Universidade Federal do ABC

Heurísticas computacionais para extração de conhecimento em problemas de maratona de programação

PGC | BCC

Mauro Mascarenhas de Araújo

O problema

Introdução

Identificando o tema e “reduzindo o escopo” do problema...

- Recuperação de informação;
- Aplicação em ambiente educacional;
- Aplicação em aprendizado de ciência da computação;
- Busca por correlações entre enunciados de problemas de maratona de programação.

Trabalho relacionado

Trabalho relacionado:

Em "*Comparative Analysis of String Similarity and Corpus-Based Similarity for Automatic Essay Scoring System on E-Learning Gamification*" é proposto um sistema de pontuação automática para redações conduzidas em plataformas de aprendizado eletrônico utilizando métodos não supervisionados.

- Similaridade do cosseno (se saiu melhor: baixa complexidade computacional e boa acurácia);
- Análise semântica latente (maior complexidade computacional e boa acurácia).

Autores: Sakti e Fauzi

Trabalho relacionado

Algoritmo baseado em PLN para sistema de perguntas e respostas

Em "*NLP Algorithm Based Question and Answering System*", o objetivo foi entender os termos de busca do usuário utilizando técnicas de PLN juntamente com um novo mecanismo de pontuação para extrair as informações relacionadas a eles.

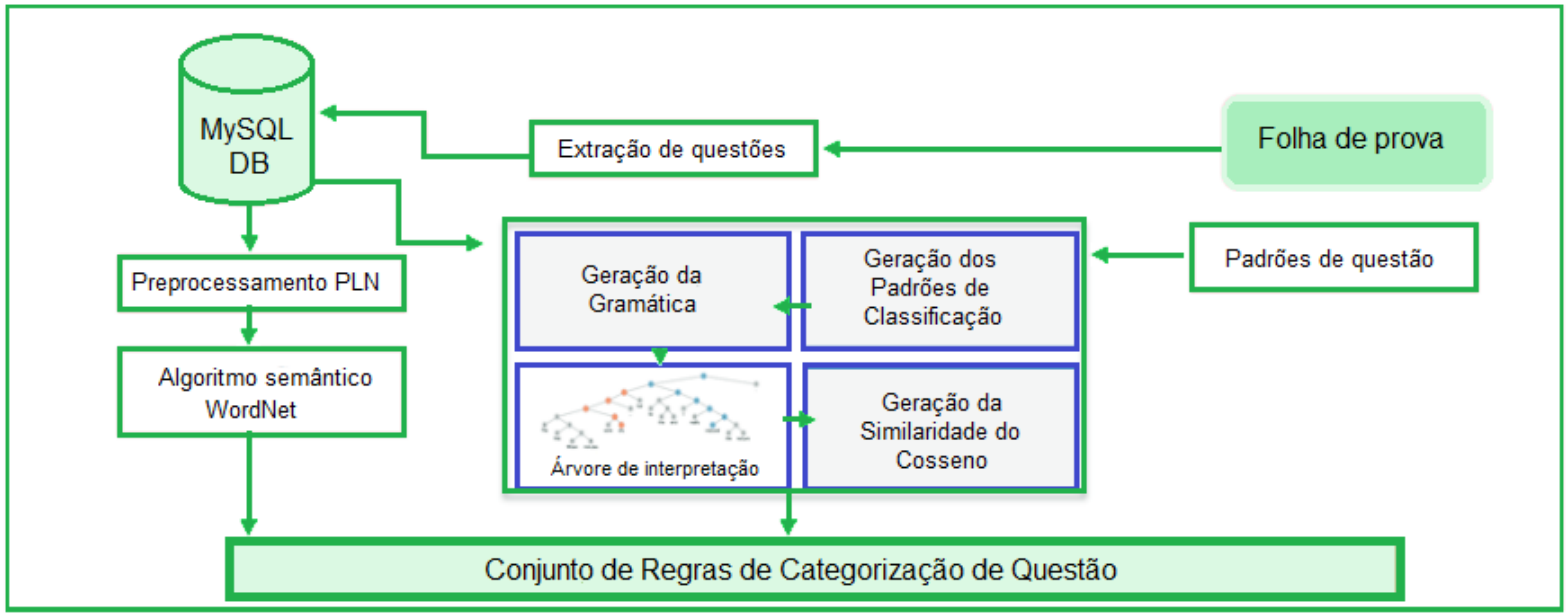
- Elaborado um sistema de FAQ para uma seguradora;
- Melhorado ao ponto de poder responder qualquer pergunta relacionada a seguro;
- Validado por três especialistas da área.

Autores: Sarkar, Rao Madasu, Baala Mithra e VRK Rao

Trabalho relacionado

Um classificador automático para questões de prova com WordNet e similaridade do cosseno.

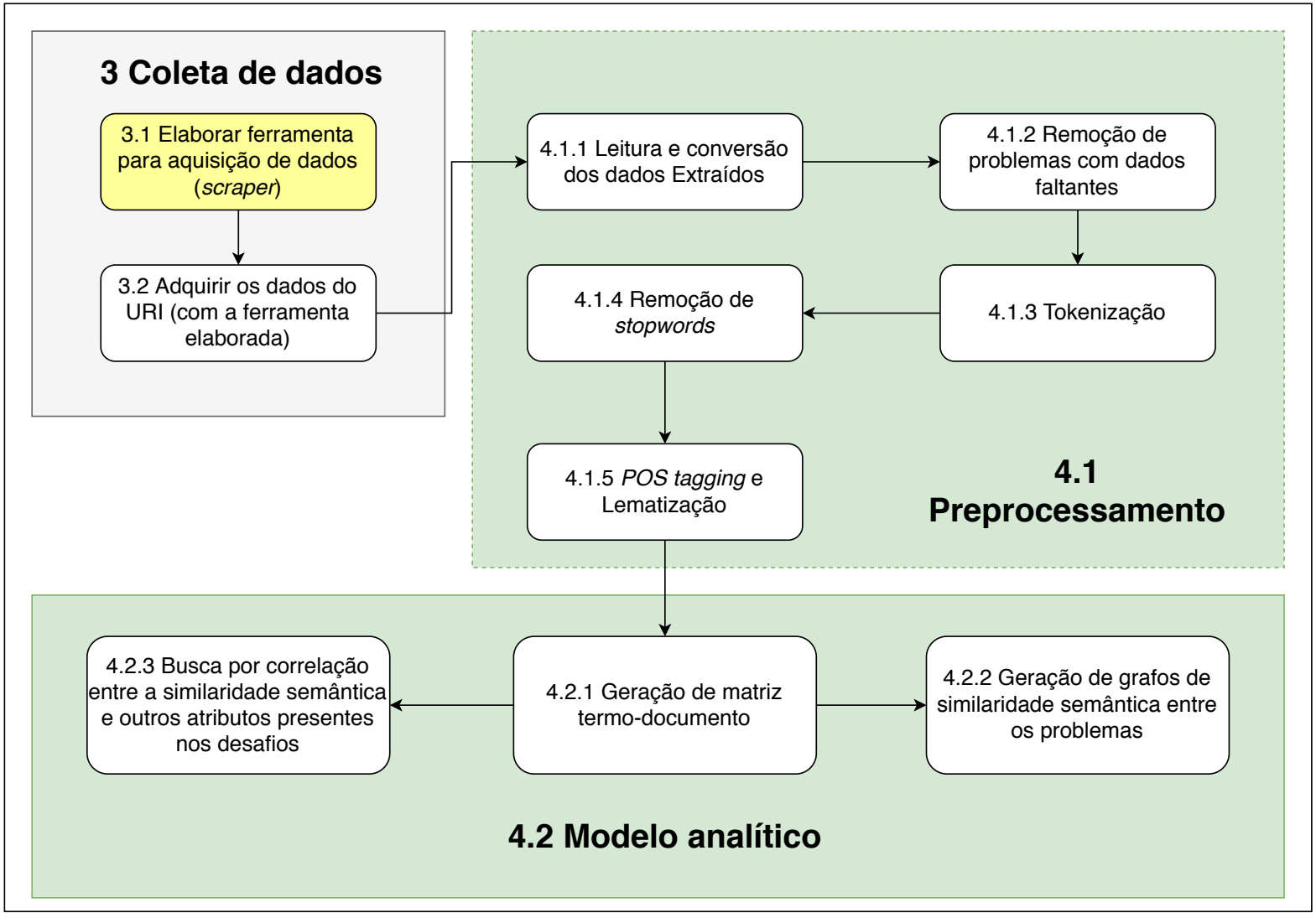
- Classificação de problemas de acordo com a taxonomia de Bloom (curva de aprendizado);
- Acurácia de aproximadamente 71%.



Fonte: Adaptado de Jayakodi, Bandara e Meedeniya (2016, p. 16)

Metodologia

Coleta de dados



Preparação

Coleta de dados

- Busca por um conjunto de dados existente na *web*: não havia o conjunto desejado;
- Especificação da fonte dos dados (URI);
- Exploração da estrutura do site;
- Elaboração do *web scraper* (Java + JSOUP);
- Aquisição dos dados utilizando a ferramenta elaborada.

Preparação

Coleta de dados: estrutura do site (problemas)

```
<div class="problem">
    <div class="description">
        <p> <!-- DESCRICAO DO PROBLEMA --> </p>
    </div>
    <h2>Input</h2>
    <div class="input">
        <p> <!-- DESCRICAO DA ENTRADA --> </p>
    </div>
    <h2>Output</h2>
    <div class="output">
        <p> <!-- DESCRICAO DA SAIDA --> </p>
    </div> <!-- ... -->
</div>
```

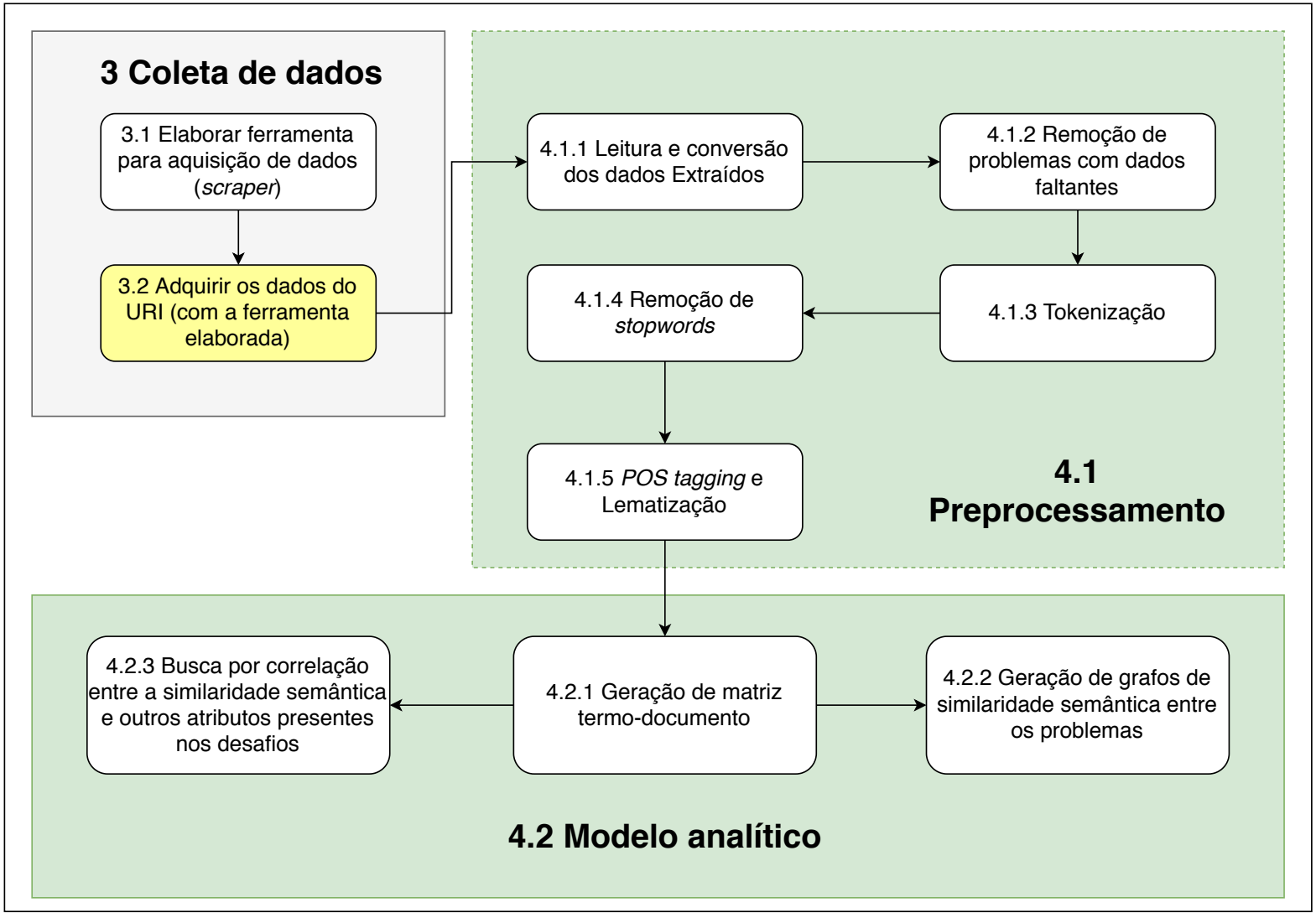
Preparação

Coleta de dados: estrutura do site (estatística)

```
<div class="st-small-box" id="problem-statistics">
    <h3>Statistics</h3>
    <p><!-- NUMERO DE VEZES RESOLVIDO --></p>
    <dl>
        <dt>Level:</dt>
        <dd><!-- NIVEL --> / 10</dd>
        <dt>Submissions:</dt>
        <dd><!-- NUMERO DE TENTATIVAS --></dd>
        <dt>Solved:</dt>
        <dd><!-- NUMERO DE VEZES RESOLVIDO --></dd>
        <dt>Ratio:</dt>
        <dd><!-- PORCENTAGEM DE ACERTO -->%</dd>
    </dl>
</div>
```

Metodologia

Coleta de dados: URI Scraper



Preparação

Coleta de dados: URI Scraper

```
Language index (valid options):
0 - Portuguese;
1 - English;
2 - Spanish;
Please, select the language of the URI problems :
```

Este é o menu inicial do *URI Scraper*...

Preparação

Coleta de dados: dados extraídos

```
{
   "output": "<informacao de dados de saida>",
   "input": "<informacao de dados de entrada>",
   "level": <numero>,
   "name": "<nome>",
   "has_images": {false/true},
   "description": "<descricao do problema>",
   "id": "<id>",
   "category": "<categoria>",

   ...
```

Preparação

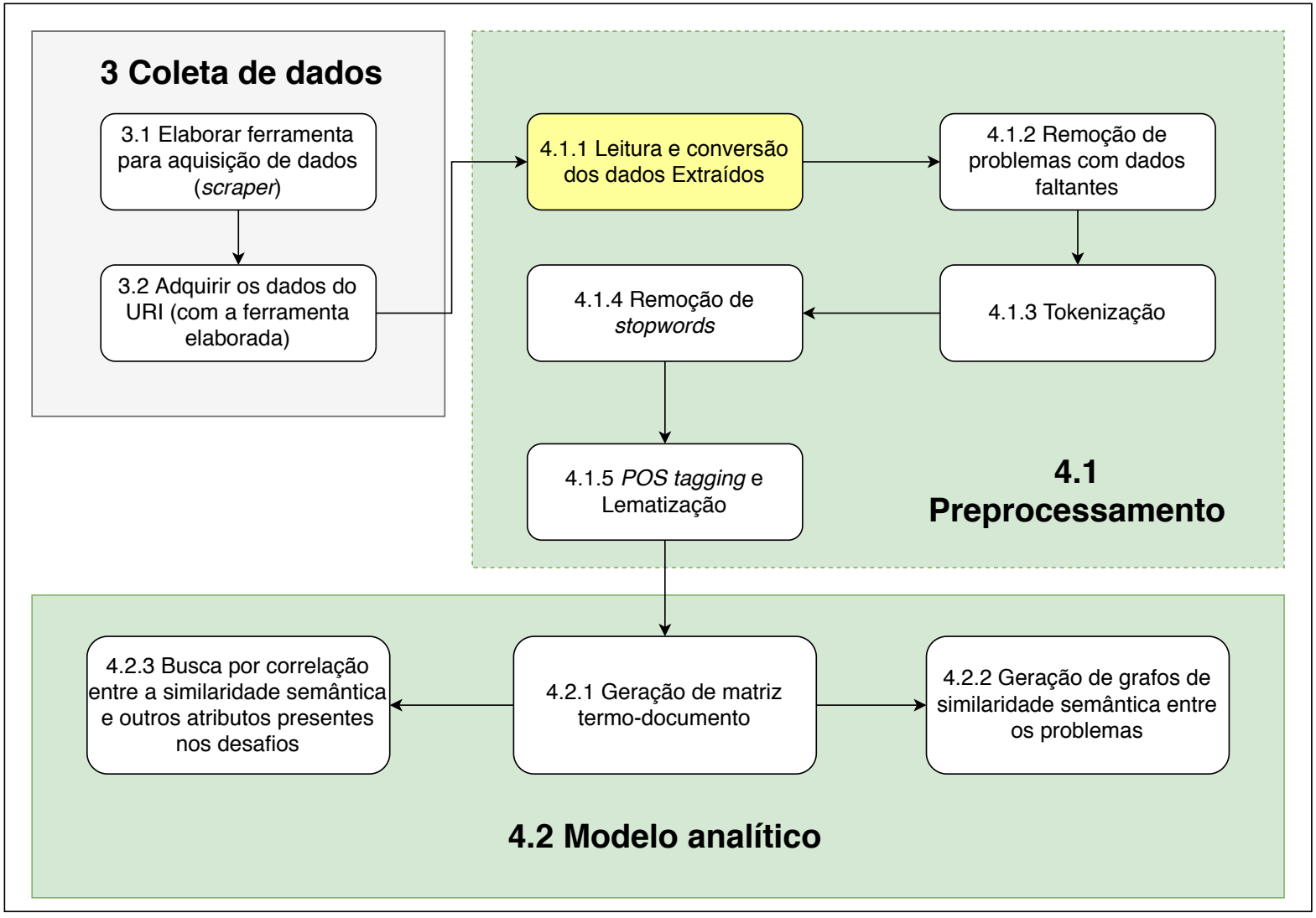
Coleta de dados: dados extraídos

```
...

  "statistics": {
    "level": "<numero> / 10",
    "submissions": <numero>,
    "solved": <numero>,
    "ratio": "<numero>%"
  }
}
```

Metodologia

Preprocessamento: Leitura e conversão dos dados



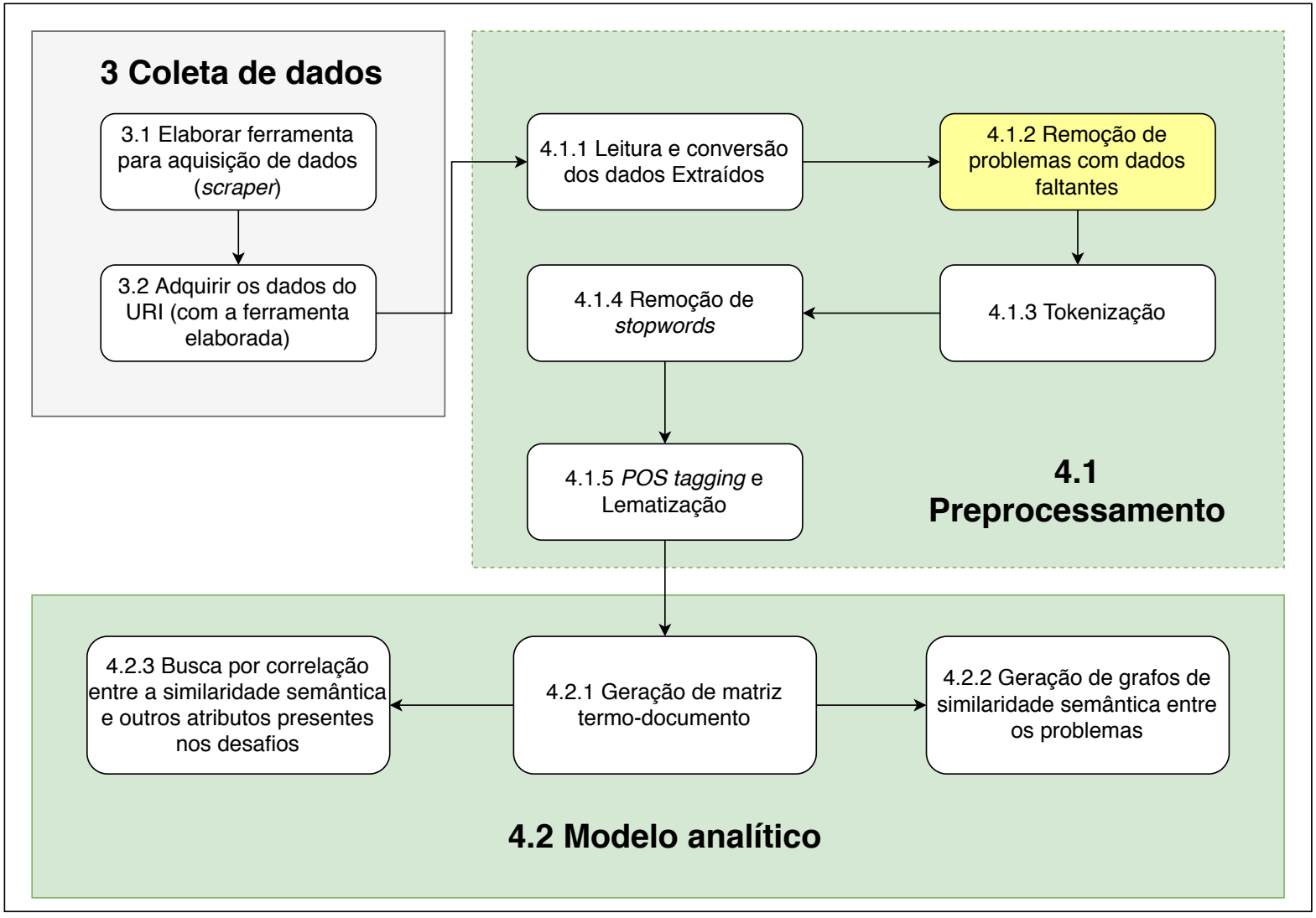
Metodologia

Preprocessamento: Leitura e conversão dos dados

A primeira etapa do processo de limpeza consiste em efetuar a leitura dos atributos de todos os problemas em uma estrutura de dados única.

Metodologia

Preprocessamento: Tratamento de dados faltantes



Metodologia

Preprocessamento: Tratamento de dados faltantes

Os desafios cujos seguintes atributos não constavam na base de dados foram removidos:

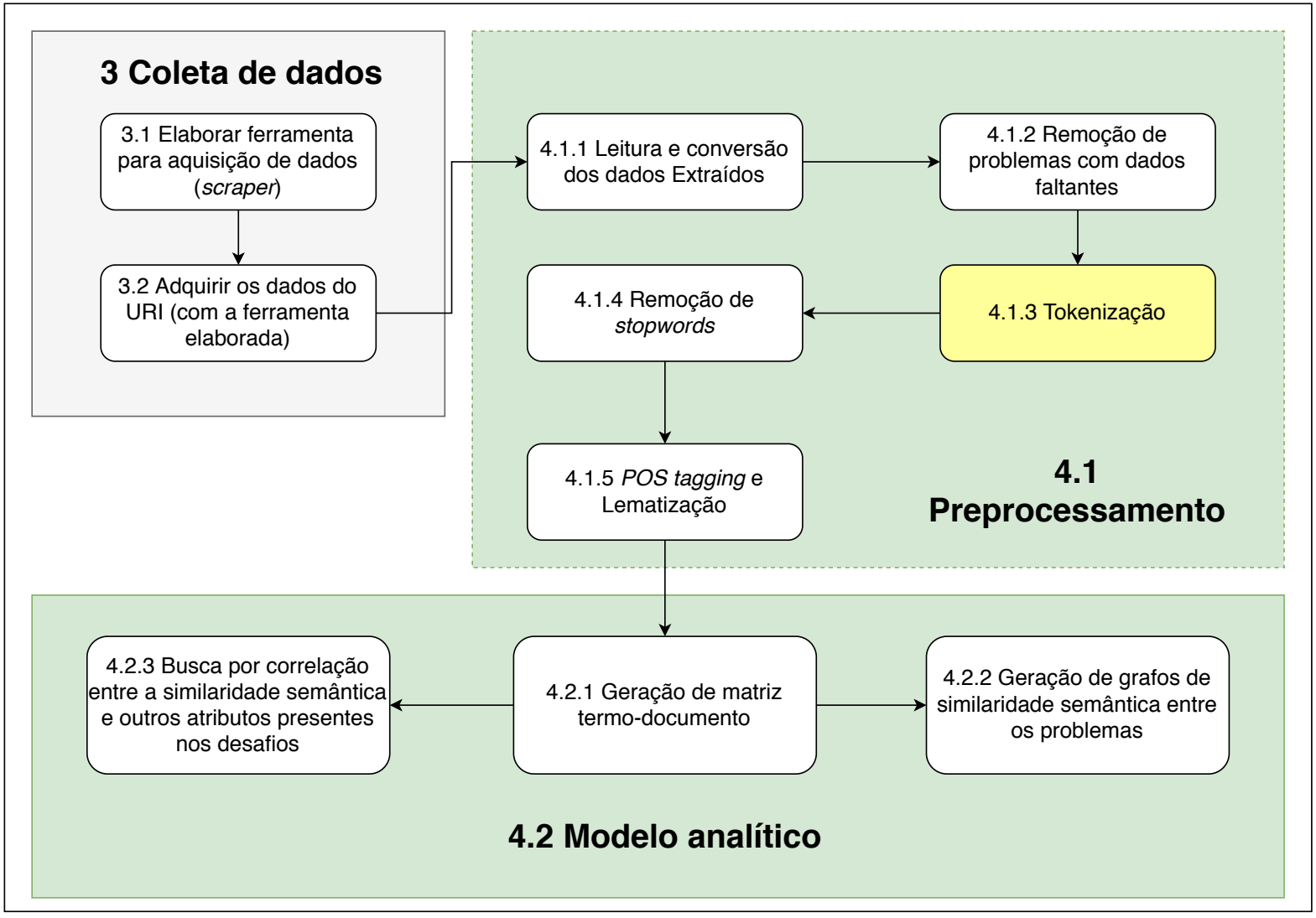
- Enunciado;
- Informações de entrada;
- Informações de saída;

Dados estatísticos

Opcionalmente, também podem ser removidos problemas sem dados estatísticos, pois, eles poderão atrapalhar a última etapa da análise.

Metodologia

Preprocessamento: Tokenização



Metodologia

Preprocessamento: Tokenização

Dada uma sequência de caracteres (*String*) e uma unidade definida de documento (Parágrafo/Frase/Oração/Palavra/etc.), tokenização (do inglês *tokenization*) é a tarefa de quebrar a sequência fornecida em pedaços (chamados *tokens*), normalmente descartando caracteres de pontuação. Exemplos:

- Dada a frase “Olá, como vai?” e a métrica “palavra” para definição de *token*, ela será quebrada da seguinte forma: [“Olá”, “como”, “vai”];
- Dada a frase “Olá, como vai?” e a métrica “letra/caractere” para definição de *token*, ela será quebrada da seguinte forma: [“O”, “l”, “á”, “c”, “o”, “m”, “o”, “v”, “a”, “i”].

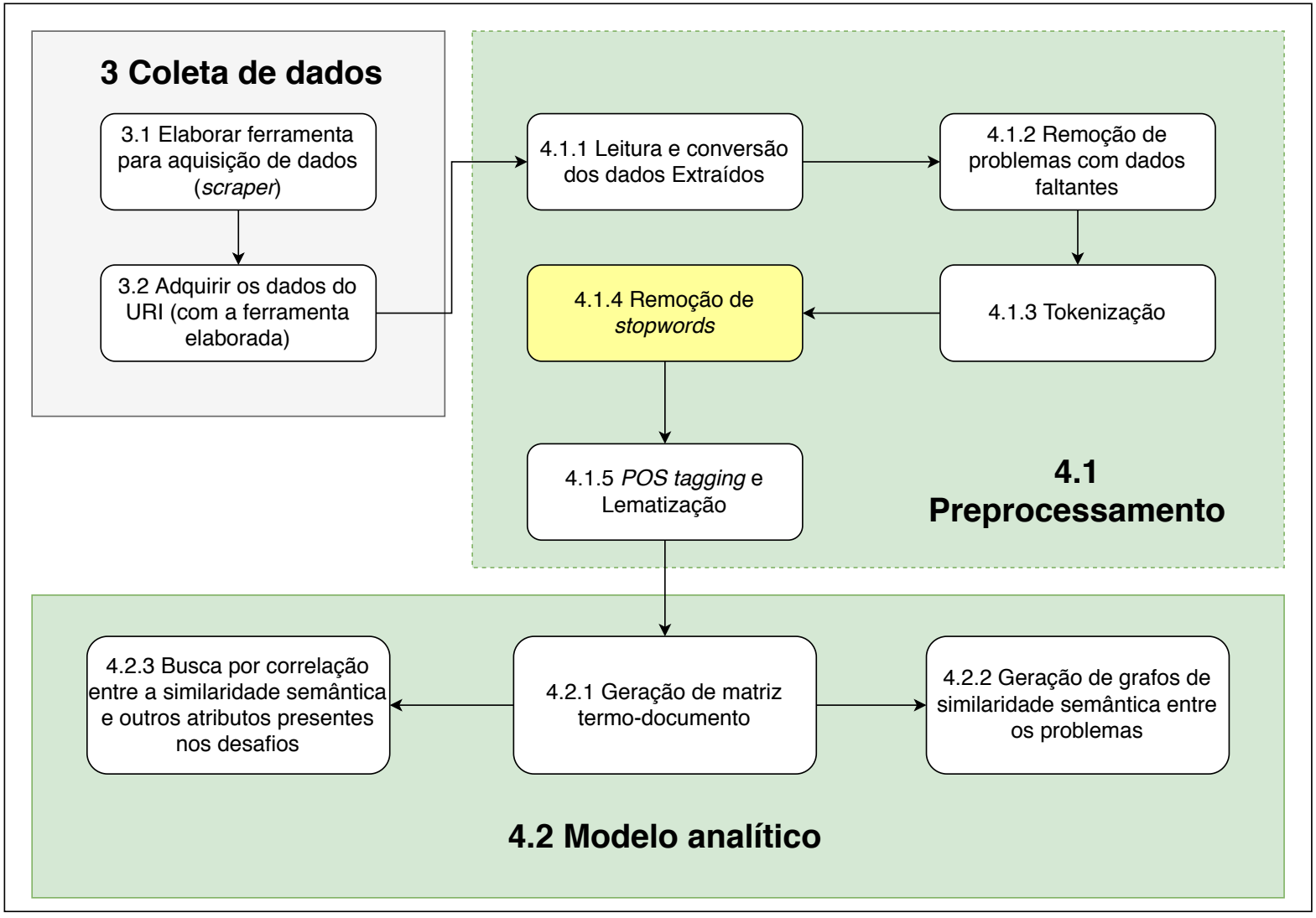
Metodologia

Preprocessamento: Tokenização

Sendo assim, a segunda etapa da normalização consiste em tokenizar tanto a descrição, quanto as informações de entrada e de saída dos problemas a nível de palavras (cada palavra é definida como um *token*).

Metodologia

Preprocessamento: Remoção de stopwords



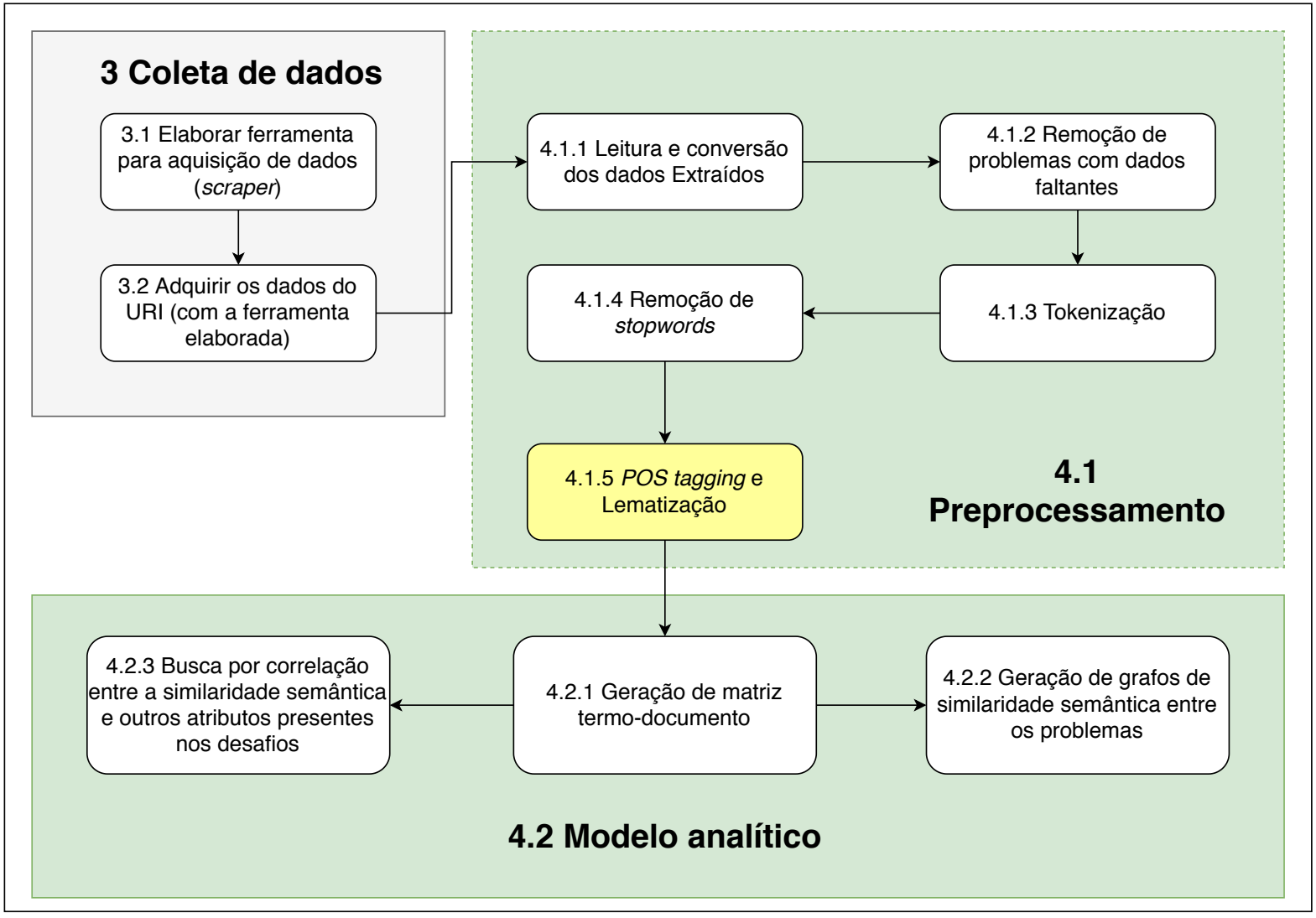
Metodologia

Preprocessamento: Remoção de stopwords

A terceira etapa, como o próprio nome diz, é onde as *stopwords* (palavras que são consideradas “irrelevantes” no contexto de recuperação de informação) são removidas do texto. Porém, como tratamento adicional, é necessário remover todas as pontuações e valores numéricos que possam acabar passando pelo processo de tokenização.

Metodologia

Preprocessamento: POS tagging e lematização



Metodologia

Preprocessamento: POS tagging e lematização

Para obter um bom resultado no processo de lematização, recomenda-se que a palavra possua uma classe gramatical atribuída, a fim de evitar ambiguidade no momento em que ela for transformada para sua forma inflexionada.

Exemplos de lematização

Verbo	Verbo inflexionado
É	Ser
Era	Ser
Seja	Ser
Sido	Ser

Metodologia

Preprocessamento: POS tagging e lemmatização

POS tagging

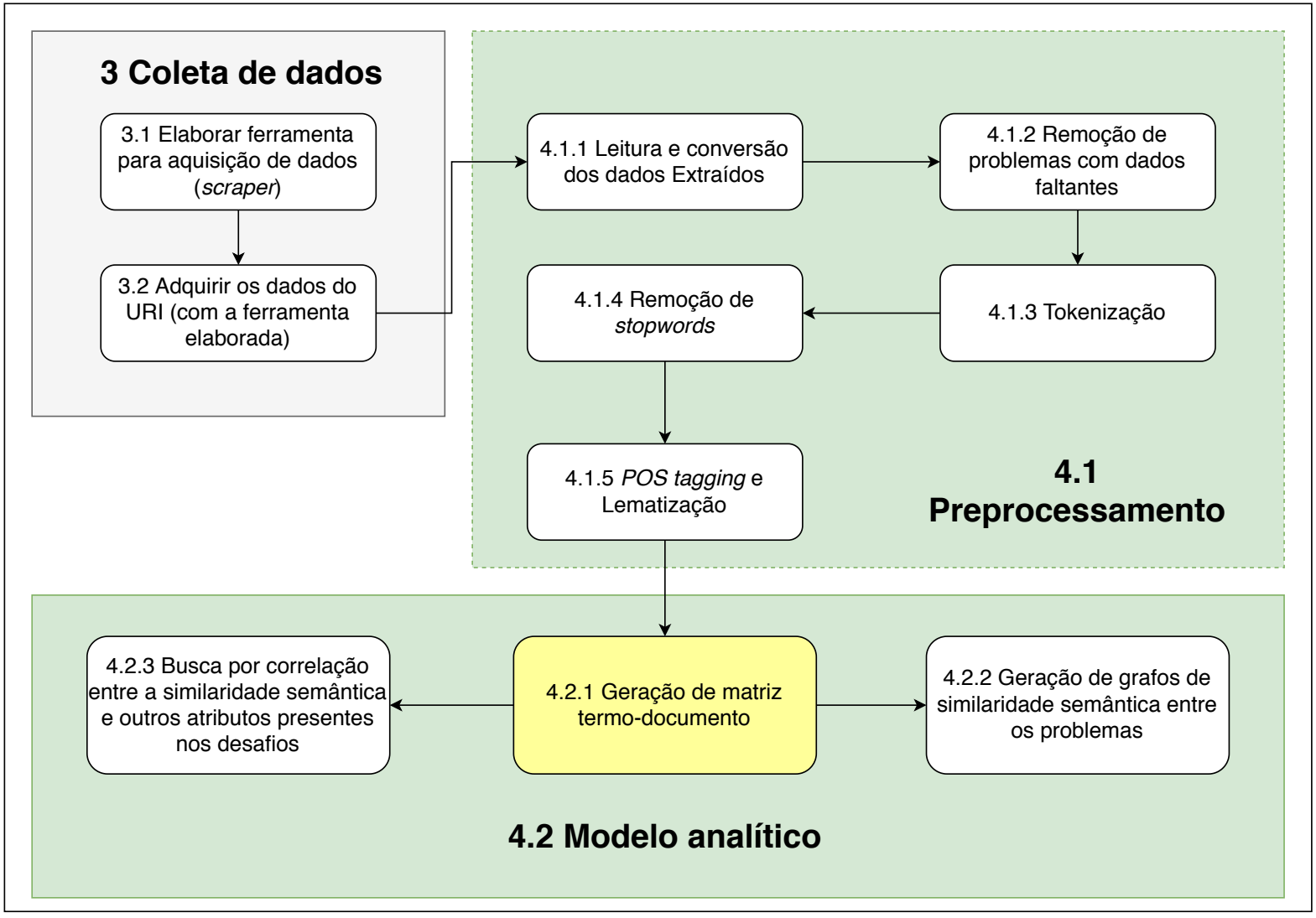
O processo que envolve aplicar uma "etiqueta" contendo a classe gramatical de cada palavra é conhecido como "*POS tagging*" (do inglês "*Part Of Speech*").

Exemplos de implementações de *taggers*:

- *Penn Treebank*;
- *WordNet*.

Metodologia

Modelo analítico: Matriz termo-documento



Metodologia

Modelo analítico: Matriz termo-documento

Para criar as matrizes termo-documento (vetorizar os textos), são necessários os seguintes passos:

1. Criar três conjuntos de palavras ($\mathscr{D}$, $\mathscr{E}$ e $\mathscr{S}$) contendo o vocabulário das descrições de todos os problemas, das informações de dados de entrada e das informações de dados de saída, respectivamente;
2. Gerar três matrizes ($M_{|\mathscr{D}|\times|\mathscr{P}|}$, $N_{|\mathscr{E}|\times|\mathscr{P}|}$ e $O_{|\mathscr{S}|\times|\mathscr{P}|}$), uma para cada conjunto, cujas colunas representam os problemas (conjunto de documentos, representados por $\mathscr{P}$) e as linhas representam as palavras dos vocabulários (termos);

Metodologia

Modelo analítico: Matriz termo-documento

3. Utilizando as matrizes geradas no passo anterior, é possível derivar três novas matrizes: $P_{|\mathscr{D}|\times|\mathscr{P}|}$, $Q_{|\mathscr{E}|\times|\mathscr{P}|}$ e $R_{|\mathscr{S}|\times|\mathscr{P}|}$, sendo estas obtidas através do cálculo da frequência do termo pelo inverso da frequência nos documentos (tf-idf).

Metodologia

Modelo analítico: Matriz termo-documento

TF-IDF

Para calcular a $tf \times idf$ (termo-frequência pelo inverso da frequência nos documentos) de cada termo, são utilizadas as seguintes equações:

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{1}$$

$$idf(t) = \ln\left(\frac{N}{df_t + 1}\right) + 1 \tag{2}$$

Metodologia

Modelo analítico: Matriz termo-documento

Exemplo de matriz termo-documento

Sejam D_1 e D_2 dois documentos:

- D_1 = “Eu amo sorvete de morango”
- D_2 = “Eles não gostam de sorvete de morango”

A matriz termo-documento para estes dois pequenos textos é apresentada na seguinte matriz:

Metodologia

Modelo analítico: Matriz termo-documento

Exemplo de matriz termo-documento

	D_1	D_2
Eu	1	0
Amo	1	0
Sorvete	1	1
De	1	2
Morango	1	1
Eles	0	1
Não	0	1
Gostam	0	1

Metodologia

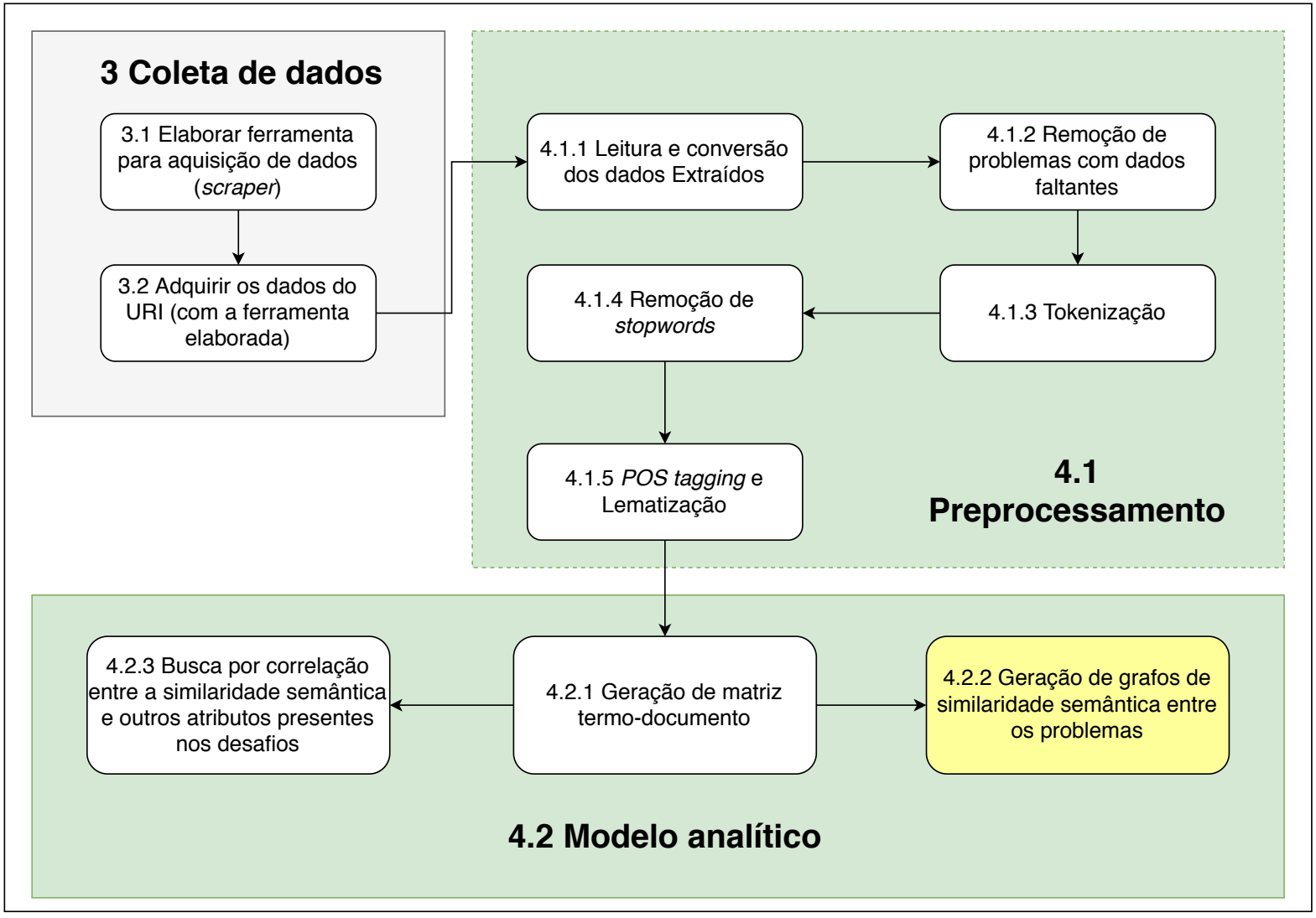
Modelo analítico: Matriz termo-documento

Exemplo de matriz tf-idf (utilizando os mesmos documentos D_1 e D_2 e as equações (1) e (2))

	D_1	D_2
Eu	0.2	0
Amo	0.2	0
Sorvete	0.1189	0.0849
De	0.1189	0.1698
Morango	0.1189	0.0849
Eles	0	0.1428
Não	0	0.1428
Gostam	0	0.1428

Metodologia

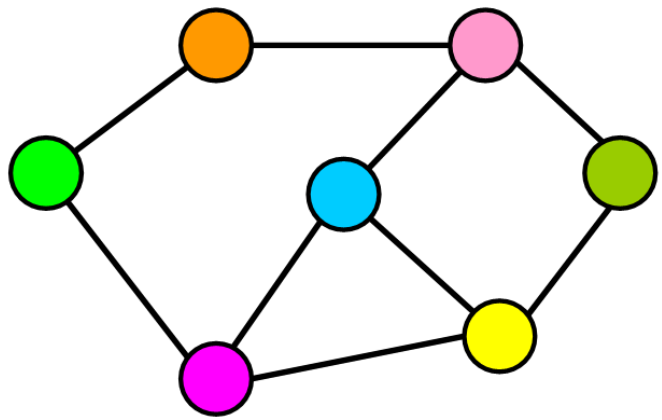
Modelo analítico: Rede de co-ocorrência



Metodologia

Modelo analítico: Rede de co-ocorrência

Para que os grafos sejam gerados, é necessário calcular a similaridade entre os problemas. Para tal, devem ser feitas $\sum_{i=1}^{p-2} (p - i)$ comparações, onde 'p' é um inteiro que representa a quantidade de problemas analisados.



Fonte: Data Structures 101: Graph A Visual Introduction for Beginners.

Metodologia

Modelo analítico: Rede de co-ocorrência

A métrica utilizada para realizar os cálculos de similaridade foi a distância do cosseno, dada pela seguinte fórmula (‘u’ e ‘v’ são os vetores-coluna obtidos da matriz termo-documento ou tf-idf):

$$d = 1 - \frac{u \cdot v}{|u|\,|v|} \tag{3}$$

Metodologia

Modelo analítico: Rede de co-ocorrência + E/S

O tamanho do vocabulário foi utilizado como métrica de "importância" para os textos, a fim de balancear os pesos que cada descrição possuirá no cálculo final da distância:

$$P_{max} = |\mathscr{D}| + |\mathscr{E}| + |\mathscr{S}| \tag{4}$$

$$P_E = \frac{|\mathscr{E}|}{P_{max}} \tag{5}$$

$$P_S = \frac{|\mathscr{S}|}{P_{max}} \tag{6}$$

Metodologia

Modelo analítico: Rede de co-ocorrência - com informações de E/S

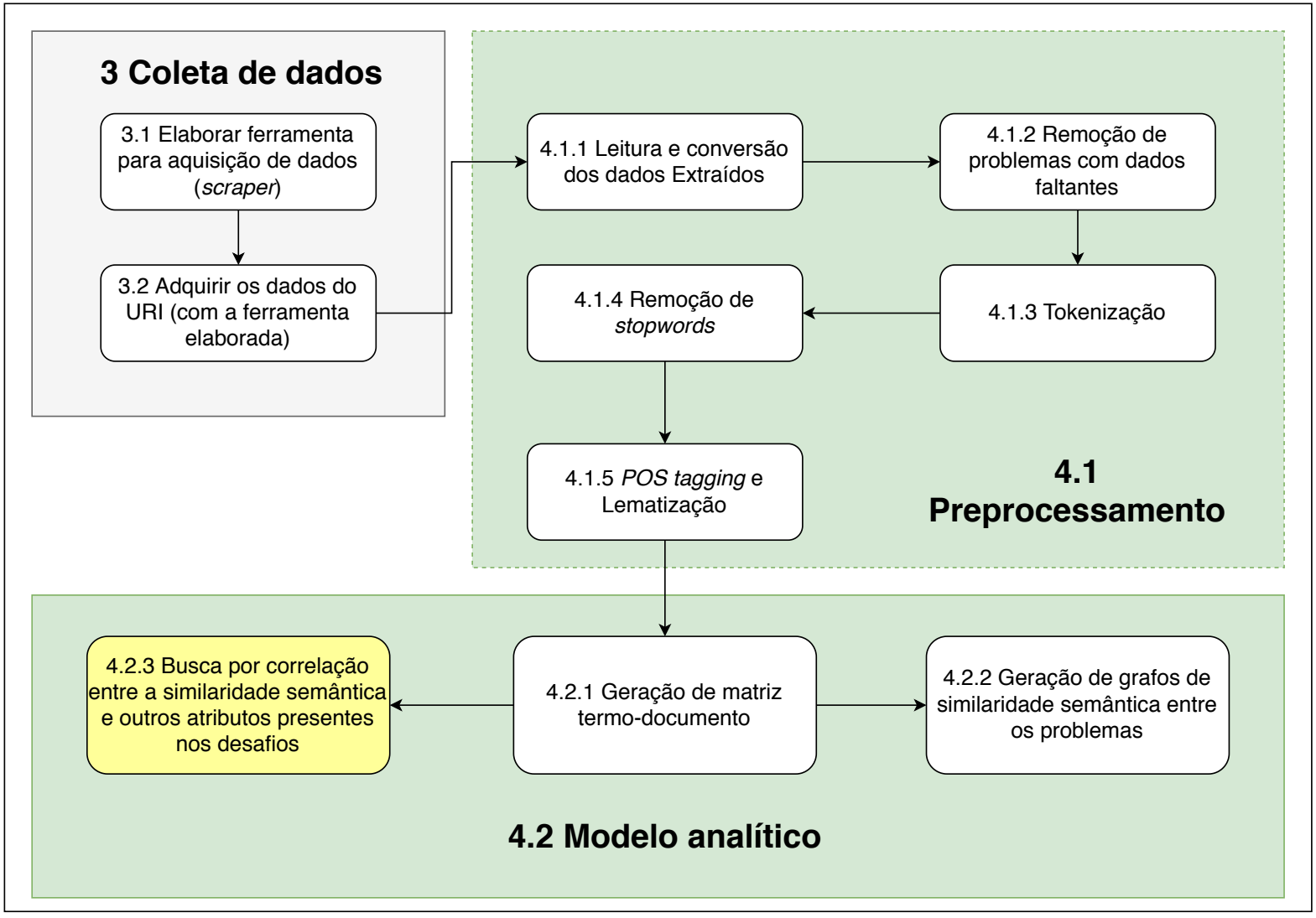
$$P_D = 1 - (P_E + P_S) \tag{7}$$

Cálculo final da distância (levando em consideração informações de E/S)

$$Dist = d(D_i, D_j) * P_D + d(E_i, E_j) * P_E + d(S_i, S_j) * P_S \tag{8}$$

Metodologia

Modelo analítico: Busca por demais correlações



Metodologia

Modelo analítico: Busca por demais correlações

A parte final da análise consiste em buscar padrões visuais que possam identificar possíveis correlações entre a similaridade semântica e três outros atributos disponíveis nos prolemas contidos no conjunto de dados analisado: categoria, nível de dificuldade e taxa de resolução.

Metodologia

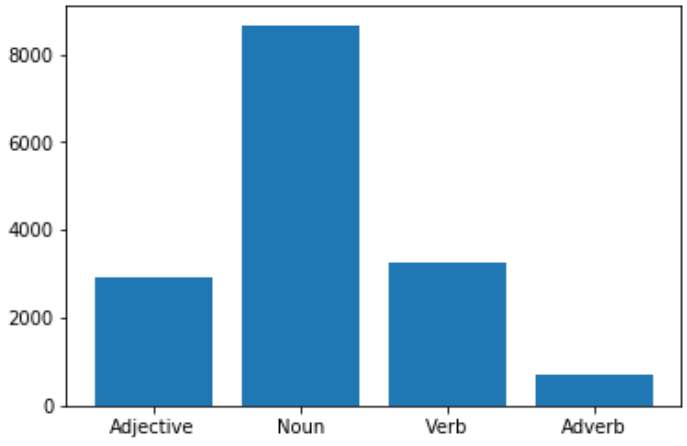
Modelo analítico: Busca por demais correlações

Como os vetores estão em uma dimensão não observável ($|\mathscr{D}|$-ésima, onde $\mathscr{D}$ é o vocabulário dos enunciados), torna-se necessário reduzi-la (para duas ou três dimensões, as quais são possíveis "plotar"). Para isso, são utilizados três métodos:

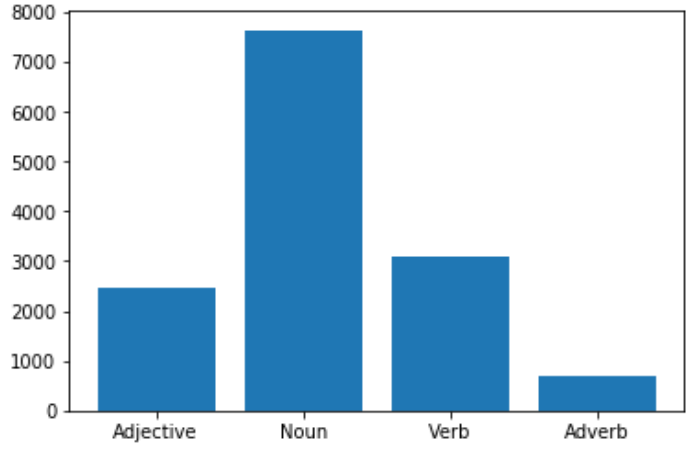
- PCA: Por ser um dos algoritmos mais utilizados para fins de redução de dimensionalidade;
- MDS: Evita sobreposição de pontos, permitindo uma visualização mais clara da distribuição dos atributos dos desafios analisados;
- SVD: Pois, ele tende a trabalhar com matrizes esparsas de forma mais eficiente.

Resultados e análises

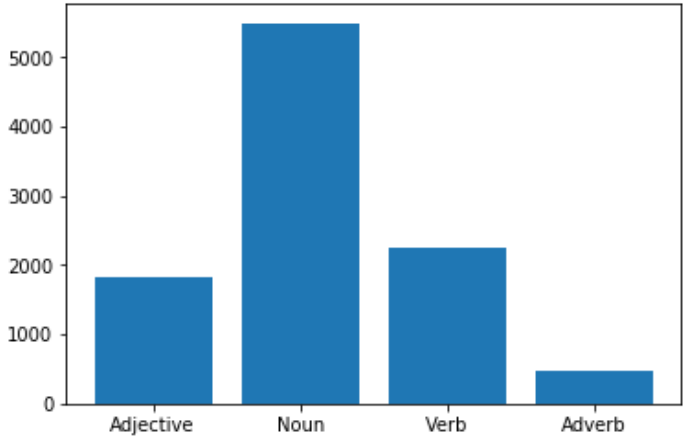
Classificações gramaticais por categoria (POS tagger)



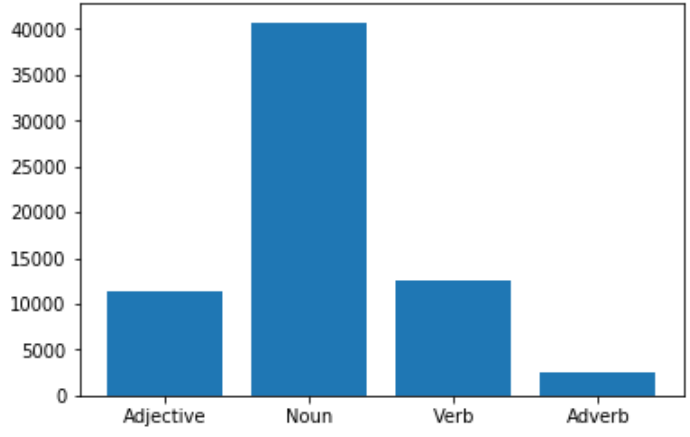
Iniciante



Estrutura de dados



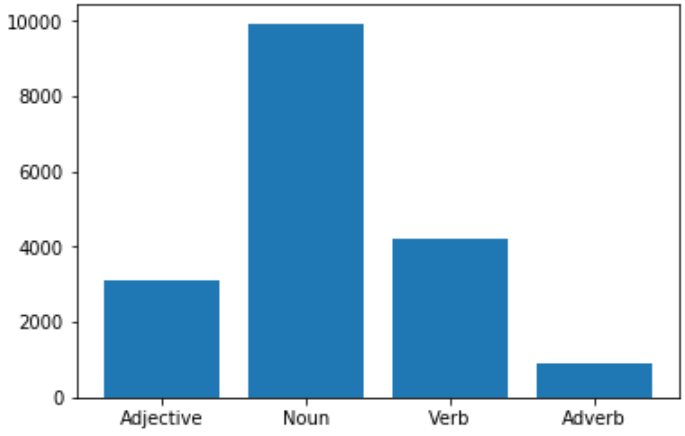
Strings



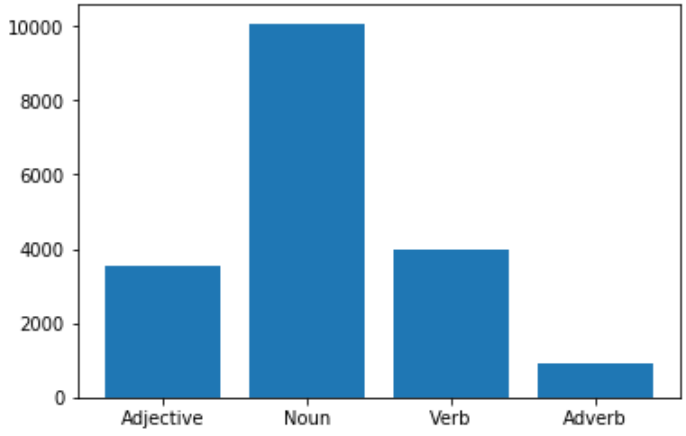
Ad-Hoc

Resultados e análises

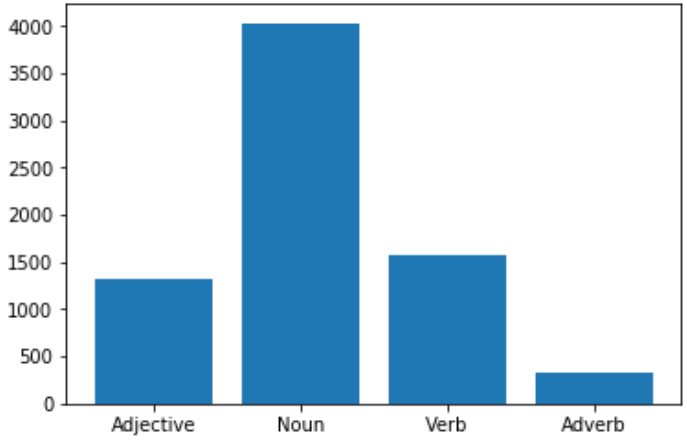
Classificações gramaticais por categoria (POS tagger)



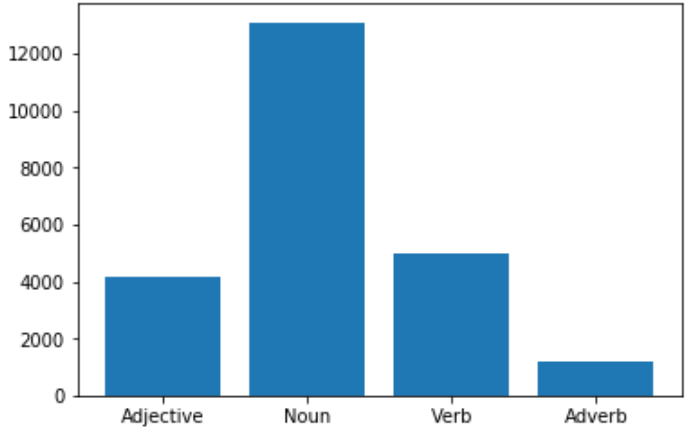
Paradigmas



Matemática



Geometria computacional



Grafos

Resultados e análises

Classificações gramaticais (POS tagger)

Possíveis causas para a grande quantidade de substantivos:

- Palavras sem classificação gramatical classificadas automaticamente como “substantivo”;
- Uso da classe “substantivo” como valor-padrão durante a conversão das *tags* do *Penn Treebank* para as quatro classes analisadas.

Resultados e análises

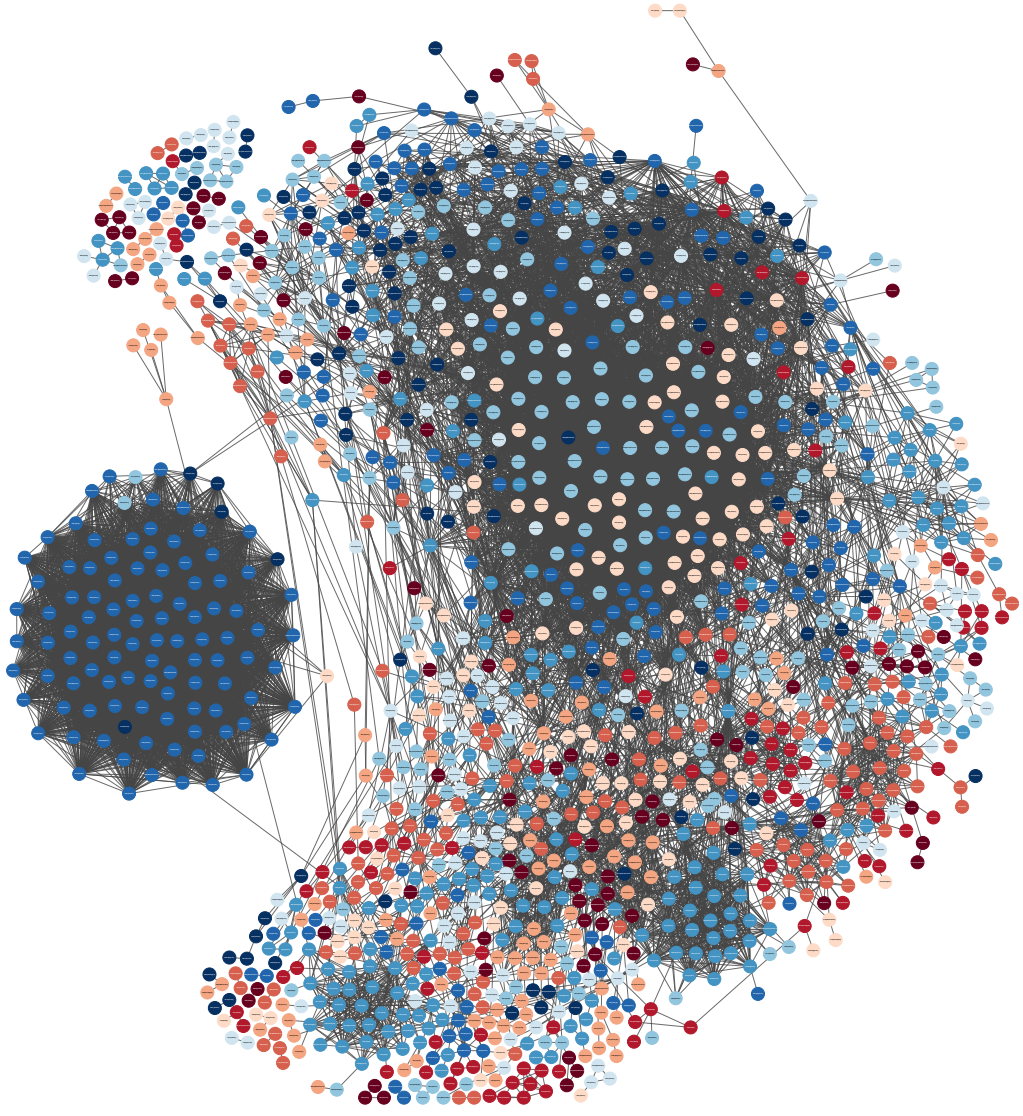
Classificações gramaticais (POS tagger)

Variação do vocabulário

É possível notar que a categoria '*ad-hoc*' é a que apresenta maior variabilidade de palavras (maior vocabulário), enquanto a 'geometria computacional' tem o menor vocabulário dentre todas elas.

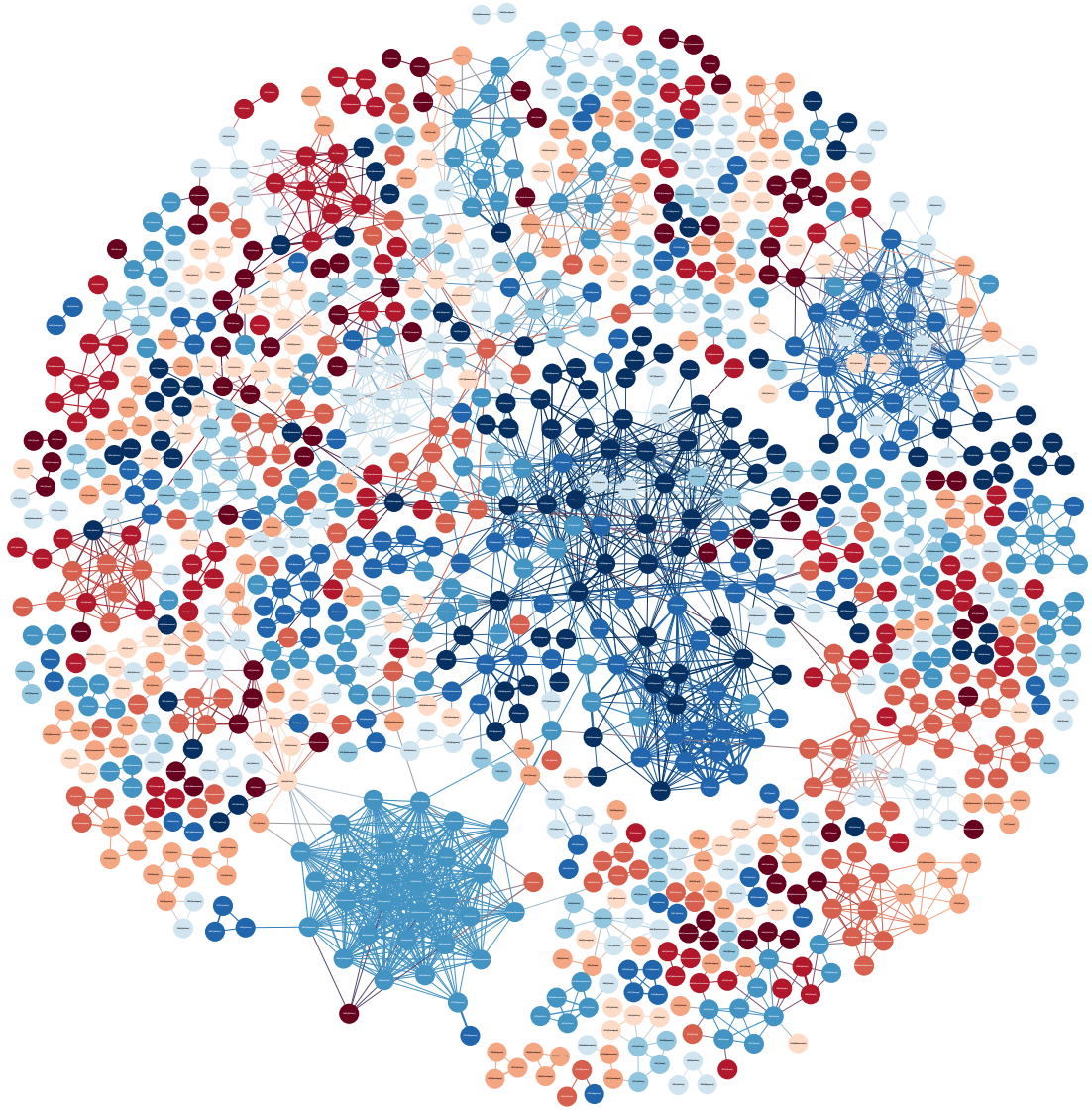
Resultados e análises

Grafos de co-ocorrência: termo-documento



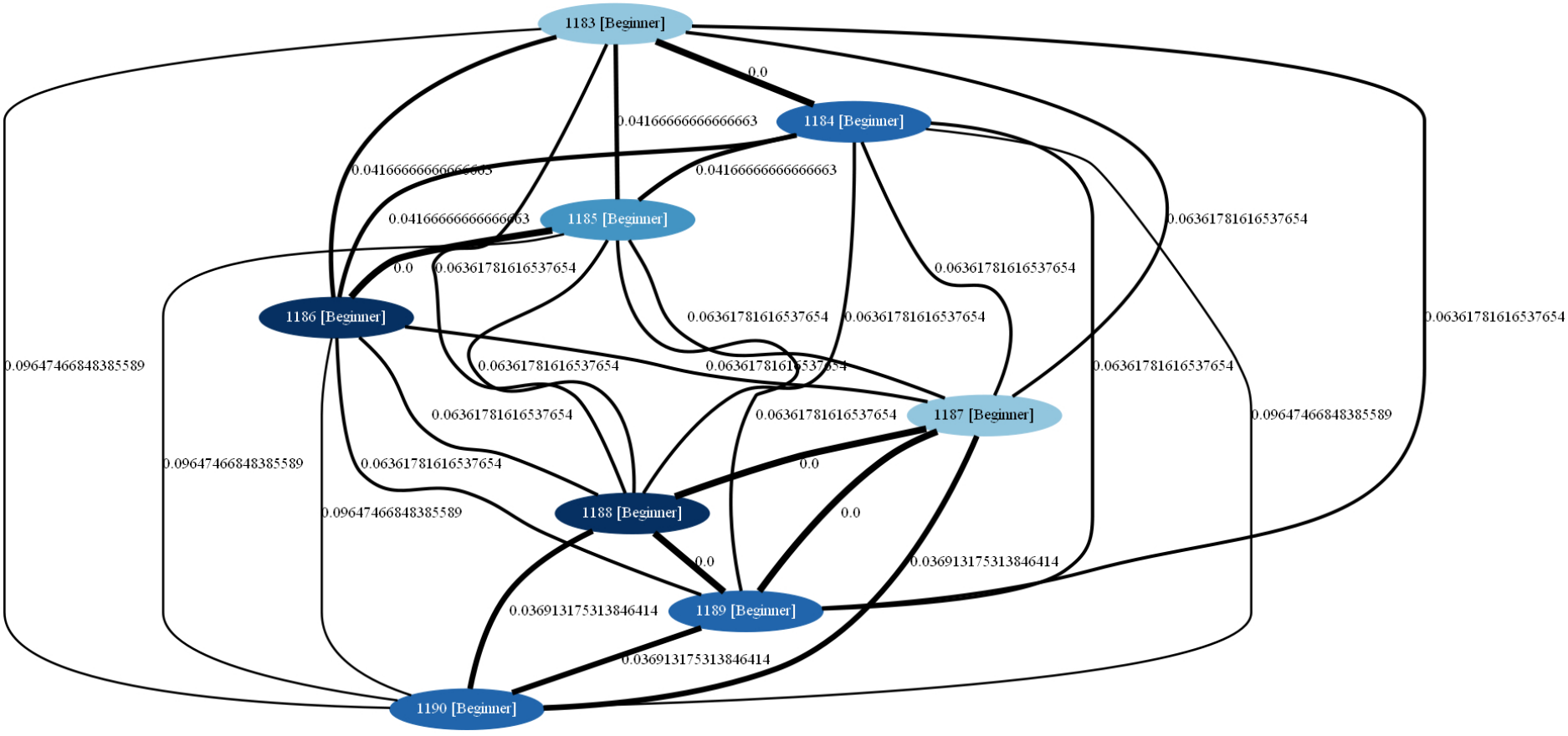
Resultados e análises

Grafos de co-ocorrência: tf-idf



Resultados e análises

Grafos de co-ocorrência: considerando apenas o enunciado



Maior componente conexa do grafo de co-ocorrência (utilizando a matriz termo-documento “padrão” e considerando similaridade superior a 0.7).

Resultados e análises

Grafos de co-ocorrência: considerando apenas o enunciado

Enunciado do problema 1183

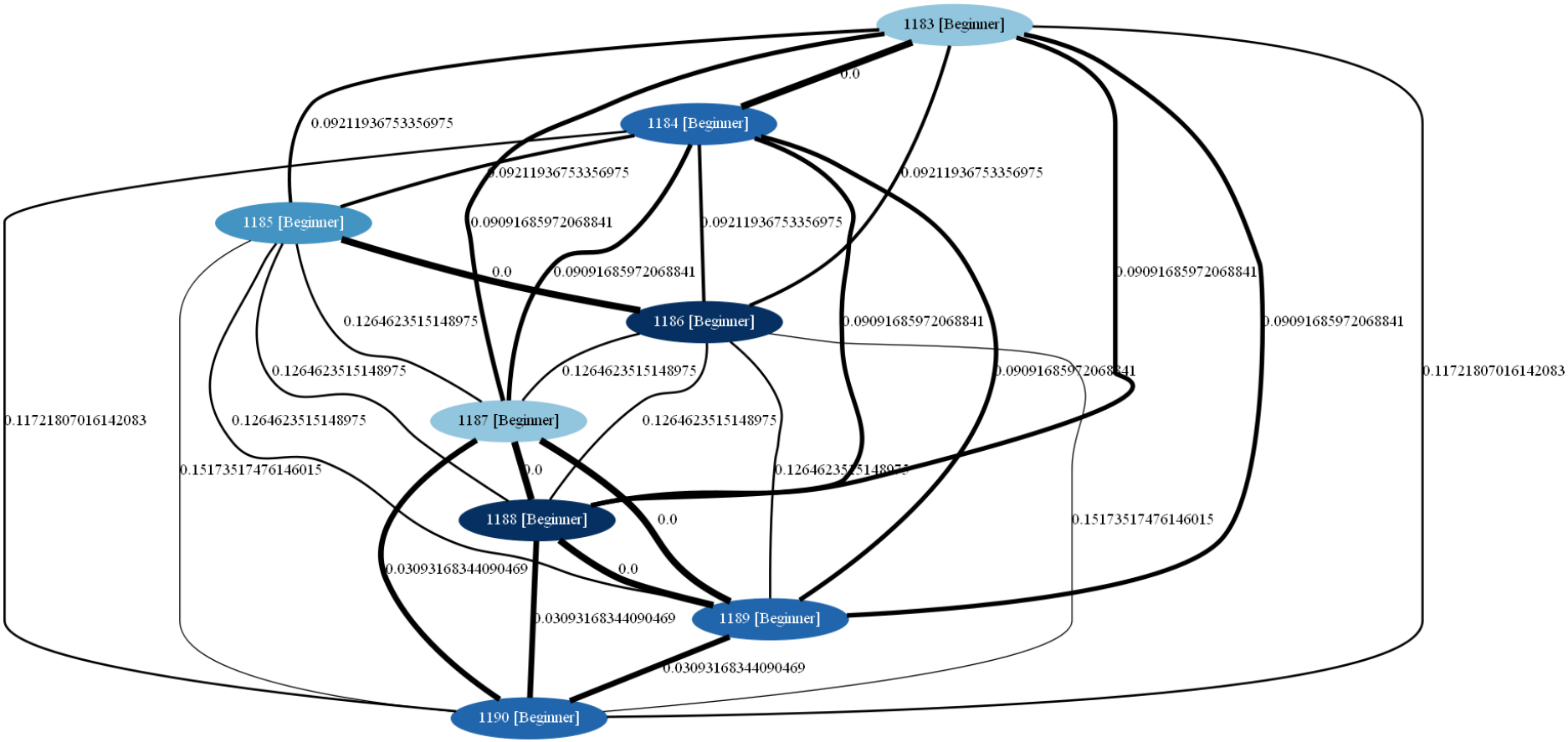
Read an uppercase character that indicates an operation that will be performed in an array **M**[12][12]. Then, calculate and print the sum or average considering only that numbers that are above the main diagonal of the array, like shown in the following figure (green area).

Enunciado do problema 1184

Read an uppercase character that indicates an operation that will be performed in an array **M**[12][12]. Then, calculate and print the sum or average considering only that numbers that are below of the main diagonal of the array, like shown in the following figure (green area).

Resultados e análises

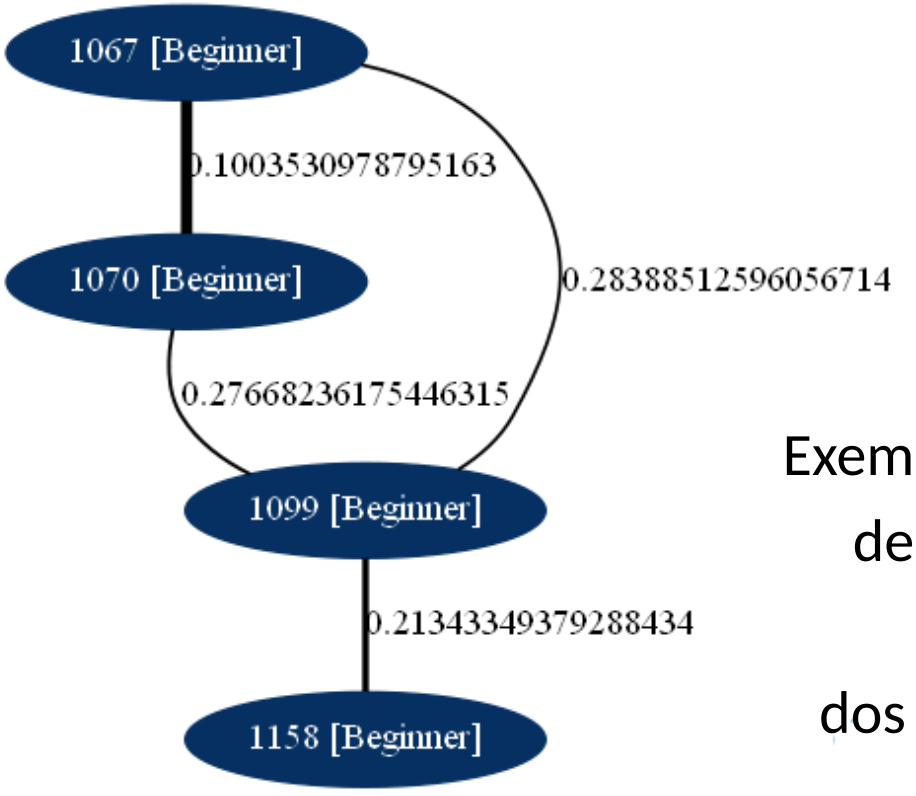
Grafos de co-ocorrência: considerando informações de entrada e de saída

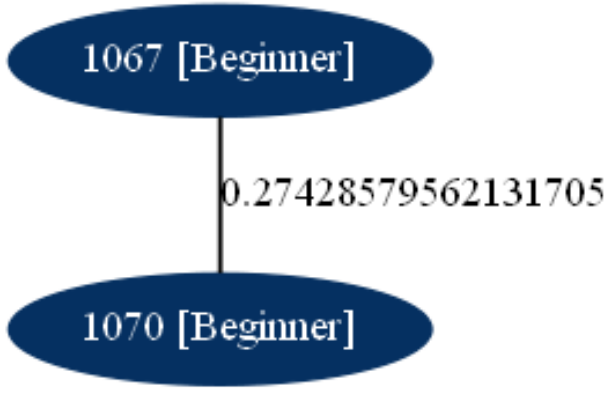


Maior componente conexa do grafo de co-ocorrência (utilizando a matriz tf-idf e considerando as descrições de E/S e com similaridade superior a 0.7).

Resultados e análises

Grafos de co-ocorrência: considerando informações de entrada e de saída





Exemplo de desafios cujos valores de similaridade foram afetados ao considerar as descrições dos dados de entrada e de saída.

Antes da inclusão: esquerda.

Depois da inclusão: direita.

Resultados e análises

Grafos de co-ocorrência: considerando informações de entrada e de saída

Descrições de entrada e de saída do problema 1067

Input

The input will be an integer value.

Output

Print all odd values between 1 and **x**, including **x** if is the case.

Descrições de entrada e de saída do problema 1070

Input

The input will be a positive integer value.

Output

The output will be a sequence of six odd numbers.

Resultados e análises

Distribuição espacial e busca por correlações

Conforme mencionado na metodologia, os seguintes atributos foram analisados a partir da dispersão espacial dos desafios:

- Categoria;
- Nível de dificuldade;
- Taxa de resolução : $(\frac{\text{‘submissões corretas’}}{\text{‘submissões’}})$.

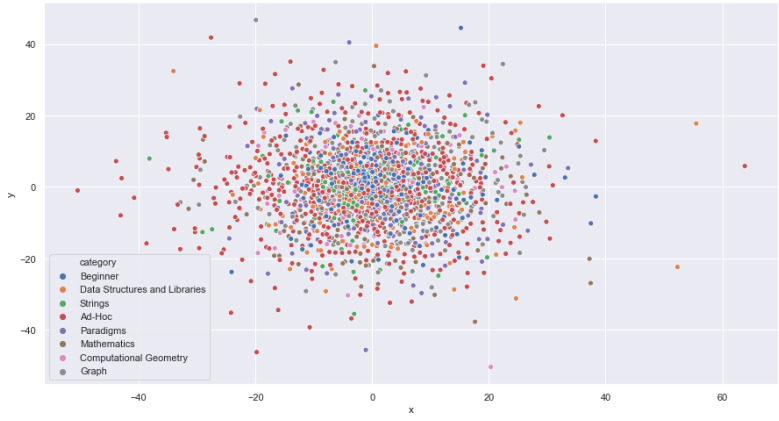
Atributos “descartados” da análise:

- “Possui imagens”;
- Submissões (total);
- Submissões corretas.

Resultados e análises

Distribuição espacial e busca por correlações - Categoria

17.B: PCA (tf-idf)



17.C: MDS (termo-documento)

17.F SVD (tf-idf)

Resultados e análises

Distribuição espacial e busca por correlações - Categoria

- Categoria “iniciante” (“*beginner*”) forma aglomerados em alguns pontos (confirmados pelo grafo);
- Alta variância na categoria “*Ad-Hoc*”.

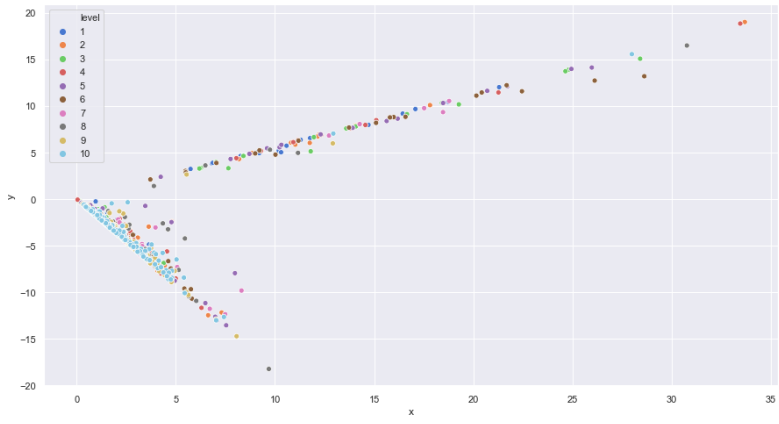
Categoria “*Ad-Hoc*”

- Muitos desafios categorizados pertencem a ela;
- Apresenta desafios que poderiam ter outra classificação.

Resultados e análises

Distribuição espacial e busca por correlações - Nível

18.A: PCA (termo-documento)



18.E: SVD (termo-documento)

18.F SVD (tf-idf)

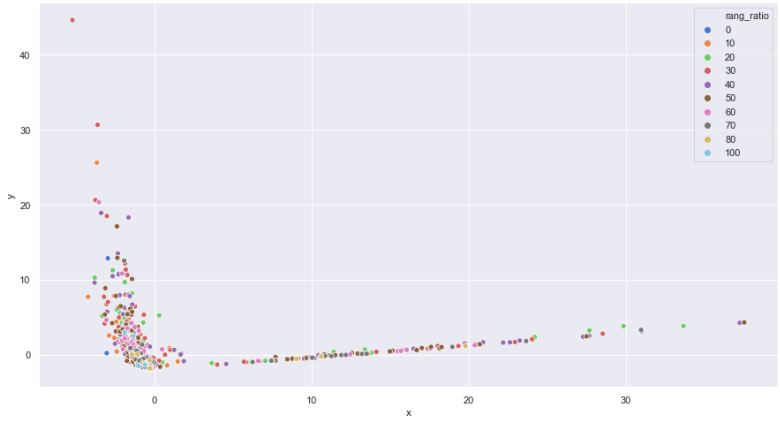
Resultados e análises

Distribuição espacial e busca por correlações - Nível

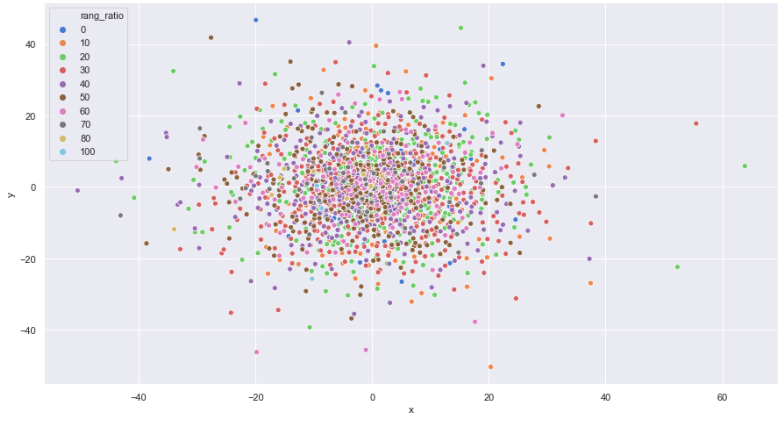
- Problemas de nível 10 “agrupados por serem diferentes”;
- Problemas mais fáceis não formam aglomerados (talvez não seja a melhor representação para eles).

Resultados e análises

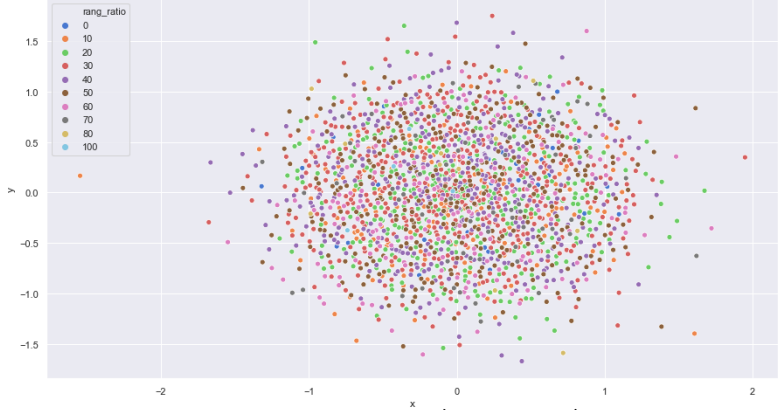
Distribuição espacial e busca por correlações - Taxa de resolução



19.A: PCA (termo-documento)



19.C: MDS (termo-documento)



19.D: MDS (tf-idf)

19.F: SVD (tf-idf)

Resultados e análises

Distribuição espacial e busca por correlações - Taxa de resolução

Cuidado!

A não existência de aglomerados (*clusters*) não implica a inexistência de similaridade semântica entre os problemas nas faixas de taxa de resolução estabelecidas.

Considerações finais

Conclusão

- Sucesso na aplicação das técnicas de PLN sobre o *corpus* analisado;
- Problemas de níveis de dificuldade próximos tendem a apresentar “tipo de vocabulário” similar;
- Descrições de entrada e de saída dos problemas são partes importantes do enunciado;
- Matriz “termo-documento padrão” apresentou melhores resultados em busca por similaridade semântica;
- Matriz “tf-idf” apresentou melhores resultados em análise de dispersão espacial.

Considerações finais

Trabalhos futuros

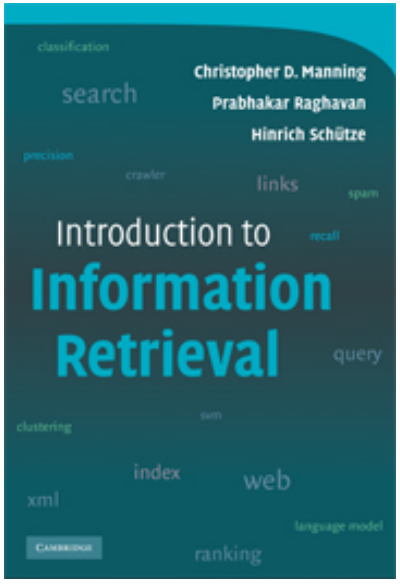
- Análise mais aprofundada do vocabulário (com todas as classificações gramaticais disponíveis);
- Elaborar um sistema de recomendação de enunciados com base na similaridade semântica ou através de aprendizado de máquina (considerando os demais atributos disponíveis);
- Analisar o significado de cada palavra no contexto dos problemas (ao invés de apenas a similaridade dos documentos).

Considerações finais

Algumas referências...

Introduction to Information Retrieval

Christopher D. Manning, Prabhakar Raghavan e Hinrich Schütze.
Cambridge University Press.
2008.



URI ONLINE JUDGE
PROBLEMS & CONTESTS

scikit learn

Considerações finais

Agradecimentos...

- Ao meu orientador, Monael Pinheiro e co-orientador, Jesús Mena-Chalco;
- Aos meus familiares e amigos, que me ajudam e apoiam;
- A todos os professores que contribuíram com minha formação;
- A todos vocês, aqui presentes :)

Enfim...

UFABC

Obrigado!