DESCOBERTA DE CONHECIMENTO: MINERAÇÃO DE DADOS E AS OLIMPÍADAS RIO 2016

Eric Chagas Farias <efarias94@gmail.com>
Marco Aurélio Schunke <marcoschunke@gmail.com> – Orientador

Universidade Luterana do Brasil (Ulbra) – Curso de Ciência da Computação – Campus Guaíba
BR 116, 5.724 – Bairro Moradas da Colina – CEP 92500-000 – Guaíba - RS

27 de junho de 2016

RESUMO

Este projeto pretende analisar o cenário atual das notícias relacionadas aos jogos olímpicos do Rio de Janeiro 2016, utilizando de descoberta de conhecimento em base de dados, busca-se, com o aumento das notícias no âmbito dos esportes olímpicos, relacionar os mais diversos temas adjacentes as olímpiadas com os jogos em sua totalidade, encontrando padrões e relações a fim de se compreender a visão dos produtores de conteúdo quanto aos jogos olímpicos.

A motivação desta análise é para poder compreender qual o cenário atual das notícias relacionadas aos jogos olímpicos que tenham conteúdos que abordam estas questões relacionadas ao desempenho do atleta, qual a quantidade de acessos a estes conteúdos em relação aos demais inseridos na página, qual é o conteúdo destas publicações a respeito de cada um dos temas relacionados e quem é o produtor deste conteúdo, qual o sentimento em relação ao conteúdo e o interesse na inter-relação destes temas.

Para se obter os resultados propostos, serão utilizadas técnicas de mineração de dados, onde serão encontrados os padrões e contabilizadas as comparações; tratamento de dados para ser possível manipular as publicações a fim de se obter dados contabilizáveis e processamento de linguagem natural para ser possível a implementação destas análises.

Palavras-Chave: Descoberta de Conhecimento, Mineração de Dados, Jogos Olímpicos.

ABSTRACT

Title: *"knowledge discovery: the Olympics and its relation to the health of athletes"*

This project aims to analyze the current scenario of news related to the Olympic Games in Rio de Janeiro 2016 using knowledge discovery in databases, is sought, with increasing news within the Olympic sports, relate the various adjacent themes the Olympics with the games in their entirety, finding patterns and relationships in order to understand the vision of content producers as the Olympics.

The motivation of this analysis is to be able to understand what the current scenario of news related to the Olympic games that have content that address these issues related to the athlete's performance, how much access to this content in relation to the other inserted into the page, which is the content of these publications about each of the related issues and who is the producer of this content, which feeling about the content and interest in the interrelation of these issues.

To achieve the proposed results, data mining techniques will be used, where standards and accounted comparisons will be found; processing data to be able to manipulate the publications order to obtain countable data and natural language processing is possible for the implementation of these analyzes.

Key-words: *Knowledge Discovery, Data Mining, Olympic Games.*

1 INTRODUÇÃO

Com as Olimpíadas de 2016 sendo realizadas no Brasil, na cidade do Rio de Janeiro, é esperado um aumento no número de informações decorrentes da divulgação deste evento através dos veículos midiáticos de todos os gêneros, como jornal, televisão e principalmente em sites de notícias e redes sociais.

Com base nestas informações, o presente trabalho pretende explorar as publicações disponibilizadas na WEB para compreender o cenário atual das notícias que abordem o tema jogos olímpicos a fim de se descobrir conhecimento bem como sua relação com outros temas relacionados a saúde dos atletas.

Para atender a este propósito é identificada a página oficial dos jogos olímpicos no Facebook, "The Olympic Games", que é uma das fanpages com maior número de seguidores em todo o mundo e também uma das principais relacionadas a esportes olímpicos, sendo pertinente para se compreender o cenário atual das publicações relacionadas a este tema.

Explorando as técnicas aprendidas no decorrer do curso, o seguinte trabalho irá fundamentar todos os tópicos necessários para a elaboração da segunda etapa do Trabalho de Conclusão de Curso, onde serão empregadas todas as técnicas do processo de Descoberta de Conhecimento em Base de Dados, a seguir descritas.

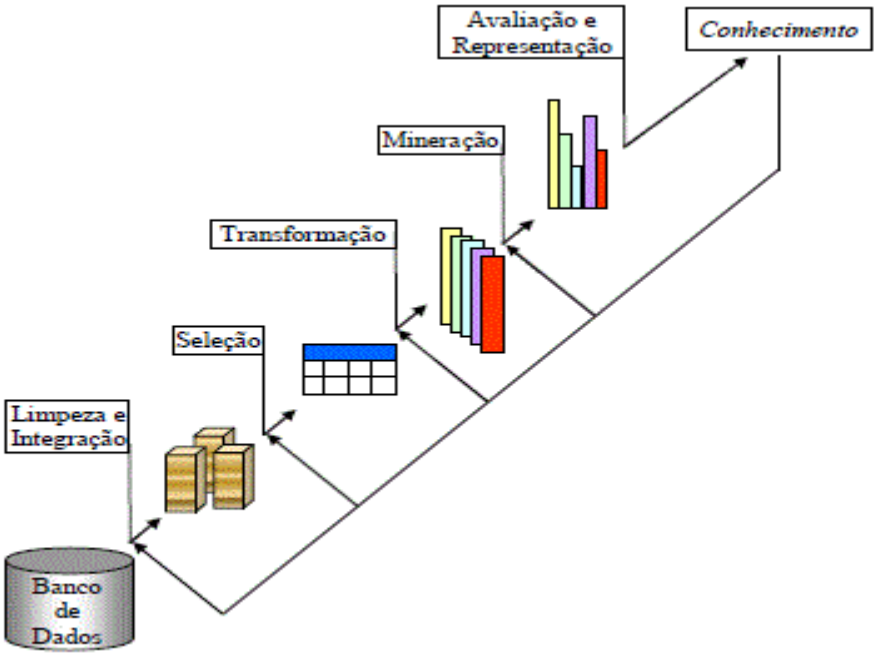
2 FUNDAMENTAÇÃO TEÓRICA

2.1 Descoberta de Conhecimento em Base de Dados

Segundo Cardoso e Machado (2008), o termo descoberta de conhecimento em base de dados (DCBD ou KDD do inglês *knowledge Discovery in databases*) consiste na técnica aplicada para se obter um resultado ao explorar um conjunto de dados; o processo de extração de conhecimento se dá nas relações e padrões encontrados nestes itens.

Como define Cardoso e Machado (2008), DCBD em um processo eficiente e não trivial quando se identifica estes padrões em conjuntos com as seguintes características: validade – os padrões encontrados devem possuir um certo grau de certeza ou probabilidade; novidade – os padrões não devem ter sidos encontrados anteriormente; utilidade potencial – os padrões devem ter uma utilidade ou possuir alguma função; Assimiláveis – tornar os padrões interpretáveis ao conhecimento humano.

Figura 1 – Fluxo do processo de KDD



Fonte: TRONCHONI, 2010, p. 3.

Como apresentado na figura 1, segundo Tronchoni, Pretto, Rosa e Lemos (2010), cada passo desta sequência consiste nos seguintes:

Limpeza dos dados: remoção de dados irrelevantes;

Integração dos dados: na combinação de múltiplas fontes de dados;

Seleção dos dados: recuperação dos dados relevantes a análise;

Transformação dos dados: tornar os dados ao formato apropriado para a mineração;

Mineração de dados: ocorre a extração de padrões através de métodos inteligentes;

Avaliação e representação do conhecimento: são utilizadas técnicas e interpretação a fim de se obter o conhecimento.

2.2 Mineração de Dados

Segundo Cardoso e Machado (2008), mineração de dados ou *data mining,* é uma das técnicas utilizadas na descoberta de conhecimento em banco de dados, sendo responsável por revelar conhecimento que está implícito em grandes quantidades de informações armazenadas nos bancos de dados. Sendo capaz de prever eventos, tendências e comportamentos futuros com base nas análises encontradas nas bases de informações, possibilitando, por exemplo, a tomada de decisão.

"Data Mining (DM) – é uma das alternativas mais eficazes para extrair conhecimento a partir de grandes volumes de dados, descobrindo relações ocultas, padrões e gerando regras para predizer e correlacionar dados, que podem ajudar as instituições nas tomadas de decisões mais rápidas ou, até mesmo, a atingir um maior grau de confiança. " (GALVÃO & MARIN, 2009).

Dados podem ser obtidos através de *web crawler*, ou rastreador online específico para o site de rede social Facebook, que segundo Rouse (2016), um crawler é um software que faz uma varredura e armazena as

informações, contidas em uma página *web,* para anexar a um catálogo de pesquisas. Os *crawlers* são programados para visitar determinados sites e atualizar as informações antes já catalogadas ou então adicionar novas páginas descobertas; sites completos ou determinados conteúdos podem ser obtidos através desta ferramenta.

2.3 Tratamento de Dados

Segundo Castro, Mattos e Simões (2009), o processo de descoberta de conhecimento se inicia nesta etapa onde se preparam os dados, como a seleção destes dados podem representar significativamente o conteúdo abordado, visando também a redução dos dados, descartando o conteúdo irrelevante.

São utilizadas atividades como recuperação da informação, mostrando modelos de coleções textuais; análise dos dados, identificando os termos do texto encontrando padrões; e a transformação dos dados, que tornam os dados utilizáveis, todas técnicas utilizadas para tornar a preparação dos dados eficiente.

Para tratamento de dados, a ferramenta Eurekha, que segundo o seu desenvolvedor, Wives (2011), é um software que faz agrupamento de documentos (em inglês *clustering*), que os agrupa com base em assuntos semelhantes e separa os diferentes; estes identificados por palavras ou conceitos presentes nos textos.

2.4 Processamento de Linguagem Natural

Processamento de linguagem natural (PLN) é a capacidade de processar objetos de natureza linguística através de sistemas computacionais, sendo uma subárea de Inteligência Artificial, que pode ser definido como a habilidade de processar dados linguísticos que possam ser utilizados por humanos; o que difere a linguística de *Corpus (LC)* como o estudo da língua em uso de grande escala com o apoio informatizado e tratamento estatístico, com o objetivo de relevar respostas com os diferentes usos de grandes conjuntos de dados de natureza linguística, na identificação de padrões de uso e combinatórias destas amostras (FINATTO, LOPES & CIRULLA, 2015).

Natural Language Toolkit (NLTK) é uma plataforma construída na linguagem de programação Python para trabalhar com o processamento de linguagem natural. Composto de um conjunto de bibliotecas de processamento de texto, é capaz de classificar, analisar, catalogar e organizar semanticamente os dados contidos nos *Corpora*.

Para se obter os dados estatísticos, podem ser utilizadas técnicas de análise das ocorrências de palavras. As funções a seguir descritas, fazem parte do NLTK e podem ser exploradas para se alcançar dados probabilísticos da ocorrência das palavras contidas nos *Corpora.*

2.4.1 N-grams

Segundo Hanai (2014), este modelo de linguagem busca capturar as características na sequência de palavras de uma linguagem e combinar sobre todas as sequencias possíveis de palavras em um vocabulário. Estas combinações podem eventualmente ser computadas e que podem ser unigram, birgram ou trigram.

2.4.2 Unigram

É a relativa ocorrência de uma palavra em um *Corpus*, HANAI (2014). Por exemplo, em um *corpus* onde é encontrada a sequência " Processamento de Linguagem Natural", Unigram evidencia "Processamento", "de", "Linguagem" e "Natural" como termos distintos e isolados.

2.4.3 Bigram

É a condição de ocorrência do primeiro elemento seguido do seu subsequente, HANAI (2014). Por exemplo, em um *corpus* onde é encontrada a sequência " Processamento de Linguagem Natural", Bigram evidencia "Processamento de", "de Linguagem" e "Linguagem Natural" como termos distintos e isolados.

2.4.4 Trigram

O modelo trigram é similar aos citados anteriormente, porém são considerados 3 elementos na sequência, HANAI (2014). Por exemplo, em um *corpus* onde é encontrada a sequência " Processamento de Linguagem Natural", trigram evidencia "Processamento de Linguagem" e "de Linguagem Natural" como termos distintos e isolados.

3 TRABALHOS RELACIONADOS

3.1 Castro et al. (2007)

Seu trabalho tem como objetivo apresentar a utilização de ferramentas de *text mining* voltadas a área da saúde, mostrando as etapas relacionadas com a descoberta de conhecimento em base de dados e processamento de dados textuais.

Utilizando a ferramenta Eurekha, foram analisados os relacionamentos entre os documentos levantados e também para o tratamento dos dados, efetuando um filtro nas documentações a fim de eliminar dados irrelevantes. A base de dados refere-se a área farmacêutica. Os formatos dos arquivos utilizados encontram-se em .txt.

Na etapa de pré-processamento, foi utilizado o subprograma Cleanner, onde foram removidas *stopwords*, acentos e números também tiveram sua fonte alterada para minúscula ou maiúscula. Na etapa de *text mining os* documentos foram armazenados em *clusters* e posteriormente foram analisados através de algoritmos próprios da ferramenta. Referente ao pós-processamento, os resultados foram mostrados em formato de gráfico e também em matriz de similaridade.

Foi realizado um estudo de caso com todos os resultados coletados, após as etapas de tratamento de dados e mineração de dados, contribuindo para o processo de descoberta de conhecimento em base de dados.

3.2 Cardoso et al. (2008)

O presente trabalho teve como objetivo desenvolver, aplicar e analisar uma ferramenta de mineração de dados para se descobrir conhecimento em base de dados, utilizando como referência a produção científica de pessoas envolvidas com pesquisa da Universidade Federal de Lavras.

Foi implementado um programa para transformar os dados contidos na plataforma Lattes, armazenando em um banco de dados Oracle. Sendo possível desenvolver uma ferramenta para descobrir o conhecimento, aplicando técnica de data mining.

Ao utilizar o Lattes Extrator, obtiveram os materiais necessários para fazer as análises, comparações

e desenvolvimento da ferramenta de mineração de dados. Utilizando o método de estudo de caso para a realização do trabalho, a etapa de seleção de dados contou com mais de mil documentos, dentre eles 575 currículos. Na etapa de pré-processamento de dados, foram eliminados os dados incongruentes, filtrando as informações incompletas e duplicadas.

Na transformação de dados, foram convertidos os dados obtidos do Lattes Extrator, em XML para documentos SQL, gerando um código de inserção que depois foi adicionado ao sistema gerenciador de banco de dados da Oracle. Na etapa de mineração de dados, foram utilizadas técnicas específicas para comparar as informações, utilizando funções do próprio gerenciador do banco de dados, a linguagem de programação PL/SQL.

A interpretação dos resultados foi feita por meio da criação de relatórios, desenvolvendo uma dissertação de mestrado apresentada ao Departamento de Administração e Economia de Universidade Federal de Lavras, onde se encontra todo o referencial teórico e os resultados apresentados de diversas formas como por exemplo gráficos resumidos e a descrição dos principais resultados.

O resultado ocorreu nas seguintes etapas: Analise de regras de associação, um exemplo onde mostrou a associação entre a quantidade de publicações realizadas por pessoas que trabalham na Universidade e as que não trabalham. Envolvendo 11 tabelas do banco de dados, e obtidas 1977 publicações, onde 55% foram publicadas por pessoas que não estavam atuando na Universidade na época da publicação.

Outro resultado foi referente as análises de regras de associação e outliers, onde buscou verificar a relação existente entre as atividades de pesquisa e o número de linhas de pesquisa envolvidas, observou-se três pessoas com um número muito superior de linhas de pesquisa em comparação com a sua atividade de pesquisa, onde obteve-se 84 pesquisas e 186 linhas de pesquisa.

Referente a análise de classificação e predição, analisou os grupos já existentes, comparando os grupos de pesquisa, ensino e direção. Em 101 publicações, resultou que a maioria das publicações foi realizada enquanto as pessoas exerciam atividades de pesquisa, em segundo de ensino e por último de direção.

3.3 Galvão et al. (2009)

O trabalho propôs uma revisão da literatura encontrada sobre a técnica de mineração de dados, em bases de dados como o Scientific Eletronic Library Online (SCIELO), o Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (LILACS) e também alguns livros sobre o tema.

Foram levantadas as informações, buscando pelos termos mineração de dados e também *data mining*, em inglês, dentre o período de 1999 a 2008. Para o estudo foram utilizadas 10 publicações do LILACS, 28 do SCIELO, excluídos 6 pois estavam duplicados.

Dividido o estudo em três tópicos, descoberta de conhecimento, tarefas do *data mining* e métodos do *data mining*: ao analisar descoberta de conhecimento em base de dados, o autor compreendeu que o processo de mineração de dados é uma das etapas da descoberta de conhecimento e não um sinônimo, se mostrando útil em diversas áreas do conhecimento. Utiliza-se de métodos estatísticos, e de inteligência artificial passando pelas etapas de seleção, pré-processamento, transformação, mineração e interpretação, sendo a mais importante a mineração dos dados.

Devendo ser eficiente, a descoberta de conhecimento deve ser genérica, podendo ser aplicável a vários tipos de dados e flexível (modificável). Referente as tarefas de *data mining,* concluiu-se que nesta etapa são definidos os algoritmos e tarefas a serem utilizadas, com dois tipos de tarefas principais, são elas: classificação e regressão: classificação consiste em encontrar características no conjunto estudado, a fim de classificar em qual classe os dados se enquadram, são eles os exemplos de Redes Neurais, Classificadores Bayesianos e Algoritmos Genéricos.

Regressão consiste em atributo numérico presente no conjunto estudado com valores reais, utilizam os métodos da estatística e Redes Neurais. Métodos de *data mining* transformam dados em conhecimento

útil, utilizando vários métodos existentes, identificando os mais utilizados, exemplos de técnicas são Árvore de Decisão, Lógica Nebulosa e estatística.

A Árvore de Decisão é um modelo que representa graficamente nós e galhos, visando descobrir a estrutura ou o modelo de um problema, associando as variáveis com seguintes atributos, chamados filhos. A lógica Nebulosa é um modelo que representa o raciocínio humano, aplicados a um ambiente de incertezas e imprecisão, podendo ser utilizados em sistemas de apoio a decisão. A lógica é representada formalizando conceitos imprecisos e também ao processar informações imprecisas, por exemplo na tomada de decisão. A estatística funciona com a interpretação dos dados, com a probabilidade da ocorrência de um evento.

3.4 Finatto et al. (2015)

O objetivo do trabalho foi demonstrar que a integração do Processamento de Linguagem Natural aliado a linguística de *Corpus* é um processo possível e que pode ser proveitoso para os envolvidos, ou seja tentar mostrar que dois campos de estudos distintos, das Ciências Exatas e Humanas, consigam contribuir para este propósito.

Os *corpora* utilizados no processo são um de medicina e o outro sobre linguística, esta escolha foi pensada visando a distinção entre o tratamento de textos de áreas relacionadas a medicina e de ciências humanas, no caso a linguística, a fim de detectar as diferenças e representá-las em forma de ontologias. Com o objetivo de subsidiar programas computacionais, visando um melhor desempenho com a linguagem científica escrita.

A etapa de tratamento de dados ocorreu para a limpeza do *corpus* onde foi convertido do formato .PDF para .TXT, ocasionando erros de grafia e também na remoção de caracteres irrelevantes, como travessões e numeração das páginas. Foi utilizada a etiquetagem de palavras para se identificar cada uma das partes do discurso, utilizando a ferramenta PALAVRAS, foram classificados os elementos do texto para que se possa processar os sintagmas nominais, que são os mais relevantes de um *corpus*.

O processamento de linguagem natural ocorreu com a ferramenta ExATOlp, onde a identificação dos sintagmas nominais ocorre de um processo com base linguística e estatística, por meio de heurísticas e cálculos do índice de relevância.

Os resultados foram mostrados por meio da hierarquia de conceitos, representados por árvores hiperbólicas, onde trazem a partir do seu centro as expressões mais relevantes do *corpus*. Foram também analisados com as ferramentas bigramas e trigramas, mostrando que a computação e o processamento de linguagem natural trabalham com dados estatísticos e organização de informação.

A linguística pode descrever as regras e padrões para que se processe os dados encontrados dentro de um *corpus*. Ao apresentar os resultados encontrados pela ferramenta de processamento de linguagem natural, concluiu-se que a cooperação foi bem-sucedida e que apesar de mostrar as informações sem a necessidade de intervenção humana, sendo geradas de maneira automática, ainda carecem de uma profunda análise que somente os especialistas sobre o assunto estão capacitados a executar.

3.5 Fonseca et al. (2016)

O trabalho teve como intuito fazer uma análise do ensino fundamental público do Brasil, utilizando como base de dados as provas aplicadas pelo Sistema de Avaliação da Educação Básica (Saeb).

Foi utilizado o processo de Descoberta de Conhecimento em Base de Dados para identificar o perfil dos professores de matemática e sua relação com a proficiência dos seus alunos.

Com o objetivo de identificar o professor que possui a classe “Maior que 65%”, onde identifica uma influência positiva relacionada ao processo de ensino-aprendizagem, foi utilizado o algoritmo Naive Bayes,

que apresenta a ocorrência da probabilidade para cada atributo, com seu respectivo valor.

Ao serem analisados os atributos, observou-se que alguns temas de conteúdo originalmente previsto, como a frequência dos alunos e a regularidade dos professores, favorecem para um desempenho positivo. Outros estudos foram feitos com base nestes dados, como quando não há carência de professor para uma disciplina com altos índices de faltas, o desempenho dos estudantes era superior.

Em questões voltadas ao futuro dos alunos, respostas como "quase todos" ou " próximo da metade", mostram que a expectativa do professor em relação ao futuro dos alunos pode afetar positivamente o desempenho dos alunos, entrando na classe" Maior que 65%". Questões relacionadas ao salário dos professores, apontam influência negativa no desempenho do aluno, onde a somas dos percentuais mais baixos, indicam mais de 50% das situações.

Trabalhos sobre o tema mostraram que o baixo salário dificulta a permanência dos professores e também a atração de novos para os cargos. Quanto a questões sobre o conhecimento dos resultados do Saeb, indicam um favorecimento para o bom desempenho do aluno, mas também não indica uma relevante influência quanto ao desempenho ruim dos alunos, mostrando que há uma relação direta com os bons resultados obtidos pelos professores que conhecem os resultados, não indicando que ao desconhecer os resultados implicaria para um baixo índice, mas que se consegue bons resultados ao conhecer sobre os resultados.

Com uma grande base de dados e ao depender da interpretação humana para as análises dos estudos, o processo de exploração não é feito com o seu máximo potencial. Ao fazer o uso efetivo dos dados por meios de ferramentas para extração de informações, podem ser definidas ações para melhoria dos resultados no processo de ensino, com a utilização de mineração de dados.

As análises destas informações contribuem para o processo de tomada de decisão e mostrar perspectivas sobre os alunos e professores, por exemplo o levantamento dos os fatores positivos para o desempenho do aluno como: Conteúdo previsto desenvolvido, assiduidade dos professores, baixo índice de falta dos professores, expectativa do desenvolvimento educacional dos professores perante os alunos e a avaliação do ensino básico. Estudos também podem ser feitos com base nos resultados que influenciam negativamente o desempenho do aluno, como: desvalorização salarial dos professores, grande índice abstêmio dos alunos e a baixa crença de que os alunos entrarão para a faculdade.

4 PROPOSTA

4.1 Objetivos

O presente trabalho tem como objetivo revelar conhecimento sobre os jogos olímpicos de 2016. Ao explorar todas as etapas do processo de descoberta de conhecimento em base de dados, como mineração de dados, utilizando o processamento de linguagem natural, pretende-se analisar o cenário atual das publicações referente as olímpiadas.

Utilizando as publicações da página oficial dos jogos olímpicos no Facebook, The Olimpic Games, será construído um *corpus* onde as técnicas do processo de descoberta de conhecimento serão aplicadas a fim de se encontrar algum padrão relevante ou não a fim de encontrar uma relação com temas relacionados a saúde dos atletas.

4.2 Tecnologias Envolvidas

Para o desenvolvimento do trabalho, serão utilizadas as seguintes tecnologias: Web Crawler para obtenção dos dados a serem analisados. Neste processo de formação do *corpus* será utilizado o Web Crawler disponibilizado pelo Facebook que possui uma API onde é possível fazer a utilização do Facebook Query Language (FQL) para extração dos conteúdos necessários das páginas da rede social.

Neste *corpus*, serão obtidos os dados da página The Olympic Games, que é a página oficial dos jogos olímpicos no Facebook, que conta com mais 10 milhões de curtidas e que conta com publicações diárias. Os dados para análise serão coletados do ano de 2010 até 2016, traçando um panorama geral das publicações pois será levado em consideração as olimpíadas de 2012, que ocorreram em Londres, sendo possível relacionar as publicações ocorridas antes do evento, durante o evento e após o evento. Foi utilizado o servidor web Apache e a linguagem PHP para conseguir a mensagem publicada, a data e o horário, a quantidade de likes e os comentários da postagem.

Para o tratamento de dados, será utilizada a ferramenta Eurekha, desenvolvida por Wives (2011), onde será feita a remoção de termos irrelevantes ou inconsistentes, para que seja possível a confiabilidade das informações geradas, por exemplo pontuação, acentuação, artigos, preposições, algumas conjunções e numerais.

Na etapa de mineração de dados, será utilizado processamento de linguagem natural. A plataforma Natural Language Toolkit (NLTK) será responsável pela interpretação do *corpus*, onde serão utilizadas técnicas de análise de ocorrência de palavras e dados probabilísticos e suas classificações.

Utilizando o modelo de linguagem N-grams, será possível conhecer qual a palavra que mais ocorre no *corpus*, as duas palavras que mais ocorrem juntas e também as três palavras que mais ocorrem juntas, utilizando Unigram, Bigram e Trigram.

Todas as etapas fazem parte do processo de descoberta de conhecimento em base de dados, utilizado como ferramenta principal e servindo como guia para o desenvolvimento do projeto. Sua etapa final é a avaliação e representação do conhecimento, onde será escalado um especialista de área para fazer a análise das informações encontradas a fim de descobrir o conhecimento interpretando os padrões e as interações destes temas.

4.3 Descrição da Proposta

Com os jogos olímpicos de 2016 sendo sediados no Brasil, espera-se um aumento no número de publicações geradas pela imprensa e desenvolvedores de conteúdo nos diversos meios de comunicação disponíveis. Levando em consideração que existem inúmeros produtores de conteúdo, pretende-se analisar o cenário atual das publicações e sua relação com temas que abordam a saúde dos atletas, sendo possível elaborar um panorama com os temas explorados e descobrir conhecimento com base nessas premissas.

5 MODELAGEM E CRONOGRAMA

5.1 Modelagem

O exemplo de modelagem que o trabalho irá apresentar é referente ao diagrama de atividade. Representado na Linguagem de Modelagem Unificada (do inglês, UML - Unified Modeling Language), segundo Barbosa e Sena (2011), pode ser utilizada na construção de projetos devido as facilidades de representação das etapas envolvidas, sendo uma linguagem de representação (modelo), apresenta em seus diagramas as etapas dos processos envolvidas no desenvolvimento de um sistema.

Separado em colunas, mostrando cada uma das 5 etapas do processo de Descoberta de Conhecimento em Base de Dados e suas devidas ações.

Referente ao Aluno, caberá escolher a base de dados a qual será extraído o conhecimento, posteriormente a etapa de Seleção, serão formatados os dados para que então seja efetuado o tratamento de dados.

Na etapa de Seleção, será examinada a base de dados em busca dos itens pré-estabelecidos, posteriormente serão armazenados estes dados relevantes.

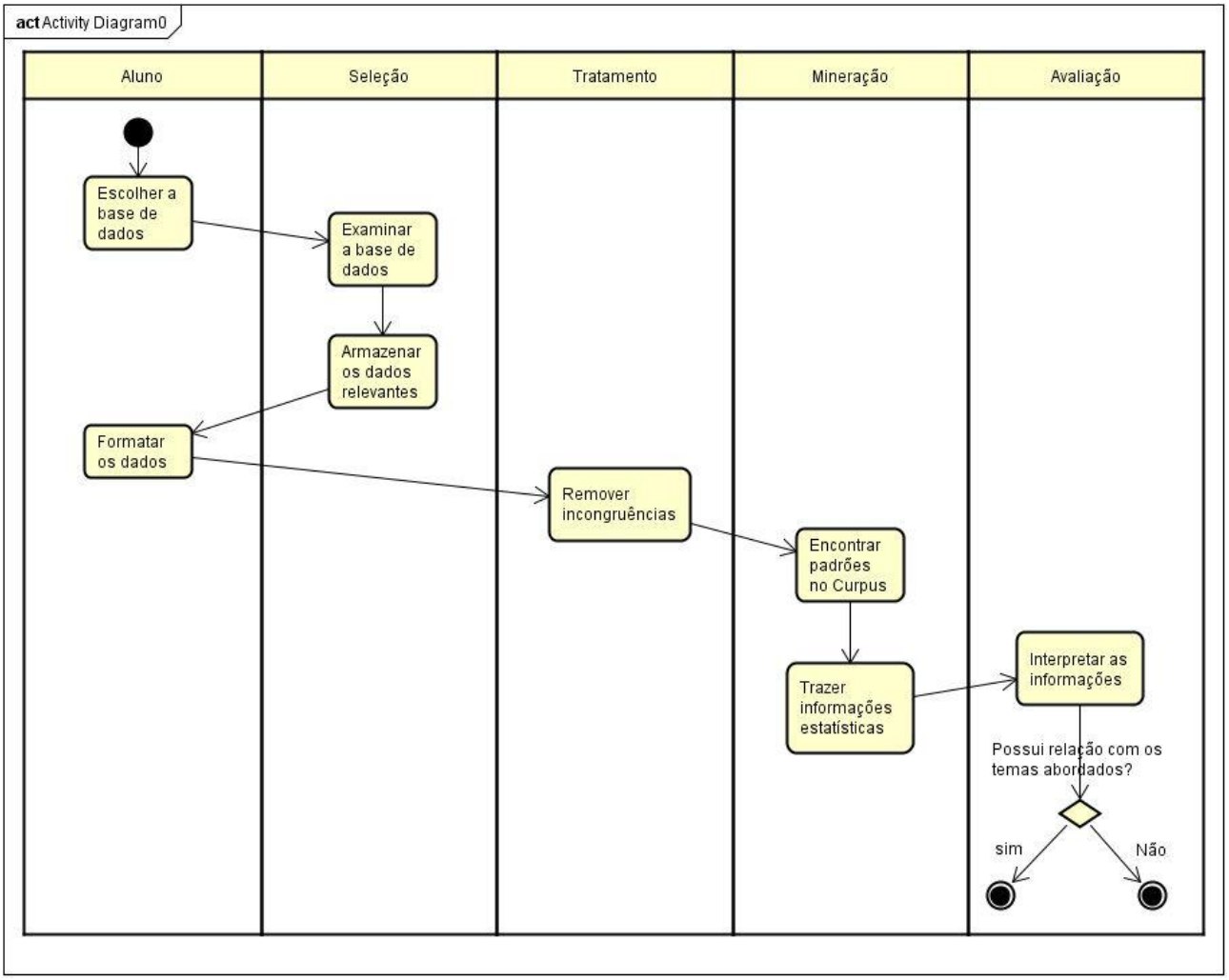
Quanto ao Tratamento, será feita uma análise no *Corpus* em busca de dados irrelevantes ao processo e os mesmos serão removidos.

Na etapa de Mineração, ocorrerá o processamento de linguagem natural para se encontrar padrões e também dados estatísticos (informações).

Referente a etapa de Avaliação, será feita a interpretação e representação do conhecimento a fim de gerar conhecimento e também será analisado se estes padrões encontrados possuem alguma relação com os temas relacionados a saúde dos atletas olímpicos, como doping, antidoping, drogas, suplementação, fair play e atleta limpo.

A figura a seguir representa um diagrama de atividade referente as etapas do projeto:

Figura 1 – Diagrama de Atividade

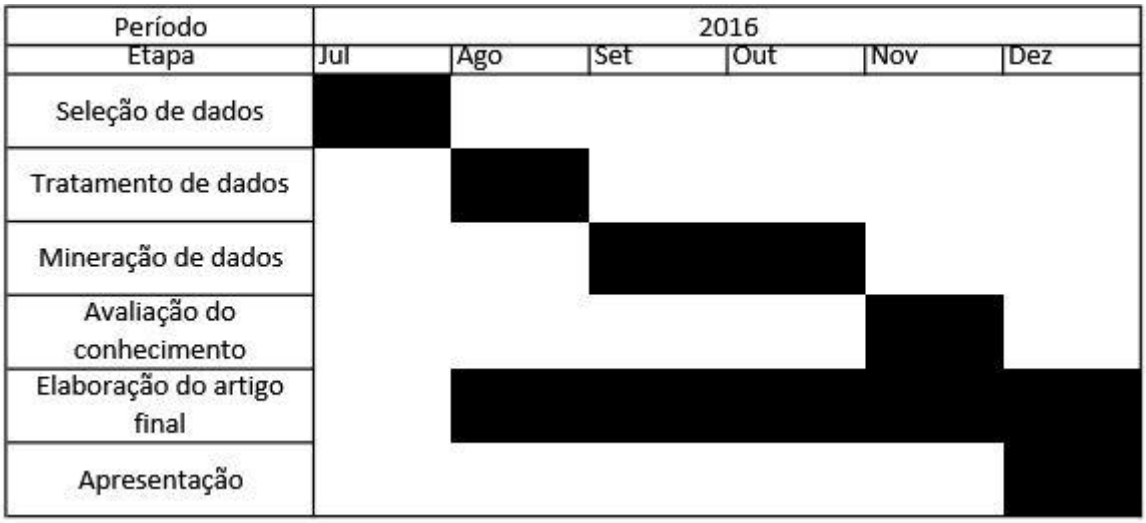


Fonte: Elaborado pelo autor

5.2 Cronograma para TCCII

Para o Trabalho de conclusão de Curso II, serão efetuadas as etapas descritas anteriormente, a imagem a seguir apresenta um cronograma com as atividades e o período do segundo semestre em que serão elaboradas:

Figura 2 – Cronograma TCCII



Fonte: Elaborado pelo autor

6 CONSIDERAÇÕES FINAIS

A presente pesquisa teve como objetivo trazer todas as referências necessárias para a elaboração de segunda etapa do Trabalho de Conclusão de Curso, onde serão aplicadas as técnicas para se descobrir conhecimento, como o processamento de linguagem natural, tratamento de dados e mineração de dados com base nos materiais analisados.

Com os resultados da análise, será possível identificar o cenário atual das publicações relacionadas ao Olímpiadas Rio 2016 sediada no brasil, pela página oficial dos jogos olímpicos no Facebook.

Serão aplicadas todas as etapas do processo de descoberta de conhecimento em base de dados, a fim de se encontrar algum padrão ou até mesmo a relação encontrada com as palavras do *corpus,* visando a geração de conhecimento ao se explorar todas as técnicas do processo.

REFERÊNCIAS

CARDOSO, Olinda Nogueira Paes; MACHADO, Rosa Teresa Moreira. **Gestão do conhecimento usando *data mining*: estudo de caso na Universidade Federal de Lavras**, Lavras: [Universidade Federal de Lavras], 2008.

TRONCHONI, Alex B; PRETTO, Carlos O; ROSA, Mauro A; LEMOS, Flávio A. Becon. **Descoberta de conhecimento em base de dados de eventos de desligamentos de empresas de distribuição**, Porto Alegre: [Pontifícia Universidade Católica do Rio Grande do Sul], 2010.

GALVÃO, Noemi Dreyer; MARIN, Heimar de Fatima. **Técnica de mineração de dados: Uma revisão da literatura**, São Paulo: [Universidade Federal de São Paulo], 2009.

CRAWLER Definition. **TechTarget, Where Serious Technology Buyers Decide.** Disponível em: <http://searchsoa.techtarget.com/definition/crawler> Acesso em: 03 de abril de 2016.

CASTRO, Flaviana dos santos; MATTOS, Merisandra Côrtes de; SIMÕES, Priscyla Waleska Targino de Azevedo. **Mineração de Textos na Saúde por Meio da Utilização de Ferramenta Eurekha**, Criciúma: [Universidade do Extremo Sul Catarinense], 2007.

EUREKHA. **Instituto de Informática da Universidade Federal do Rio Grande do Sul.** Disponível em: <http://www.inf.ufrgs.br/~wives/wiki/doku.php?id=eurekha> Acesso em: 03 de abril de 2016.

FINATTO, Maria José Bocorny; LOPES, Lucelene; CIRULLA, Alena. **Processamento de Linguagem Natural, Linguística de *Corpus* e Estudos Linguísticos: uma parceria bem-sucedida**, Uberlândia: [Universidade Federal de Uberlândia], 2015.

NLTK 3.0 documentation. **Natural Language Toolkit**. Disponível em: < http://www.nltk.org/> Acesso em: 03 de abril de 2016.

HANAI, Tuka Al. **Lexical and Language Modeling of Diacritics and Morphemes in Arabic Automatic Spech Recognition**, Cambridge: [Massachusetts Institute of Technology], 2014.

FONSECA, Stella Oggioni da; NAMEN, Anderson Amendoeira. **Mineração em bases de dados do INEP: Uma análise exploratória para nortear melhorias no sistema educacional brasileiro**, rio de Janeiro: [Universidade do Estado do Rio de Janeiro], 2016.

APACHE About. **Apache HTTP Server Project.** Disponível em: < https://httpd.apache.org/> Acesso em: 15 de maio de 2016.

PHP Documentation. **PHP :Hypertext Preprocessor.** Disponível em: <http://php.net/> Acesso em: 15 de maio de 2016.

BARBOSA, Eduardo Batista de Moraes; SENA, Galeno Jose de. **Data Information System to Promote the Organization Data of Collections – Modeling Considerations by the Unified Modeling Language (UML)** São Paulo: [Universidade Estadual Paulista], 2011.