Material Complementar

Tiradentes no TripAdvisor - O que se fala sobre essa simpática cidade histórica

Antonio Pedro Santos Alves
antoniopedrosantosalves@gmail.com
UFSJ
São João del-Rei, Minas Gerais, Brasil

Lucas Félix
UFMG
Belo Horizonte, Minas Gerais, Brasil

Carlos M. G. Barbosa
UFSJ
São João del-Rei, Minas Gerais, Brasil

Vinícius da Fonseca Vieira
UFSJ
São João del-Rei, Minas Gerais, Brasil

Carolina Ribeiro Xavier
UFSJ
São João del-Rei, Minas Gerais, Brasil

1 PRÉ-PROCESSAMENTO

A etapa de 'Junção de nomes próprios' é totalmente ligada ao conteúdo da base de *reviews* que possuímos. Por ser uma cidade histórica e extremamente ligada ao catolicismo dos séculos XVIII e XIX, Tiradentes possui muitas atrações, hotéis e restaurantes que trazem consigo nomes de santos católicos. Portanto, visando evitar perder a referência para esses lugares - que podem ser importantes dentro do conjunto de documentos - estabelecemos esta etapa de pré-processamento, salvando o nome de santos que aparecem com um *underscore*. Assim, a Serra de São José, com esta etapa, é passada para as próximas etapas como "Serra de São_José". Esse processo difere do uso de bigramas e trigramas (n-gramas) pois só é feito sobre nomes de santos católicos. A aplicação de n-gramas sobre todo texto, piorou o resultado qualitativo obtido no final, por isso o filtro sobre os santos.

Na 'Remoção de entidades desnecessárias', fazemos uso de um recurso amplamente utilizado na literatura, Reconhecimento de Entidades Nomeadas (ou NER - Named Entity Recognition), cujo objetivo é categorizar pessoas, lugares, organizações e outras entidades de interesse no texto [**?**]. Com a utilização do NER foi possível avaliar que nos *reviews* realizados algumas entidades não agregavam valor para uma categorização textual. Essas entidades são: datas (exemplo: *visitei na* ***quarta-feira***), porcentagens (exemplo: *a atração está restrita a* ***10%*** *de sua capacidade em dias chuvosos*, números ordinais e cardinais (exemplo: *fui o* ***primeiro*** *a entrar em todos os* ***cinco*** *museus que visitei*) e pessoas (exemplo: *o* ***dono da pousada*** *era muito simpático*). As entidades identificadas dentro destas categorias foram descartadas, dado que as mesmas não agregam para a coesão dos tópicos gerados.

Por fim, em 'Remoção de *stopwords* e outros componentes gramaticais' realizamos uma etapa que é de *praxe* quando se fala em mineração de texto, que é a remoção de *stopwords* e componentes gramaticais. *Stopwords*, diferente das entidades desconsideradas em *ii* são **palavras irrelevantes** para uma análise textual, como pronomes e preposições. Por não existir uma lista completa com estas palavras, utilizamos *stopwords* customizadas, baseadas na nossa expertise sobre a base e a lista de palavras utilizadas no artigo [**?**]. Todas as palavras contidas nessas listas são removidas. Além destas, também removemos os seguintes componentes gramaticais {pontuação, verbos, adjetivos, advérbios}, categorizados segundo o léxico definido no Universal Part-of-Speech tags (ou *Universal Part-of-Speech tags*[1]).

2 FIGURAS

Neste artigo, possuímos algumas imagens que contribuem para a compreensão do trabalho como um todo. Nesse sentido, vamos detalhar cada uma das que estão disponíveis neste repositório.

A iniciar pela figura intitulada 'Pipeline de Execução'. Nela é possível visualizar as etapas realizadas na coleta de dados sobre o Tripadvisor.

Trazemos também 3 figuras intituladas 'Reviews de Atrações, Hoteis, Restaurantes em Tiradentes', que são resumos quantitativos sobre a base construída sobre o Tripadvisor para este paper, isto é, quantos dados relacionados a um review no Tripadvisor conseguimos extrair.

Com milhares de reviews coletados, fomos capazes de determinar os termos mais importantes desses reviews através da modelagem de tópicos (MT). Os termos resultantes da MT, desconsiderando em quais tópicos estes termos foram agrupados, foram utilizados para gerar uma nuvem de palavras (ou wordcloud), disponível no arquivo intitulado 'WordCloud Tiradentes'.

Com a Análise de Sentimento (AS), pudemos verificar o sentimento associado aos reviews realizados em cada uma das categorias recuperadas e verificar se o sentimento varia de acordo com regiões específicas de Tiradentes, como mostra as figuras intituladas 'Mapa de Sentimento Associado aos Atrações, Hotéis, Restaurantes de Tiradentes'.

Com a AS, ainda foi possível verificar o quão discrepante é o sentimento dos visitantes de Tiradentes para as notas (no Tripadvisor, são estrelas) dadas em cada um de seus reviews. Essa discrepância pode ser visualizada por um histograma ou por um gráfico de barras. No primeiro caso, temos as figuras intituladas 'Histograma sobre as notas dadas e o sentimento associado em Atrações, Hotéis, Restaurantes de Tiradentes'. Nelas observa-se o comportamento das notas dadas (barras e linhas azul) e do sentimento obtido (barras e linhas laranja) em cada possibilidade de nota. Observa-se que as barras escuras ocorrem devido a sobreposição das barras azul e laranja. No segundo caso, temos gráficos de barras que trazem um comparativo quantitativo de notas dadas em cada uma das possibilidades (1 a

[1] https://universaldependencies.org/docs/u/pos/

In: XVII Workshop de Trabalhos de Iniciação Científica (WTIC 2020), São Luís, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2020.

ISSN 2596-1683

5) e quantos reviews, baseado no sentimento, se enquadram nesse espectro de possibilidades.