Distinctions in Amazon Product Advertising and Pricing

Christina Chen | Justine Schabel MIDS Spring 2021



Problem Statement and Goals

- Expose the oversaturation of marketing jargon on products
- Analyze products cohesively, from titles to ingredients
- Understand the relationship between terms and product price
- Evaluate whether advanced NLP techniques perform better compared to manual text analysis
- Adapt a paper's method and apply it to a new domain or dataset



Freedman & Jurafsky

Authenticity in America Class Distinctions in Potato Chip Advertising

- 12 potato chip bags: 6 lower-priced, 6 high-priced
- Authenticity naturalness, locality, historicity
- **Health** particularly bad ingredients
- **Falutin** flesch-kincaid, tokens, word commonality
- **Distinction** comparator, uniqueness, linguistic negation















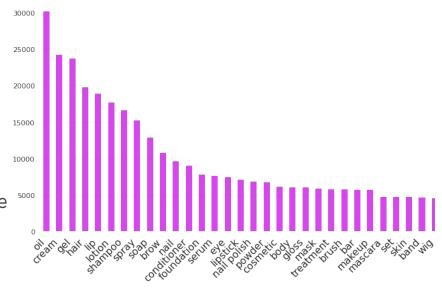






Data

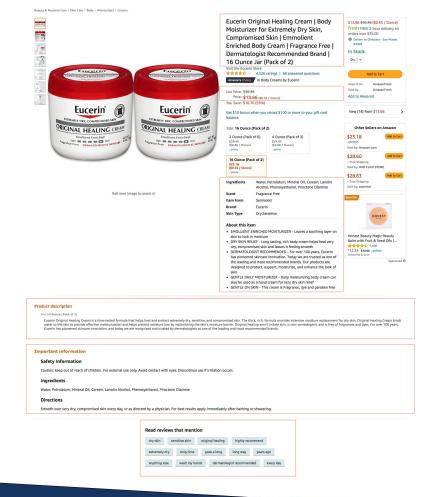
- Hugging Face Amazon Reviews (Beauty and Personal Care)
- Jungle Scout Amazon Beauty Products
- Amazon Web Scrape
- California's Chemicals in Cosmetics Database (CDPH)





Data

- Lotion & Hair products
- Extracted from Hugging Face and Jungle Scout product_title
- Used ids to scrape and concatenate title, price, about, description, information, and tags
- Used BeautifulSoup, Selenium, requests-HTML, and Tor
- Scraped 33880 unique Amazon pages





Data

Lotion

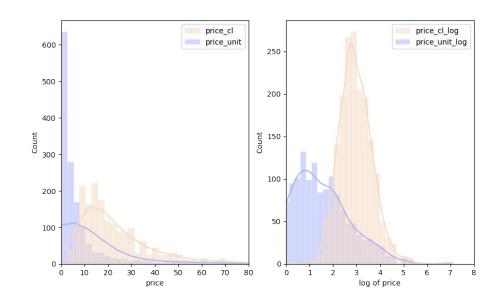
lotions, creams, moisturi(z/s)ers

- 2511 unique products with price
- median price: \$3.42 / oz

Hair

shampoos, conditioners

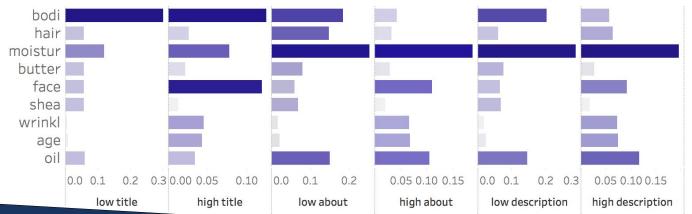
- 2464 unique products with price
- median price: \$1.72 / oz





Freedman and Jurafsky Analysis I

	Metric	Lotion	Hair	Eucerin example
Authenticity	TF-IDF	Higher \$: FDA/Physician tested Lower \$: bias in locality	Higher \$: Recommended	dermatologist recommend
Health	Number of Toxic Chemicals	Higher Priced Higher (slightly)	Same	mineral oil





TF-IDF example for lotion

Freedman and Jurafsky Analysis I

	Metric	Lotion	Hair	Eucerin
Falutin	Flesch-Kincaid	Lower Priced Higher	Lower Priced Higher	63.4
	Readability	Lower Priced Higher	Higher Priced Higher	-84.84
	Lexicon Count	Lower Priced Higher	Lower Priced Higher	154
	Marketing Terms	Lower Priced Higher	Lower Priced Higher	41
Distinction	Uniqueness/Comparator Terms (Method I: Word2Vec)	Lower Priced Higher	Higher Priced Higher	0
	Uniqueness/Comparator Terms (Method II: GloVe)	Lower Priced Higher	Lower Priced Higher	19
	Sentiment	Lower Priced Higher (overall)	Lower Priced Higher (overall)	0.9947



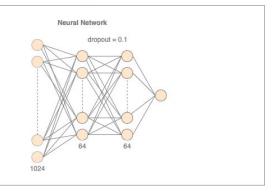
Models

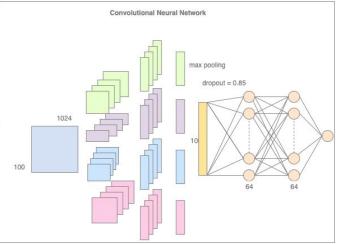
Model

0000 00 1024 BERT Ō Input receiv panten prov radiant color shine in prod condition fl ozradiant color shine 00 prod prod nourish hair prov nutrient antioxid hair strong damag clean condit one step hydrat colortr 1024 hair allday last shine power prov formula hydrat nourish natur shine touchabl soft hair bring shine colortr hair DistilBERT 00 00 1024 RoBERTa) () 1024

Embeddings

ELMO







Results

Hair

	NN	CNN	NN	CNN
Baseline	MAE: 6.47602 MSE: 164.18031		MAE: 1.341023 MSE: 6.54206	
ELMO	MAE: 5.603	MAE: 3.849	MAE: 1.467	MAE: 1.298
	MSE: 109.130	MSE: 39.555	MSE: 6.371	MSE: 7.214
DistilBERT	MAE: 5.761	MAE: 3.844	MAE: 1.423	MAE: 1.283
	MSE: 125.258	MSE: 39.430	MSE: 6.645	MSE: 7.352
RoBERTa	MAE: 6.929	MAE: 3.638	MAE: 1.415	MAE: 1.270
	MSE: 150.332	MSE: 35.470	MSE: 8.206	MSE: 7.013
BERT	MAE: 6.730	MAE: 3.791	MAE: 1.371	MAE: 1.293
	MSE: 325.299	MSE: 37.524	MSE: 6.581	MSE: 7.488



Thank You!

Christina Chen

christinachen@ischool.berkeley.edu

Justine Schabel

justinetaylor@ischool.berkeley.edu

