PINK TEXT

Inspired by Pink Tax and Pant Pockets ... Using Sentiment Analysis on Amazon Reviews for Women's, Men's, and Unisex Pants

ARE POCKETS A FEATURE IN POSITIVE & NEGATIVE REVIEWS?

Authors: Akemi Vuong, Trinity Renee B., Vivian Lam, Yafira Martinez, Yannelly Mercado

PRESENTATION HIGHLIGHTS

FOCUS AREAS

- Pink Tax and Pockets in Pants -> Pink Text
- Data Processing & Analysis
- Natural Language Processing (NLP)
- Training & Testings
- Interesting Results
- Actionable Insights
- In the future...

DEFINING PINK TAX

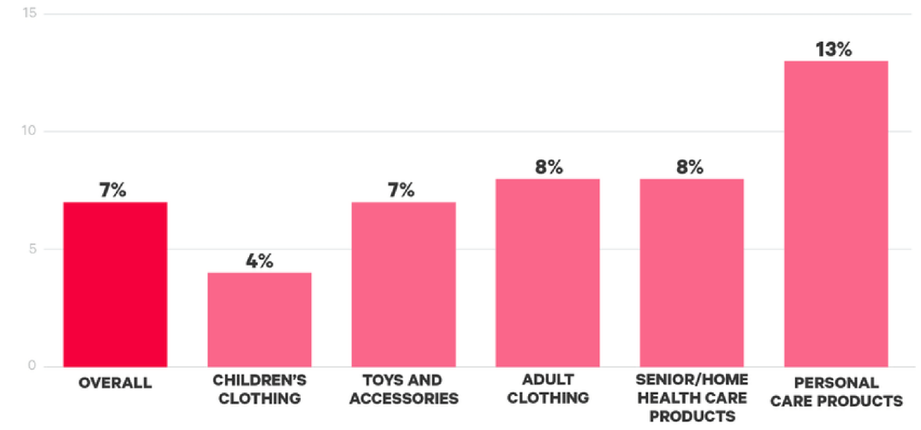
WHAT IS PINK TAX?
WHO DOES IT AFFECT?

The Pink Tax refers to products that are specifically marketed towards women and are more expensive than the same products marketed towards men. It's the extra amount women pay for everyday products that men use. The term “pink tax” was popularized around the mid-1990s, when the Gender Tax Repeal Act of 1995 passed in California, prohibiting price discrimination on services.

PRODUCTS IMPACTED BY THE PINK TAX

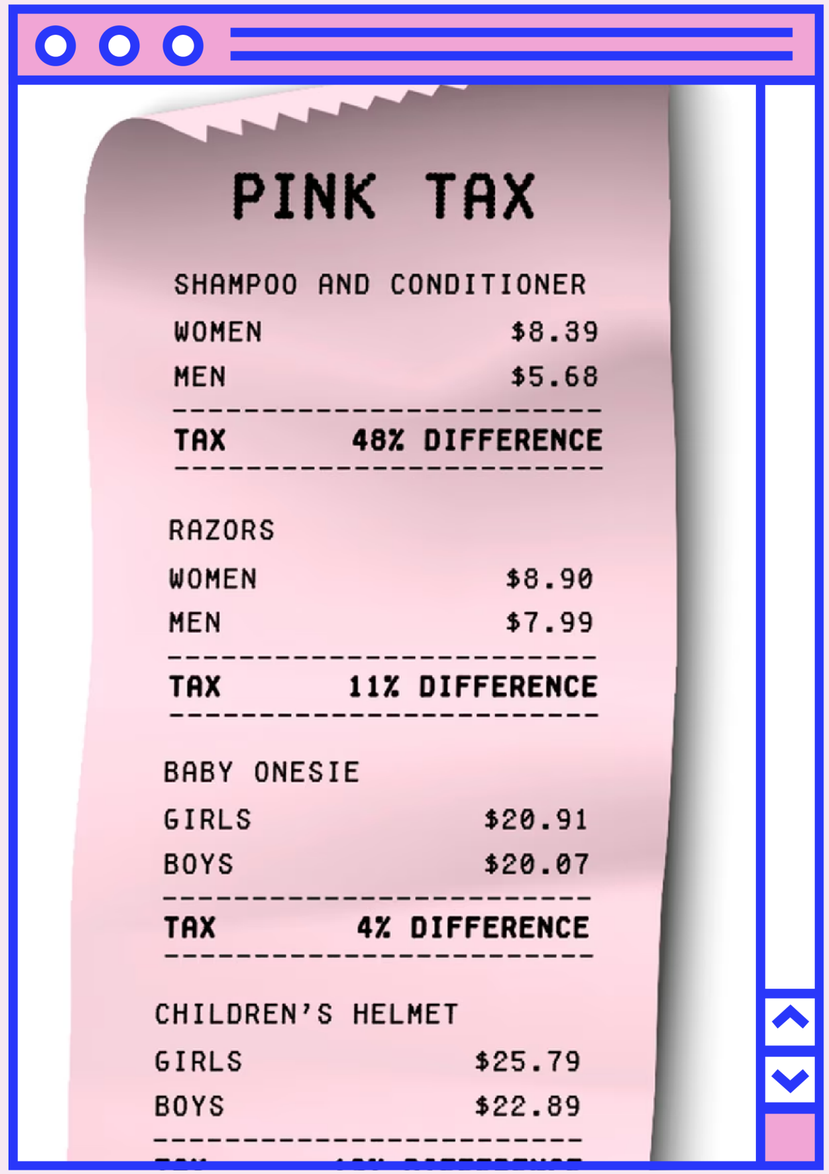
The Pink Tax: Higher Prices on Women's Products

Average markup by category



ValuePenguin

Source: NYC Dept. of Consumer Affairs, "From Cradle to Cane: The Cost of Being a Female Consumer" (Dec. 2015). Findings based on comparisons of products with similar male and female versions for sale at NYC retailers.



PINK TEXT

- THERE HAS BEEN DISCUSSIONS ABOUT HOW POCKETS IN WOMEN'S PANTS ARE TOO SMALL
- WOMEN'S FAST FASHION ARE CHEAPLY MADE

- WE WANT TO SEE IF SENTIMENT ANALYSIS CAN REVEAL HOW CONSUMERS FEEL ABOUT PANTS MADE FOR WOMEN, MEN, AND ALL PEOPLE.

POCKETS???

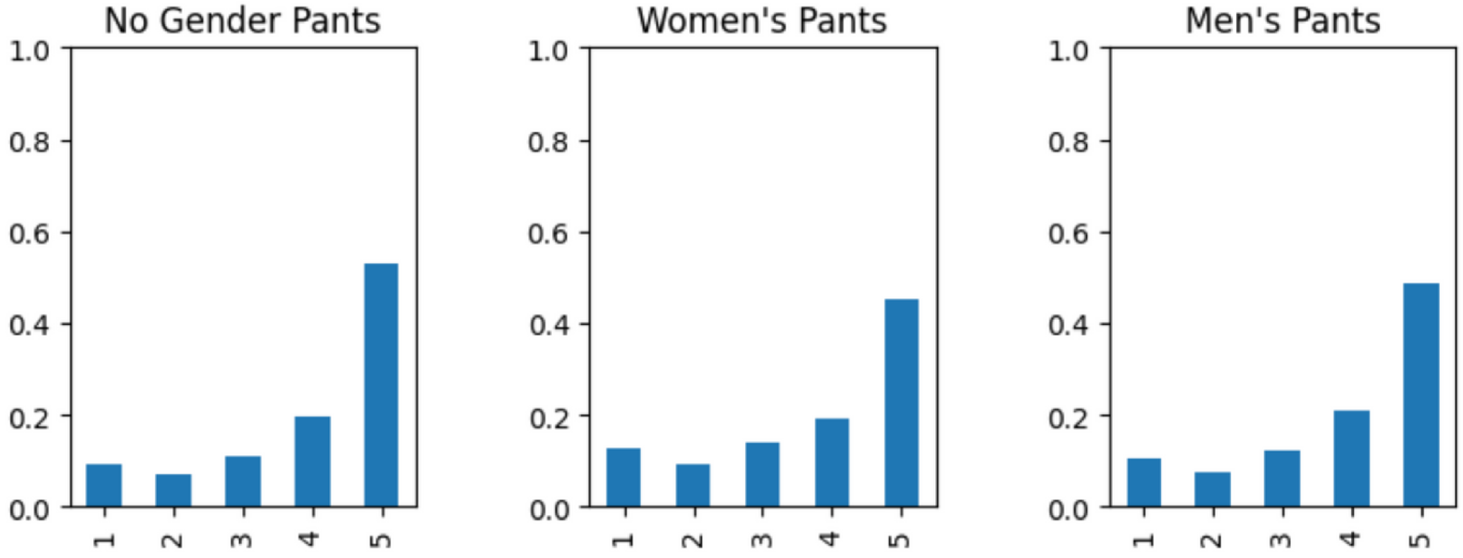
DATA PROCESSING AND ANALYSIS

AMAZON APPAREL REVIEWS:

PANTS BY GENDER

AMAZON RATING STARS

PANTS: RATINGS BY GENDER



- From the distribution of Star Ratings for Pants by Gender in the graphs above, we can see that Women's Pants have more Low Reviews (Rating 1) compared to Men's Pants. WHY?

- Interesting! Pants without Gender in the Product Title have more High Star Ratings and less Low Star Ratings than pants made for Women or Men.

The datasets were separated by ratings, then a column was inserted to label each review as positive or negative.

POSITIVE REVIEWS:

Ratings with more than 3 Stars

NEGATIVE REVIEWS:

Ratings with less than 3 Stars

NEUTRAL REVIEWS:

We will apply our findings to Reviews with Ratings equal to 3 Stars in the future.

Positive = Ratings > 3

Negative = Ratings < 3

Python's Natural Language Toolkit NLTK

Pre-Processing

tokenization to separate text into words and sentences, removing stop words, etc.

Normalization

convert similar words into uniform sequences, using lemmatization and stemming.

Training and Testing

Naive Bayes and Support Vector Machine (SVM) used to classify words in reviews as negative and positive.

NLTK

NLTK's Naive Bayes classifier is used in Python to classify text data.

DATA PREPARATION

Cleaning and preprocessing:

"Bag of Words" used to divide indicator words for classifier, list of sentences joined as one string. All words lower case, stop words removed.

TRAINING MODEL

NLTK Frequency Distribution used to label each word with its count. The positive and negative word lists were combined into one set, then shuffled and split for

PREDICTING TEST DATA

classifying test data as positive or negative, with 75% to 81% accuracy. NLTK Naive Bayes model showed the most informative features, or how likely a feature led to a positive or negative review.

TRAINING AND TESTING

```
[ ] # randomly shuffled labled words (combined dictionary of positive/negative feature dict)
    import random
    random.shuffle(labeledwords)

    # create train and test set
    train_set, test_set = train_test_split(labeledwords, test_size=0.2, random_state=25)
    classifier = nltk.NaiveBayesClassifier.train(train_set)
```

```
[ ] # test classifier with examples

    #neg
    print(classifier.classify(word_features('I do not like this pants because they are ugly')))

    #pos
    print(classifier.classify(word_features('These pants fit well')))
```

```
neg
pos
```

```
[ ] #calculate accuracy of classifer
    # accuracy calculated dividing number of
    # correct predictions made by a model divided by the total number of predictions made
    # training set has POSITIVE or NEGATIVE tags, classified with labels

    print(nltk.classify.accuracy(classifier, test_set))
    #test_set
```

```
0.8065728341076797
```

```
[ ] # shows how likely feature leads to High/Positive or Low/Negative review classification
    # Output is a list

    # from NLTK's most informative features of text classifer for the Naïve Bayes Classifier:

    classifier.show_most_informative_features(10)
```

```
Most Informative Features
                   loves = True              pos : neg    =     65.5 : 1.0
                     fee = True              neg : pos    =     64.1 : 1.0
                   worst = True              neg : pos    =     53.2 : 1.0
                  poorly = True              neg : pos    =     49.9 : 1.0
                    poor = True              neg : pos    =     38.3 : 1.0
                 unhappy = True              neg : pos    =     31.4 : 1.0
                  refund = True              neg : pos    =     29.4 : 1.0
             compliments = True              pos : neg    =     29.2 : 1.0
               pantyhose = True              neg : pos    =     28.7 : 1.0
                   sucks = True              neg : pos    =     25.9 : 1.0
```

SAMSUNG INNOVATION CAMPUS • MAY 2023

WORDCLOUD:

The word cloud displays the most frequent words in each positive and negative review dataset for pants in each gender category.

POS_WORDCLOUD = WORDCLOUD().GENERATE(POS_ALLTEXT)

NEG_WORDCLOUD = WORDCLOUD().GENERATE(NEG_ALLTEXT)

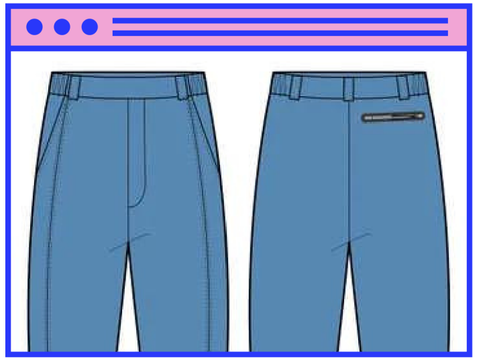
Results

In our project, we wanted to see if there are differences in reviews for pants marketed based on gender.

Most Informative Features

NLTK's Naive Bayes classifier shows how likely certain words led to a Positive or Negative classification.

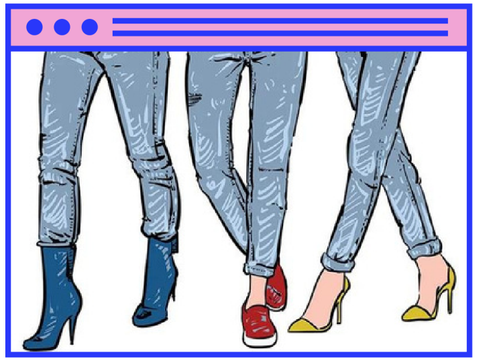
```
classifier.show_most_informative_features()
```



1. PANTS WITHOUT GENDER

HOLES
UNCOMFORTABLE
STITCHING

WARM
UNIQUE
COMFY



2. WOMEN'S PANTS

POORLY
SMELL
INCORRECT

COMPLIMENTS
VIBRANT
COMFY

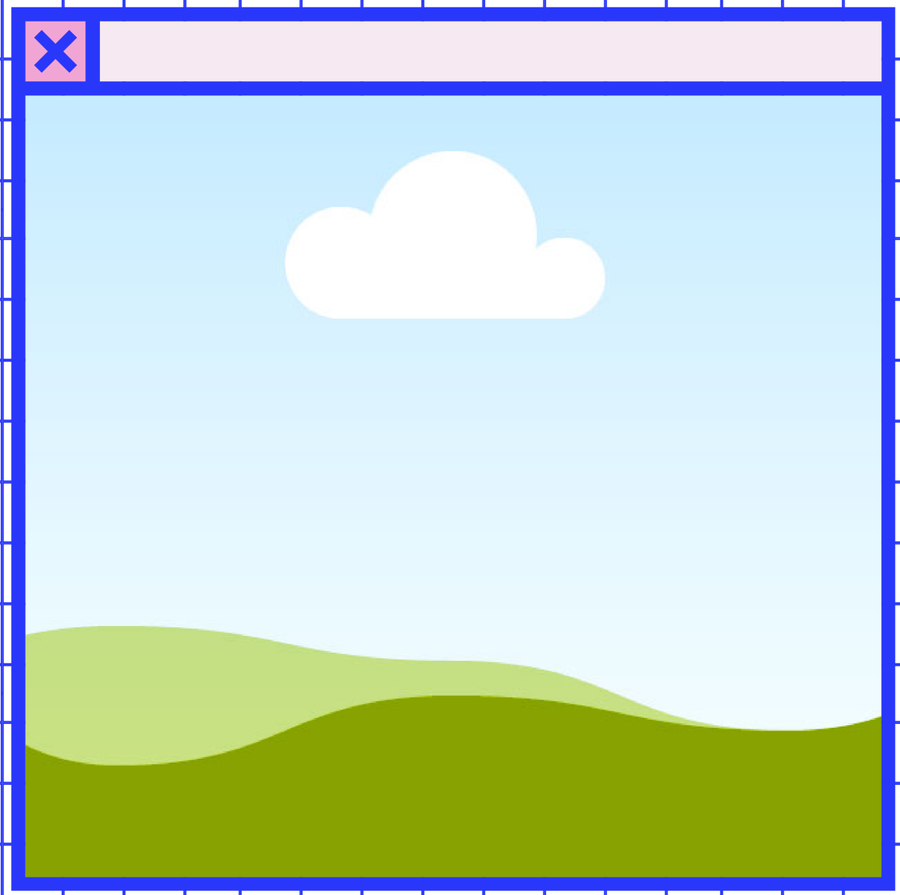
3. MEN'S PANTS

WEAK
SAGGING
RIPPED

STYLISH
LIGHTWEIGHT
COMFY

Pre-processing techniques:

- Tokenization
- Stop words
- Part of Speech (POS) Tagging
- Integer Encoding
- Padding
- One-Hot Encoding



Most Frequent Words

for all pants

for all genders

THE MOST LIKELY FEATURE IN POSITIVE REVIEWS

MATERIAL

QUALITY

COMFY

FIT

FABRIC

Women Pants v Men Pants

STEPS IN THIS PROJECT:

- THROUGHOUT THE PROJECT, AKEMI COMPILED THE DETAILS BEHIND THIS PROCESS, WROTE THE PROJECT PAPER, AND SUMMARIZED THE DETAILS BEHIND EACH STEP IN OUR TEAM'S PROJECT PAPER.

- THROUGHOUT THE PROJECT, YAFIRA SUMMARIZED THE FINDINGS VISUALLY, DESIGNED THIS PRESENTATION, AND FOUND THE SOURCE OF DATA INSTRUMENTAL TO EXTRACTING THE DATASET NECESSARY FOR THIS PROJECT.

- FIRST, WE USE VIVIAN'S APPROACH WITH NAIVE BAYES CLASSIFIER AS APPLIED HERE TO GENDERLESS PANTS TO FIND FEATURES ASSOCIATED WITH POSITIVE (HIGH) AND NEGATIVE (LOW) RATINGS IN AMAZON REVIEWS FOR WOMEN'S AND MEN'S PANTS .

- NEXT, WE USE YANNELLY'S APPROACH WITH THE WORD CLOUD TO FIND THE MOST FREQUENT WORDS TO APPEAR IN POSITIVE (HIGH) AND NEGATIVE (LOW) RATINGS ON AMAZON FOR WOMEN'S AND MEN'S PANTS.

- LATER, WE WILL USE TRINITY'S APPROACH WITH THE SUPPORT VECTOR MACHINE (SVM) MODEL TO REVISE THE ACCURACY OF THE POSITIVE AND NEGATIVE REVIEW DISTINCTIONS, AND THEN USE THE NAIVE BAYES MODEL AGAIN TO EXTRACT INFORMATION ABOUT THE DATA.

ACTIONABLE INSIGHTS

- RECOMMEND PANTS MADE WITH QUALITY MATERIAL AND FABRIC
- FOCUS ON COMFORT AND FIT IN PANTS
- EXPAND UNISEX PANTS AS A CATEGORY FOR NONBINARY, WOMEN, AND MEN AS CUSTOMERS.

COMFY + POCKETS

IMPLEMENTATION CHALLENGES

DATA COLLECTION

Amazon Reviews datasets were available, but did not have gender as a category.

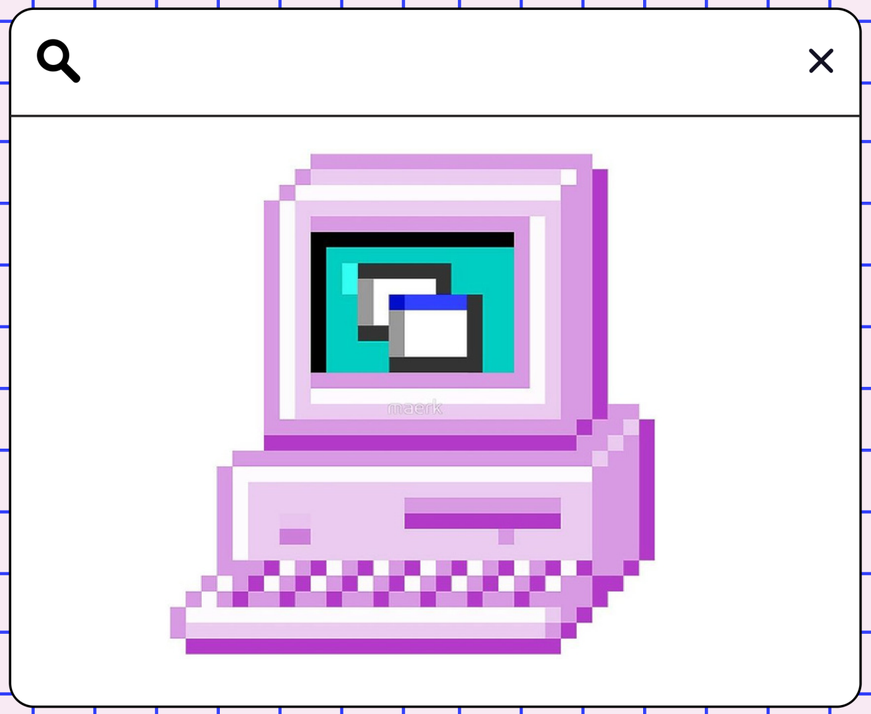
WEBSCRAPING

Webscraping would take too long, and is a resource-intensive task - CPU and memory.

FINALLY!!!

An existing Amazon Reviews dataset that Yafira linked allowed Vivian to use RegEx to extract a subset of data containing "pants" and gender in the product title.

SOLUTION



Resources

HUGGINGFACE DATASET: AMAZON_US_REVIEWS

https://huggingface.co/datasets/amazon_us_review

SENTIMENT CLASSIFICATION WITH NLTK

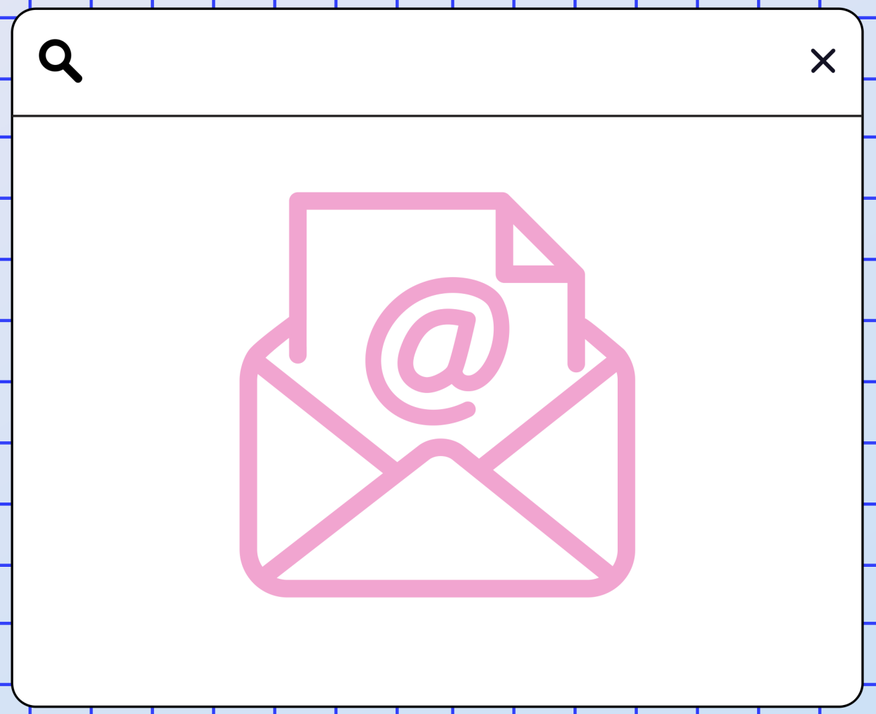
https://github.com/ashleylizg/nlp-tutorial/blob/main/main.ipynb

SAMSUNG AI COURSE

https://sic.edc.org/

SAMSUNG INNOVATION CAMPUS • MAY 2023
SAMSUNG INNOVATION CAMPUS • MAY 2023

Q&A



Contact

VIVIAN LAM

vivnlamb.ybc@gmail.com

TRINITY RENEE B.

tmr.pros@gmail.com

AKEMI VUONG

vuongakemi@gmail.com

YANNELLY MERCADO

yannellym@gmail.com

YAFIRA MARTINEZ

yfr.mrtnz@gmail.com