A Produção de Conhecimento Sociolinguístico

por meio da Coleção e Mineração de Dados Digitais

Camila Lívio

University of Georgia

June 27, 2023

Outline

Introdução

Apresentação

Estrutura

Manipulação e Armazenamento

Construindo Sentido

Análise de Bigramas

Conclusão

Conclusão

Objetivos

1. Apresentação do *RCor* (Reviews Corpus)
2. Análise de bigramas plotados contra a classificação de produtos
3. Trabalho com perguntas sociolinguísticas – especificamente o uso e variação da intensificação – com base em dados digitais

O *RCor*

- O RCor foi desenhado e pensado para o estudo de um fenômeno linguístico: a intensificação
- Tese de doutorado (2 tarefas)
- Intensificação: aumentar a qualidade ou propriedade de outra palavra (Tagliamonte 2008, Bolinger 1972), como em *super chique*, *muito bom* e *interessantíssimo*
- Corpus especializado (Brezina 2018)
- Dados adquiridos entre junho de 2020 e junho de 2021, por meio de *web scraping* utilizando o R Core Team (2021)
- Amazon Brasil: representatividade no ecommerce (Jindau and Liu 2008)
- Legalidade da obtenção desses dados: consulta com Office of Research, Termos de Serviço, entendimento sobre o uso desses dados de maneira justa e sem nenhum acarretamento negativo para os usuários, revisão de trabalhos que utilizam a Amazon como fonte de dados para mineração

O *RCor* - cont.

- Medidas de conformidade ética: anonimização total de usuários, reprodução parcial de exemplos, uso não comercial dos dados
- *Fair use*: trabalho de pesquisa
- "It is legal to scrape publicly available data such as product ratings, review descriptions, or the number of reactions to a particular review. You just need to be careful with personal data, specifically the reviewer's name and avatar, as these could be used to identify the user" (*Is web scraping legal?* by Ondra Urban)

Estrutura do *RCor*

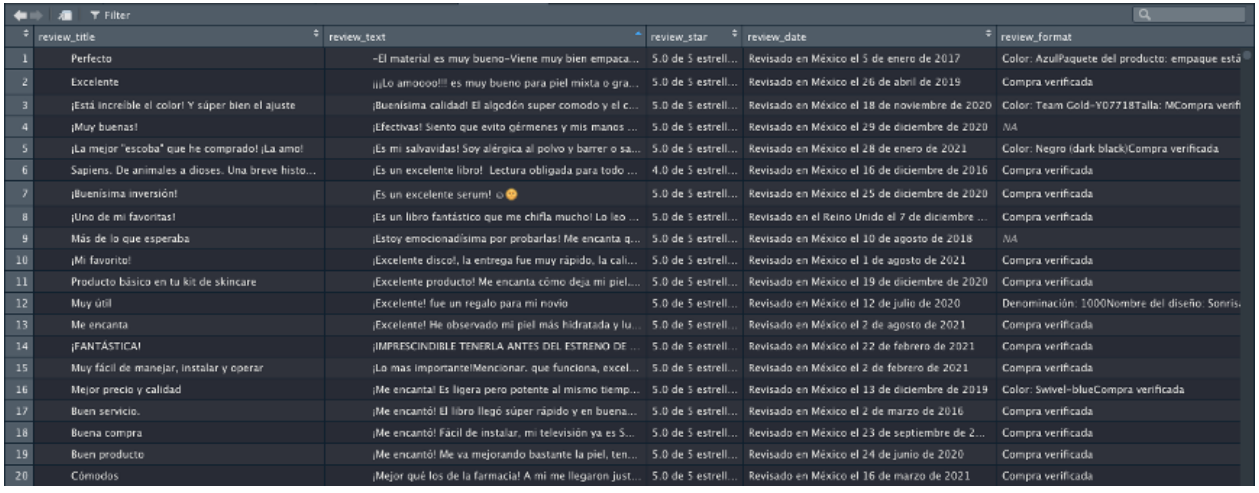
- ID do produto (asin): título, texto, avaliação, produto, categoria do produto, preço e status da compra
- Representatividade: Top sellers de cada categoria disponível no website
- 50 páginas de avaliações por produto
- Limpeza: filtrar avaliações somente em pt-br, verificar a existência de duplicatas
- **Vantagem:** aquisição é relativamente rápida, grande quantidade de material, dados organizados em formato .csv
- **Desvantagem:** problemas no código e avaliações falsas/simuladas (*My life has changed forever!*, by Vazquez 2017)

Manipulação, Mineração e Armazenamento

- Durante a escrita da tese: RStudio (e pacotes como: Dplyr, Quanteda, Tidytext, Tidyverse)
- Ciência aberta
- FAIR: findable, accessible, interoperable and reusable
- Plano para o armazenamento e manutenção
- Criação de uma interface (website) vs. Repositório (OSF)

Too often people assume that digital content will last forever, immateriality and reproducibility encouraging the false impression that anything uploaded to the cloud is safely stored for posteriority. In reality, we lose an estimated quarter of working links every seven years and digital files can quickly become incomprehensible due to the swift churn of technological obsolescence. Sustainable culture includes building archives that will allow people to explore their cultural heritage for years to come (Taylor 2014: 16)

O RCor



	review_title	review_text	review_star	review_date	review_format
1	Perfecto	-El material es muy bueno-Viene muy bien empaca...	5.0 de 5 estrell...	Revisado en México el 5 de enero de 2017	Color: AzulPaquete del producto: empaque está
2	Excelente	¡¡¡Lo amooo!!! es muy bueno para piel mixta o gra...	5.0 de 5 estrell...	Revisado en México el 26 de abril de 2019	Compra verificada
3	¡Está increíble el color! Y súper bien el ajuste	¡Buenísima calidad! El algodón super comodo y el c...	5.0 de 5 estrell...	Revisado en México el 18 de noviembre de 2020	Color: Team Gold-Y07718Talla: MCompra verif
4	¡Muy buenas!	¡Efectivas! Siento que evito gérmenes y mis manos ...	5.0 de 5 estrell...	Revisado en México el 29 de diciembre de 2020	NA
5	¡La mejor "escoba" que he comprado! ¡La amo!	¡Es mi salvavidas! Soy alérgica al polvo y barrer o sa...	5.0 de 5 estrell...	Revisado en México el 28 de enero de 2021	Color: Negro (dark black)Compra verificada
6	Sapiens. De animales a dioses. Una breve histo...	¡Es un excelente libro! Lectura obligada para todo ...	4.0 de 5 estrell...	Revisado en México el 16 de diciembre de 2016	Compra verificada
7	¡Buenísima inversión!	¡Es un excelente serum!	5.0 de 5 estrell...	Revisado en México el 25 de diciembre de 2020	Compra verificada
8	¡Uno de mi favoritas!	¡Es un libro fantástico que me chifla mucho! Lo leo ...	5.0 de 5 estrell...	Revisado en el Reino Unido el 7 de diciembre ...	Compra verificada
9	Más de lo que esperaba	¡Estoy emocionadísima por probarlas! Me encanta q...	5.0 de 5 estrell...	Revisado en México el 10 de agosto de 2018	NA
10	¡Mi favorito!	¡Excelente disco!, la entrega fue muy rápido, la cali...	5.0 de 5 estrell...	Revisado en México el 1 de agosto de 2021	Compra verificada
11	Producto básico en tu kit de skincare	¡Excelente producto! Me encanta cómo deja mi piel....	5.0 de 5 estrell...	Revisado en México el 19 de diciembre de 2020	Compra verificada
12	Muy útil	¡Excelente! fue un regalo para mi novio	5.0 de 5 estrell...	Revisado en México el 12 de julio de 2020	Denominación: 1000Nombre del diseño: Sonris.
13	Me encanta	¡Excelente! He observado mi piel más hidratada y lu...	5.0 de 5 estrell...	Revisado en México el 2 de agosto de 2021	Compra verificada
14	¡FANTÁSTICA!	¡IMPRESCINDIBLE TENERLA ANTES DEL ESTRENO DE ...	5.0 de 5 estrell...	Revisado en México el 22 de febrero de 2021	Compra verificada
15	Muy fácil de manejar, instalar y operar	¡Lo mas importante!Mencionar. que funciona, excel...	5.0 de 5 estrell...	Revisado en México el 2 de febrero de 2021	Compra verificada
16	Mejor precio y calidad	¡Me encanta! Es ligera pero potente al mismo tiemp...	5.0 de 5 estrell...	Revisado en México el 13 de diciembre de 2019	Color: Swivel-blueCompra verificada
17	Buen servicio.	¡Me encantó! El libro llegó súper rápido y en buena...	5.0 de 5 estrell...	Revisado en México el 2 de marzo de 2016	Compra verificada
18	Buena compra	¡Me encantó! Fácil de instalar, mi televisión ya es S...	5.0 de 5 estrell...	Revisado en México el 23 de septiembre de 2...	Compra verificada
19	Buen producto	¡Me encantó! Me va mejorando bastante la piel, ten...	5.0 de 5 estrell...	Revisado en México el 24 de junio de 2020	Compra verificada
20	Cómodos	¡Mejor qué los de la farmacia! A mí me llegaron just...	5.0 de 5 estrell...	Revisado en México el 16 de marzo de 2021	Compra verificada

Figure: RCor-MS

A Produção de Conhecimento Sociolinguístico

- O corpus foi desenhado para estudar a intensificação no PB
- Especificamente, avaliações/revisões são consideradas textos altamente avaliativos (Athanasiadou 2007, Hunston 2010) = "boca a boca" escrito
- Gênero textual, com estrutura e convenções
- Uma característica essencial das avaliações é o uso da intensificação: prefixos, sufixos e bases lexicais
- Já que a intensificação é abundante em tal gênero textual, o que se pode dizer sobre a sua função social nesses textos?
- **Análise:** bigramas e avaliação estrela

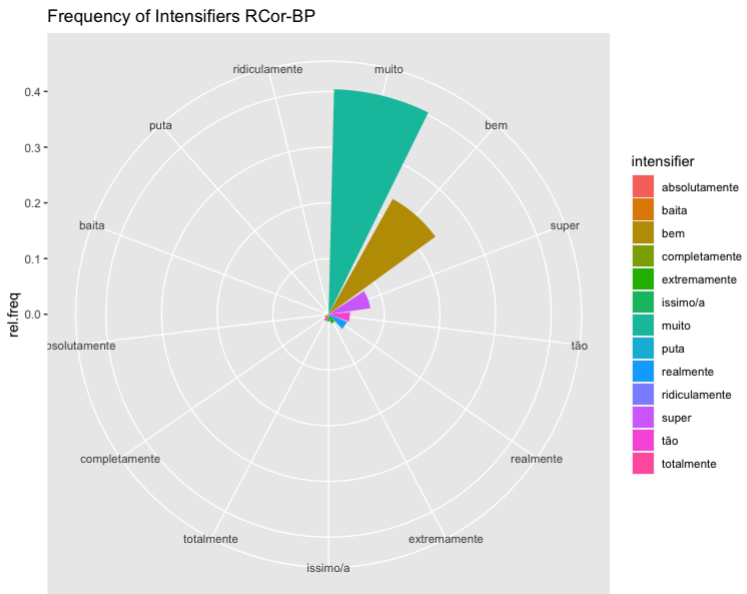


Figure: Relative frequency of intensifiers in the RCor-BP

Bigramas e Avaliação por Estrela

- Tendência é ter mais avaliações 4-5 estrelas que 1-2; distribuição em J (Hu et al. 2009)
- 4-5 estrelas: bigramas mostram o uso de linguagem positiva e uso de intensificadores lexicais em posição pré-adjetiva. Intensificação é mais presente em avaliações positivas.
- 1-2 estrelas: bigramas mostram o uso de linguagem descritiva e uso mais frequente da negação
- Avaliação é um gênero avaliativo e a intensificação é vastamente utilizada porque cumpre função epistêmica. Demonstra o comprometimento da usuária com relação produto. O uso de intensificadores mais canônicos (em contraste com expressões/palavras tabu) é um reflexo da convenção desse gênero textual, bem como a formação de uma persona online confiável (Mackiewicz 2008)
- Para demonstrar confiabilidade, faz-se uso da intensificação
- *Digital self-expression* (Weninger and Li 2022): intensificação é utilizada para comunicar uma postura e influenciar comportamentos

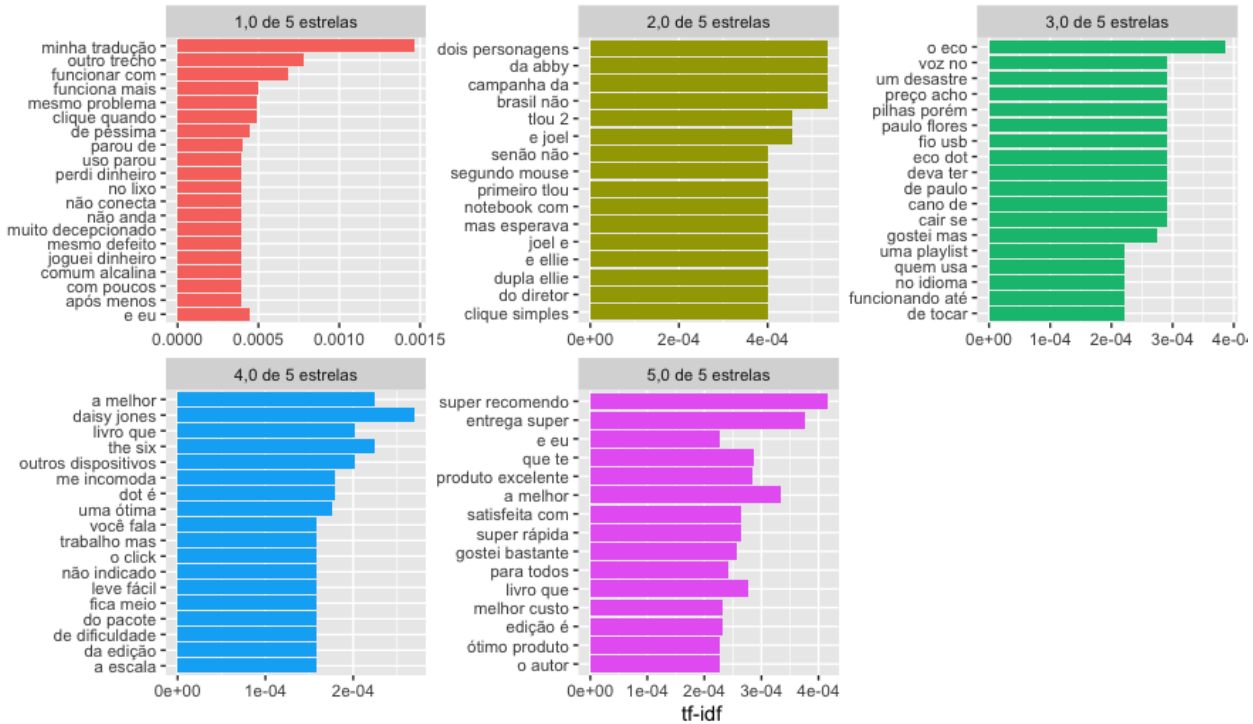


Figure: Bigramas plotados contra a classificação de produtos

Conclusão

- Mandatos federais nos EUA e abundância de dados digitais
- Discussão sobre acesso aberto e infraestrutura para ciência aberta
- A criação e manutenção de corpus, bem como a construção de saberes dele depreendido, requer o trabalho em times (Nguyen 2020)
- Análise de padrões linguísticos e criação de sentidos em espaços digitais

Email para receber a lista completa de referências e perguntas:
camila.emidio25@uga.edu
@camlivio