Universidade de São Paulo
Instituto de Matemática e Estatística
Bachalerado em Ciência da Computação

Arthur Sakayan Vieira de Melo

Construção de uma base de dados contendo fatos da Amazônia Azul

São Paulo
2021

Construção de uma base de dados contendo fatos da Amazônia Azul

Monografia final da disciplina
MAC0499 – Trabalho de Formatura Supervisionado.

Supervisor: Prof. Dr. Denis Deratani Mauá

São Paulo
2021

Agradecimentos

Gostaria de agradecer:
Minha familía, por sempre ter me apoiado;
Meus Professores, por me ensinar tudo isto e assim, permitir que eu tenha uma excelente carreira;
Os monitores, por terem resolvido minhas dúvidas.

Resumo

Dentro aqui deste trabalho primeiramente apresentamos o conceito da Amazônia Azul, uma região do Oceano Atlântico economicamente muito valiosa, e mostramos a necessidade de conscientizar o público geral sobre o assunto. Também explicamos a necessidade de extrair fatos para ajudar na conscientização do público leigo, e justificamos o uso de ferramentas especializadas para localizar e extrair informações de textos, juntamente dos objetivos deste trabalho.

Prosseguindo, apresentamos os fundamentos teóricos necessários para a compreensão do trabalho, e também uma descrição detalhada do que foi feito nele.

Finalmente, concluímos com impressões finais sobre o trabalho em si.

Palavras-chave: extração de informações, processamento de linguagem natural, aprendizado de máquina.

Abstract

Within this capstone project, we firstly present the concept of the Blue Amazon, an economically viable region within the Atlantic Ocean, and we show why it is necessary to raise public awareness regarding the subject. We also explain why we must use information extraction in order to perform the aforementioned task and also we justify the use of specialized software tools to locate and extract information from texts and perform the beforementioned goals.
Moving on, we present the theoretical concepts behind the mentioned tasks (to ensure a full comprehension of this project) and a detailed description of what has been done within this project.
Finally, we conclude the project with some final remarks regarding the project as a whole.

Keywords: information extraction, natural language processing, machine learning.

Sumário

Lista de Figuras **xi**

Lista de Tabelas **xiii**

1 Introdução **1**

1.1 Motivação 1

1.2 Justificativa 1

1.3 Objetivos 2

2 Fundamentos Teóricos **3**

2.1 Extração de informações 3

2.2 Classificação 4

2.3 Regressão Logística 4

2.4 Relações 5

2.5 PoS Tagging 5

2.6 NP Chunking 5

2.7 Restrições do ReVerb 5

2.8 Funcionamento do Reverb 6

3 Descrição do Trabalho **9**

3.1 Construção do Corpus 9

3.2 Processamento e Construção dos Fatos 10

3.3 Resultados 10

4 Impressões Finais **17**

Referências Bibliográficas **19**

Lista de Figuras

2.1 Um exemplo típico de execução do ReVerb 6

3.1 Aqui nesta imagem, palavras maiores aparecem mais frequentemente que palavras menores 11
3.2 Aqui podemos ter uma idéia melhor da distribuição das palavras no banco de dados 12
3.3 Aqui podemos ter uma idéia melhor sobre que tipos de assuntos são abordados no banco de dados 13
3.4 Frequência das palavras do segundo argumento de cada fato 14
3.5 Frequência de pares de palavras 15
3.6 Distribuição das palavras no Texto Coletado 15
3.7 Frequência do conjunto reduzido de palavras, ordenado decrescentemente . . 16

Lista de Tabelas

3.1 A distribuição das palavras dentro das relações 14

Capítulo 1

Introdução

1.1 Motivação

Existe uma região muito além da Costa do Atlântico, desconhecida de muitos Brasileiros (vide (Bandini, 2017)), no qual o Brasil e apenas o Brasil pode explorar os recursos[1].

Tal região possui inúmeros recursos minerais e biológicos, juntamente de um grande potencial econômico, energético, turístico e navegacional. Também vale a pena mencionar que é lá que se encontram as reservas do pré-sal, juntamente de outros recursos minerais de níquel, cobre, cobalto e manganês (Bandini, 2017).

Devido a esta grande e incalculável diversidade e riqueza de recursos, juntamente de seu tamanho significativo, tal região foi nomeada Amazônia Azul, em comparação com a floresta Amazônica, localizada ao norte do Brasil.

Porém, devido à inerente dificuldade de vigilância do local, existem várias ameaças a extração de recursos locais, especialmente os de natureza econômica (vide https://www.marinha.mil.br/secirm/amazoniaazul (Acessado em 17/12/2021)). Assim, mostra-se necessário a conscientização do público sobre tal tema.

Uma maneira de conscientizar o público é por exemplo, sumarizar as informações existentes que no momento se encontram dispersas na Internet através de vários sites. Ao agregar as informações dispersas, é possível assim reduzir o esforço necessário para que o público geral consiga ter conhecimento do tema. Ao resumir tal informação agregada existente e espalhada na internet, podemos apresentar ela de uma maneira de mais fácil compreensão comparada a meramente exibir todas as informações existentes conforme elas estão apresentadas.

Também, outro possível uso de poder sumarizar as informações é a possibilidade de que elas possam ser usadas para um sistema de perguntas e respostas. Tal sistema teria a vantagem de poder conter uma interface mais amigável para o público geral, além de poder explicar melhor o conceito para aqueles que não o conhecem.

1.2 Justificativa

Devido ao vasto número de informações sobre a Amazônia Azul[2], torna-se imprático a construção de um banco de dados para conter informações sobre a Amazônia Azul a partir da coleta manual de dados. Idealmente, para uma melhor compreensão das informações,

[1] https://www.un.org/Depts/los/clcs_new/submissions_files/bra04/bra_exec_sum.pdf (Acessado em 17/12/2021).

[2] conforme visto em https://www.google.com/search?hl=en&q=amaz%C3%B4nia%20azul (Acessado em 17/12/2021), existem aproximadamente 27.100.000 sites contendo as palavras "Amazônia Azul"dentro da Internet.

é desejável que elas estejam em forma de fatos relacionados ao assunto, assim facilitando a compreensão daqueles que desconhecem o assunto. Assim, necessita-se da utilização de formas automatizadas de extrair informações de textos no formato de fatos anteriormente citado.

Felizmente, dentro da vasta área de Ciências da Computação, tal tarefa já é conhecida e foi conceituada formalmente (Cowie e Lehnert, 1996), e assim, é de se esperar que existam vários algoritmos (e também programas) que também sejam capazes de resolver o problema de extrair informações de textos.

Atualmente, existem, por exemplo, técnicas que identificam entidades (Nguyen e Verspoor, 2019), que extraem informações de quaisquer domínios (Banko *et al.*, 2007; Fader *et al.*, 2011), que extraem certos padrões sintáticos de textos (Levy *et al.*, 2017), que conseguem extrair certos padrões sintáticos em domínios específicos (Wu e Weld, 2010) e que constroem bases de conhecimento (ou seja, uma tecnologia que contém várias informações agregadas, elas podendo estar estruturadas ou não) (Winn *et al.*, 2019; Zhang, 2015).

1.3 Objetivos

O programa ReVerb (Fader *et al.*, 2011) realiza extração de informações, isto é, a tarefa de buscar por determinadas informações contidas dentro de um documento contendo somente conteúdo textual. Adicionalmente, ele também não está restrito a um único e específico domínio. Ele foi projetado especialmente para poder funcionar em qualquer tipo de texto. A razão por trás de sua criação foi construir um sistema que tivesse um melhor desempenho comparado aos que já existiam. Ao ser executado, ele busca por determinados padrões no conjunto de texto(s) da entrada e exibe-os ao usuário. Para a realização da coleta de informações no formato de fatos, tal programa foi utilizado devido a sua facilidade de uso. Um exemplo de execução do programa segue abaixo:

Entrada: "Bananas are an excellent source of potassium".
Saída: "Bananas, are an excellent source of, potassium",
"Bananas, be source of, potassium"

Vale a pena notar que o ReVerb aceita textos apenas em inglês. Mesmo que a segunda saída seja gramaticalmente incorreta, ela tem uma função: servir para a consulta entre relações que foram encontradas no texto. Assim, mostra-se necessário guardar apenas o mínimo necessário para evitar uma possível sobrecarga de espaço supérfluo.
Neste trabalho, foram coletados vários textos da internet e eles serviram como entrada para o ReVerb. Com a saída do programa, foi construído um banco de dados. Tal banco de dados poderá ajudar a informar melhor o público geral sobre a Amazônia Azul, assim muito provavelmente cativando-o e motivando-o a dar maior importância a tal assunto.
Também é relevante mencionar que dentro deste Trabalho, estamos byscando extrair especificamente fatos (trechos como por exemplo, "Fulano é um sujeito") do texto.
Adicionalmente, este texto contém os fundamentos teóricos necessários para um bom entendimento do ReVerb, juntamente da descrição em detalhes de como foi construído um banco de dados e uma discussão de seus resultados, finalizando com impressões finais sobre o Trabalho em si.

Capítulo 2

Fundamentos Teóricos

2.1 Extração de informações

A Extração de Informações é essencialmente a tarefa de buscar informações específicas em vastos volumes de documentos contendo texto (Álvarez, 2007; Cowie e Lehnert, 1996). A área de extração de informação tem como objetivo principal buscar e extrair informações que sejam relevantes a um objetivo específico em um ou mais documentos usando software e técnicas de Processamento de Linguagem Natural. Também, busca-se estruturar os dados para análises futuras (Grishman, 1997). Usualmente, tal tarefa costuma ser feita da seguinte forma:
Primeiramente, extraímos pequenas informações do texto de entrada através de uma análise local do texto usando Processamento de Linguagem Natural. Após isto, juntamos as informações pequenas em informações maiores (ou até mesmo novas informações) seguindo regras de inferência. Finalmente, as informações relevantes ao domínio específico do problema são então estruturadas para o formato desejado da saída.

Em relação ao contexto geral da Extração de Informações, podemos classificar os documentos de texto existentes em três categorias: Estruturado, Semi-Estruturado, e finalmente, Não-Estruturado (Álvarez, 2007).

Em documentos Estruturados, as informações disponíveis estão apresentadas ordenadamente. Devido a essa regularidade do documento, a extração torna-se significantemente menos trabalhosa. Exemplos de documentos deste tipo são formulários preenchidos e arquivos CSV [1], que contém valores separados por vírgula (ou outro delimitador a escolha do usuário).

Agora, em documentos Semi-Estruturados, apenas alguns dos dados seguem algum padrão, com o restante do documento estando fora dele. Um bom exemplo de documentos deste tipo são os resumos de monografias. As palavras-chave seguem um padrão, no caso a separação por ponto e vírgula, ao passo que a descrição propriamente dita não possui nenhuma garantia de seguir algum tipo de formatação além daquelas inerentes à linguagem no qual ela foi escrita.

Por fim, em documentos Não-Estruturados, nenhuma informação disponível segue algum tipo de padrão ou ordenação. Devido a esta falta de regularidade, torna-se necessário o uso de técnicas especiais e especializadas (Cowie e Lehnert, 1996; Soderland, 1999). Conforme foram feitas conferências sobre o tema (Grishman e Sundheim, 1996), percebeu-se a necessidade de criar boas métricas para avaliar os resultados obtidos. Dentre os vários existentes, aqui serão citados dois importantes: Precisão e Cobertura (Sundheim, 1992). Definimos a precisão como a razão da quantidade de informações extraídas corretamente sobre o número

[1] https://en.wikipedia.org/wiki/Comma-separated_values (Acessado em 17/12/2021).

de todas extrações, e a cobertura como sendo a razão das informações extraídas corretamente sobre o número de extrações classificadas como positivas.[2]

2.2 Classificação

O problema de classificação consiste em, essencialmente, mapear elementos de um espaço de entradas X, geralmente R^n, e um elemento em um espaço de saídas Y, em geral um conjunto de classes definidas. Como entrada do problema, temos um conjunto de pares (x, y), com x sendo um elemento de X e y um elemento de Y. Assumimos que os pares originam de acordo com alguma distribuição conjunta $\mathrm{P}(X, Y)$ fixa, que é desconhecida por nós. Por fim, temos uma função de perda $L : Y \text{ x } Y \rightarrow R^n \geq 0$, que avalia a qualidade de uma estimativa y' quando a classe é y. O problema de aprendizado de um classificador consiste em obter uma função $f : X \rightarrow Y$ de uma classe de hipóteses H que maximiza a perda esperada:

$$m = arg\ min\ E[\ L(f(X), Y)\ ] \tag{2.1}$$

onde $E[\ .\ ]$ indica o valor esperado, no caso da função L sobre as variáveis aleatórias X e Y. Isso em geral é feito otimizando o risco empírico, definido como a média da função de perda no conjunto de dados observados. A qualidade de generalização da função, isto é, quão bem a função consegue aproximar o minimizador da equação acima, depende da classe de hipóteses e do algoritmo de otimização. A classe de hipóteses busca um equilíbrio entre expressividade e tratabilidade. Classes com funções muito simples são mais fáceis de otimizar e conseguem ser generalizadas a partir de poucos dados, porém representam apenas fenômenos simples. Já classes com funções muito complexas conseguem representar fenômenos complexos, porém possuem alto custo computacional e demandam quantidades grandes de dados. O algoritmo de otimização ou de aprendizado pode compensar parte desse dilema, produzindo funções mais complexas de acordo com a quantidade de dados, a complexidade do fenômeno sendo representado (como medido através da perda empírica nos dados) e o tempo de computação (com mais tempo resultando em classificadores melhores, em média). Duas classes notáveis de hipóteses são as funções lineares de limiar rígido (mais conhecidos como classificadores logísticos) e as redes neurais, que compõem sequencialmente classificadores logísticos de maneira sofisticada, permitindo o compartilhamento de pesos entre componentes distintas e técnicas de otimização. O principal algoritmo de aprendizado de tais modelos é o de descida do gradiente, que realiza uma busca gulosa no espaço de parâmetros do modelo. Embora o algoritmo garante apenas encontrar um ótimo local, técnicas modernas permitem encontrar soluções satisfatórias na prática para um grande número de aplicações.

2.3 Regressão Logística

A regressão Logística é uma sub-categoria do problema de classificação. Nela, buscamos modelar a probabilidade de um evento pertencer a uma de duas possíveis categorias[3]. No caso específico do ReVerb, queremos modelar a probabilidade de uma extração ser confiável. Para calcularmos tal probabilidade, nós primeiro computamos o produto escalar de um vetor de hipóteses w com o vetor de conjunto de dados de entrada x. Tal produto será denominado p. A probabilidade que buscamos é igual a $\theta(p)$, aonde θ(x) é uma função com a fórmula

[2] https://en.wikipedia.org/wiki/Precision_and_recall (Acessado em 17/12/2021).
[3] https://en.wikipedia.org/wiki/Logistic_regression (Acessado em 21/12/2021).

abaixo:

$$\theta(x) = \frac{e^x}{1 + e^x} \tag{2.2}$$

Tal função é denominada de sigmóide devido ao formato semelhante à letra S. Como o contra-domínio de θ é o intervalo [0, 1], podemos seguramente usá-lo para interpretar como probabilidades incertas (Abu-Mostafa *et al.*, 2012).

2.4 Relações

No contexto de Extração de Informações, relações são conjuntos de palavras. Normalmente são compostos de dois substantivos (podendo ter ou não advérbios adjacentes) com um verbo entre eles. No contexto deste trabalho, os substantivos são chamados de argumentos e o verbo é chamado de relação.

2.5 PoS Tagging

PoS Tagging é essencialmente, o problema de classificar cada palavra de uma frase em uma componente morfossintática (verbo, substantivo, advérbio, adjetivo, etc.) correspondente. Devido a ambiguidade existente nas palavras de uma frase, dado que existem palavras que ocupam duas categorias morfossintáticas (chamadas de homônimos), PoS Tagging pode também ser considerado um problema de desambiguação. Neste processo, temos como entrada uma sequência de palavras, juntamente de um conjunto de rótulos de palavras, e como saída uma sequência de rótulos, cada um correspondendo a uma palavra (Jurafsky e Martin, 2020). Um exemplo:
Entrada: Estes monges sãos são devotos de São Benedito.
Saída: Determinante, Substantivo, Adjetivo, Verbo, Substantivo, Preposição, Substantivo, Substantivo.
Aqui vemos que a palavra "são"possui três sentidos: Sadio, a terceira pessoa do plural do verbo ser e a denominação de Santo.

2.6 NP Chunking

Chunking é o processo de identificar, classificar e agrupar uma frase em seus sintagmas correspondentes (Jurafsky e Martin, 2020). Um sintagma é considerado uma das unidades linguísticas que unidas, formam uma frase[4]. Como entrada, temos uma frase (em português, inglês, etc.) e como saída temos a mesma frase, só que com as classificações de sintagma. Um exemplo:
Entrada: O voo matutino do Rio de Janeiro chegou.
Saída: [$_{Sintagma\ Nominal}$ O voo matutino] [$_{Sintagma\ Preposicional}$ do] [$_{Sintagma\ Nominal}$ Rio de Janeiro] [$_{Sintagma\ Verbal}$ chegou]

2.7 Restrições do ReVerb

Para garantir que o ReVerb tenha um bom funcionamento, foram introduzidas duas restrições: uma sintática e uma léxica. A restrição sintática tem como objetivo filtrar re-

[4] https://www.todamateria.com.br/sintagma/ (Acessado em 21/12/2021).

lações resultantes incoerentes como por exemplo, a frase "The Mark 14 was central to the torpedo scandal of the fleet"retornar como resultado "was central torpedo". A restrição léxica tem como objetivo eliminar relações extremamente específicas, como por exemplo "The Obama administration is offering only modest greenhouse gas reduction targets at the conference"retornar como resultado "is offering only modest greenhouse gas reduction targets at". Definimos a restrição sintática nos seguintes termos: toda relação com mais de uma palavra tem que começar com um verbo, terminar com uma preposição, e ser uma sequência contínua de palavras dentro da frase. Definimos a restrição léxica nos seguintes termos: uma frase do tipo (substantivo1, verbo, substantivo2), chamada de frase de relação binária (pois lida com apenas dois substantivos) tem que pelo menos um (ou mais, se preciso for) de pares de substantivos distintos (que se encaixem no tipo de frase descrita anteriormente) dentro do texto de entrada (Fader *et al.*, 2011).

2.8 Funcionamento do Reverb

O propósito principal do ReVerb é justamente realizar a extração de informações de textos que não estejam restritos a um domínio específico. O programa tem como entrada um documento de texto (ou também ele pode ler textos da entrada padrão), e sua saída é um conjunto de informações conforme visto abaixo:

```
$ echo "Bananas are an excellent source of potassium." |
    ./reverb -q | tr '\t' '\n' | cat -n
 1  stdin
 2  1
 3  Bananas
 4  are an excellent source of
 5  potassium
 6  0
 7  1
 8  1
 9  6
10  6
11  7
12  0.9999999997341693
13  Bananas are an excellent source of potassium .
14  NNS VBP DT JJ NN IN NN .
15  B-NP B-VP B-NP I-NP I-NP I-NP I-NP O
16  bananas
17  be source of
18  potassium
```

Figura 2.1: *Um exemplo típico de execução do ReVerb*

O software ReVerb usa o software OpenNLP da fundação Apache para poder realizar as tarefas de PoS Tagging e NP Chunking de uma frase. O OpenNLP usa um modelo probabilístico não especificado para a tarefa de PoS Tagging[5] e um modelo de entropia máxima para a tarefa de NP Chunking[6]. Explicando mais sobre o modelo de entropia máxima: Ele é um método para o problema da classificação anteriormente descrito, no qual selecionamos o modelo com maior entropia nos que satisfaz as restrições do problema. O princípio é

[5] https://opennlp.apache.org/docs/1.9.4/manual/opennlp.html#tools.postagger (Acessado em 21/12/2021).

[6] https://opennlp.apache.org/docs/1.9.4/manual/opennlp.html#tools.chunker (Acessado em 21/12/2021).

que, mesmo que o objetivo principal de algoritmos classificadores seja de reduzir a entropia cruzada, ainda é desejável manter o máximo de incerteza dentro do modelo[7]. Tal modelo é usado para auxiliar na escolha de qual sintagma específico aplicar a uma parte da frase na tarefa de NP Chunking.

Também, abaixo segue um exemplo ilustrativo do funcionamento do ReVerb:
Entrada: "From an economic point of view , it is worth noting that approximately 95% of our foreign trade is carried out by sea"
Saída:"approximately 95% of our foreign trade, is carried out, by sea"

Algo importante a se notar é que o ReVerb foi projetado para funcionar apenas com textos que estejam em inglês. De acordo com artigo escrito pelos autores do software (Fader *et al.*, 2011), o algoritmo de extração de relações funciona da seguinte maneira: Dado uma frase f, e para cada verbo v dentro dessa frase, procuramos a maior sequência de palavras s que comece dentro de f e que satisfaça as restrições sintática e léxica anteriormente definidas. Se houver uma intersecção entre relações, elas serão mescladas. Para cada palavra identificada como sendo uma possível relação anteriormente, procuramos um substantivo mais próximo à esquerda e a direita da palavra para poder construir uma relação. Para determinar se a relação satisfaz a restrição léxica, usamos um dicionário contendo várias frases de relação retiradas da internet, sendo que tal dicionário é o conjunto de frases de relação que levam pelo menos 20 pares de argumentos no conjunto de extrações. Os argumentos (que geralmente são substantivos) das frases são identificados heuristicamente seguindo os passos descritos anteriormente. Para flexibilizar a extração, nós removemos inflexões, verbos auxiliares, adjetivos e advérbios. Como tal algoritmo tem uma precisão baixa, nós usamos uma função de confiança baseada em regressão logística para poder estabelecer um parâmetro mínimo para uma relação ser considerada correta. A função de confiança também é usada para medir quão confiável é uma extração. Como o ReVerb usa aprendizado supervisionado apenas para dar pontuações na função de confiança, ele consegue usar menos exemplos de treinamento que outros programas do tipo.

[7] https://opennlp.apache.org/docs/1.9.4/manual/opennlp.html#opennlp.ml.maxent (Acessado em 21/12/2021).

Capítulo 3

Descrição do Trabalho

3.1 Construção do Corpus

Primeiramente, para que se possa realizar a construção do banco de dados, é preciso obter um conjunto de documentos de textos, chamado de corpus. Para tal, foi realizada uma busca em sites como o Google e DuckDuckGo por páginas que continham o termo Amazônia Azul, assim obtendo um primeiro conjunto de dados consistindo de dez textos escritos em português e um em espanhol. Os textos foram obtidos dos seguintes sites:

- Comando Geral do Corpo de Fuzileiros da Marinha do Brasil[1].
- Estadão[2].
- Secretaria da Comissão Interministerial para os Recursos do Mar[3].
- Ecycle[4]
- Portal de Revistas USP[5]
- Inter Press Service[6]
- Diretoria do Patrimônio Histórico e Documentação da Marinha[7]
- Igui Ecologia[8]

[1] https://www.marinha.mil.br/cgcfn/amazonia_azul
[2] https://politica.estadao.com.br/blogs/fausto-macedo/amazonia-azul-em-risco/
[3] https://www.marinha.mil.br/secirm/amazoniaazul
[4] https://amp.ecycle.com.br/component/content/article/67-dia-a-dia/6740-amazonia-azul.html
[5] https://www.revistas.usp.br/revusp/article/view/139265/134606
[6] http://www.ipsnoticias.net/2015/04/amazonia-azul-la-nueva-frontera-de-recursos-naturales-en-brasil/
[7] http://www.redebim.dphdm.mar.mil.br/vinculos/000006/0000060a.pdf
[8] https://www.iguiecologia.com/amazonia-azul/

- Correio Brasiliense[9]
- EngeMarinha[10]
- Gazeta do Povo[11]

Todos estes sites foram acessados pela última vez em 17/12/2021.
O processo de criação do corpus foi relativamente fácil, pois na maioria dos textos apenas poucos ajustes foram feitos. Porém, houve uma exceção: um documento em PDF da revista USP, por ser o único que continha quebra de palavras. Para corrigir isto foi usado o recurso de buscar e substituir do editor de texto, ao passo que todos os outros links tiveram mínimos ajustes. Assim, foi construído um corpus de um conjunto de documentos de texto, usando o ReVerb para extrair conjuntos de fatos na forma de triplas de frases, com uma frase (chamada de relação) formando uma ligação entre as outras duas (chamadas de entidades).

3.2 Processamento e Construção dos Fatos

Após a obtenção dos dados, foi necessário buscar um programa que fosse capaz de realizar a tarefa proposta de extrair as informações do texto e com elas, construir um banco de dados. Vale a pena notar que primeiramente, o programa DeepDive, da Universidade de Stanford foi usado para a tarefa. Porém ele acabou sendo descartado, pois seu consumo de memória RAM acabou sendo muito maior que o esperado. Outro programa utilizado foi o JNERE de Dat Quoc Nguyen (Nguyen e Verspoor, 2019), que também foi descartado pois sua saída era de difícil compreensão. Uma coisa relevante a se notar é que o ReVerb foi construído e planejado para rodar apenas com textos em inglês, logo assim foi necessário traduzir o conjunto de textos previamente citado. Para traduzir os textos de tal modo que eles pudessem produzir resultados desejáveis no ReVerb, foi utilizado o site Google Tradutor[12]. Vale a pena notar que por praticidade, o programa ReVerb foi rodado em um único documento, que continha todos os textos traduzidos e concatenados. Devido às imperfeições da extração de informações do ReVerb, foi necessário percorrer e vasculhar a saída do programa em busca de resultados imperfeitos e incompletos e assim corrigi-los para que o resultado final estivesse com muito boa qualidade.

3.3 Resultados

No final de tudo,após a revisão do arquivo, os resultados obtidos foram convertidos em um arquivo .json[13], com 497 relações no formato de tripla (argumento1, relação, argumento2), 55.8 kilobytes de tamanho, 6.223 palavras no total, 1492 palavras distintas, e considerando que todas elas foram vasculhadas em busca de imperfeições, todas estão gramaticalmente corretas. Algumas podem não ter relevância ao tema, dado que os textos extraídos por vezes

[9] https://www.correiobraziliense.com.br/brasil/2020/12/4897118-meio-ambiente-amazonia-azul-e-um-verdadeiro-teso
html
[10] https://engemarinha.com.br/voce-sabe-o-que-e-amazonia-azul-conheca-o-conceito-e-sua-importancia/
[11] https://www.gazetadopovo.com.br/republica/amazonia-azul-brasil-potencia-militar-atlantico/
[12] disponível em https://translate.google.com (Acessado em 17/12/2021).
[13] disponível aqui em https://linux.ime.usp.br/~arthursak/mac0499/baseDados.json (Acessado em 17/12/2021).

abordam outros assuntos além da Amazônia Azul, mas tais relações permanecem dentro do banco de dados pois eles podem ter algum futuro uso. Um outro dado relevante é que o conjunto de 11 textos usado para construir o banco de dados tem 18.454 palavras no total e 4.003 palavras distintas.

Um exemplo de relação dentro do arquivo segue abaixo:
ent1: The Brazilian Exclusive Economic Zone,
rel: is an oceanic area of approximately,
ent2: 4.5 million km^2

Para melhor demonstrar o processo de busca e correção realizado, um exemplo é exibido abaixo:

Antes:
ent1: It
rel: is the capacity of
ent2: sea water

Depois:
ent1: The solubility pump
rel: is the capacity of
ent2: sea water to maintain a certain amount of dissolved carbon dioxide.

A maioria das correções foi semelhante à vista acima. Para uma melhor compreensão dos dados, várias figuras foram construídas. Primeiramente, exibimos uma distribuição das palavras do banco de dados. Devido à vasta quantidade de palavras, aquelas que possuem frequência baixa não estão representadas.

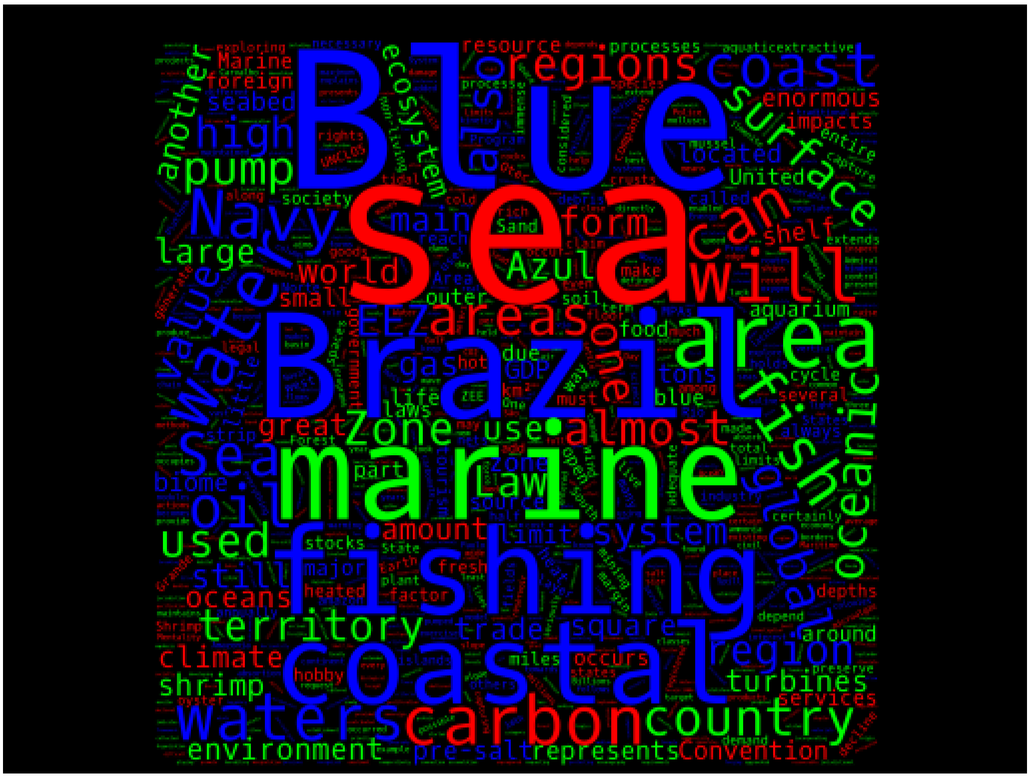

Figura 3.1: *Aqui nesta imagem, palavras maiores aparecem mais frequentemente que palavras menores*

Pode-se reparar que palavras centrais ao tema da Amazônia Azul predominam em maioria bastante significantiva dentro do banco de dados, com o restante do conteúdo estando disperso em quantidades menores. Segue-se uma outra representação das palavras dentro do banco de dados, desta vez ordenado em ordem descrescente de ocorrências.

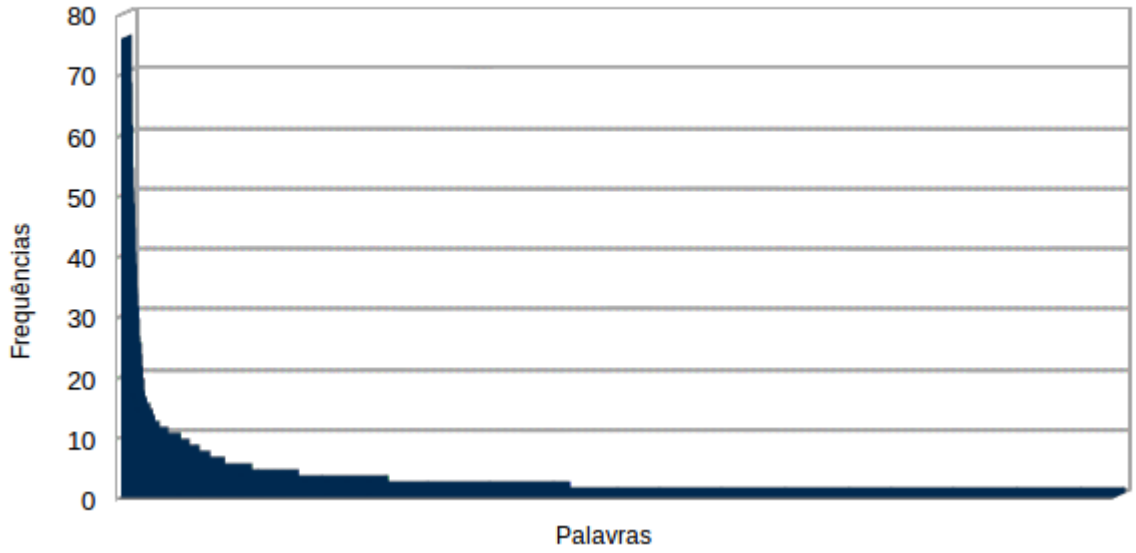

Figura 3.2: *Aqui podemos ter uma idéia melhor da distribuição das palavras no banco de dados*

Após, vemos uma representação semelhante à primeira para as palavras dos assuntos que o banco de dados aborda.

Figura 3.3: *Aqui podemos ter uma idéia melhor sobre que tipos de assuntos são abordados no banco de dados*

Abaixo disto, vemos uma tabela contendo as palavras mais frequentes dentro das relações do banco de dados. Palavras com menos de dez ocorrências não foram incluídas.

Palavra	Frequência
is	130
of	73
to	61
the	50
are	49
in	48
has	38
by	29
for	22
be	17
was	16
not	16
a	16
will	15
on	13
have	13
can	13
used	11
from	11
been	11
with	10
also	10

Tabela 3.1: *A distribuição das palavras dentro das relações*

Vale a pena notar que a maioria das palavras nas relações são verbos, ao passo que substantivos compõem a maioria das palavras em menor quantidade. Para se ter uma melhor idéia da distribuição das palavras dentro do banco de dados, mais imagens foram construídas.

Figura 3.4: *Frequência das palavras do segundo argumento de cada fato*

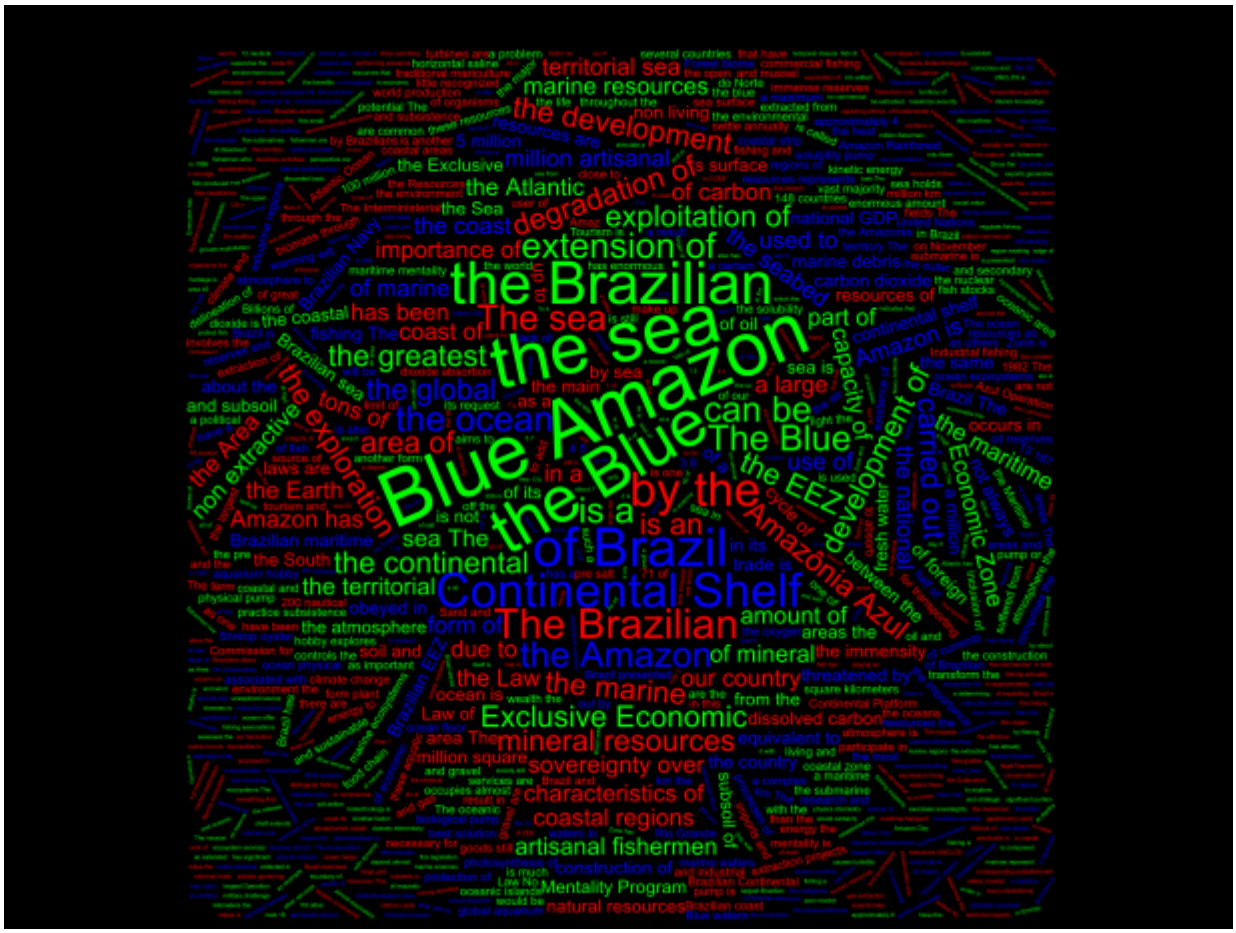

Figura 3.5: *Frequência de pares de palavras*

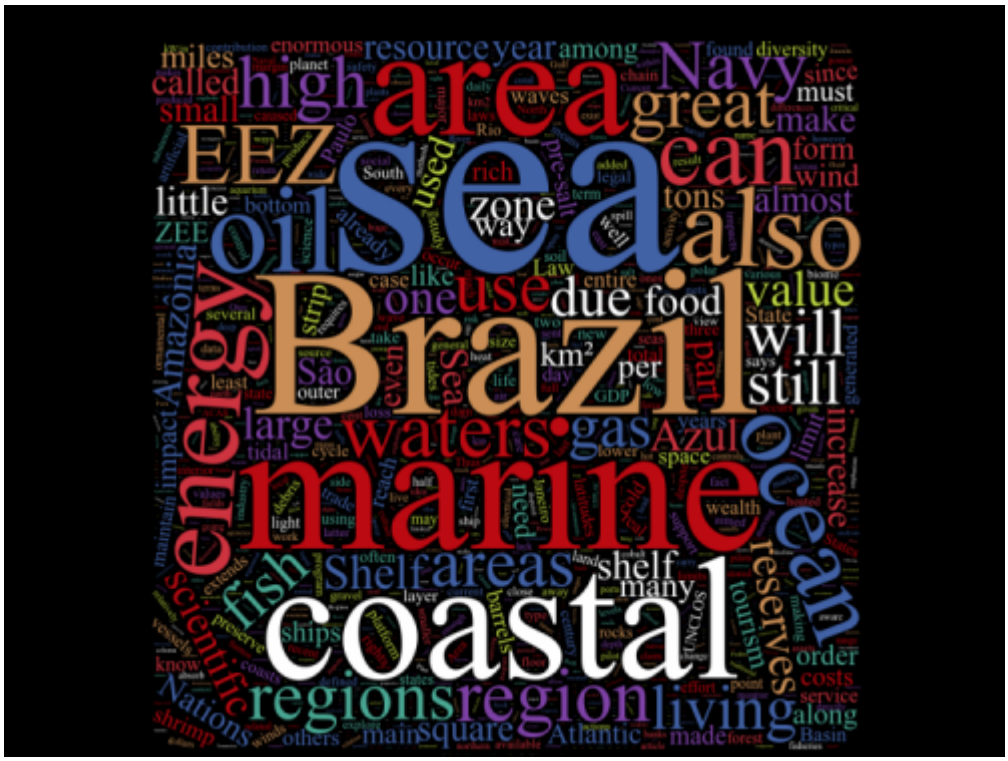

Figura 3.6: *Distribuição das palavras no Texto Coletado*

Também, devido ao fato de que cada palavra foi contada individualmente ao calcular a frequência das palavras, houve casos em que palavras aproximadamente semelhantes (como por exemplo, biotechnology e biotechnological, ou jurisdiction e jurisdictional, e também ocean e oceanical) foram contadas como sendo duas palavras distintas, quando na realidade elas possuem um grau de semelhança entre si. Logo assim, mostrou-se necessário a construção de um conjunto reduzido, no qual todas as palavras que possuem o mesmo tipo de estrutura central mas com alguma variação (conforme descrito antes) fossem agrupadas em uma palavra só. Abaixo, vemos a distribuiçao da frequência das palavras deste novo conjunto, ordenadas pelo número de ocorrências.

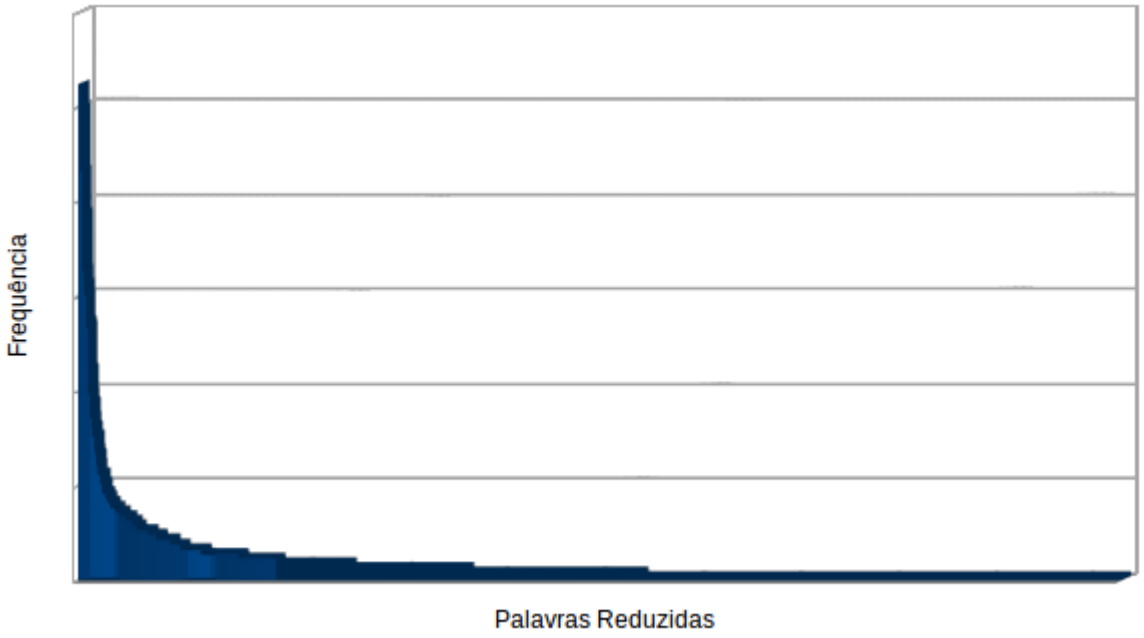

Figura 3.7: *Frequência do conjunto reduzido de palavras, ordenado decrescentemente*

Como muitas palavras aparecem com baixa frequência, é de se esperar que mesmo agrupando várias palavras distintas em apenas uma, assim aumentando a frequência, a forma do gráfico não irá mudar significantemente. A média da frequência das palavras do banco de dados é 2,8 e o desvio padrão é igual a 5.2. Agora, no conjunto reduzido de palavras, Conforme visto nos gráficos e imagens, vemos que muitas palavras estão dispersas em baixas quantidades, enquanto palavras relevantes ao tema aparecem com maior frequência. Assim, nota-se que o banco de dados possui uma grande e significativa variedade dentro de seu conteúdo (ainda que ele esteja restrito ao tema da Amazônia Azul, isso é esperado dentro deste trabalho).

Capítulo 4

Impressões Finais

Este trabalho também está inserido em um contexto maior: O Centro de Inteligência Artificial da USP[1] está planejando usar bancos de dados para o intuito de contruir um programa que seja capaz de responder perguntas dos usuários (tal programa é denominado de agente conversacional). Assim, com tal banco de dados providenciado aqui dentro deste trabalho, será possível que ele seja usado para educar e informar o público leigo sobre a Amazônia Azul.
Em melhores detalhes, o Centro de Inteligência Artificial tem como objetivo concreto desenvolver uma estrutura para agentes conversacionais que consegue lidar com vários tipos de interações em um domínio específico, como: perguntas, discussões, explicações, inferências e planejamentos. Atualmente, o grupo está construindo um especialista conversacional focado na Amazônia Azul para mostrar as capabilidades da estrutura[2], com um objetivo futuro de ampliar os assuntos possíveis.
Assim, a base de dados construída ao longo deste Trabalho irá servir como um apoio para tal agente conversacional. Exemplificando um caso de uso deste projeto (juntamente de como este Trabalho pode ajudar neste caso): Se alguém fizer uma pergunta ao agente conversacional, ele irá buscar a base de dados por fatos relevantes a ela. Evidentemente, dependendo do tipo da pergunta, fatos distintos serão buscados. Após a busca terminar, o agente conversacional vai então determinar qual fato que se poderia melhor satisfazer a pergunta. Depois, o agente conversacional construirá a resposta e a exibirá ao usuário.

Conforme descrito anteriormente, houve vários casos em que o ReVerb não conseguia realizar uma extração completa. Um Exemplo Segue Abaixo:
Frase: The Brazilian exclusive economic zone is a belt of sea which extends from twelve to two hundred nautical miles.
Extração parcial realizada pelo ReVerb: The Brazilian exclusive economic zone, is a belt of, sea.
Extração correta: The Brazilian exclusive economic zone, is a belt of, sea which extends from twelve to two hundred nautical miles.
Logo, existem alguns aspectos do ReVerb que são passíveis de melhora. Apesar de neste trabalho ter o ReVerb como o principal programa para realizar a tarefa proposta, existem várias outras alternativas, conforme mencionado na introdução. Apesar das falhas mencionadas, o uso do ReVerb foi satisfatório. Vale a pena mencionar que existem muito mais sites do que os 11 escolhidos para compôr o Banco de Dados, porém como foi possível construir um banco de dados de significativo tamanho com apenas as páginas escolhdas, os sites adicionais não

[1] https://c4ai.inova.usp.br (Acessado em 17/12/2021).
[2] Conforme visto no site, acessado em 30/01/2022.

foram levados em consideração. Mesmo com este trabalho estando concluído, ainda é possível que ele possa ser usado no futuro para a construção de agentes conversacionais. Caso algum futuro aluno queira continuar este trabalho, é recomendado que primeiro de tudo busque-se melhorar a performance do ReVerb. Neste trabalho, pode-se aprender um pouco mais sobre Extração de Informações e um pouco sobre Processamento de Linguagem Natural. Quando este trabalho foi escolhido dentre os vários propostos, estas duas áreas da computação acima citadas eram desconhecidas pelo autor. Conforme visto no capítulo 2, agora se possui um conhecimento no mínimo básico sobre elas.

Referências Bibliográficas

Abu-Mostafa *et al.***(2012)** Yaser S Abu-Mostafa, Malik Magdon-Ismail e Hsuan-Tien Lin. *Learning from data*, volume 4. AMLBook New York, NY, USA:. Citado na pág. 5

Álvarez(2007) Alberto Cáceres Álvarez. *Extração de informação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem*. Tese de Doutorado, Universidade de São Paulo. Citado na pág. 3

Bandini(2017) Castro Bandini. A amazônia azul: recursos e preservação. doi: 10.11606/issn.2316-9036.v0i113p7-26. URL https://www.revistas.usp.br/revusp/article/view/139265. Citado na pág. 1

Banko *et al.***(2007)** Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead e Oren Etzioni. Open information extraction from the web. Em *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, páginas 2670–2676, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. Citado na pág. 2

Cowie e Lehnert(1996) James R. Cowie e Wendy G. Lehnert. Information extraction. *Commun. ACM*, 39:80–91. Citado na pág. 2, 3

Fader *et al.***(2011)** Anthony Fader, Stephen Soderland e Oren Etzioni. Identifying relations for open information extraction. Em *Proceedings of the Conference of Empirical Methods in Natural Language Processing (EMNLP '11)*, Edinburgh, Scotland, UK. Citado na pág. 2, 6, 7

Grishman(1997) Ralph Grishman. Information extraction: Techniques and challenges. doi: 10.1007/3-540-63438-x_2. Citado na pág. 3

Grishman e Sundheim(1996) Ralph Grishman e Beth Sundheim. Message Understanding Conference- 6: A brief history. Em *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. URL https://aclanthology.org/C96-1079. Citado na pág. 3

Jurafsky e Martin(2020) Daniel Jurafsky e James Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 3. Citado na pág. 5

Levy *et al.***(2017)** Omer Levy, Minjoon Seo, Eunsol Choi e Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115*. Citado na pág. 2

Nguyen e Verspoor(2019) Dat Quoc Nguyen e Karin Verspoor. End-to-end neural relation extraction using deep biaffine attention. Em *Proceedings of the 41st European Conference on Information Retrieval*. Citado na pág. 2, 10

Soderland(1999) Stephen Soderland. Learning information extraction rules for semi-structured and free text. *Mach. Learn.*, 34(1-3):233–272. ISSN 0885-6125. doi: 10.1023/A:1007562322031. URL https://doi.org/10.1023/A:1007562322031. Citado na pág. 3

Sundheim(1992) Beth M. Sundheim. Overview of the fourth message understanding evaluation and conference. ISBN 1558602739. URL https://doi.org/10.3115/1072064.1072066. Citado na pág. 3

Winn *et al.***(2019)** John Winn, John Guiver, Sam Webster, Yordan Zaykov, Martin Kukla e Dany Fabian. Alexandria: Unsupervised high-precision knowledge base construction using a probabilistic program. Em *Automated Knowledge Base Construction (AKBC)*. Citado na pág. 2

Wu e Weld(2010) Fei Wu e Daniel S. Weld. Open information extraction using wikipedia. Em *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, páginas 118–127, USA. Association for Computational Linguistics. Citado na pág. 2

Zhang(2015) Ce Zhang. *DeepDive: a data management system for automatic knowledge base construction*. Tese de Doutorado, The University of Wisconsin-Madison. Citado na pág. 2