Creating Descriptions for Amazon Products

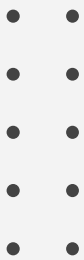By Linus Jen

01

PROBLEM STATEMENT
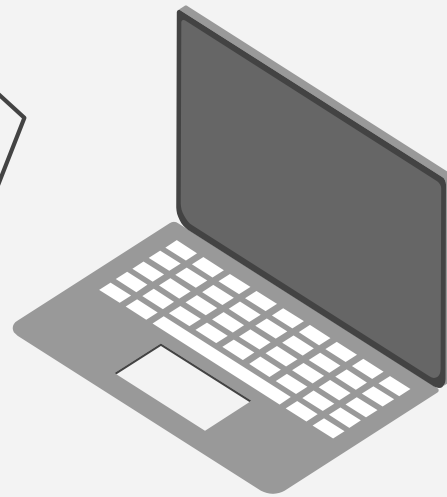
Deriving Meaningful Insights

Given a dataset of product IDs, titles, descriptions and summaries in bullet form, product type IDs, and product lengths, what valuable information can be extracted from the text?

02

SOLUTION

Using product titles, we can use language models to create descriptions for products that don't have descriptions!

WHY?

Improve Vendor Relations

By creating a base description, we can inspire sellers what to add to their descriptions

Increase Traffic to Products

Descriptions with specific keywords help search engines identify and recommend products. >50% of descriptions were empty, and ~34% didn't have any associated bullet points

03

TEHCNIQUES & ANALYSIS

IDEAL SOLUTION

**Fake Title:
Large, Insulating Glass Mug**

01 **Extract keywords from title**
Use spaCy to extract keywords,
such as:
"Insulating", "glass", "mug"

02 **Use GPT(-J) to Create Description**
"This glass mug is great for both
hold and cold drinks! Perfect for all
seasons"

GOOD EXAMPLE

Title: Constructing the Nation-State: International Organization and Prescriptive Action (Contributions in Sociology) (Product ID = 40305)

Keywords: nation, state, international, organization, prescriptive, action, contributions, sociology

Description: "Building nations is one of the fundamental questions of sociology. While the nation-state has been the primary focus of social scientists for over a century, questions about how to construct a nation have now moved to the center of sociological research. This book is an attempt to answer those questions, with an attempt to understand the nature of nations and the role of sociological theory in the construction of the nation-state."

Correctly extracts keywords

Recognizes that product is a book, and creates a summary (Could be hallucinating!)

BAD EXAMPLE

Title: Silence Adult Coloring Book: Self Stress Management and Healing in the Himalayas: 5 (Coloring Medition) (Product ID = 40305)

Keywords: silence, adult, book, self, stress, management, healing, himalayas

Description: "silence adult coloring book: self stress management and healing in the himalayas: 5 (coloring medition)

silence adult coloring book: self stress management and healing in the himalayas: 5 (coloring medition) ..."

Extracts *too many* "key"words

Model ends up repeating the title over and over again

04

LIMITATIONS & LEARNINGS

LIMITATIONS

- Large dataset: >2 million rows -> for project, limited dataset to 1K products

- LLMs are *large*! Pre-trained models + weights require lots of RAM

- Lack of other data (customer reviews, number of purchases, etc.) make it hard to clearly define a target variable

LEARNINGS

- Tested different models (GPT-2, GPT-J, attempted LLaMA)

- Explored non-sentiment analysis and non-QA tasks