Relatório Técnico (50%)

Pad: Padrão (o trabalho contém todos os itens requeridos): 10,0
Ling: Linguagem (o texto está bem escrito, correto gramaticalmente e as ideias bem expressas): 8,0
Mod:Modelo (o modelo está bem descrito e sua instanciação feita adequadamente):8.0
Res: Resultados (os resultados foram devidamente descritos, inclusive utilizando gráficos): 10.0
Aval: Avaliação (os resultados foram adequadamente apresentados): 10,0

Implementação (50%)

Com: Comentários: (o código está devidamente documentado): 10,0
Cod: Codificação: (o código está bem escrito e roda adequadamente): 10,0
Result: Resultados: (os resultados foram produzidos adequadamente, inclusive utilizando gráficos): 10,0

Nota: 9.6

Relatório Técnico
sobre Processamento de Linguagem Natural

Thiago M. de Sousa *Luana G. B. Martins*
Ruan C. Rodrigues

Technical Report - RT-INF_000-19 - Relatório Técnico
July - 2019 - Julho

Instituto de Informática
Universidade Federal de Goiás
www.inf.ufg.br

Relatório Técnico
sobre Processamento de Linguagem Natural

Thiago M. de Sousa
thiagomontelesofc@gmail.com

Luana G. B. Martins
luanagbmartins@gmail.com

Ruan C. Rodrigues
ruanchaves93@gmail.com

Abstract. *This report describes the decisions, resolution process and results of the proposed Artificial Intelligence problem. The challenge of developing a tool that allows text comparison using Natural Language Processing methods was introduced.*

Keywords: Technical Report, Natural Language Processing.

Resumo. *Este relatório descreve quais foram as decisões, processo de resolução e resultados do problema proposto na matéria de Inteligência Artificial. Foi introduzido o desafio de desenvolver uma ferramenta que permita a comparação entre textos utilizando métodos de Processamento de Linguagem Natural.*

Palavras-Chave: Relatório Técnico,Processamento de Linguagem Natural .

1 Introdução

Este Relatório Técnico consiste na documentação de uma estratégia assumida para se obter um processo de tomada de decisões, buscando resolver o problema de desenvolver um sistema utilizando técnicas de Processamento de Linguagem Natural (PLN) que seja capaz de analisar a similaridade entre dois textos pequenos.

O processo de desenvolvimento será realizado utilizando técnicas de PLN, ~~a qual seu~~ objetivo é fornecer aos computadores a capacidade de entender e criar textos na forma da língua humana.

O problema consiste em receber um dataset com as respostas do estudo dirigido feito ao decorrer do curso de inteligência artificial e analisar a similaridade entre as respostas de outros alunos do curso.

No restante deste documento estão definidas a forma abordada na base de dados (Seção 2), descrição geral da solução proposta contendo ~~nela~~ a descrição geral do modelo utilizado e dos dados selecionados para o modelo (Seção 3), dos resultados obtidos (Seção 4), das propostas para como utilizar os resultados obtidos (Seção 5), conclusões finais (Seção 6) e referências.

2 Descrição da base de dados

Os dados consistem em três arquivos no formato csv (Comma-separated values), referente às respostas dos estudos dirigidos feitos ao decorrer da matéria de Inteligência Artificial ~~feito na Universidade Federal de Goiás~~. cada linha representa um aluno que cursou a matéria e cada coluna a resposta de uma determinada questão.

3 Descrição da solução

Nesta seção serão discutidos os princípios do funcionamento de um Processamento de Linguagem Natural (PLN) e como foi aplicado para resolver o problema proposto.

3.1 Descrição do modelo utilizado

3.1.1 Processamento de Linguagem Natural

Também conhecido como PLN, processamento de linguagem natural é o responsável pelo estudo da comunicação humana em sistemas computacionais, sendo ela uma subárea da inteligência artificial e linguística. Seus estudos foram propostos inicialmente na década de 1950, quando Alan Turing publicou o artigo *Computing machinery and Intelligence*, onde foi proposto um teste para criar um critério para verificar a inteligência das máquinas, hoje é conhecido como o Teste de Turing.[2]

3.1.2 Processos e abordagens para PLN

Alguns dos objetivos para PLN está na compreensão da língua natural, fazendo o computador tanto extrair o sentido da linguagem escrita e falada, como também ser capaz de manter uma interação nas mesmas formas. Para isso foram definidos alguns níveis de processamento e abordagens para auxiliar na aplicação das técnicas.

- Pré-processamento
  Com o objetivo de moldar a língua, de modo que a máquina possa compreendê-la é essencial a fase de pré-processamento, ela é constituída pela :

  – Normalização: Trata-se do processo de tokenização, transformando letras maiúsculas em minúsculas, remoção de caracteres especiais. O processo de tokenização tem como objetivo separar as palavras ou sentenças em unidades. Um exemplo de texto tokenização lexicalmente seria: [3]

  Esta é uma sentença.

  ['esta', 'é', 'uma', sentença', '.']

  A tokenização sentencial seria :

  Esta é a primeira sentença. Esta é a segunda. Esta é a terceira!

  ['Esta é a primeira sentença.', 'Esta é a segunda.', 'Esta é a terceira!']

  – Remoção de Stopwords: O processo denominado *stopwords* consiste em remover palavras que têm uma alta frequência como, "é", "de", "a" , "o", "que" entre outros. Essas palavras na maioria das vezes são irrelevantes para construção de um modelo. [3]

  – Remoção de numerais: Outro tipo de informação que não adicionam informações relevantes para semântica são símbolos como "kg", "R$" entre outros. [3]

  – Correção Ortográfica: A correção é feita para tratar um dataset que contém erros de digitação, esse erros são prejudiciais quando for gerado novos *tokens*, aumentando os resultados de erros no sistema final.

  - Stemização e Lematização: O processo denominado stemização(*stemming*) consiste em reduzir uma palavra ao seu radical. Um exemplo é a palavra meninas que será reduzida a menin, assim como meninos ou menininho, outro exemplo é a palavra gato que irá ser reduzido a gat. No caso de verbos o processo de lematização é aplicada, assim é retirado o infinitivo do verbo. Como exemplo as palavras tiver, tenho, tinha, tem são forma do mesmo lema ter. Esse processo é feita com o objetivo de reduzir o vocabulário e abstrair informações. [3]

- Processamento:

  - Fonológico: Identifica os sons que formam as palavras, geralmente é usada para reconhecimento da linguagem falada.
  - Morfológico: estudo sobre a natureza das palavras e sua composição.
  - Lexical: Processamento responsável por interpretar o significado de cada palavra.
  - Sintático: Estuda a composição da frase.
  - Semântico: Processamento que tem o objetivo de compreender o significado da frase.
  - Pragmático: Processamento que visa buscar os significados além das palavras.

- Tipos de Abordagem:
  A abordagem se refere ao tratamento que os algoritmos dão ao processamento.[4]

  - Simbólica: Essa abordagem se baseia nas regras linguísticas sem ambiguidades e bem estruturadas.
  - Estatística: Utiliza modelos matemáticos para deduzir o uso correto dos níveis de processamento.
  - Conexionista: Semelhante a abordagem estatística, também desenvolve modelos matemáticos, entretanto, ele faz utilizando aprendizado estatístico e teorias de representação do conhecimento.
  - Híbrida: Ela é a combinação entre todas as abordagens citados acima, sendo decidido conforme a estratégia para o problema proposto.

Além disso, um termo muito usado em PLN é o Corpus Linguístico, trata-se de um conjunto de documentos (ou frases) de uma determinada língua e que serve como base de análise. Outra estratégia usada é o TF-IDF (*Term Frequency - inverse document frequency*), que consiste em uma estratégia baseada em estatística onde é medida um peso que é usada para avaliar a importância de uma palavra para um documento em uma coleção de documentos. O valor tf-idf de uma palavra aumenta proporcionalmente conforme o número de vezes que a palavra aparece em um documento, no entanto, esse valor é equilibrado pela frequência da palavra no corpus.[1]

TF(Frequência do termo): A soma do número de vezes que um determinado termo aparece em cada documento após a pesquisa é denominado frequência do termo.

IDF(Inverso da frequência nos documentos): O IDF é responsável por equilibrar o peso de um termo, uma vez que, algumas palavras podem ser comuns no documento e assim sendo atribuído um alto peso ao seu termo, para isso, o inverso da frequência do termo é adicionado para diminuir o peso dos termos que mais ocorrem no conjunto de textos selecionados, ao mesmo tempo que aumenta o peso daqueles que ocorrem raramente.

3.2 Descrição do modelo final

O nosso modelo incluiu as seguintes etapas de pré-processamento:

- Separar palavras em tokens a partir dos espaços;
- Remover a pontuação e as *stop words*;
- Remover palavras que só ocorrem uma única vez;
- Efetuar o processo de *lemmatization*, isto é, reduzir as palavras à sua forma básica (o lema) usando como auxílio informações sobre o vocabulário da linguagem e a análise morfológica das palavras (o que vai além das heurísticas simples do processo de *stemming*).

Em seguida, associamos um vetor TF-IDF a cada documento, sendo TF a frequência do termo no documento em questão e IDF o inverso da quantidade de documentos nos quais esse termo ocorre em nosso corpus.

E sendo assim, calculamos uma matriz de similaridades que consiste na distância de cossenos entre todos os vetores TF-IDF de todos os documentos tomados dois a dois, isto é, no cosseno do ângulo que existe entre cada par de vetores multidimensionais. Esta medida de similaridade de cossenos é tomada por nosso modelo como sendo a similaridade existente entre dois documentos quaisquer.

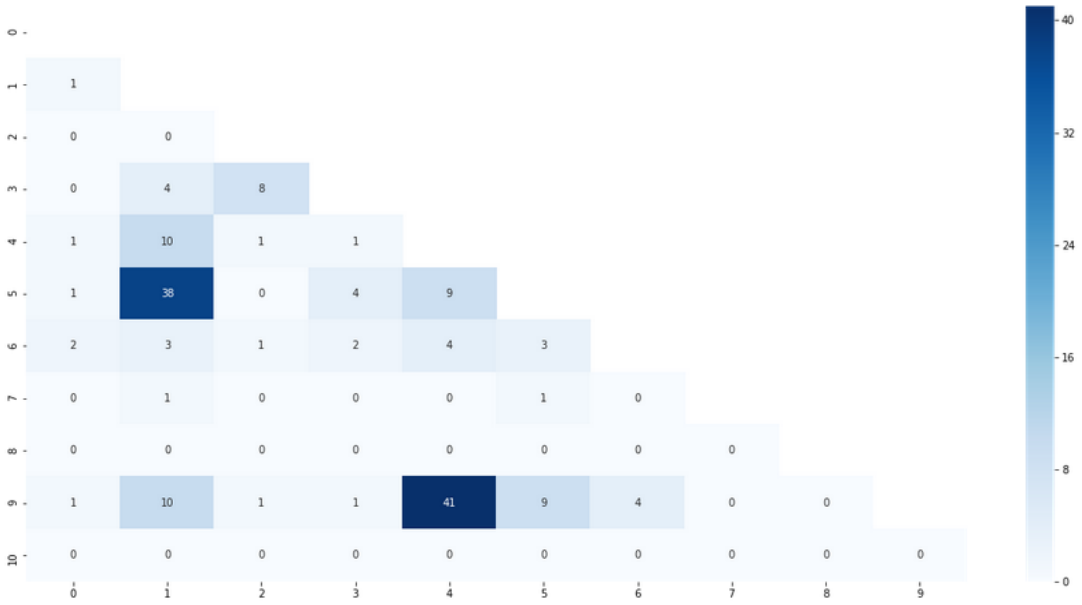
Como o *corpus* consiste de resumos feitos a partir de uma única referência, existe uma alta repetição de termos entre os documentos, dado que foram retirados de uma mesma fonte. Sendo assim, compreendemos que o TF-IDF seria uma medida satisfatória para a resolução do problema, dado que, no caso deste *corpus* específico, simplesmente observar quantos termos dois documentos compartilham em comum é uma boa medida da similaridade entre eles.

4 Resultados obtidos

O algoritmo foi executado sobre três datasets: ED05, ED08 e ED09.

Para cada dataset, fizemos a seguinte pergunta: "Dado um par de alunos, em quantas respostas ao questionário eles atingiram mais de 90% de similaridade entre si?". O resultado está exibido nos heatmaps abaixo, onde todos os alunos estão representados por índices no eixo x, e os mesmos alunos estão representados pelos mesmos índices no eixo y. Uma célula do heatmap corresponde a quantas respostas os dois alunos correspondentes à célula nos eixos x e y responderam com mais de 90% de similaridade entre si.

Para a atividade ED05:



A partir deste heatmap, constatamos uma alta semelhança entre os alunos 9 e 4. Observamos uma seleção aleatória de cinco de suas respostas abaixo:

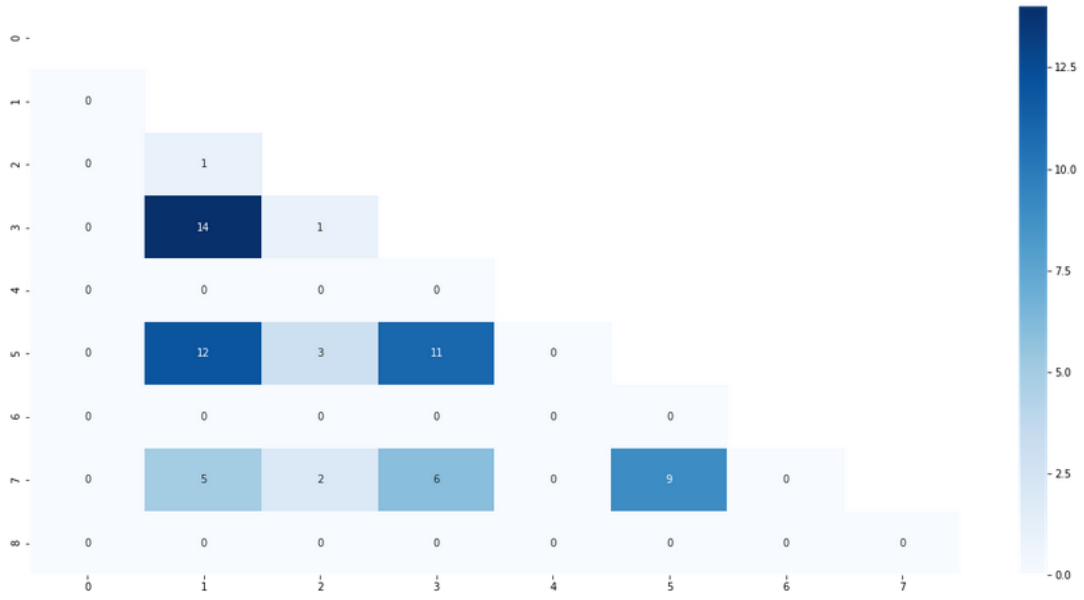
	9	**4**
Resposta 22	A principal limitação das redes de uma camada, como as redes perceptron e adaline, é que elas conseguem classificar apenas objetos que são linearmente separáveis.	A principal limitação das redes de uma camada, como as redes perceptron e adaline, é que elas conseguem classificar apenas objetos que são linearmente separáveis.
Resposta 39	O número de neurônios em uma camada intermediária de uma RNA depende de vários fatores, como: número de exemplos de treinamento; quantidade de ruído presente nos exemplos; complexidade da função a ser aprendida; distribuição estatística dos dados de treinamento.	O número de neurônios em uma camada intermediária de uma RNA depende de vários fatores, como: número de exemplos de treinamento; quantidade de ruído presente nos exemplos; complexidade da função a ser aprendida; distribuição estatística dos dados de treinamento.
Resposta 5	Apesar de os neurônios biológicos possuírem um tempo de execução normalmente da ordem de 10^(-3) segundos, o cérebro é capaz de realizar diversas tarefas (como reconhecimento de padrões, percepção e controle motor) várias vezes mais rapidamente que o mais rápido computador digital existente na atualidade.	Apesar de os neurônios biológicos possuírem um tempo de execução normalmente da ordem de 10^(-3) segundos, o cérebro é capaz de realizar diversas tarefas (como reconhecimento de padrões, percepção e controle motor) várias vezes mais rapidamente que o mais rápido computador digital existente na atualidade.
Resposta 19	O teorema de convergência de uma rede perceptron diz que se é possível classificar um conjunto de entradas linearmente, uma rede perceptron fará a classificação.	O teorema de convergência de uma rede perceptron diz que se é possível classificar um conjunto de entradas linearmente, uma rede perceptron fará a classificação.
Resposta 20	As principais diferenças entre as duas redes é que a rede adaline utiliza uma função de ativação linear e, assim, leva a magnitude do erro em consideração na hora de ajustar os pesos na rede.	As principais diferenças entre as duas redes é que a rede adaline utiliza uma função de ativação linear e, assim, leva a magnitude do erro em consideração na hora de ajustar os pesos na rede.

As respostas foram idênticas por todo o questionário, além do que está mostrado na amostra acima.

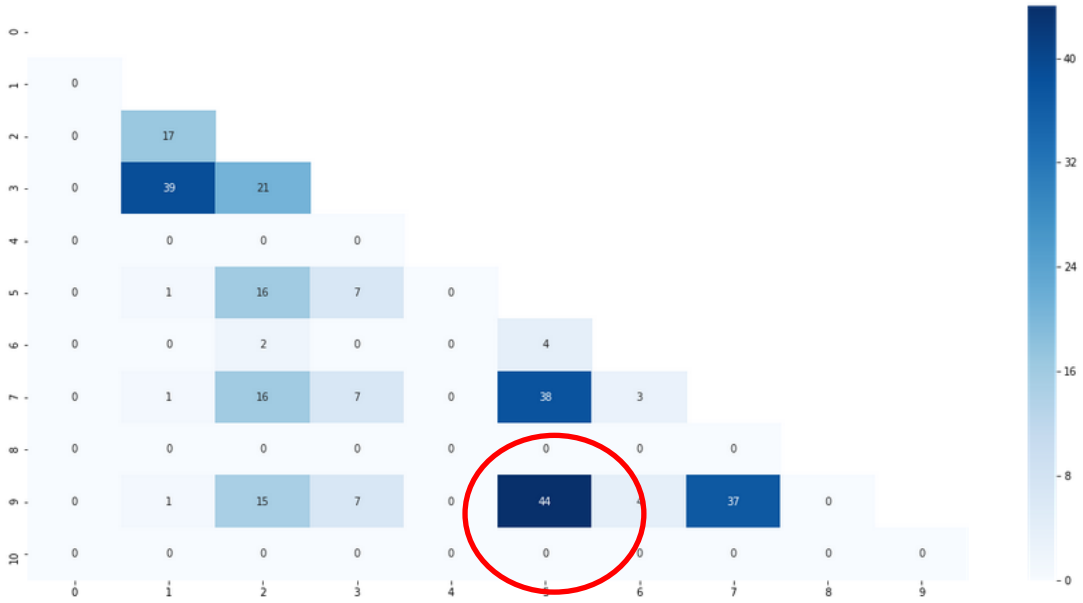
O mesmo ocorre entre os alunos 5 e 1:

	5	1
Resposta 37	Plotando um gráfico com as taxas de erros dos dados de treinamento e validação e examinando. Quando a taxa de erro de validação começar a subir, significa que a rede parou de aprender.	Plotando um grafico com as taxas de erros dos dados de treinamento e validação e examinando. Quando a taxa de erro de validação começar a subir, significa que a rede parou de aprender.
Resposta 17	O valor da taxa de aprendizado define a magnitude do ajuste feito no valor de cada peso. Valores altos fazem com que as variações sejam grandes,enquanto taxas pequenas implicam poucas variações nos pesos.	O valor da taxa de aprendizado define a magnitude do ajuste feito no valor de cada peso. Valores altos fazem com que as variações sejam grandes,enquanto taxas pequenas implicam poucas variações nos pesos.
Resposta 7	A arquitetura de uma RNA está relacionada ao tipo e número de unidades de processamento e à forma como os neurônios estão conectados.	A arquitetura de uma RNA está relacionada ao tipo e numero de unidades de processamento e a forma como os neurônios estão conectados.
Resposta 1	O fundamento natural das redes neurais artificiais é baseado na complexa estrutura biológica dos seres humanos. Especialmente, a redes neurais artificiais foram inspiradas no funcionamento do sistema nervoso, com o objetivo de simular a capacidade de aprendizado do cérebro humano na aquisição de conhecimento.	O fundamento natural das redes neurais artificiais é baseado na complexa estrutura biológica dos seres humanos. Especialmente, a redes neurais artificiais foram inspiradas no funcionamento do sistema nervoso, com o objetivo de simular a capacidade de aprendizado do cérebro humano na aquisição de conhecimento.
Resposta 32	Como os valores dos erros são conhecidos apenas para os neurônios da camada de saída, o erro para os neurônios das camadas intermediárias precisa estimado. O algoritmo back-propagation propõe uma maneira de estimar o erro dos neurônios das camadas intermediárias utilizando os erros observados nos neurônios da camada anterior. O erro de um neurônio de uma dada camada intermediária é estimado como a soma dos erros dos neurônios da camada seguinte, cujos terminais de entrada estão conectados a ele, ponderados pelo valor do peso associado a essas conexões.	Como os valores dos erros são conhecidos apenas para os neurônios da camada de saída, o erro para os neurônios das camadas intermediárias precisa estimado. O algoritmo back-propagation propõe uma maneira de estimar o erro dos neurônios das camadas intermediárias utilizando os erros observados nos neurônios da camada anterior. O erro de um neurônio de uma dada camada intermediária é estimado como a soma dos erros dos neurônios da camada seguinte, cujos terminais de entrada estão conectados a ele, ponderados pelo valor do peso associado a essas conexões.

Para a atividade ED08:



Para a atividade ED09:



Nestes dois últimos heatmaps, pudemos constatar que alguns alunos (representados por colunas) exibiram alta similaridade de respostas com outros alunos (representados por linhas). Neste último heatmap, por exemplo, podemos constatar que o aluno 5 exibiu alta similaridade com os alunos 9 e 7. Inspecionando visualmente os alunos, conseguimos constatar que respostas idênticas, exatas ou com paráfrases, foram enviadas por estes conjuntos de alunos.

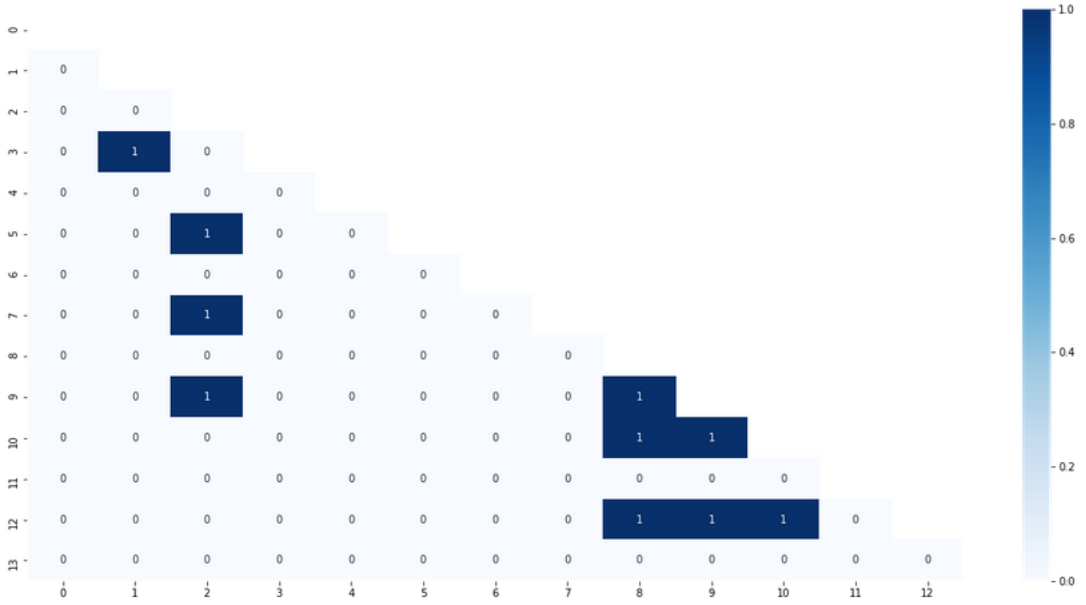
Observe, por exemplo, uma amostra aleatória das respostas do primeiro e segundo aluno, mostrados no último entre os heatmaps acima:

	2	1
Resposta 26	Uma das principais diferenças entre linguagens naturais e linguagens formais como C++ é que uma sentença em uma linguagem natural pode ter mais de um significado. Isso é ambigüidade - o fato de que uma sentença pode ser interpretada de maneiras diferentes, dependendo de quem está falando, do contexto em que ela é falada e de vários outros fatores.	UMA DAS PRINCIPAIS DIFERENÇAS ENTRE LINGUAGENS NATURAIS E LINGUAGENS FORMAIS COMO C++ É QUE UMA SENTENÇA EM UMA LINGUAGEM NATURAL PODE TER MAIS DE UM SIGNIFICADO. ISSO É AMBIGÜIDADE - O FATO DE QUE UMA SENTENÇA PODE SER INTERPRETADA DE MANEIRAS DIFERENTES, DEPENDENDO DE QUEM ESTÁ FALANDO, DO CONTEXTO EM QUE ELA É FALADA E DE VÁRIOS OUTROS FATORES.
Resposta 28	A ambiguidade semântica ocorre quando uma sentença tem mais de um significado possível - geralmente como resultado de uma ambigüidade sintática.	A AMBIGUIDADE SEMÂNTICA OCORRE QUANDO UMA SENTENÇA TEM MAIS DE UM SIGNIFICADO POSSÍVEL - GERALMENTE COMO RESULTADO DE UMA AMBIGÜIDADE SINTÁTICA.
Resposta 38	O conceito principal usado na recuperação de informação é conhecido como TF-IDF, (Frequência de Termo - Frequência de Documento Inversa). Geralmente, um valor TF-IDF é calculado para cada conjunto de palavras e os valores resultantes são colocados em um vetor, que representa um documento ou uma parte do texto (como uma consulta).\n\nA freqüência inversa do documento (IDF) de uma palavra W é calculada da seguinte forma:\n \nIDF(W) = log |D| / DF(W)	O CONCEITO PRINCIPAL USADO NA RECUPERAÇÃO DE INFORMAÇÃO É CONHECIDO COMO TF-IDF, (FREQUÊNCIA DE TERMO - FREQUÊNCIA DE DOCUMENTO INVERSA). GERALMENTE, UM VALOR TF-IDF É CALCULADO PARA CADA CONJUNTO DE PALAVRAS E OS VALORES RESULTANTES SÃO COLOCADOS EM UM VETOR, QUE REPRESENTA UM DOCUMENTO OU UMA PARTE DO TEXTO (COMO UMA CONSULTA).\n\nA FREQÜÊNCIA INVERSA DO DOCUMENTO (IDF) DE UMA PALAVRA W É CALCULADA DA SEGUINTE FORMA:\n\nIDF(W) = LOG |D| / DF(W)
Resposta 41	Em geral, para a maioria das técnicas de recuperação de informação, a precisão e a rechamada estão em oposição umas às outras, o que significa que quando a precisão do sistema aumenta, isso acontece às custas do recall e vice-versa. Isso é intuitivo: a única maneira de obter 100% de recordação na maioria das situações do mundo real é ficar muito relaxado sobre quais documentos são classificados. Em outras palavras, uma grande quantidade de documentos deve ser classificada como relevante para garantir que todos os documentos relevantes sejam encontrados. Inevitavelmente, isso significará que alguns documentos irrelevantes serão encontrados também.	EM GERAL, PARA A MAIORIA DAS TÉCNICAS DE RECUPERAÇÃO DE INFORMAÇÃO, A PRECISÃO E A RECHAMADA ESTÃO EM OPOSIÇÃO UMAS ÀS OUTRAS, O QUE SIGNIFICA QUE QUANDO A PRECISÃO DO SISTEMA AUMENTA, ISSO ACONTECE ÀS CUSTAS DO RECALL E VICE-VERSA. ISSO É INTUITIVO: A ÚNICA MANEIRA DE OBTER 100% DE RECORDAÇÃO NA MAIORIA DAS SITUAÇÕES DO MUNDO REAL É FICAR MUITO RELAXADO SOBRE QUAIS DOCUMENTOS SÃO CLASSIFICADOS. EM OUTRAS PALAVRAS, UMA GRANDE QUANTIDADE DE DOCUMENTOS DEVE SER CLASSIFICADA COMO RELEVANTE PARA GARANTIR QUE TODOS OS DOCUMENTOS RELEVANTES SEJAM ENCONTRADOS. INEVITAVELMENTE, ISSO SIGNIFICARÁ QUE ALGUNS DOCUMENTOS IRRELEVANTES SERÃO ENCONTRADOS TAMBÉM.

No mesmo heatmap, também pudemos observar os alunos 6 e 5, que conforme foi indicado pelo heatmap, pouco possuem em comum:

	6	5
Resposta 1	Linguagens Naturais são as linguagens utilizada por nós humanos para nos comunicarmos. Sua principal diferença com as Linguagens Formais é que a linguagem natural é ambígua, o que significa que uma dada sentença pode ter mais de um significado possível, e em alguns casos o significado correto pode ser muito difícil de determinar. As linguagens formais são quase sempre projetadas para garantir que essa ambiguidade não ocorra, assim o computador não precisa tomar uma decisão arbitrária sobre qual interpretação trabalhar.	As línguas naturais são as línguas usadas pelos humanos para comunicação (entre outras funções). Eles são distintamente diferentes das linguagens formais, como C ++, Java e PROLOG. Uma das principais diferenças, é que as línguas naturais são ambíguas, o que significa que uma determinada sentença pode ter mais de um significado possível e, em alguns casos, o significado correto pode ser muito difícil de determinar. As linguagens formais são quase sempre projetadas para garantir que a ambigüidade não possa ocorrer. Portanto, um determinado programa escrito em C ++ pode ter apenas uma interpretação. Isto é claramente desejável porque, caso contrário, o computador teria que tomar uma decisão arbitrária sobre qual interpretação trabalhar.
Resposta 12	BNF é utilizado como uma notação para as gramáticas de linguagens de programação, e também como notação para representar partes de gramáticas de linguagens naturais. Um BNF é um conjunto de regras de derivação, escritas como:\n \n<símbolo> ::= <expressão com símbolos>\n \nOnde <símbolo> é um não terminal, e a expressão consiste em sequências de símbolos e/ou sequências separadas pela barra vertical, '|', indicando uma escolha. Esta notação indica as possibilidades de substituição para símbolo da esquerda. Símbolos que nunca aparecem no lado esquerdo são ditos terminais.	São regras gramaticais, frequentemente escritas no que é conhecido como forma Backus – Naur (também conhecida como forma normal de Backus - ambos os nomes são abreviados como BNF). O BNF é amplamente utilizado por cientistas da computação para definir linguagens formais, como C ++ e Java. Também podemos usá-lo para definir a gramática de uma linguagem natural.
Resposta 6	Esta etapa envolve a aplicação das regras da gramática da linguagem que está sendo usada. A sintaxe determina o papel de cada palavra em uma sentença e, assim, permite que um sistema de computador converta sentenças em uma estrutura que possa ser mais facilmente manipulada.	Esta etapa envolve a aplicação das regras da gramática da linguagem que está sendo usada. A sintaxe determina o papel de cada palavra em uma sentença e, assim, permite que um sistema de computador converta sentenças em uma estrutura que possa ser mais facilmente manipulada.

Abaixo também representamos um mapa que exibe, para uma pergunta específica, se dois alunos exibiram mais de 50% de similaridade entre si em sua resposta a aquela pergunta. O número 1 indica que ocorreu mais de 50% de similaridade; o número 0 indica que ela não ocorreu.



	Unnamed: 0	,",Resposta 1"
0	0	0,"0,"
1	1	1,"1,""AS LÍNGUAS NATURAIS SÃO AS LÍNGUAS USADAS PELOS HUMANOS PARA COMUNICAÇÃO (ENTRE OUTRAS FUNÇÕES). ELES SÃO DISTINTAMENTE DIFERENTES DAS LINGUAGENS FORMAIS, COMO C ++, JAVA E PROLOG. UMA DAS PRINCIPAIS DIFERENÇAS, É QUE AS LÍNGUAS NATURAIS SÃO AMBÍGUAS, O QUE SIGNIFICA QUE UMA DETERMINADA SENTENÇA PODE TER MAIS DE UM SIGNIFICADO POSSÍVEL E, EM ALGUNS CASOS, O SIGNIFICADO CORRETO PODE SER MUITO DIFÍCIL DE DETERMINAR. AS LINGUAGENS FORMAIS SÃO QUASE SEMPRE PROJETADAS PARA GARANTIR QUE A AMBIGÜIDADE NÃO POSSA OCORRER. PORTANTO, UM DETERMINADO PROGRAMA ESCRITO EM C ++ PODE TER APENAS UMA INTERPRETAÇÃO. ISTO É CLARAMENTE DESEJÁVEL PORQUE, CASO CONTRÁRIO, O COMPUTADOR TERIA QUE TOMAR UMA DECISÃO ARBITRÁRIA SOBRE QUAL INTERPRETAÇÃO TRABALHAR."""
2	2	2,"2,""As linguagens naturais são as linguagens utilizadas pelos seres humanos para comunicação, a principal diferença das linguagens naturais e linguagens formais é a ambiguidade. As linguagens naturais são ambíguas de forma que uma sentença pode ter mais de um significado possível e nas linguagens formais temos que as mesmas são projetadas para evitar ao máximo a ambiguidade."""
3	3	3,"3,""AS LÍNGUAS NATURAIS SÃO AS LÍNGUAS USADAS PELOS HUMANOS PARA COMUNICAÇÃO (ENTRE OUTRAS FUNÇÕES). ELES SÃO DISTINTAMENTE DIFERENTES DAS LINGUAGENS FORMAIS, COMO C ++, JAVA E PROLOG. UMA DAS PRINCIPAIS DIFERENÇAS, É QUE AS LÍNGUAS NATURAIS SÃO AMBÍGUAS, O QUE SIGNIFICA QUE UMA DETERMINADA SENTENÇA PODE TER MAIS DE UM SIGNIFICADO POSSÍVEL E, EM ALGUNS CASOS, O SIGNIFICADO CORRETO PODE SER MUITO DIFÍCIL DE DETERMINAR. AS LINGUAGENS FORMAIS SÃO QUASE SEMPRE PROJETADAS PARA GARANTIR QUE A AMBIGUIDADE NÃO POSSA OCORRER. PORTANTO, UM DETERMINADO PROGRAMA ESCRITO EM C ++ PODE TER APENAS UMA INTERPRETAÇÃO. ISTO É CLARAMENTE DESEJÁVEL PORQUE, CASO CONTRÁRIO, O COMPUTADOR TERIA QUE TOMAR UMA DECISÃO ARBITRÁRIA SOBRE QUAL INTERPRETAÇÃO TRABALHAR."""
4	4	4,"4,""Linguagens naturais são os idiomas padrões utilizados pelos seres humanos para"
5	5	5,comunicação. Elas se diferenciam das linguagens formais (linguagens de programação)
6	6	6,"primeiramente pois são ambíguas, i.e., uma frase pode ter diversos significados"
7	7	7,"possíveis, enquanto as linguagens formais não podem conter ambiguidade."""
8	8	8,"5,""As línguas naturais são as línguas usadas pelos humanos para comunicação (entre outras funções). Eles são distintamente diferentes das linguagens formais, como C ++, Java e PROLOG. Uma das principais diferenças, é que as línguas naturais são ambíguas, o que significa que uma determinada sentença pode ter mais de um significado possível e, em alguns casos, o significado correto pode ser muito difícil de determinar. As linguagens formais são quase sempre projetadas para garantir que a ambigüidade não possa ocorrer. Portanto, um determinado programa escrito em C ++ pode ter apenas uma interpretação. Isto é claramente desejável porque, caso contrário, o computador teria que tomar uma decisão arbitrária sobre qual interpretação trabalhar."""
9	9	9,"6,""Linguagens Naturais são as linguagens utilizada por nós humanos para nos comunicarmos. Sua principal diferença com as Linguagens Formais é que a linguagem natural é ambígua, o que significa que uma dada sentença pode ter mais de um significado possível, e em alguns casos o significado correto pode ser muito difícil de determinar. As linguagens formais são quase sempre projetadas para garantir que essa ambiguidade não ocorra, assim o computador não precisa tomar uma decisão arbitrária sobre qual interpretação trabalhar."""
10	10	10,"7,""As línguas naturais são as línguas usadas pelos humanos para comunicação. Uma das principais diferenças, é que as línguas naturais são ambíguas, o que significa que uma determinada sentença pode ter mais de um significado possível e, em alguns casos, o significado correto pode ser muito difícil de determinar. As linguagens formais são quase sempre projetadas para garantir que a ambiguidade não possa ocorrer."""
11	11	11,"8,"
12	12	12,"9,""As línguas naturais são as línguas usadas pelos humanos para comunicação (entre outras funções). Eles são distintamente diferentes das linguagens formais, como C ++, Java e PROLOG. Uma das principais diferenças, é que as línguas naturais são ambíguas, o que significa que uma determinada sentença pode ter mais de um significado possível e, em alguns casos, o significado correto pode ser muito difícil de determinar. As linguagens formais são quase sempre projetadas para garantir que a ambigüidade não possa ocorrer. Portanto, um determinado programa escrito em C ++ pode ter apenas uma interpretação. Isto é claramente desejável porque, caso contrário, o computador teria que tomar uma decisão arbitrária sobre qual interpretação trabalhar."""
13	13	13,"10,""São as linguagens usadas para comunicação entre seres humanos. A diferença é que suas regras mudam naturalmente com o tempo, e suas sentenças possuem ambiguidade."""
14	14	14,"11,""São linguagens usadas por humanos para comunicação, elas se diferenciam das formais como c, java etc."""

5 Como a solução proposta pôde resolver o problema

O nosso algoritmo foi capaz de constatar tanto alunos que enviaram respostas idênticas quanto alunos que enviaram respostas que eram paráfrases umas das outras, bem como detectar alunos que enviaram respostas bastante distintas entre si.

Sendo assim, constatamos que o algoritmo TF-IDF pode ser uma medida eficiente para determinar a similaridade entre resumos escritos a partir de um livro em comum quando não existe um adversário ao modelo, ou seja, quando os escritores dos resumos não estão trabalhando deliberadamente para ofuscar o fato de que realizaram uma cópia ou uma paráfrase aproximada.

A introdução de um adversário exigiria modelos mais complexos, como por exemplo o uso de *word embeddings* em arquiteturas de aprendizado profundo, e/ou o uso de uma WordNet. Porém, na ausência de um adversário inteligente, o nosso modelo alcançou um desempenho bastante satisfatório, conforme demonstram os resultados obtidos neste *corpus*.

6 Conclusões

Concluímos que modelos estatísticos simples, como o TF-IDF, podem atingir resultados satisfatórios em problemas básicos de processamento de linguagem natural, especialmente na detecção de paráfrases e cópias. Isso é relevante, pois, dado o baixo uso de recursos computacionais deste método, é possível utilizá-lo em larga escala em bancos de dados e aplicações especializadas em busca (tais como, por exemplo, *Apache Solr* ou *Elasticsearch*).

Porém, existem casos onde tais modelos podem falhar. Existem textos que são bastante semelhantes pois falam sobre o mesmo assunto, porém possuem poucas palavras em comum. Nestes casos, o modelo TF-IDF se torna inadequado para medir a similaridade entre eles, e começa a ser necessário utilizar modelos que levam em conta também as redes semânticas que existem entre palavras e seus sinônimos e termos relacionados.

Referências

[1] **IF-IDF**. https://en.wikipedia.org/wiki/Tfidf,.

[2] **Natural language processing**. https://en.wikipedia.org/wiki/Natural_language_processing,.

[3] Rodrigues, J. **O que é o Processamento de Linguagem Natural?** https://medium.com/botsbrasil/o-que-é-o-processamento-de-linguagem-natural-49ece9371cff,.

[4] STEFANINI. **PNL**. https://stefanini.com/pt-br/trends/artigos/o-que-e-processamento-de-linguagem-natural, último acesso em Agosto de 2019.