Universidade do Estado do Amazonas
Escola Superior de Tecnologia
Pós-graduação Lato Sensu em Ciência de Dados
Programação para Ciência de Dados (Turma 02)

31 de outubro de 2020

Defesa: 28/11/2020 *Trabalho Prático*

Especificação do Trabalho

1 Contextualização

Com o advento da Web 2.0, os usuários *online* estão sob os holofotes, pois possuem grande poder. Eles podem influenciar potencialmente vários indivíduos e organizações com seus comentários, análises, classificações e opiniões *online*. Por esse motivo, especialistas em Ciência de Dados têm um papel essencial na análise do conteúdo gerado por usuários [1].

Comentários de usuários são uma parte significativa da imagem organizacional, pois medeiam a experiência. Os usuários gostam de compartilhar suas experiências boas ou ruins, mas também consultam sobre outras experiências semelhantes de outros usuários. As coisas que ouvem os afetam e refletem em seus comentários. Conforme os usuários publicam mais informações sobre um determinado tópico, eles intensificam o quadro dominante da história. Quanto mais os usuários estão envolvidos na comunicação sobre um determinado tópico, mais as suas avaliações produzirão uma avaliação mais criteriosa de um produto, serviço, organização ou pessoa. O número total de comentários de usuários representados em um determinado site captura uma imagem online momentânea.

Surge aí um problema de consistência porque as avaliações do produto, serviço, organização ou pessoa não são igualmente positivas, negativas ou neutras o tempo todo. No entanto, ao publicar seus comentários e análises, as ações combinadas dos usuários facilitam a inteligência coletiva. Essa difusão de informações, sua multiplicação de opiniões individuais e consequentemente a criação de uma imagem online dão suporte ao conceito de inteligência coletiva. James Surowiecki enfatiza que os grupos mais inteligentes são aqueles que consistem "*em pessoas com perspectivas diversas que são capazes de se manter independentes umas das outras*", mesmo que algum membro do grupo seja irracional ele não fará o grupo todo menos inteligente [2].

Diante desse cenário, surgiu uma área de pesquisa, chamada de análise de sentimentos ou mineração de opinião, que vem despertando enormemente o interesse de pesquisadores e da indústria. A análise de sentimentos pode ser vista como uma tarefa de processamento de linguagem natural (PNL) que visa analisar opiniões, sentimentos e emoções expressas em dados não estruturados. Uma tarefa inicial é identificar quais sentenças dentro de um texto são opinativas e quais são meramente factuais. Somente o primeiro tipo interessa na análise de sentimentos. Uma outra

tarefa comum nesta área de pesquisa é a classificação de polaridade, que consiste em classificar o sentimento geral presente em um documento ou frase. Normalmente, esta tarefa é simplificada classificando um texto ou uma frase em 3 classes: positiva, negativa ou neutra.

Considerando o contexto descrito acima, o objetivo das equipes consiste em analisar os comentários postados por usuários sobre restaurantes, com o intuito de realizar uma análise exploratória e visualização de dados, passando também pelas etapas de limpeza e organização. Mais especificamente, as equipes devem explorar os comentários postados pelos usuários, pois arguimos que seja um tipo de mensagem apropriado para se compreender a maneira pela qual os clientes enxergam os estabelecimentos comerciais. Os dados já foram coletados e serão disponibilizados para os alunos.

As equipes serão compostas de **3 alunos** e devem trabalhar de maneira organizada e sistemática para produzir os seguintes artefatos:

- **Relatório Técnico**. Um notebook com o texto e o código-fonte apresentado na ferramenta Google COLAB. Este texto deve ser construído na forma de parágrafos, utilizando um encadeamento lógico, com figuras e tabelas que a equipe considerar apropriadas para esclarecer os questionamentos avaliativos. O código deve estar organizado em blocos (células) de maneira a facilitar o entendimento.
- **Comentários Anotados.** Um arquivo no formato CSV com os comentários anotados. O procedimento para gerar esse arquivo será detalhado na Seção 2.4.2.
- **Repositório no GitHub**. De caráter público (privado durante o desenvolvimento, público na ocasião da entrega), incluindo todos os integrantes da equipe e contendo as evidências de código e de progresso, registrando o trabalho dos integrantes;

2 Detalhamento das Atividades

2.1 Análise de Sentimentos

A Figura 1 apresenta um exemplo de comentário real postado em outubro de 2020 no TripAdvisor sobre o restaurante Coco Bambu.

O primeiro passo na análise é dividir o texto em sentenças. Em seguida, é identificar quais sentenças são opinativas e quais são apenas factuais. Por exemplo, na primeira sentença "*Conheço o restaurante em várias cidades do Brasil*", o cliente apenas informa um fato. Portanto, não há qualquer grau de subjetividade. Já na terceira sentença "*Atendimento bom, ambiente bonito*", o cliente declara estar muito satisfeito com o atendimento e com o ambiente do restaurante.

Além das informações que podem ser extraídas do texto, é possível usar outras informações contidas no comentário. Por exemplo, o local de origem do cliente, a data em que o cliente frequentou o restaurante, a avaliação geral (número de estrelas), o tipo de dispositivo usado para enviar o comentário, entre outras. Essas outras informações são bastante relevantes e podem ser utilizadas para definir o perfil dos clientes.

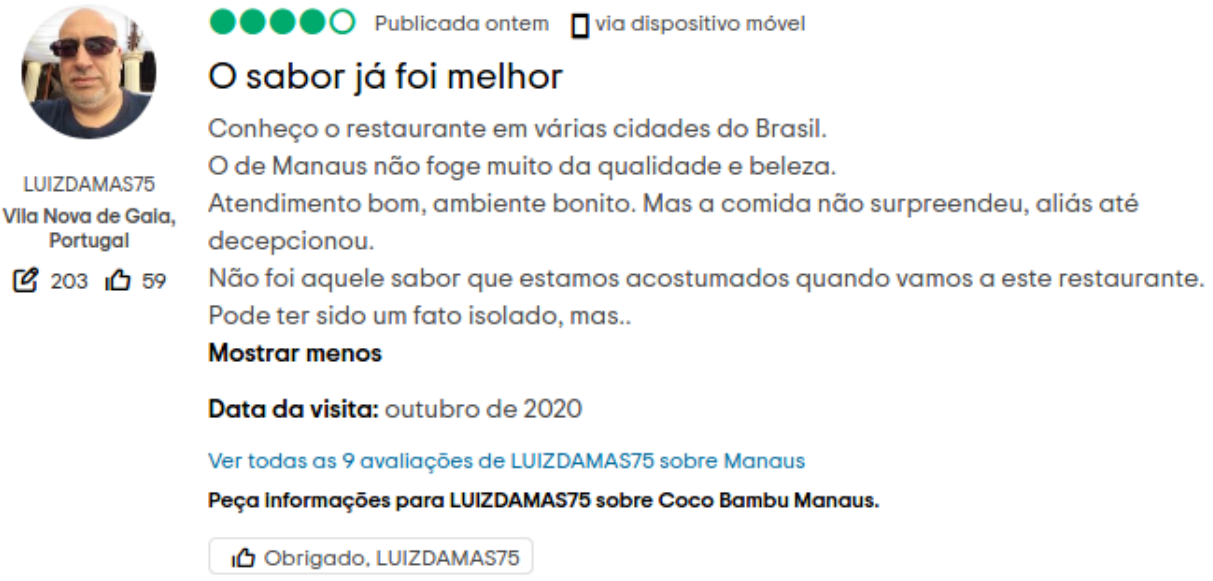

Figura 1: Comentário real.

2.2 Base de Dados

A base de dados a ser utilizada pelas equipes encontra-se disponível no site do professor[1]. Esta base de dados contém **100.000** comentários coletados do site TripAdvisor. Cada linha do arquivo contém um comentário e outros dados.

2.3 Visão Geral dos Dados

A primeira tarefa que as equipes devem realizar é apresentar um resumo com as estatísticas dos dados do *dataset* (base de dados). A apresentação deste tipo de informação é relevante para que outras pessoas possam ter uma visão geral dos dados. As estatísticas podem ser apresentadas através de tabelas e/ou gráficos. Por exemplo, as equipes podem apresentar as características dos textos, tais como a quantidade média de sentenças e palavras por comentários, a distribuição das avaliações (notas), etc.

2.4 Análise de Sentimentos

A análise de sentimentos neste trabalho terá foco em duas tarefas (tarefas 2 e 3). A tarefa 2 consiste na identificação das sentenças subjetivas (opinativas) e a tarefa 3 consiste no cálculo de polaridade das sentenças subjetivas. A descrição detalhada do modo de executar essas duas tarefas podem ser encontradas no endereço `https://gist.github.com/tmelo-uea/162d766210377d96ab108b2e481e4aad`. Lá existe um código pronto, escrito pelo próprio professor, que pode ser usado para executar as duas

[1] `http://tiagodemelo.info/datasets/dataset-v2.dat`

tarefas. As equipes deverão rodar os dois métodos em todo o *dataset*. Os resultados alcançados devem ser apresentados e analisados. A seguir, são definidas as etapas dessa análise.

2.4.1 Etapa I

Inicialmente os alunos devem apresentar a distribuição dos comentários do *dataset*. Mostrar a distribuição de sentenças factuais e subjetivas, assim como a distribuição de comentários positivos e negativos, são exemplos de análise.

2.4.2 Etapa II

As equipes deverão avaliar a eficência do método proposto. O primeiro passo é gerar o *ground-truth* (gabarito). Para isto, as equipes deverão selecionar 200 sentenças aleatoriamente da base de dados e identificar se cada uma das sentenças são subjetivas (valor 1) ou objetivas (valor 0). As sentenças subjetivas deverão ser classificadas em positivas (valor 1), negativas (valor -1) ou neutras (valor 0). As sentenças anotadas deverão ser gravadas em um arquivo no formato CSV e enviadas para o professor juntamente com os demais arquivos do projeto. Note que o arquivo CSV deverá ter três colunas.

O mesmo conjunto de sentenças deverá ser submetido ao método de análise de sentimentos apresentado no início da seção e o resultado deverá ser avaliado. As métricas[2] utilizadas para avaliação são precisão (P), revocação (R) e medida-F (F_1). Seja A o conjunto de respostas corretas, de acordo com um conjunto de referência (gabarito), e seja B o conjunto de respostas produzidas pelo método que está sendo avaliado. Nós definimos precisão (P), revocação (R) e medida-F (F_1) como:

$$P = \frac{|A \cap B|}{|B|} \quad R = \frac{|A \cap B|}{|A|} \quad F_1 = \frac{2 \times (P \times R)}{(P + R)}$$

A seguir, são apresentadas três diferentes questões adicionais. Cada equipe deverá escolher, pelo menos, uma das três questões propostas.

Questão 1 *O método utilizado para cálculo da subjetividade e polaridade é baseado em tradução de texto do próprio pacote TextBlob. Será que se usarmos outra ferramenta para tradução o resultado seria melhor? Considere, por exemplo, o pacote Google Trans*[3]*.*

Questão 2 *O método utilizado para cálculo da subjetividade e polaridade precisou traduzir o texto original e isso diminuiu a acurácia final do método. Será que se usarmos uma ferramenta que faça isso diretamente em Português teríamos um resultado melhor? Considere, por exemplo, a ferramenta iFeel desenvolvida por professores da Universidade Federal de Minas Gerais (UFMG)*[4]*.*

[2] As métricas serão explicadas com mais detalhes em sala de aula pelo professor.
[3] https://pypi.org/project/googletrans
[4] http://blackbird.dcc.ufmg.br:1210

Questão 3 *Os dados analisados vieram apenas de uma fonte de dados. Se a mesma análise fosse aplicada aos dados de uma outra fonte, o resultado seria semelhante? Considere outras fontes como Instagram, Twitter, sites especializados, etc.*

2.5 Visão Temporal

Deve-se fazer uma análise temporal das opiniões que formam o *dataset*. Pode-se considerar o intervalo temporal de dias, semanas ou meses. A escolha do intervalo de tempo ficará a cargo das equipes. Exemplo de análise temporal: "*existe um dia da semana em que ocorra o maior número de postagens?*"

2.6 Visão Geográfica

Deve-se fazer uma análise geográfica (espacial) dos comentários que formam o *dataset*. Existem algumas colunas no *dataset* que trazem a informação das localizações como, por exemplo, país, estado e cidade. Exemplo de análise geográfica: "*os usuários de estados (regiões) diferentes são mais ou menos rigorosos na avaliação dos restaurantes?*" Essa análise ainda pode ser realizada em diversos níveis de área (cidade, estado ou região). Além de apresentar a distribuição das opiniões dos usuários por região, a equipe deverá fazer uma análise dessa distribuição. Por exemplo, apresentar as razões (ou hipóteses) da ocorrência dessa distribuição.

Recomenda-se utilizar o pacote GeoPandas[5] para essa tarefa. As informações geográficas como latitude e longitude podem (devem) ser obtidas na Web. Deve-se investigar como automatizar esse processo.

2.7 Tecnologias e Sugestões

Para a realização desta tarefa, é obrigatório o uso da linguagem de programação Python 3 (e superiores) e das bibliotecas Pandas e NumPy. Para exibição dos gráficos, as equipes podem usar as bibliotecas MatPlotLib[6] ou Searborn[7]. Outras bibliotecas complementares também podem ser usadas, especialmente na análise exploratória.

3 Critérios de Avaliação

Os critérios de avaliação leverão em conta a organização do repositório, a qualidade do código produzido, a completude das tarefas solicitadas, a documentação, a quantidade textual do relatório em termos de utilização da norma culta, coesão, coerência, o respeito aos prazos e a colaboração da equipe na elaboração do projeto.

[5] https://geopandas.org
[6] https://matplotlib.org
[7] https://seaborn.pydata.org

A apresentação do trabalho será ainda avaliada pelos demais colegas da turma. Cada aluno deverá avaliar e dar uma nota aos trabalhos dos demais colegas. A média aritmética dessas notas irá compor 20% da nota final e a nota professor irá compor 60% da nota final. Os 20% restantes serão avaliados pela participação do aluno em sala de aula (conforme descrito no plano de ensino).

Durante a especificação do trabalho foram apresentadas três questões. Os alunos **devem** escolher e executar, pelo menos, uma das três questões.

4 Datas

O trabalho deverá ser enviado para o professor até o dia 27 de novembro através do Google Classroom. A defesa (apresentação) do trabalho será no dia 28 de novembro. As equipes terão 10 minutos para a realização da apresentação. A escolha da ordem das equipes será feita no dia da apresentação.

Referências

[1] H. Jakopović. Detecting the online image of "average" restaurants on tripadvisor. *Media Studies*, 7(13), 2016.

[2] J. Surowiecki. *The wisdom of crowds*. Anchor, 2005.