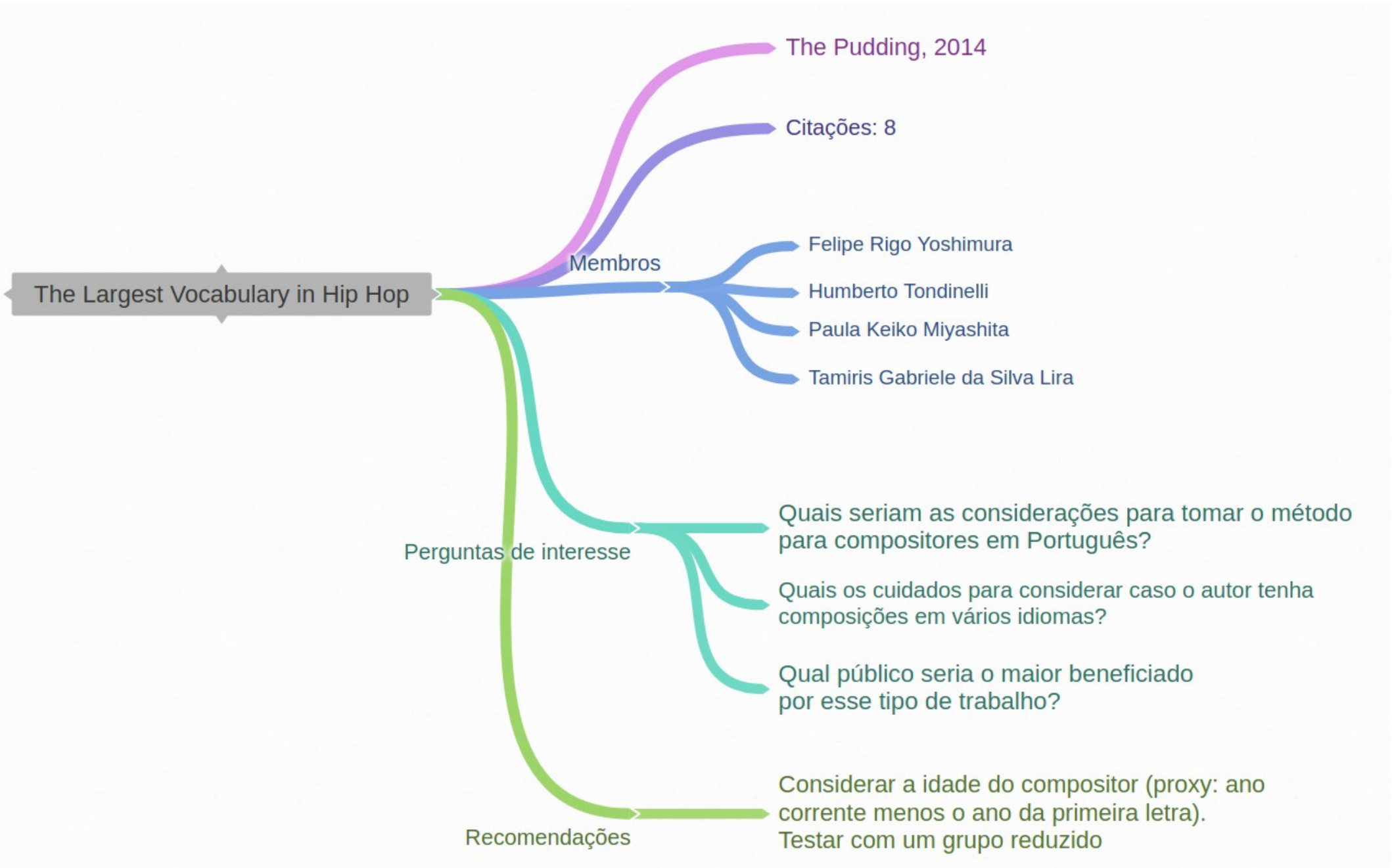
Data analyst Matt Daniels returns with an updated infographic report that showcases prominent names and the strength of their vocabulary. Once again using Rap Genius' database for his investigation, Daniels added newer lyrics for the artists already present and added an additional 75 artists to the list, including Lil Uzi Vert, Lil Yachty, Migos, and 21 Savage.

Each artist's first 35,000 lyrics (3 to 5 studio albums and EPs) are compared to for unique words, so veterans with large catalogs like JAY-Z can be compared to newer artists like Drake. Lil Uzi Vert and NF tie for the least amount of unique words in their lyrics at less than 2,650. Other newer artists, such as Lil Baby, 21 Savage and Rich the Kid also use a little number of unique words, all tying at 2,675-3,050. "Since the original release, there's now a notable trend of fewer unique words among newer artists," Daniels explains.

More veteran rappers like JAY-Z, Mobb Deep and Method Man cap off around the 4,175-4,925 area. Meanwhile, known lyricists like Jedi Mind Tricks, MF DOOM, and of course Aesop Rock top the list at 6,050-6,425+, with the ladder once again topping the list—this time sharing the spotlight with Busdriver.



O objetivo do trabalho é mensurar a diversidade lexical de cada compositor através de conceitos básicos de PLN.


The Largest Vocabulary in Hip Hop
The Pudding, 2014
Citações: 8
Membros
Felipe Rigo Yoshimura
Humberto Tondinelli
Paula Keiko Miyashita
Tamiris Gabriele da Silva Lira
Perguntas de interesse
Quais seriam as considerações para tomar o método para compositores em Português?
Quais os cuidados para considerar caso o autor tenha composições em vários idiomas?
Qual público seria o maior beneficiado por esse tipo de trabalho?
Recomendações
Considerar a idade do compositor (proxy: ano corrente menos o ano da primeira letra). Testar com um grupo reduzido

Artigo

A Rule-Based Approach to Implicit Emotion Detection in Text

Orizu Udochukwu(✉) and Yulan He

School of Engineering and Applied Science, Aston University, Birmingham, UK
{orizuus,y.he9}@aston.ac.uk

Abstract. Most research in the area of emotion detection in written text focused on detecting explicit expressions of emotions in text. In this paper, we present a rule-based pipeline approach for detecting implicit emotions in written text without emotion-bearing words based on the OCC Model. We have evaluated our approach on three different datasets with five emotion categories. Our results show that the proposed approach outperforms the lexicon matching method consistently across all the three datasets by a large margin of 17–30 % in F-measure and gives competitive performance compared to a supervised classifier. In particular, when dealing with formal text which follows grammatical rules strictly, our approach gives an average F-measure of 82.7 % on "Happy", "Angry-Disgust" and "Sad", even outperforming the supervised baseline by nearly 17 % in F-measure. Our preliminary results show the feasibility of the approach for the task of implicit emotion detection in written text.

Keywords: Implicit emotions · OCC model · Emotion detection · Rule-based approach

1 Introduction

Human emotions are defined as subjective feelings and thoughts, and is a short episode that is coordinated by the brain [4]. Emotions exist in various forms and Ekman [2] made a strong compelling case for the six basic emotion categories. In Natural language Processing (NLP), emotion detection focuses on categorising a piece of text into an emotion category. The expression of emotion in written text is through the use of words and most often emotion-bearing words such as "happy". However, emotions can be adequately expressed without the use of emotion-bearing words. For example, given two sentences "The outcome of my exam makes me happy." and "I passed my exam.", both sentences express the emotion of happiness, with the first expressing it explicitly and the second implying it. Most research in the area of emotion detection focuses on explicit emotion detection [6,9]. Implicit emotion detection is a much more difficult task and the approaches which rely on emotion lexicons are inapplicable here. Although it is possible to train supervised classifiers from annotated data, acquiring sufficient annotated data for training requires heavy manual effort.

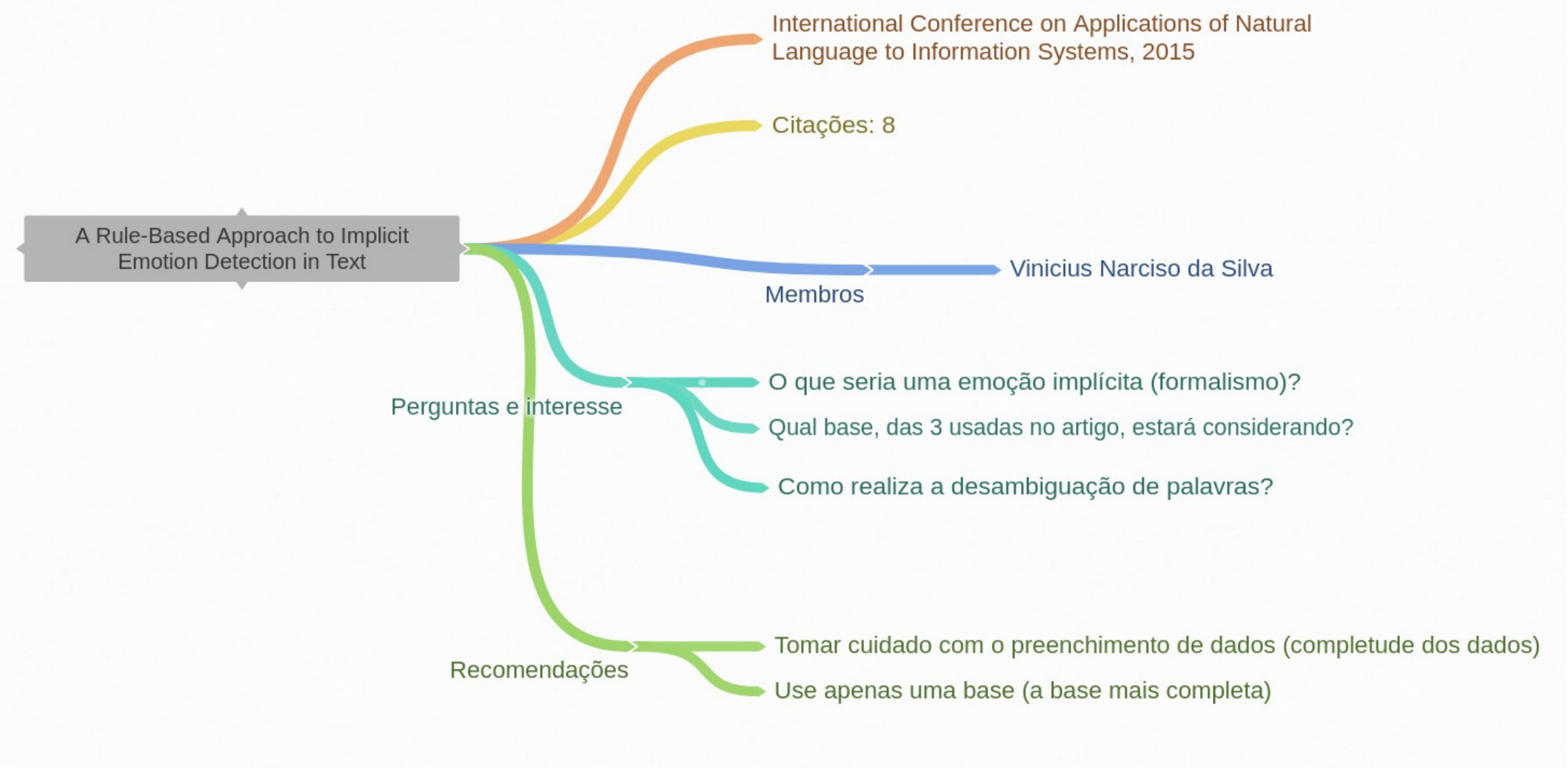
C. Biemann et al. (Eds.): NLDB 2015, LNCS 9103, pp. 197–203, 2015.
DOI: 10.1007/978-3-319-19581-0_17

O objetivo do trabalho é a criação de um método para detectar emoções implícitas:

Eu passei na UFABC -> Alegria

Ganhei o dia -> Alegria

Meu time perdeu -> Tristeza


A Rule-Based Approach to Implicit Emotion Detection in Text
International Conference on Applications of Natural Language to Information Systems, 2015
Citações: 8
Membros
Vinicius Narciso da Silva
Perguntas e interesse
O que seria uma emoção implícita (formalismo)?
Qual base, das 3 usadas no artigo, estará considerando?
Como realiza a desambiguação de palavras?
Recomendações
Tomar cuidado com o preenchimento de dados (completude dos dados)
Use apenas uma base (a base mais completa)

Artigo

BraSNAM - III Brazilian Workshop on Social Networks Analysis and Mining

Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013

Tiago C. de França[1], Jonice Oliveira,

Programa de Pós-Graduação em Informática da Universidade Federal do Rio de Janeiro (PPGI-UFRJ) – RJ – Brasil

tcruz.franca@ufrj.br, jonice@dcc.ufrj.br

Abstract. *The sentiment analysis of citizens is possible by using suitable techniques of analyzes applied to a massive database which is composed by messages provided by persons on Web. The goal of this paper is to analyze the opinion about protests that occurred in Brazil in 2013. For this, a database composed by tweets written in Brazilian Portuguese was used. This database was pre-processed for the corpus' creation. We observed that polarity (agreement or disagreement with the protests) of these messages and the final results have shown that the majority of messages are agreement ones.*

Resumo. *A análise de sentimento da população de um país é possível através da aplicação de técnicas adequadas sobre uma grande massa de dados formada por mensagens disponibilizadas pelas pessoas na Web. Este trabalho tem como objetivo analisar o sentimento acerca dos protestos que ocorreram no Brasil entre os meses de Junho e Agosto de 2013. Para tanto, foi criada uma base de tweets escritos em português brasileiro. Essa base foi pré-processada para criação do corpus de mensagens com menos ruídos. Esse corpus foi analisado para extração do sentimento presente nas mensagens. Observou-se a polaridade (apoio ou repúdio aos protestos) expressa nos tweets. Os dados foram analisados e o resultado final demonstrou que a maioria das mensagens apoiaram os protestos.*

1. Introdução

As mídias digitais da Web são fonte de uma vasta quantidade de informação disponibilizada em diferentes idiomas. Atualmente, as redes sociais na Web tem sido foco de diferentes tipos de estudo. Redes sociais *online* são redes formadas a partir da interação entre pessoas, grupos ou instituições motivadas por interesses ou objetivos comuns que se relacionam através de mídias digitais. É imensa a quantidade de usuários publicando informações de diferentes tipos e em diferentes idiomas nessas redes [Gonçalves et al. 2012; Nascimento et al. 2012].

Dentre as ferramentas que permitem a criação de redes sociais, está o Twitter[1], um *microblog* que permite que as pessoas divulguem qualquer tipo de informação quase em tempo real para todos aqueles ligados à sua rede. As publicações nessa plataforma são limitadas a um número pequeno de caracteres. Essa característica obriga que os usuários expressem sua opinião, sentimento ou qualquer informação através de mensagens curtas [Nascimento et al. 2012]. O Twitter possui mais de 200 milhões de usuários que geram aproximadamente 110 milhões de *tweets* por dia. Esses *tweets* possuem opiniões, informações pessoais ou sobre eventos em geral [Naaman e Boase 2010]. Por esse motivo, o Twitter tem sido visto como uma importante fonte

[1] https://twitter.com

O objetivo do trabalho é analisar a opinião das pessoas que utilizaram a plataforma Twitter durante os protestos que ocorreram em 2013.

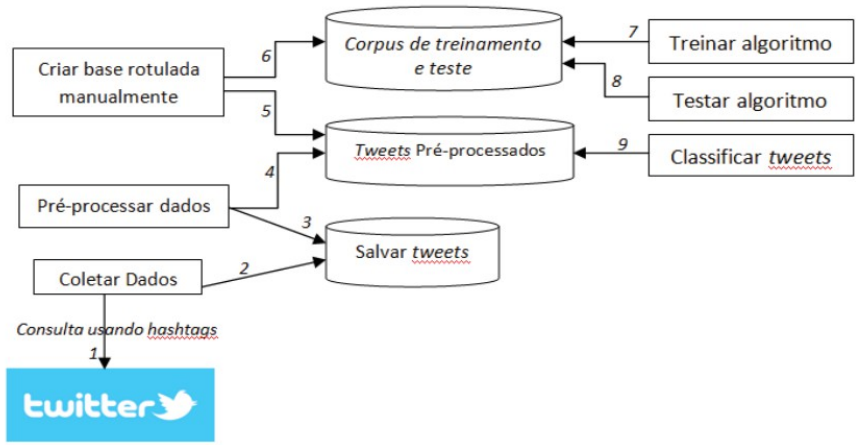
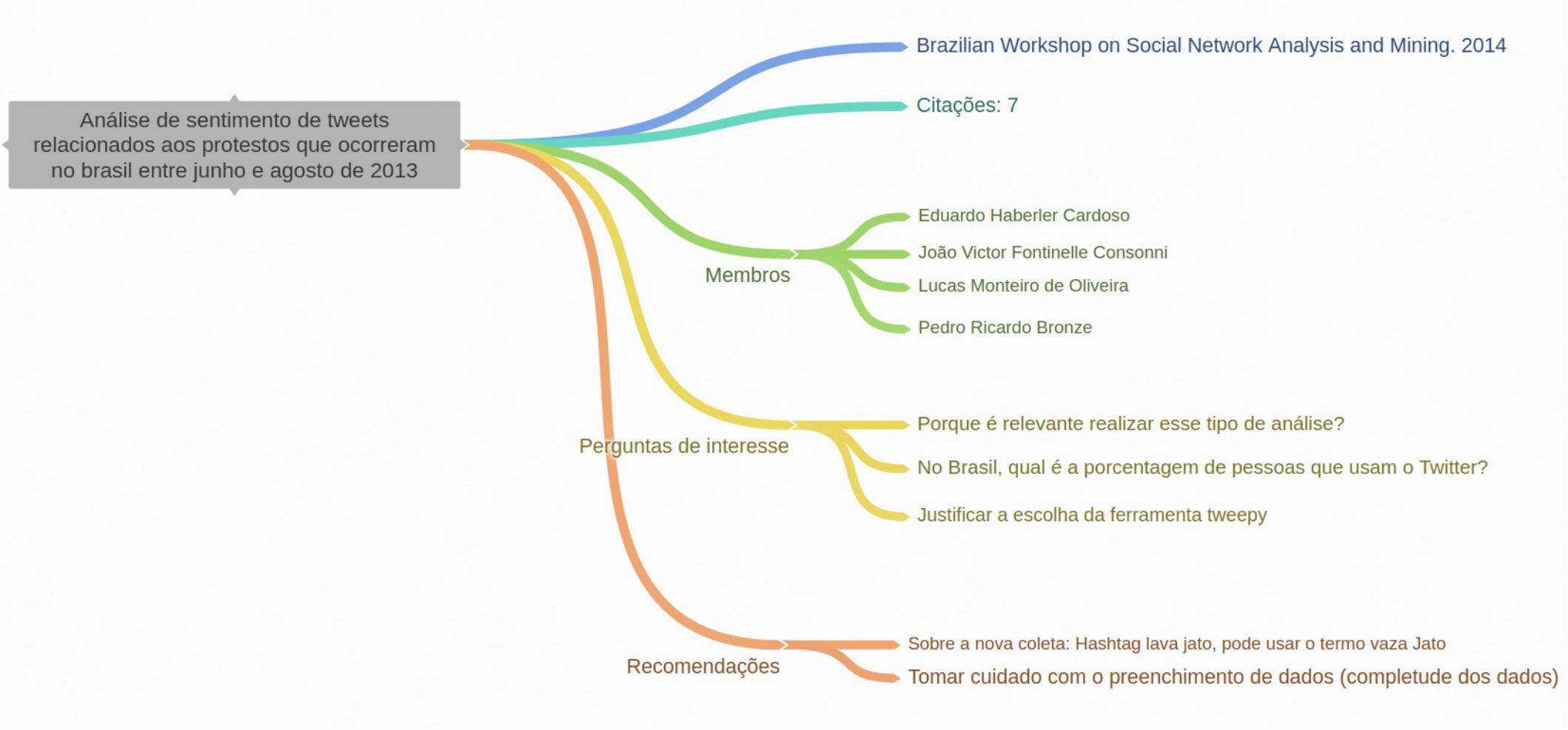


Figura 1 - Etapas da Análise de Polaridade dos *tweets*


Análise de sentimento de tweets relacionados aos protestos que ocorreram no brasil entre junho e agosto de 2013
Brazilian Workshop on Social Network Analysis and Mining. 2014
Citações: 7
Membros
Eduardo Haberler Cardoso
João Victor Fontinelle Consonni
Lucas Monteiro de Oliveira
Pedro Ricardo Bronze
Perguntas de interesse
Porque é relevante realizar esse tipo de análise?
No Brasil, qual é a porcentagem de pessoas que usam o Twitter?
Justificar a escolha da ferramenta tweepy
Recomendações
Sobre a nova coleta: Hashtag lava jato, pode usar o termo vaza Jato
Tomar cuidado com o preenchimento de dados (completude dos dados)

Artigo

Comparative Study of CNN and RNN for Natural Language Processing

Wenpeng Yin[†], **Katharina Kann**[†], **Mo Yu**[‡] and **Hinrich Schütze**[†]
[†]CIS, LMU Munich, Germany
[‡]IBM Research, USA
{wenpeng,kann}@cis.lmu.de, yum@us.ibm.com

Abstract

Deep neural networks (DNNs) have revolutionized the field of natural language processing (NLP). Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), the two main types of DNN architectures, are widely explored to handle various NLP tasks. CNN is supposed to be good at extracting position-invariant features and RNN at modeling units in sequence. The state-of-the-art on many NLP tasks often switches due to the battle of CNNs and RNNs. This work is the first systematic comparison of CNN and RNN on a wide range of representative NLP tasks, aiming to give basic guidance for DNN selection.

1 Introduction

Natural language processing (NLP) has benefited greatly from the resurgence of deep neural networks (DNNs), due to their high performance with less need of engineered features. There are two main DNN architectures: convolutional neural network (CNN) (LeCun et al., 1998) and recurrent neural network (RNN) (Elman, 1990). Gating mechanisms have been developed to alleviate some limitations of the basic RNN, resulting in two prevailing RNN types: long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and gated recurrent unit (GRU) (Cho et al., 2014).

Generally speaking, CNNs are hierarchical and RNNs sequential architectures. How should we choose between them for processing language? Based on the characterization "hierarchical (CNN) vs. sequential (RNN)", it is tempting to choose a CNN for classification tasks like sentiment classification since sentiment is usually determined by some key phrases; and to choose RNNs for a sequence modeling task like language modeling as it requires flexible modeling of context dependencies. But current NLP literature does not support such a clear conclusion. For example, RNNs perform well on document-level sentiment classification (Tang et al., 2015); and Dauphin et al. (2016) recently showed that gated CNNs outperform LSTMs on language modeling tasks, even though LSTMs had long been seen as better suited. In summary, there is no consensus on DNN selection for any particular NLP problem.

This work compares CNNs, GRUs and LSTMs systematically on a broad array of NLP tasks: sentiment/relation classification, textual entailment, answer selection, question-relation matching in Freebase, Freebase path query answering and part-of-speech tagging.

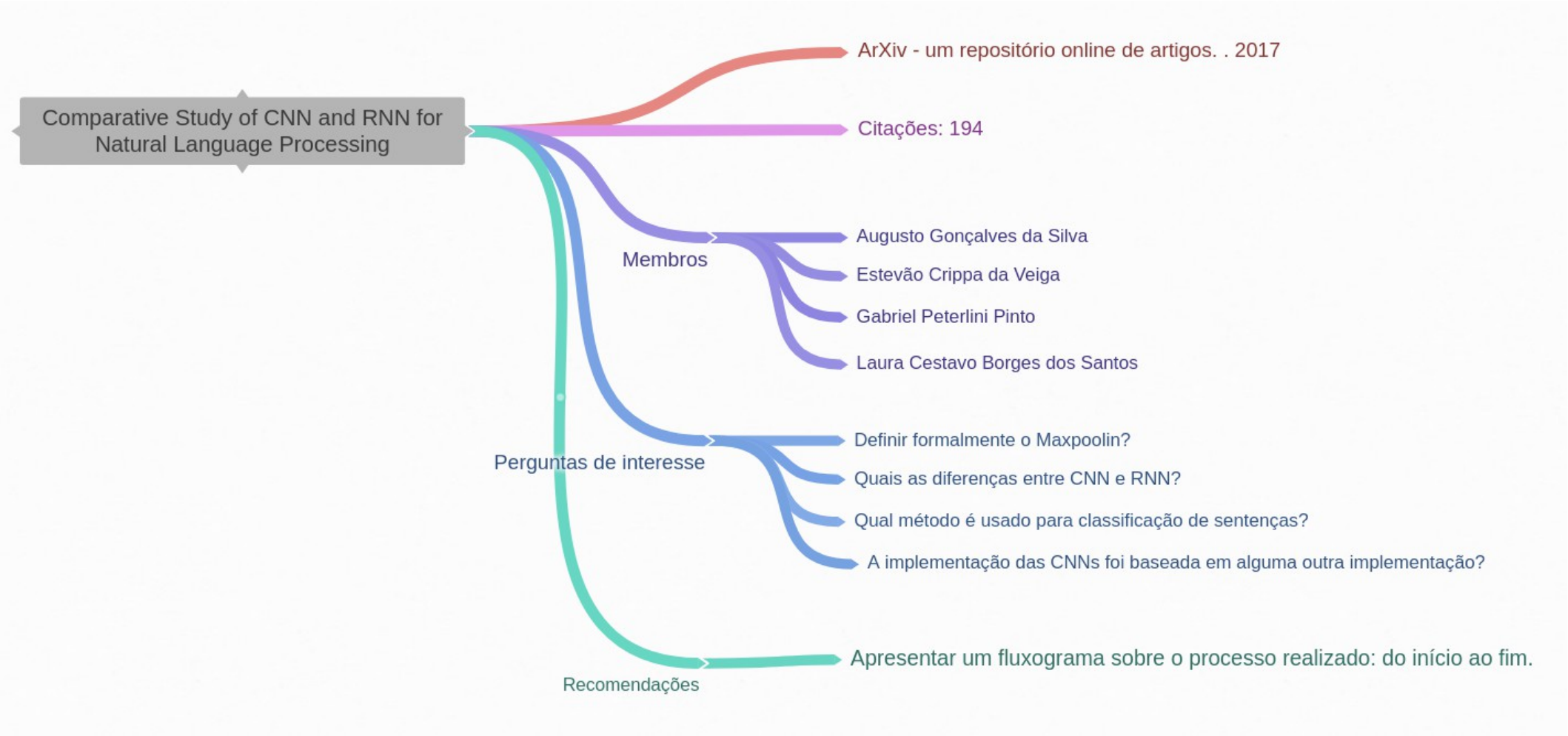
Our experiments support two key findings. (i) CNNs and RNNs provide complementary information for text classification tasks. Which architecture performs better depends on how important it is to *semantically understand the whole sequence*. (ii) Learning rate changes performance relatively smoothly, while changes to hidden size and batch size result in large fluctuations.

2 Related Work

To our knowledge, there has been no systematic comparison of CNN and RNN on a large array of NLP tasks.

Vu et al. (2016) investigate CNN and basic RNN (i.e., no gating mechanisms) for relation classification. They report higher performance of CNN than RNN and give evidence that CNN and RNN provide complementary information: while the RNN computes a weighted combination of all words in the sentence, the CNN extracts the most informative ngrams for the relation and only considers their resulting activations.

O objetivo do trabalho é apresentar elementos quantitativos para comparar redes neurais artificiais **profundas** com redes neurais **convolucionais** no contexto de PLN.


Comparative Study of CNN and RNN for Natural Language Processing
ArXiv - um repositório online de artigos. . 2017
Citações: 194
Membros
Augusto Gonçalves da Silva
Estevão Crippa da Veiga
Gabriel Peterlini Pinto
Laura Cestavo Borges dos Santos
Perguntas de interesse
Definir formalmente o Maxpoolin?
Quais as diferenças entre CNN e RNN?
Qual método é usado para classificação de sentenças?
A implementação das CNNs foi baseada em alguma outra implementação?
Recomendações
Apresentar um fluxograma sobre o processo realizado: do início ao fim.

Artigo

A Survey on Hate Speech Detection using Natural Language Processing

Anna Schmidt
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
anna.schmidt@lsv.uni-saarland.de

Michael Wiegand
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
michael.wiegand@lsv.uni-saarland.de

Abstract

This paper presents a survey on hate speech detection. Given the steadily growing body of social media content, the amount of online hate speech is also increasing. Due to the massive scale of the web, methods that automatically detect hate speech are required. Our survey describes key areas that have been explored to automatically recognize these types of utterances using natural language processing. We also discuss limits of those approaches.

1 Introduction

Hate speech is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic (Nockleby, 2000). Examples are (1)-(3).[1]

(1) Go fucking kill yourself and die already useless ugly pile of shit scumbag.
(2) The Jew Faggot Behind The Financial Collapse
(3) Hope one of those bitches falls over and breaks her leg

Due to the massive rise of user-generated web content, in particular on social media networks, the amount of hate speech is also steadily increasing. Over the past years, interest in online hate speech detection and particularly the automatization of this task has continuously grown, along with the societal impact of the phenomenon. Natural language processing focusing specifically on this phenomenon is required since basic word filters do not provide a sufficient remedy: What is considered a hate speech message might be influenced by aspects such as the domain of an utterance, its discourse context, as well as context consisting of co-occurring media objects (e.g. images, videos, audio), the exact time of posting and world events at this moment, identity of author and targeted recipient.

This paper provides a short, comprehensive and structured overview of automatic hate speech detection, and outlines the existing approaches in a systematic manner, focusing on feature extraction in particular. It is mainly aimed at NLP researchers who are new to the field of hate speech detection and want to inform themselves about the state of the art.

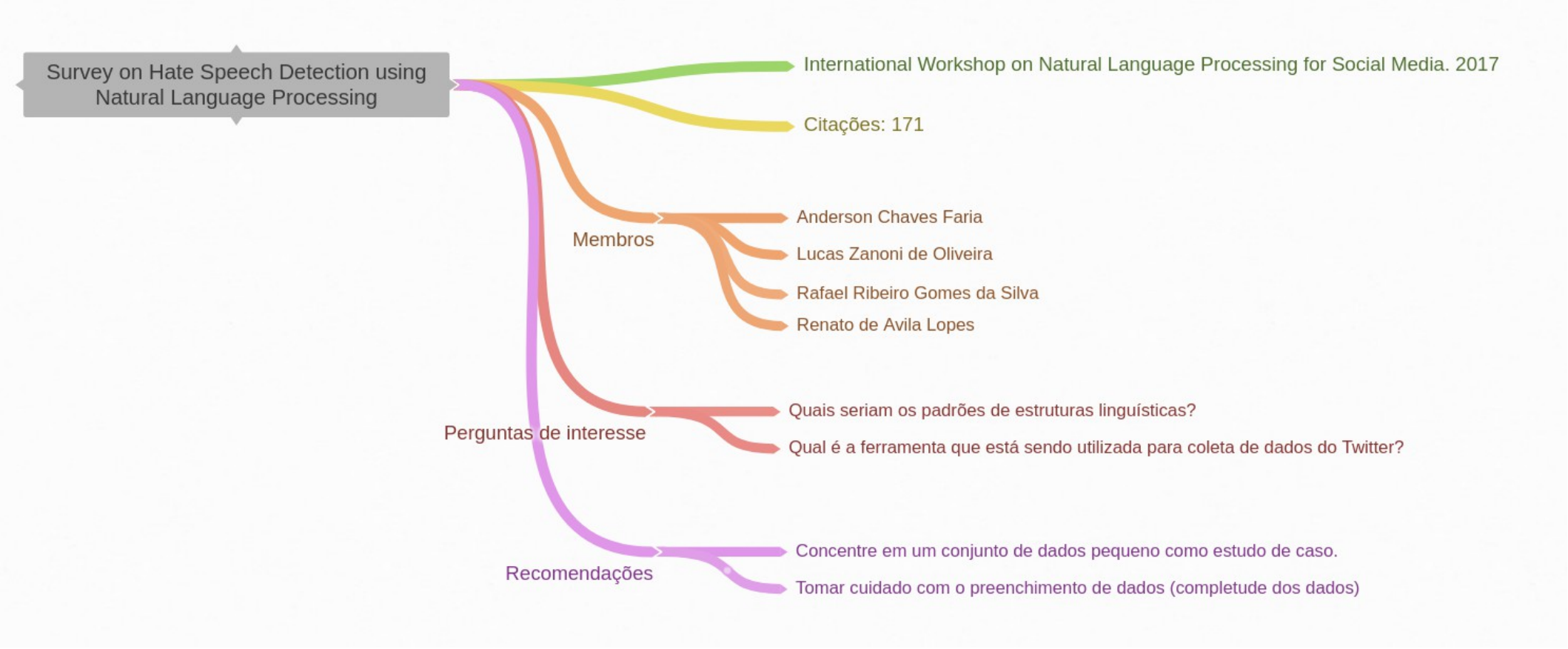
2 Terminology

In this paper we use the term ***hate speech***. We decided in favour of using this term since it can be considered a broad umbrella term for numerous kinds of insulting user-created content addressed in the individual works we summarize in this paper. *Hate speech* is also the most frequently used expression for this phenomenon, and is even a legal term in several countries. Below we list other terms that are used in the NLP community. This should also help readers with finding further literature on that task.

In the earliest work on hate speech, Spertus (1997) refers to *abusive* messages, *hostile* messages or *flames*. More recently, many authors have shifted to employing the term *cyberbullying* (Xu et al., 2012; Hosseinmardi et al., 2015; Zhong et al., 2016; Van Hee et al., 2015; Dadvar et al., 2013; Dinakar et al., 2012). The actual term *hate speech* is used by Warner and Hirschberg (2012), Burnap and Williams (2015), Silva et al. (2016), Djuric et al. (2015), Gitari et al. (2015), Williams and Burnap (2015) and Kwok and Wang (2013). Further,

[1]The examples in this work are included to illustrate the severity of the hate speech problem. They are taken from actual web data and in no way reflect the opinion of the authors.

O objetivo do trabalho é a criação de um método para identificar discurso de ódio usando como fonte de dados a Plataforma Twitter.


Survey on Hate Speech Detection using Natural Language Processing
International Workshop on Natural Language Processing for Social Media. 2017
Citações: 171
Membros
Anderson Chaves Faria
Lucas Zanoni de Oliveira
Rafael Ribeiro Gomes da Silva
Renato de Avila Lopes
Perguntas de interesse
Quais seriam os padrões de estruturas linguísticas?
Qual é a ferramenta que está sendo utilizada para coleta de dados do Twitter?
Recomendações
Concentre em um conjunto de dados pequeno como estudo de caso.
Tomar cuidado com o preenchimento de dados (completude dos dados)

Artigo

Fast and accurate sentiment classification using an enhanced Naive Bayes model.

Vivek Narayanan[1], Ishan Arora[2], Arjun Bhatia[3]

Department of Electronics Engineering,
Indian Institute of Technology (BHU), Varanasi, India

[1]vivek.narayanan.ece09@iitbhu.ac.in
[2]ishan.arora.ece09@iitbhu.ac.in
[3]arjun.bhatia.ece09@iitbhu.ac.in

Abstract. We have explored different methods of improving the accuracy of a Naive Bayes classifier for sentiment analysis. We observed that a combination of methods like effective negation handling, word n-grams and feature selection by mutual information results in a significant improvement in accuracy. This implies that a highly accurate and fast sentiment classifier can be built using a simple Naive Bayes model that has linear training and testing time complexities. We achieved an accuracy of 88.80% on the popular IMDB movie reviews dataset. The proposed method can be generalized to a number of text categorization problems for improving speed and accuracy.

Keywords :- Sentiment classification, Negation Handling, Mutual Information, Feature Selection, n-grams

1 Introduction

Among the most researched topics of natural language processing is sentiment analysis. Sentiment analysis involves extraction of subjective information from documents like online reviews to determine the polarity with respect to certain objects. It is useful for identifying trends of public opinion in the social media, for the purpose of marketing and consumer research. It has its uses in getting customer feedback about new product launches, political campaigns and even in financial markets [14]. It aims to determine the attitude of a speaker or a writer with respect to some topic or simply the contextual polarity of a document. Early work in this area was done by Turney and Pang ([2], [7]) who applied different methods for detecting the polarity of product and movie reviews.

Sentiment analysis is a complicated problem but experiments have been done using Naive Bayes, maximum entropy classifiers and support vector machines. Pang et al. found the SVM to be the most accurate classifier in [2]. In this paper we present a supervised sentiment classification model based on the Naïve Bayes algorithm.

O objetivo do trabalho é a criação de um método para Classificação de sentimento usando o modelo básico de Naive Bayes.


Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model
International Conference on Intelligent Data Engineering and Automated Learning. 2013
Citações: 193
Membros
Denildo Veloso Braga
Igor Neres Trindade
Lucas Ferraz Nicolau
Ramon Neres Teixeira Jardim
Perguntas de interesse
Quais as características do conjunto de dados considerado?
Quais as técnicas, inéditas, que não tratamos em aula, mas que são utilizados neste trabalho?
Recomendações
Estimar o tempo computacional da proposta
Como validar os resultados obtidos?

Artigo

Recognizing Emotion Presence in Natural Language Sentences

Isidoros Perikos and Ioannis Hatzilygeroudis

School of Engineering, Department of Computer Engineering & Informatics
University of Patras, 26500 Patras, Hellas, Greece
{perikos,ihatz}@ceid.upatras.gr

Abstract. Emotions constitute a key factor in human communication. Human emotion can be expressed through various mediums such as speech, facial expressions, gestures and textual data. A quite common way for people to communicate with each other and with computer systems is via written text. In this paper we present an emotion detection system used to automatically recognize emotions in text. The system takes as input natural language sentences, analyzes them and determines the underlying emotion being conveyed. It implements a keyword-based approach where the emotional state of a sentence is constituted by the emotional affinity of the sentence's emotional words. The system uses lexical resources to spot words known to have emotional content and analyses sentence structure to specify their strength. Experimental results indicate quite satisfactory performance.

Keywords: Sentiment Analysis, Emotion Recognition, Affective Computing, Human Computer Interaction, Natural Language Processing.

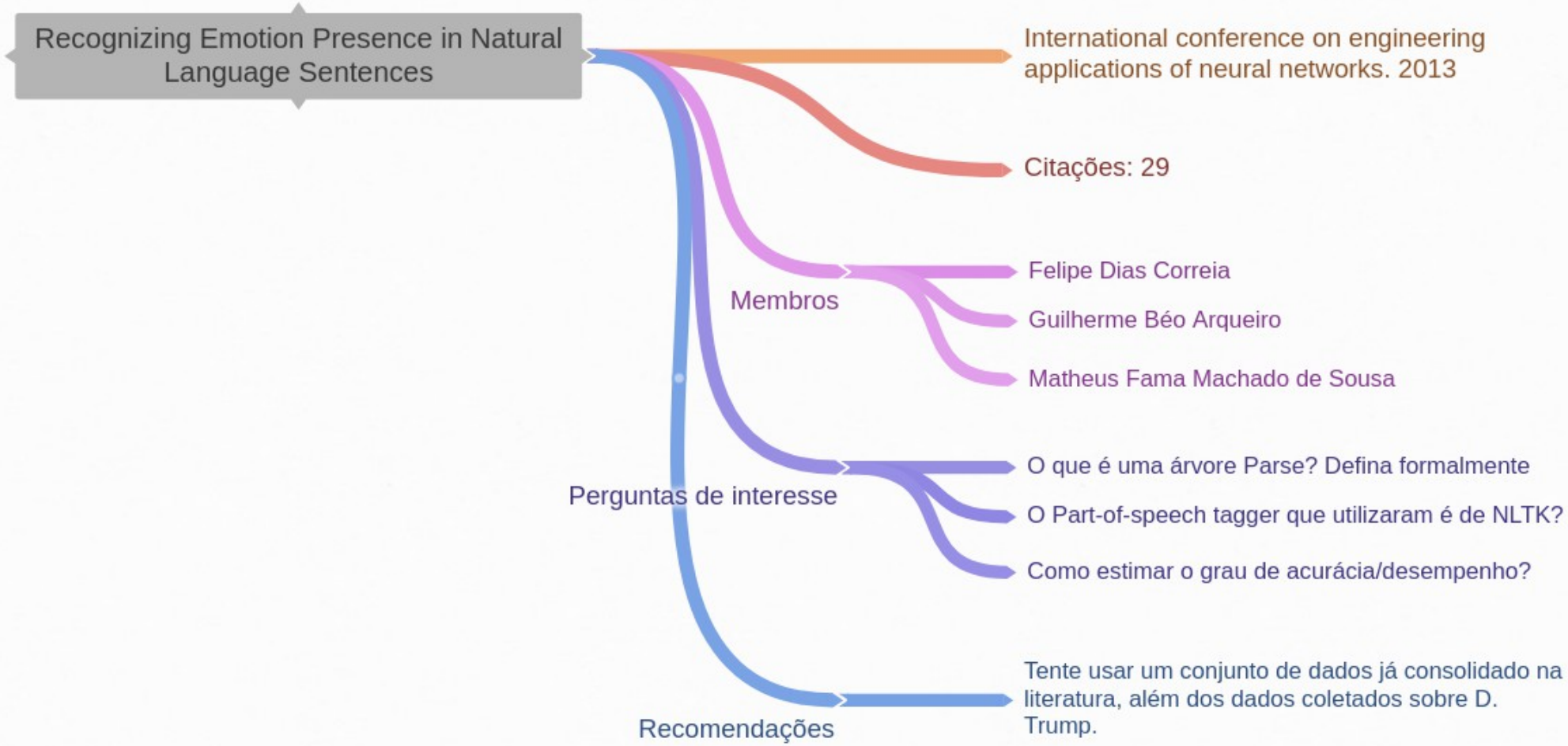
1 Introduction

Computer systems are increasingly involved in almost all aspects of everyday life. As the field of artificial intelligence matures and grows, it enhances the capabilities and the functionality of computer systems. A fundamental aspect of computer systems concerns the way that human interact and communicate with them. It becomes more and more important to be able to interact with them in a natural way, similar to the way we interact with other humans.

Emotions constitute a key factor of human nature, which colors the way of human communication. The role of emotions in human computer interaction was initially investigated by Picard, who introduced the concept of affective computing [12], indicating the importance of emotions in human computer interaction and drawing a direction for interdisciplinary research from areas such as computer science, cognitive science and psychology. The aim of affective computing is to enable computers to recognize and express emotions and bridge the gap between the emotional human and the computer by developing computational systems that recognize and adapt to the user's emotional states [3]. Automatically recognizing and responding to a user's affective states can enhance the quality of the interaction, thereby making a computer

L. Iliadis, H. Papadopoulos, and C. Jayne (Eds.): EANN 2013, Part II, CCIS 384, pp. 30–39, 2013.

O objetivo do trabalho é a criação de um sistema reconhecedor de emoções em textos.


Recognizing Emotion Presence in Natural Language Sentences
International conference on engineering applications of neural networks. 2013
Citações: 29
Membros
Felipe Dias Correia
Guilherme Béo Arqueiro
Matheus Fama Machado de Sousa
Perguntas de interesse
O que é uma árvore Parse? Defina formalmente
O Part-of-speech tagger que utilizaram é de NLTK?
Como estimar o grau de acurácia/desempenho?
Recomendações
Tente usar um conjunto de dados já consolidado na literatura, além dos dados coletados sobre D. Trump.

Artigo

BraSNAM - III Brazilian Workshop on Social Networks Analysis and Mining

Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013

Tiago C. de França[1], Jonice Oliveira,

Programa de Pós-Graduação em Informática da Universidade Federal do Rio de Janeiro (PPGI-UFRJ) – RJ – Brasil

tcruz.franca@ufrj.br, jonice@dcc.ufrj.br

Abstract. *The sentiment analysis of citizens is possible by using suitable techniques of analyzes applied to a massive database which is composed by messages provided by persons on Web. The goal of this paper is to analyze the opinion about protests that occurred in Brazil in 2013. For this, a database composed by tweets written in Brazilian Portuguese was used. This database was pre-processed for the corpus' creation. We observed that polarity (agreement or disagreement with the protests) of these messages and the final results have shown that the majority of messages are agreement ones.*

Resumo. *A análise de sentimento da população de um país é possível através da aplicação de técnicas adequadas sobre uma grande massa de dados formada por mensagens disponibilizadas pelas pessoas na Web. Este trabalho tem como objetivo analisar o sentimento acerca dos protestos que ocorreram no Brasil entre os meses de Junho e Agosto de 2013. Para tanto, foi criada uma base de tweets escritos em português brasileiro. Essa base foi pré-processada para criação do corpus de mensagens com menos ruídos. Esse corpus foi analisado para extração do sentimento presente nas mensagens. Observou-se a polaridade (apoio ou repúdio aos protestos) expressa nos tweets. Os dados foram analisados e o resultado final demonstrou que a maioria das mensagens apoiaram os protestos.*

1. Introdução

As mídias digitais da Web são fonte de uma vasta quantidade de informação disponibilizada em diferentes idiomas. Atualmente, as redes sociais na Web tem sido foco de diferentes tipos de estudo. Redes sociais *online* são redes formadas a partir da interação entre pessoas, grupos ou instituições motivadas por interesses ou objetivos comuns que se relacionam através de mídias digitais. É imensa a quantidade de usuários publicando informações de diferentes tipos e em diferentes idiomas nessas redes [Gonçalves et al. 2012; Nascimento et al. 2012].

Dentre as ferramentas que permitem a criação de redes sociais, está o Twitter[1], um *microblog* que permite que as pessoas divulguem qualquer tipo de informação quase em tempo real para todos aqueles ligados à sua rede. As publicações nessa plataforma são limitadas a um número pequeno de caracteres. Essa característica obriga que os usuários expressem sua opinião, sentimento ou qualquer informação através de mensagens curtas [Nascimento et al. 2012]. O Twitter possui mais de 200 milhões de usuários que geram aproximadamente 110 milhões de *tweets* por dia. Esses *tweets* possuem opiniões, informações pessoais ou sobre eventos em geral [Naaman e Boase 2010]. Por esse motivo, o Twitter tem sido visto como uma importante fonte

[1] https://twitter.com

O objetivo do trabalho é analisar a opinião das pessoas que utilizaram a plataforma Twitter durante os protestos que ocorreram em 2013.

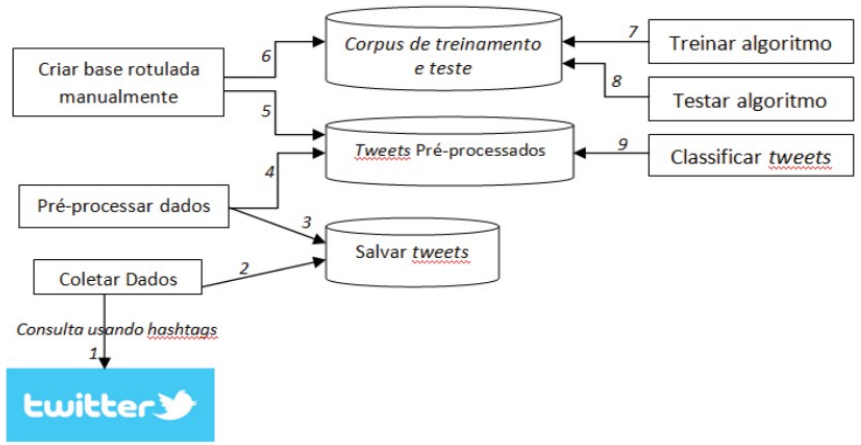
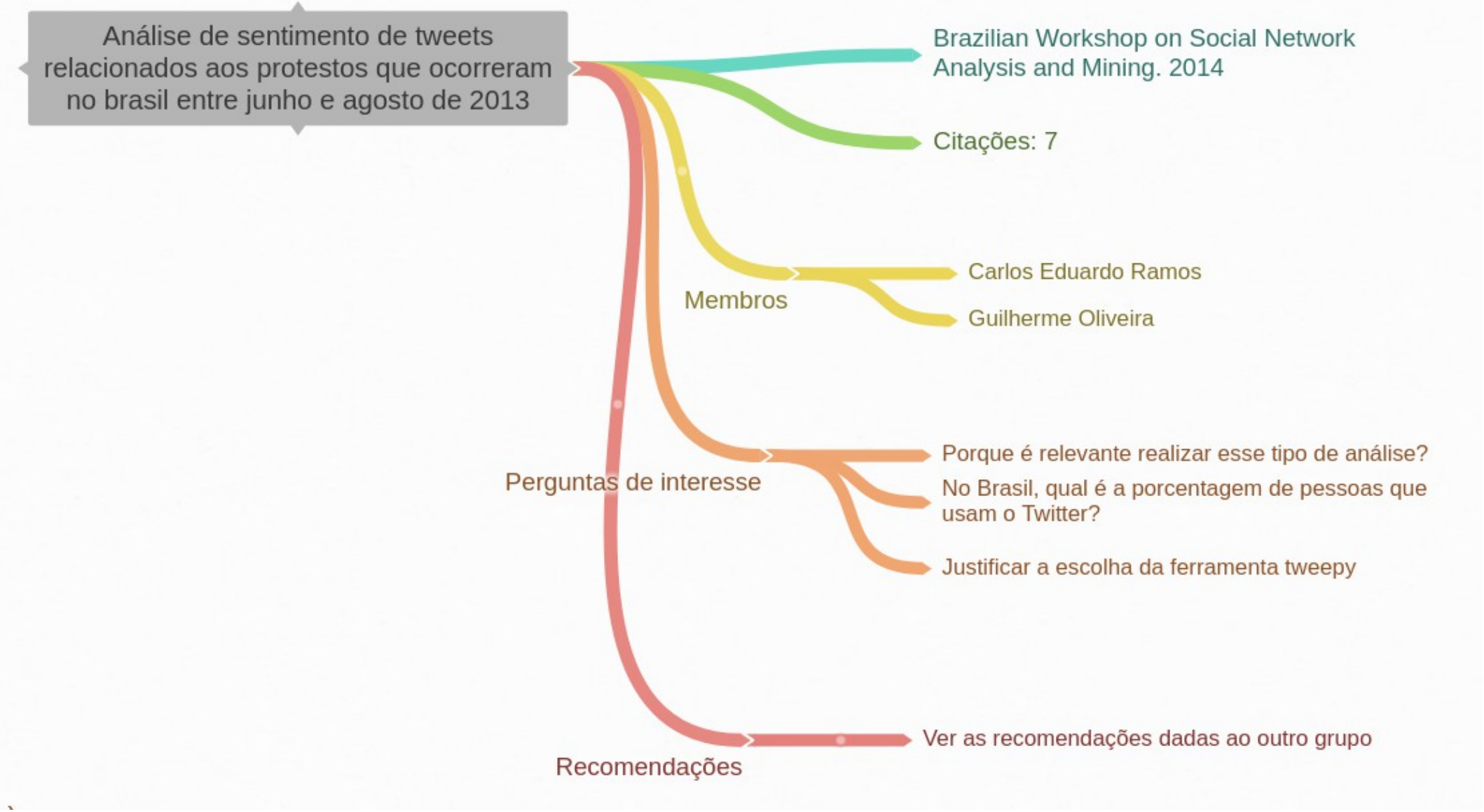


Figura 1 - Etapas da Análise de Polaridade dos *tweets*


Análise de sentimento de tweets relacionados aos protestos que ocorreram no brasil entre junho e agosto de 2013
Brazilian Workshop on Social Network Analysis and Mining. 2014
Citações: 7
Membros
Carlos Eduardo Ramos
Guilherme Oliveira
Perguntas de interesse
Porque é relevante realizar esse tipo de análise?
No Brasil, qual é a porcentagem de pessoas que usam o Twitter?
Justificar a escolha da ferramenta tweepy
Recomendações
Ver as recomendações dadas ao outro grupo

Artigo

SEQ2SQL: GENERATING STRUCTURED QUERIES FROM NATURAL LANGUAGE USING REINFORCEMENT LEARNING

Victor Zhong, Caiming Xiong, & Richard Socher
Salesforce Research
Palo Alto, CA
{vzhong,cxiong,rsocher}@salesforce.com

ABSTRACT

Relational databases store a significant amount of the world's knowledge. However, users are limited in their ability to access this knowledge due to a lack of understanding of query languages such as SQL. We propose Seq2SQL, a deep neural network for translating natural language questions to corresponding SQL queries. Our model leverages the structure of SQL queries to reduce the output space of generated queries. Moreover, it uses rewards from in-the-loop query execution over the database to learn a policy to generate unordered parts of the query, which are less suitable for optimization via cross entropy loss. In addition, we release WikiSQL, a dataset of 87673 hand-annotated examples of questions and SQL queries distributed across 26521 tables from Wikipedia. This dataset is required to train Seq2SQL and is an order of magnitude larger than comparable datasets. By applying policy-based reinforcement learning with a query execution environment to WikiSQL, Seq2SQL outperforms a state-of-the-art semantic parser by Dong & Lapata (2016), improving execution accuracy from 35.9% to 60.3% and logical form accuracy from 23.4% to 49.2%.

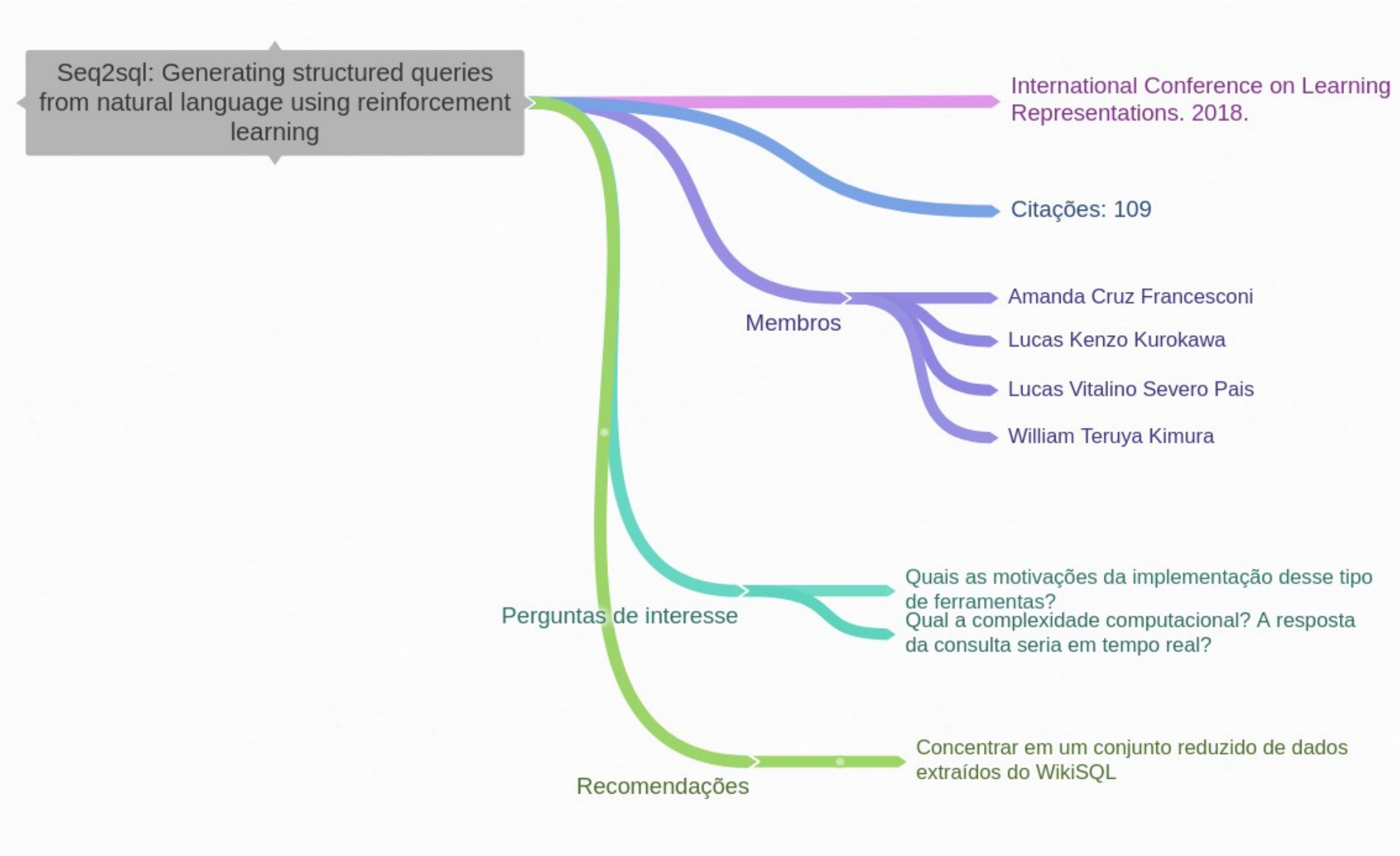
1 INTRODUCTION

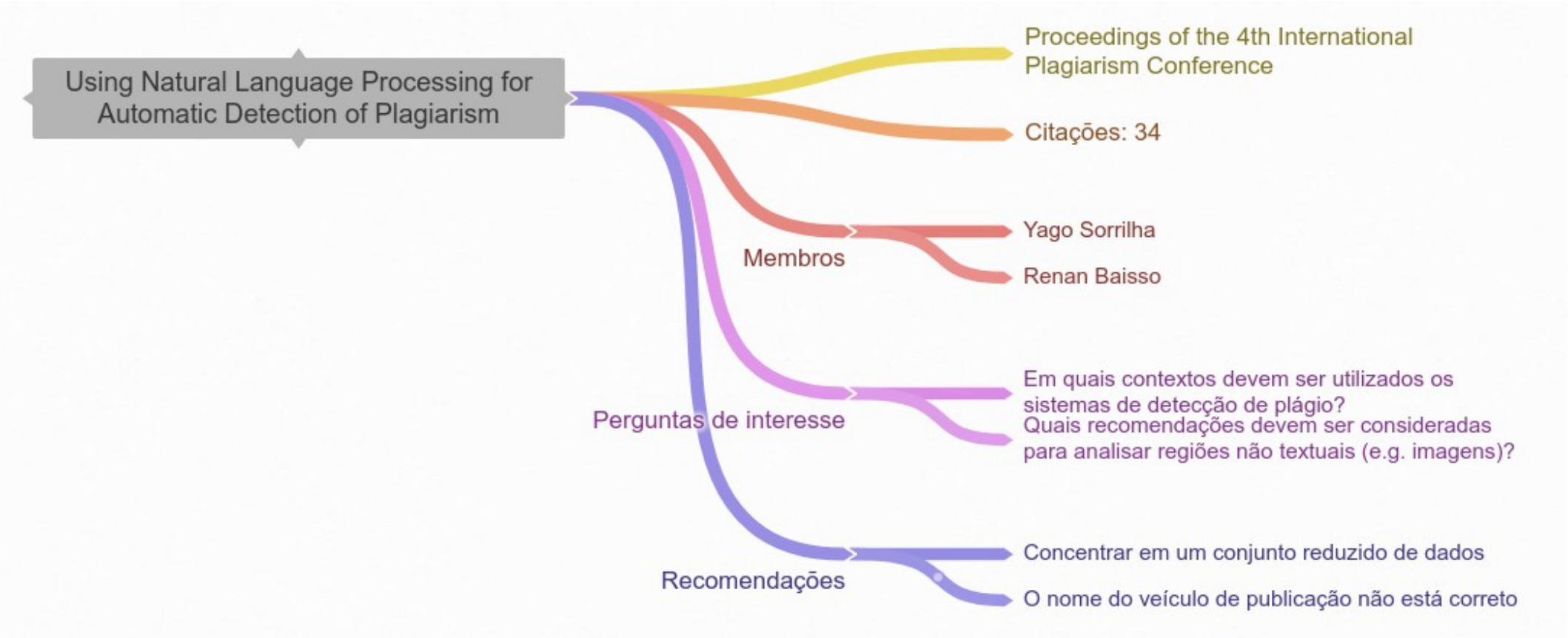
Relational databases store a vast amount of today's information and provide the foundation of applications such as medical records (Hillestad et al., 2005), financial markets (Beck et al., 2000), and customer relations management (Ngai et al., 2009). However, accessing relational databases requires an understanding of query languages such as SQL, which, while powerful, is difficult to master. Natural language interfaces (NLI), a research area at the intersection of natural language processing and human-computer interactions, seeks to provide means for humans to interact with computers through the use of natural language (Androutsopoulos et al., 1995). We investigate one particular aspect of NLI applied to relational databases: translating natural language questions to SQL queries.

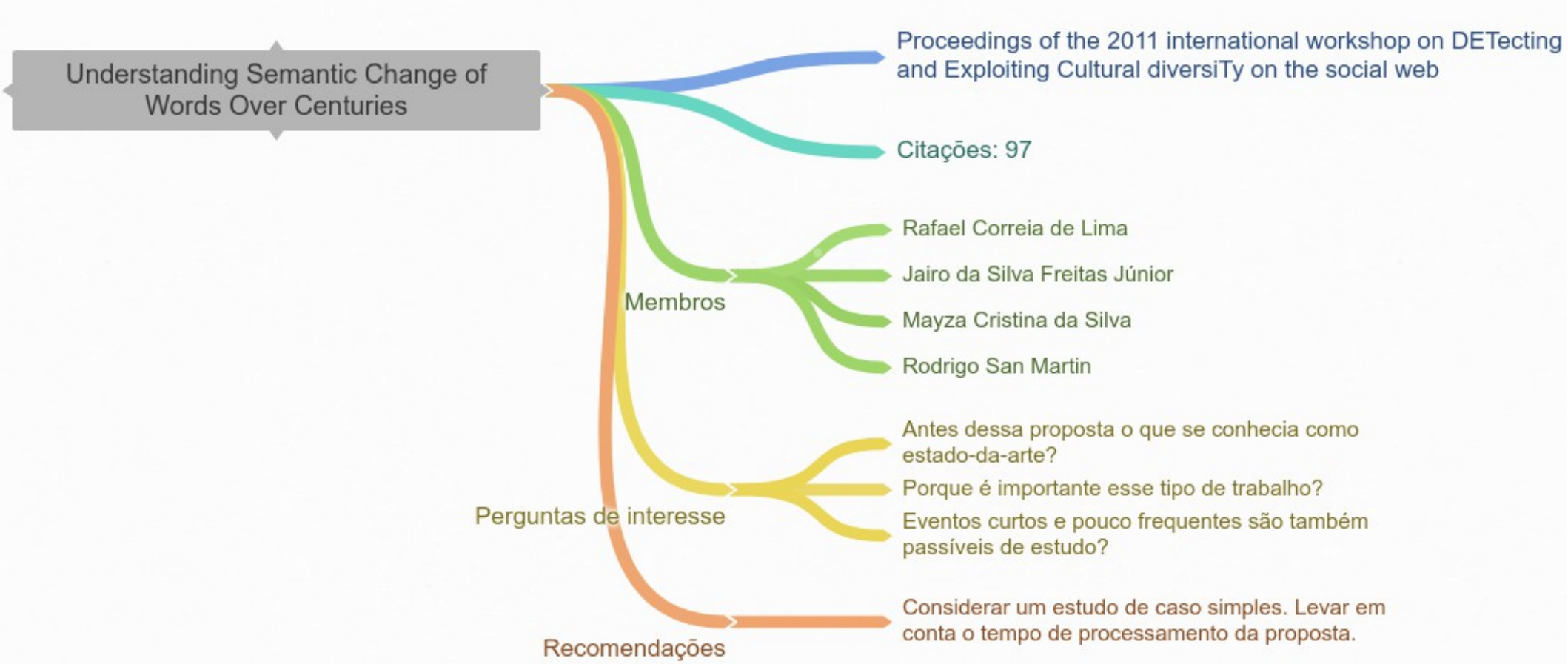
Our main contributions in this work are two-fold. First, we introduce Seq2SQL, a deep neural network for translating natural language questions to corresponding SQL queries. Seq2SQL, shown in Figure 1, consists of three components that leverage the structure of a SQL query to reduce the output space of generated queries. Moreover, it uses policy-based reinforcement learning (RL) to generate the conditions of the query, which are unsuitable for optimization using cross entropy loss due to their unordered nature. We train Seq2SQL using a mixed objective combining cross entropy losses and RL rewards from in-the-loop query execution on a database. These characteristics allow Seq2SQL to achieve state-of-the-art results on query generation.

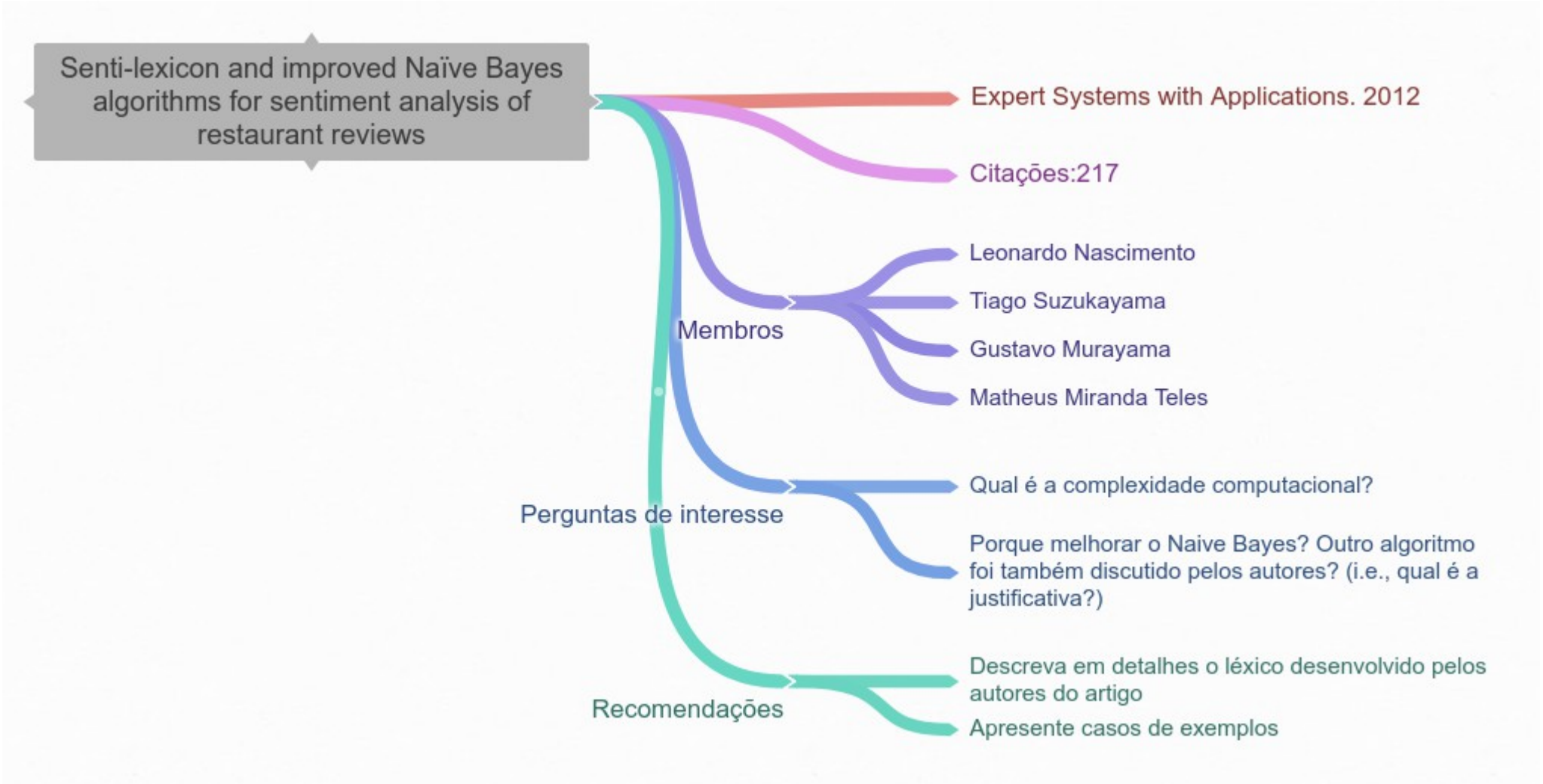
Second, we release WikiSQL, a corpus of 87673 hand-annotated instances of natural language questions, SQL queries, and SQL tables extracted from 26521 HTML tables from Wikipedia. WikiSQL is an order of magnitude larger than previous semantic parsing datasets that provide logical forms along with natural language utterances. We release the tables used in WikiSQL both in raw JSON format as well as in the form of a SQL database. Along with WikiSQL, we release a query execution engine for the database used for in-the-loop query execution to learn the policy. On Wik-

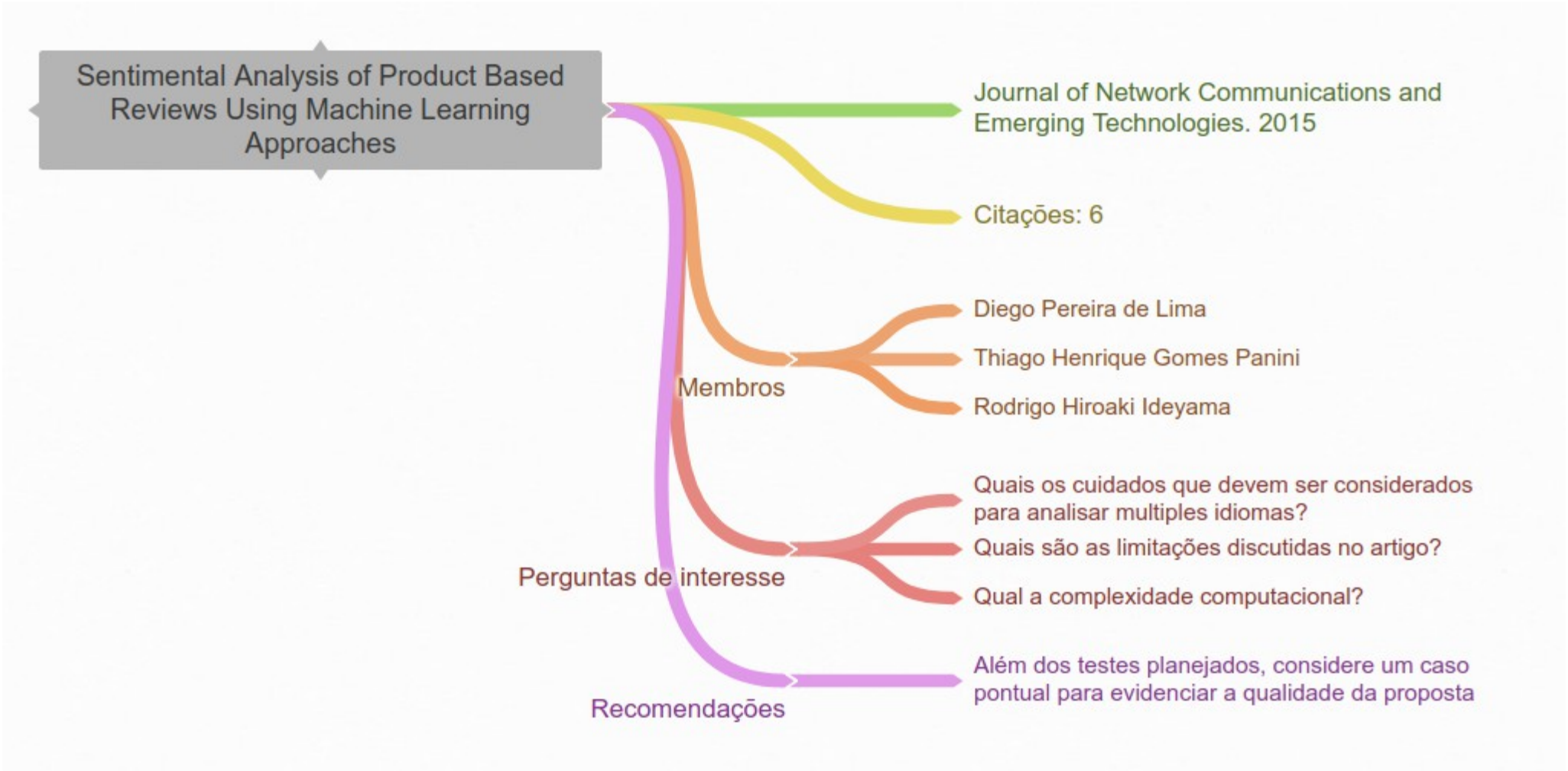
O objetivo do trabalho é criar um método para traduzir textos em linguagem natural para instruções SQL.

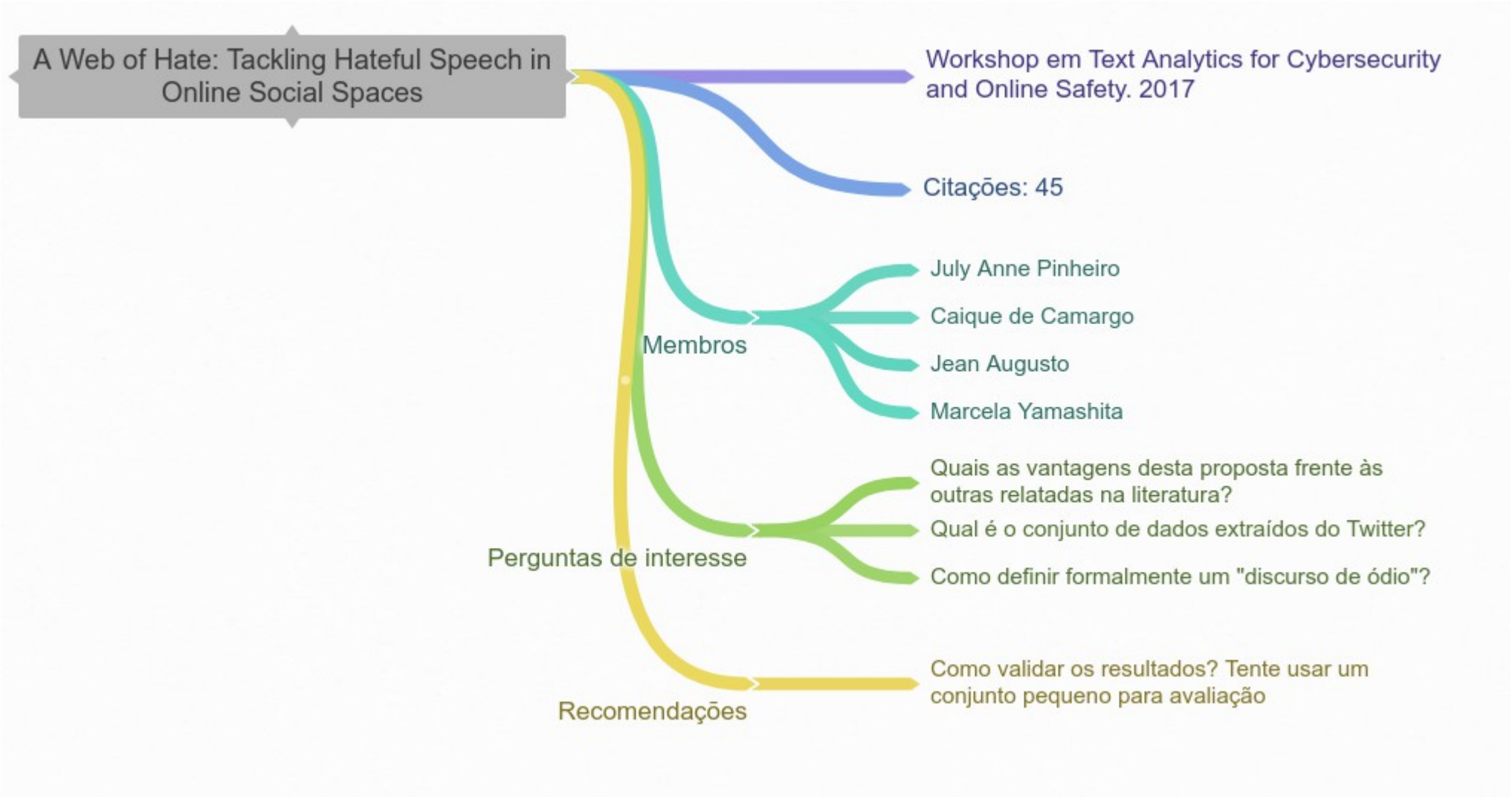

Seq2sql: Generating structured queries from natural language using reinforcement learning
International Conference on Learning Representations. 2018.
Citações: 109
Membros
Amanda Cruz Francesconi
Lucas Kenzo Kurokawa
Lucas Vitalino Severo Pais
William Teruya Kimura
Perguntas de interesse
Quais as motivações da implementação desse tipo de ferramentas?
Qual a complexidade computacional? A resposta da consulta seria em tempo real?
Recomendações
Concentrar em um conjunto reduzido de dados extraídos do WikiSQL

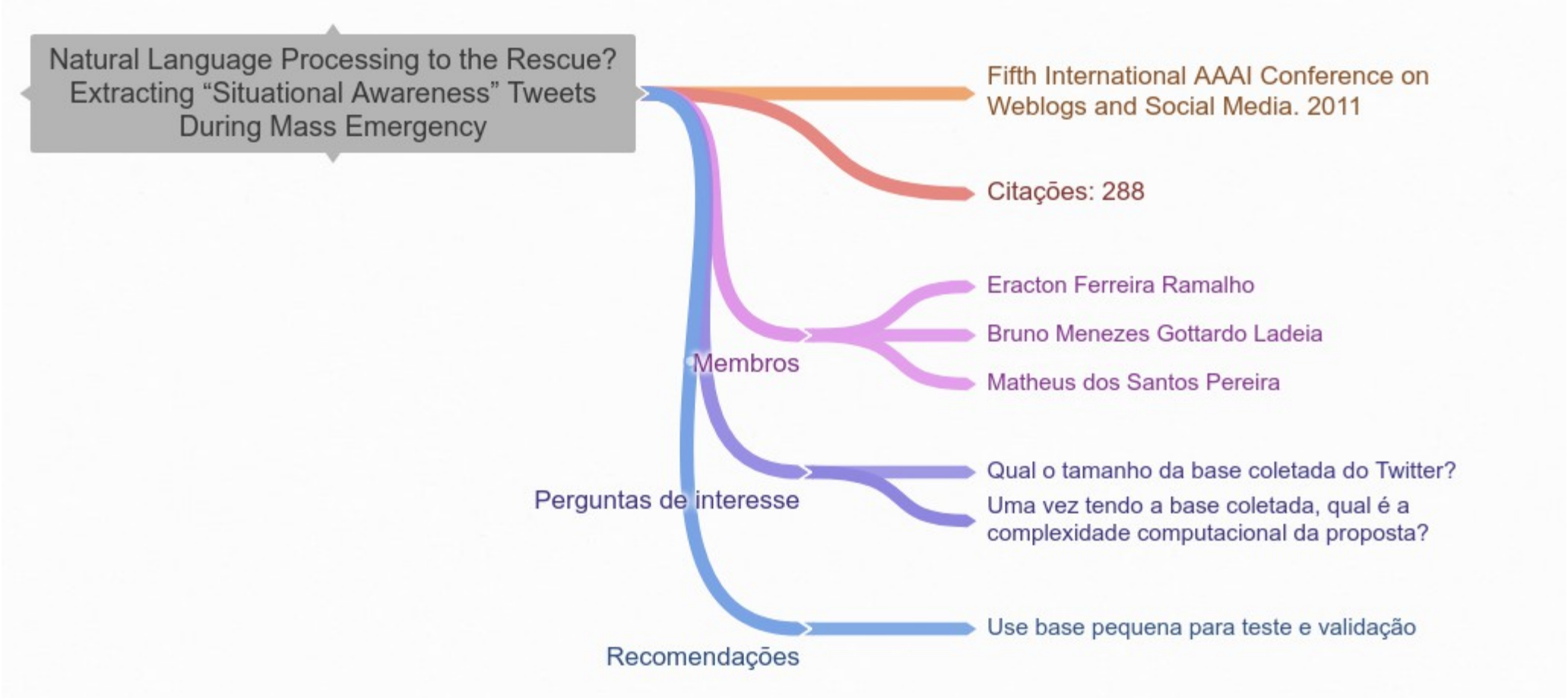

Using Natural Language Processing for Automatic Detection of Plagiarism
Proceedings of the 4th International Plagiarism Conference
Citações: 34
Membros
Yago Sorrilha
Renan Baisso
Perguntas de interesse
Em quais contextos devem ser utilizados os sistemas de detecção de plágio?
Quais recomendações devem ser consideradas para analisar regiões não textuais (e.g. imagens)?
Recomendações
Concentrar em um conjunto reduzido de dados
O nome do veículo de publicação não está correto

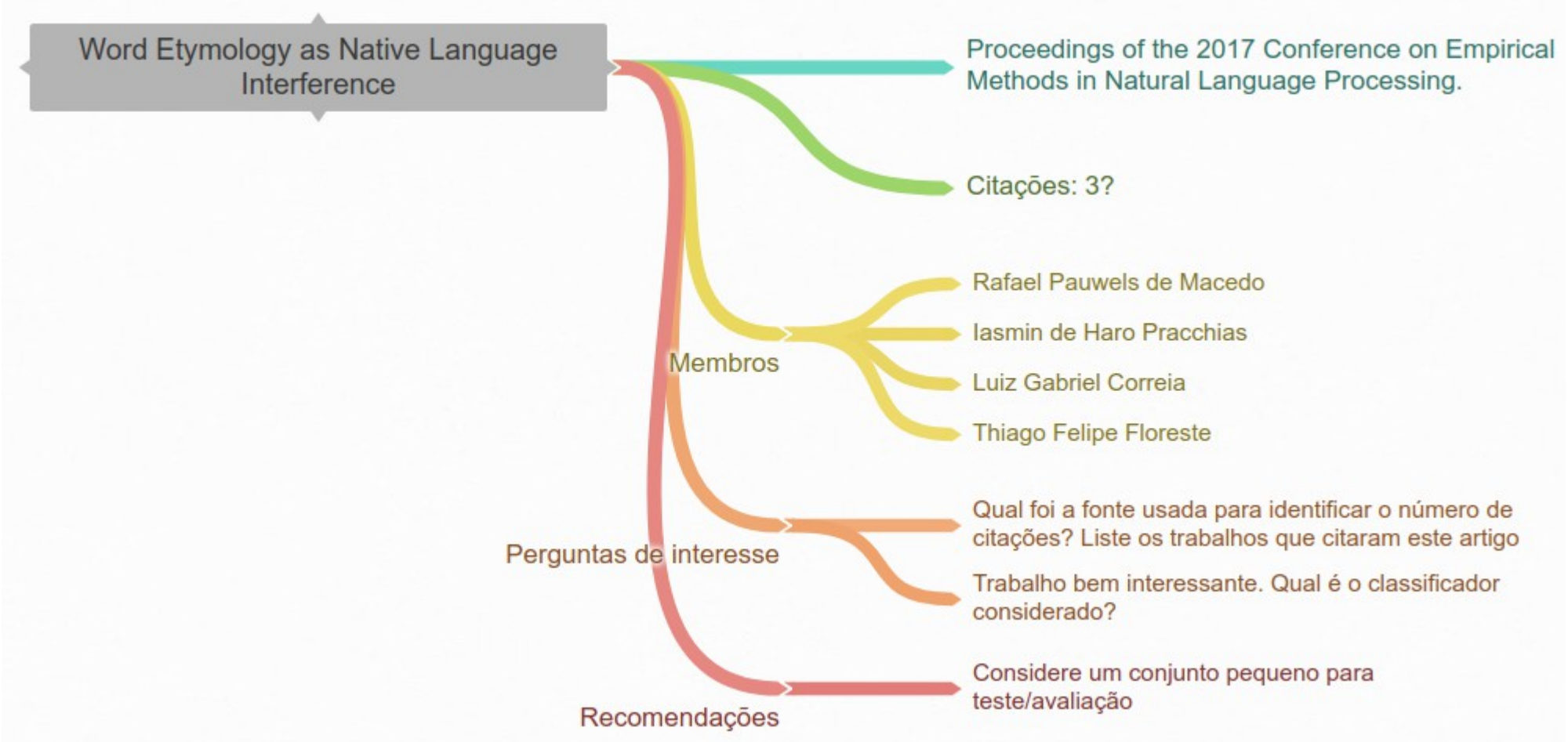

Understanding Semantic Change of Words Over Centuries
Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web
Citações: 97
Membros
Rafael Correia de Lima
Jairo da Silva Freitas Júnior
Mayza Cristina da Silva
Rodrigo San Martin
Perguntas de interesse
Antes dessa proposta o que se conhecia como estado-da-arte?
Porque é importante esse tipo de trabalho?
Eventos curtos e pouco frequentes são também passíveis de estudo?
Recomendações
Considerar um estudo de caso simples. Levar em conta o tempo de processamento da proposta.

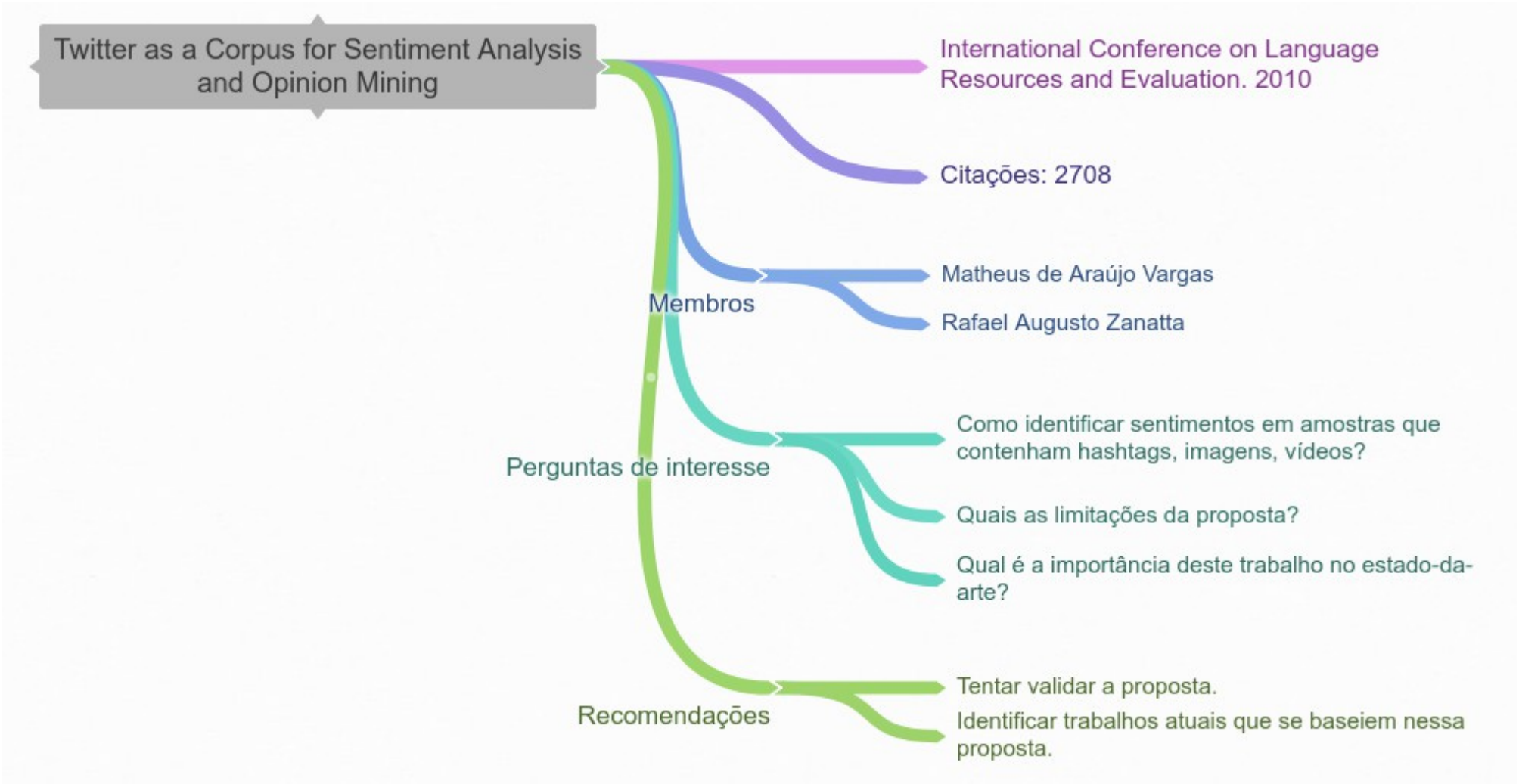

Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews
Expert Systems with Applications. 2012
Citações:217
Membros
Leonardo Nascimento
Tiago Suzukayama
Gustavo Murayama
Matheus Miranda Teles
Perguntas de interesse
Qual é a complexidade computacional?
Porque melhorar o Naive Bayes? Outro algoritmo foi também discutido pelos autores? (i.e., qual é a justificativa?)
Recomendações
Descreva em detalhes o léxico desenvolvido pelos autores do artigo
Apresente casos de exemplos

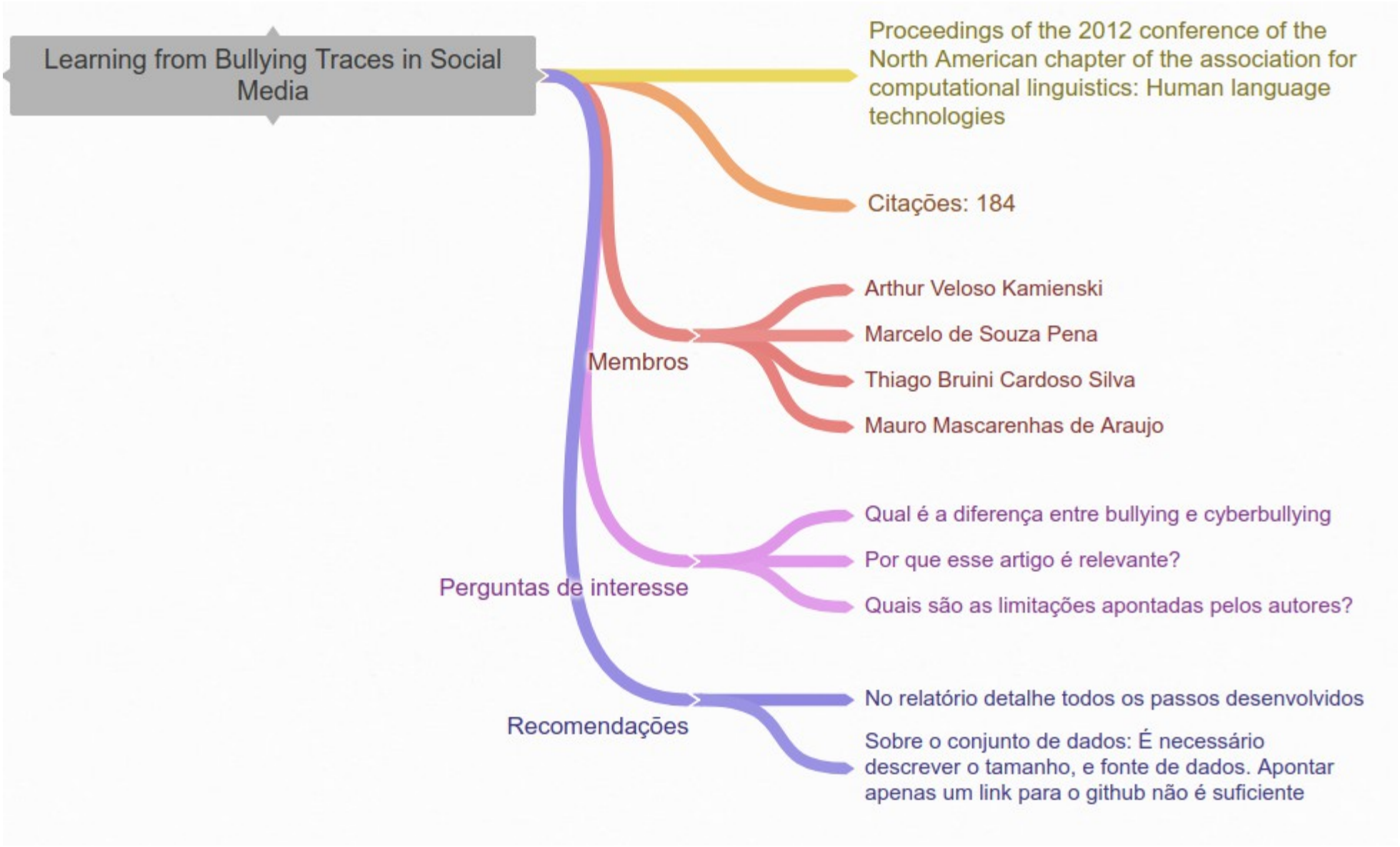

Sentimental Analysis of Product Based Reviews Using Machine Learning Approaches
Journal of Network Communications and Emerging Technologies. 2015
Citações: 6
Membros
Diego Pereira de Lima
Thiago Henrique Gomes Panini
Rodrigo Hiroaki Ideyama
Perguntas de interesse
Quais os cuidados que devem ser considerados para analisar multiples idiomas?
Quais são as limitações discutidas no artigo?
Qual a complexidade computacional?
Recomendações
Além dos testes planejados, considere um caso pontual para evidenciar a qualidade da proposta

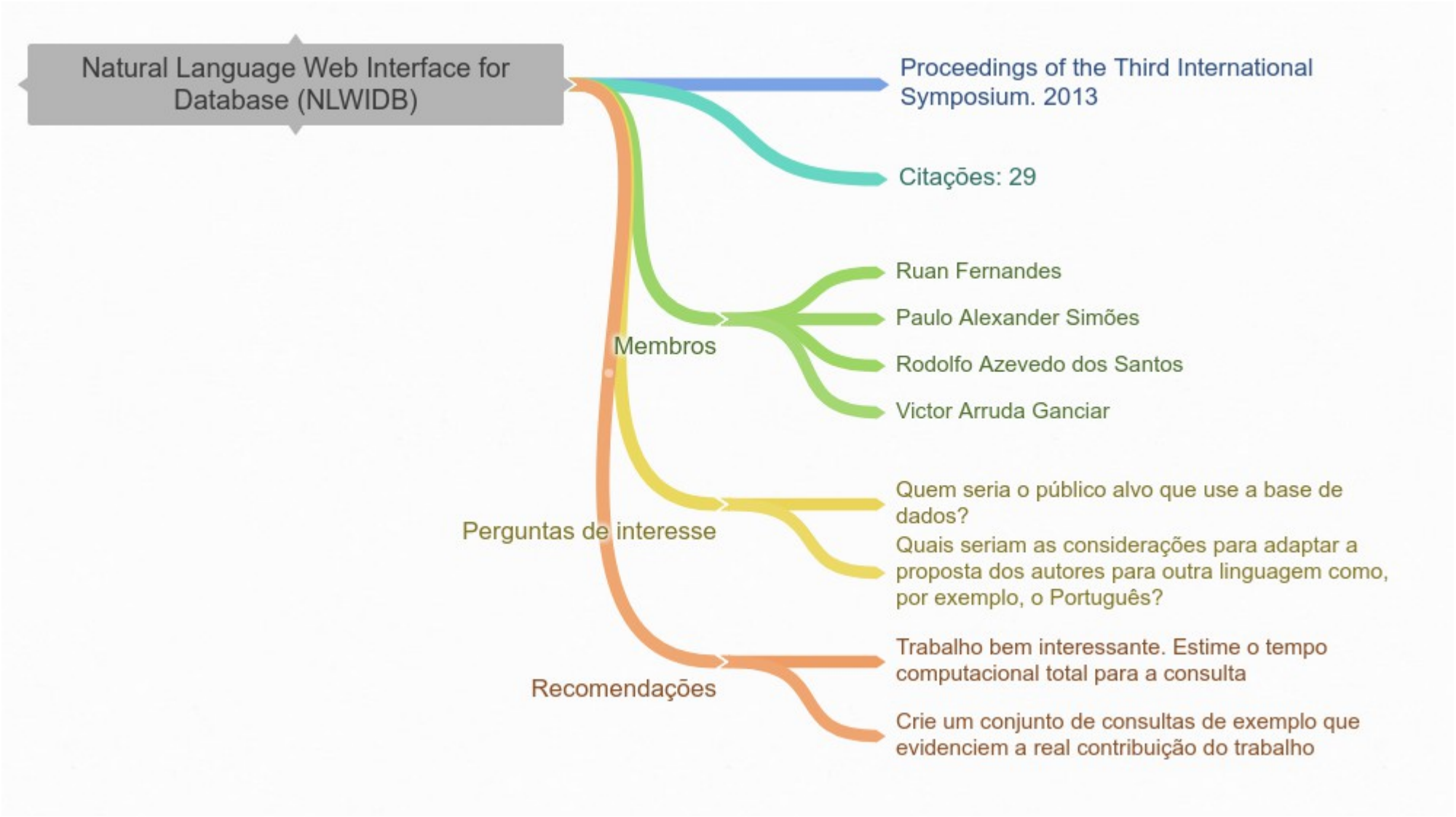

A Web of Hate: Tackling Hateful Speech in Online Social Spaces
Workshop em Text Analytics for Cybersecurity and Online Safety. 2017
Citações: 45
Membros
July Anne Pinheiro
Caique de Camargo
Jean Augusto
Marcela Yamashita
Perguntas de interesse
Quais as vantagens desta proposta frente às outras relatadas na literatura?
Qual é o conjunto de dados extraídos do Twitter?
Como definir formalmente um "discurso de ódio"?
Recomendações
Como validar os resultados? Tente usar um conjunto pequeno para avaliação

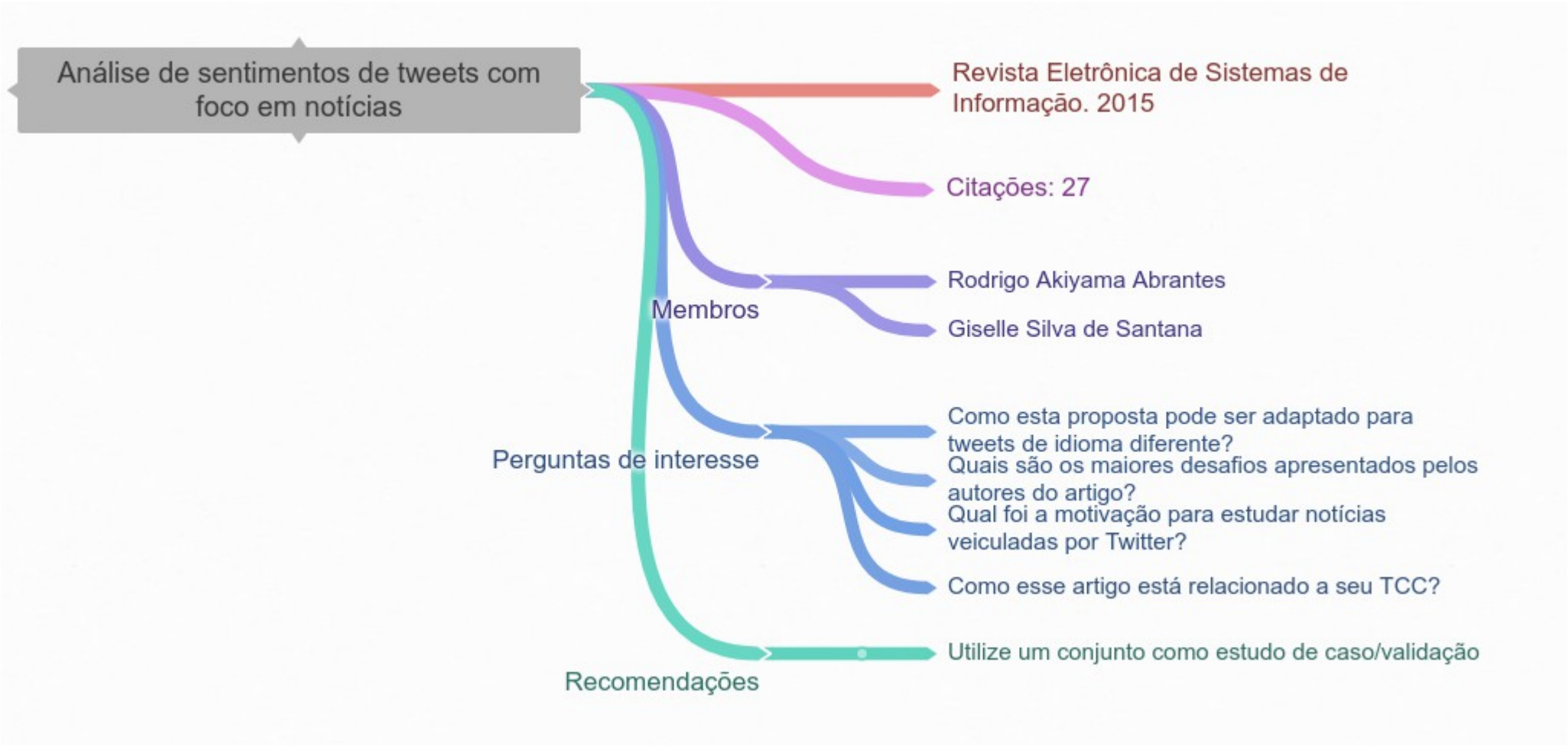

Natural Language Processing to the Rescue? Extracting “Situational Awareness” Tweets During Mass Emergency
Fifth International AAAI Conference on Weblogs and Social Media. 2011
Citações: 288
Membros
Eracton Ferreira Ramalho
Bruno Menezes Gottardo Ladeia
Matheus dos Santos Pereira
Perguntas de interesse
Qual o tamanho da base coletada do Twitter?
Uma vez tendo a base coletada, qual é a complexidade computacional da proposta?
Recomendações
Use base pequena para teste e validação

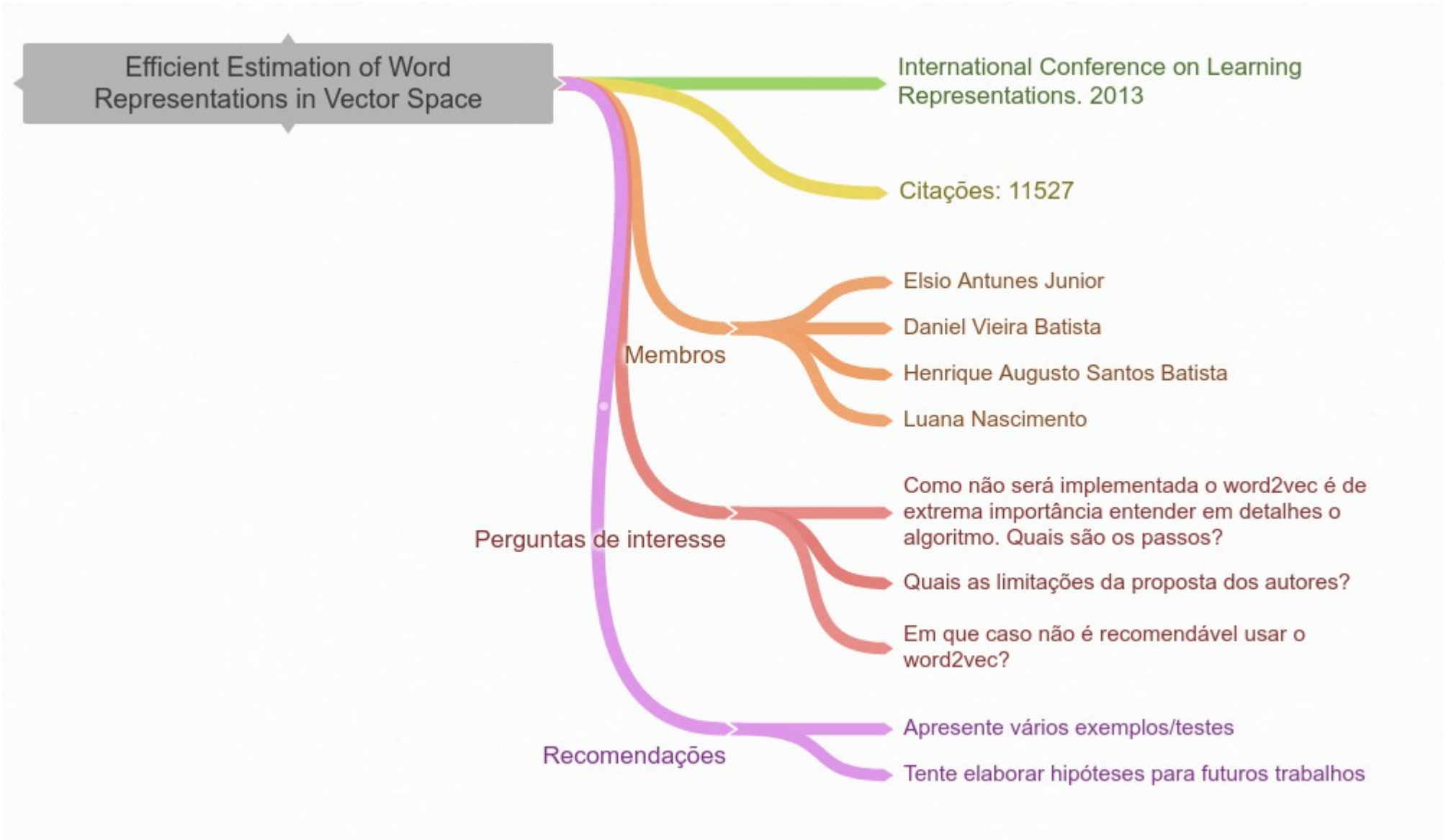

Word Etymology as Native Language Interference
Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.
Citações: 3?
Membros
Rafael Pauwels de Macedo
Iasmin de Haro Pracchias
Luiz Gabriel Correia
Thiago Felipe Floreste
Perguntas de interesse
Qual foi a fonte usada para identificar o número de citações? Liste os trabalhos que citaram este artigo
Trabalho bem interessante. Qual é o classificador considerado?
Recomendações
Considere um conjunto pequeno para teste/avaliação

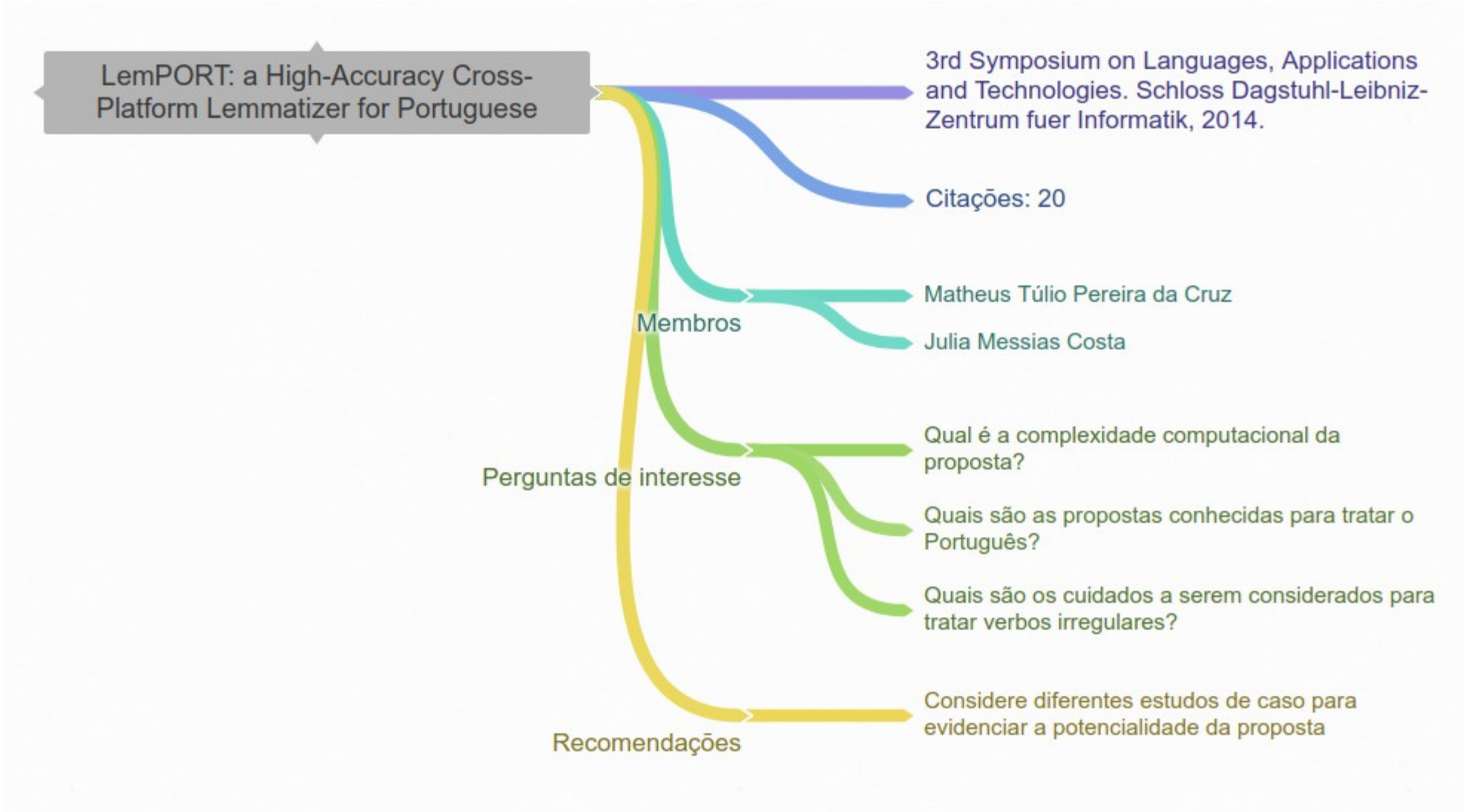

Twitter as a Corpus for Sentiment Analysis and Opinion Mining
International Conference on Language Resources and Evaluation. 2010
Citações: 2708
Membros
Matheus de Araújo Vargas
Rafael Augusto Zanatta
Perguntas de interesse
Como identificar sentimentos em amostras que contenham hashtags, imagens, vídeos?
Quais as limitações da proposta?
Qual é a importância deste trabalho no estado-da-arte?
Recomendações
Tentar validar a proposta.
Identificar trabalhos atuais que se baseiem nessa proposta.


Learning from Bullying Traces in Social Media
Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies
Citações: 184
Membros
Arthur Veloso Kamienski
Marcelo de Souza Pena
Thiago Bruini Cardoso Silva
Mauro Mascarenhas de Araujo
Perguntas de interesse
Qual é a diferença entre bullying e cyberbullying
Por que esse artigo é relevante?
Quais são as limitações apontadas pelos autores?
Recomendações
No relatório detalhe todos os passos desenvolvidos
Sobre o conjunto de dados: É necessário descrever o tamanho, e fonte de dados. Apontar apenas um link para o github não é suficiente


Natural Language Web Interface for Database (NLWIDB)
Proceedings of the Third International Symposium. 2013
Citações: 29
Membros
Ruan Fernandes
Paulo Alexander Simões
Rodolfo Azevedo dos Santos
Victor Arruda Ganciar
Perguntas de interesse
Quem seria o público alvo que use a base de dados?
Quais seriam as considerações para adaptar a proposta dos autores para outra linguagem como, por exemplo, o Português?
Recomendações
Trabalho bem interessante. Estime o tempo computacional total para a consulta
Crie um conjunto de consultas de exemplo que evidenciem a real contribuição do trabalho


Análise de sentimentos de tweets com foco em notícias
Revista Eletrônica de Sistemas de Informação. 2015
Citações: 27
Membros
Rodrigo Akiyama Abrantes
Giselle Silva de Santana
Perguntas de interesse
Como esta proposta pode ser adaptado para tweets de idioma diferente?
Quais são os maiores desafios apresentados pelos autores do artigo?
Qual foi a motivação para estudar notícias veiculadas por Twitter?
Como esse artigo está relacionado a seu TCC?
Recomendações
Utilize um conjunto como estudo de caso/validação


Efficient Estimation of Word Representations in Vector Space
International Conference on Learning Representations. 2013
Citações: 11527
Membros
Elsio Antunes Junior
Daniel Vieira Batista
Henrique Augusto Santos Batista
Luana Nascimento
Perguntas de interesse
Como não será implementada o word2vec é de extrema importância entender em detalhes o algoritmo. Quais são os passos?
Quais as limitações da proposta dos autores?
Em que caso não é recomendável usar o word2vec?
Recomendações
Apresente vários exemplos/testes
Tente elaborar hipóteses para futuros trabalhos


LemPORT: a High-Accuracy Cross-Platform Lemmatizer for Portuguese
3rd Symposium on Languages, Applications and Technologies. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2014.
Citações: 20
Membros
Matheus Túlio Pereira da Cruz
Julia Messias Costa
Perguntas de interesse
Qual é a complexidade computacional da proposta?
Quais são as propostas conhecidas para tratar o Português?
Quais são os cuidados a serem considerados para tratar verbos irregulares?
Recomendações
Considere diferentes estudos de caso para evidenciar a potencialidade da proposta

Perguntas para pensar/responder

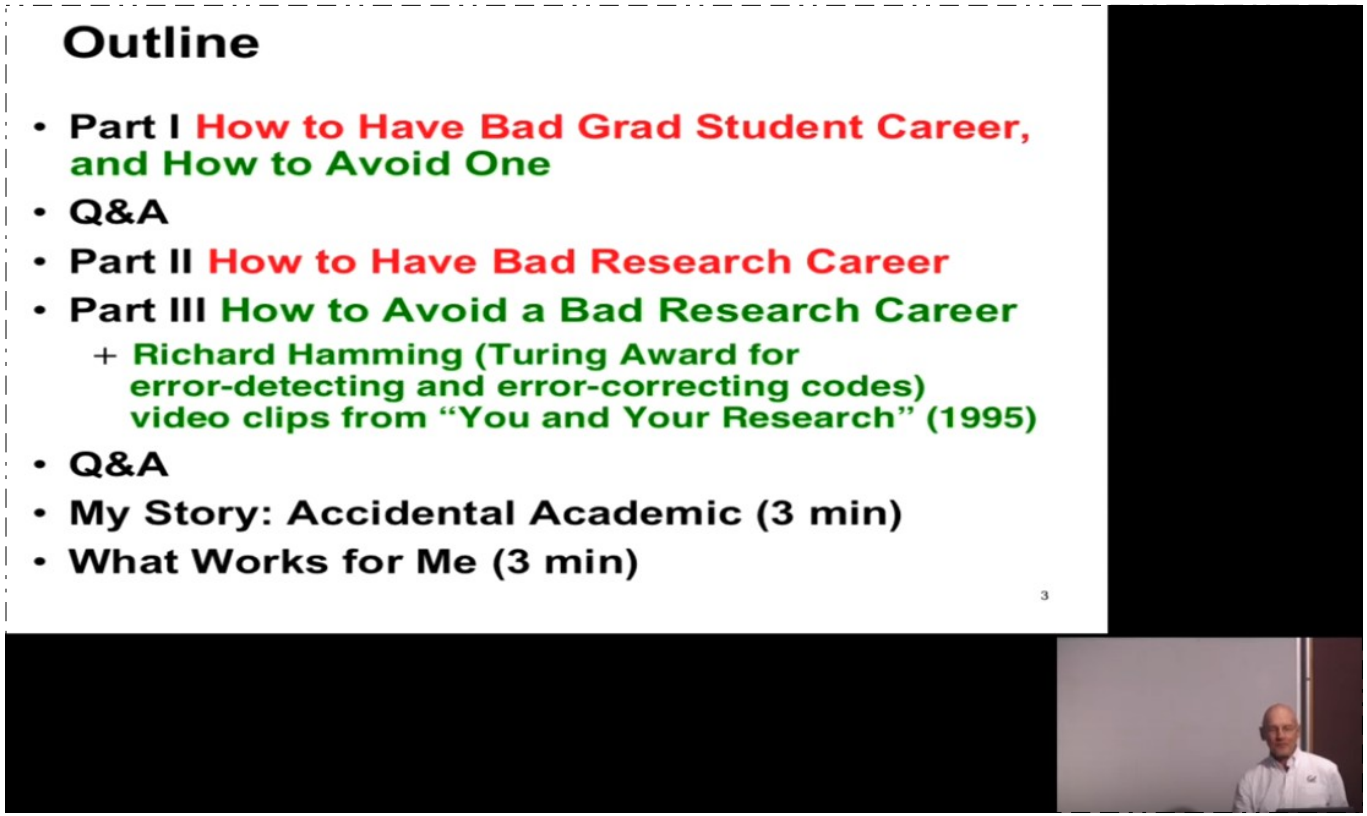
Para a apresentação se prepare para as seguintes perguntas:

- Qual é a complexidade computacional da sua implementação?
- Qual foi o maior desafio sobre sua implementação do artigo?
- Quem é o autor(a) principal? Qual a instituição de pesquisa? O que ele está realizando atualmente?
- Como foi a distribuição das tarefas do grupo?

Para finalizar

Comentários finais...

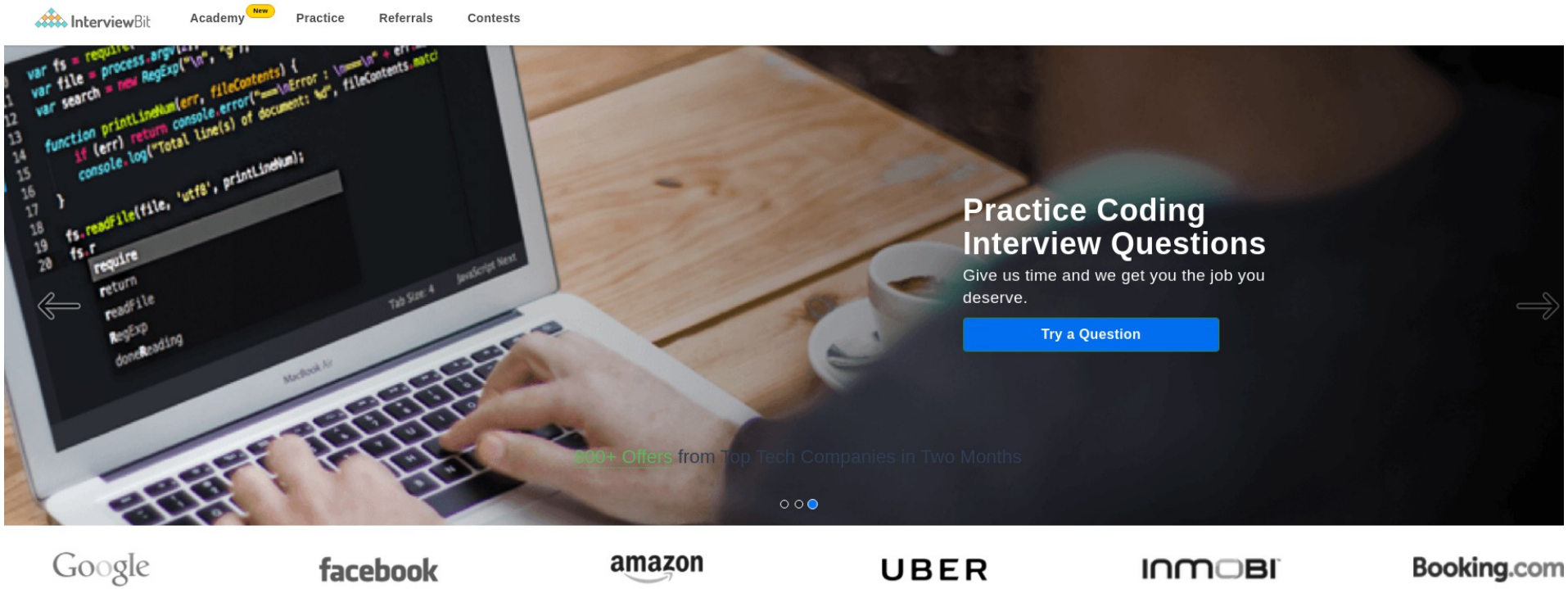
- No recesso: Assista o seguinte vídeo:
  How to have a bad career in research/academia
  (Prof. David Patterson)
  https://www.youtube.com/watch?v=Pbdo-ozuOug



Comentários finais...

- Existe uma plataforma, utilizada para contratações de estágio em empresas internacionais (Google, Uber, Facebook).

  https://www.interviewbit.com/problems/loop_cmpl/



LOOP_CMPL

What is the time, space complexity of following code :

```
int a = 0, b = 0;
for (i = 0; i < N; i++) {
    a = a + rand();
}
for (j = 0; j < M; j++) {
    b = b + rand();
}
```

Assume that rand() is O(1) time, O(1) space function.

- **O(N * M) time, O(1) space**
- **O(N + M) time, O(N + M) space**
- **O(N + M) time, O(1) space**
- **O(N * M) time, O(N + M) space**
- **O(N * M) time, O(N * M) space**

Comentários finais...

- Como ler um artigo científico?

Três passagens!

How to Read a Paper

S. Keshav
David R. Cheriton School of Computer Science, University of Waterloo
Waterloo, ON, Canada
keshav@uwaterloo.ca

ABSTRACT

Researchers spend a great deal of time reading research papers. However, this skill is rarely taught, leading to much wasted effort. This article outlines a practical and efficient *three-pass method* for reading research papers. I also describe how to use this method to do a literature survey.

Categories and Subject Descriptors: A.1 [Introductory and Survey]

General Terms: Documentation.

Keywords: Paper, Reading, Hints.

1. INTRODUCTION

Researchers must read papers for several reasons: to review them for a conference or a class, to keep current in their field, or for a literature survey of a new field. A typical researcher will likely spend hundreds of hours every year reading papers.

Learning to efficiently read a paper is a critical but rarely taught skill. Beginning graduate students, therefore, must learn on their own using trial and error. Students waste much effort in the process and are frequently driven to frustration.

For many years I have used a simple approach to efficiently read papers. This paper describes the 'three-pass' approach and its use in doing a literature survey.

2. THE THREE-PASS APPROACH

The key idea is that you should read the paper in up to three passes, instead of starting at the beginning and plowing your way to the end. Each pass accomplishes specific goals and builds upon the previous pass: The *first* pass gives you a general idea about the paper. The *second* pass lets you grasp the paper's content, but not its details. The *third* pass helps you understand the paper in depth.

2.1 The first pass

The first pass is a quick scan to get a bird's-eye view of the paper. You can also decide whether you need to do any more passes. This pass should take about five to ten minutes and consists of the following steps:

1. Carefully read the title, abstract, and introduction
2. Read the section and sub-section headings, but ignore everything else
3. Read the conclusions
4. Glance over the references, mentally ticking off the ones you've already read

At the end of the first pass, you should be able to answer the *five Cs*:

1. *Category*: What type of paper is this? A measurement paper? An analysis of an existing system? A description of a research prototype?
2. *Context*: Which other papers is it related to? Which theoretical bases were used to analyze the problem?
3. *Correctness*: Do the assumptions appear to be valid?
4. *Contributions*: What are the paper's main contributions?
5. *Clarity*: Is the paper well written?

Using this information, you may choose not to read further. This could be because the paper doesn't interest you, or you don't know enough about the area to understand the paper, or that the authors make invalid assumptions. The first pass is adequate for papers that aren't in your research area, but may someday prove relevant.

Incidentally, when you write a paper, you can expect most reviewers (and readers) to make only one pass over it. Take care to choose coherent section and sub-section titles and to write concise and comprehensive abstracts. If a reviewer cannot understand the gist after one pass, the paper will likely be rejected; if a reader cannot understand the highlights of the paper after five minutes, the paper will likely never be read.

2.2 The second pass

In the second pass, read the paper with greater care, but ignore details such as proofs. It helps to jot down the key points, or to make comments in the margins, as you read.

1. Look carefully at the figures, diagrams and other illustrations in the paper. Pay special attention to graphs. Are the axes properly labeled? Are results shown with error bars, so that conclusions are statistically significant? Common mistakes like these will separate rushed, shoddy work from the truly excellent.
2. Remember to mark relevant unread references for further reading (this is a good way to learn more about the background of the paper).