Annexure

Table of Contents

Overview

In this challenge, you are presented with titles pulled from listings of items for sale on eBay.

In item titles, sellers can include any information they consider relevant. Titles are usually not sentences but rather a sequence of keywords: nouns, adjectives, dimensions, and model numbers. They may contain spelling errors, words that are not common in our everyday vocabulary or even meaningless words.

Below are a few examples of eBay titles:

- Prada Saffiano Double Zip Tote
- Authentic GUCCI Jackie Handbag leather [Used]
- Chanel Leo Lion Flap Bag Chevron Lambskin Medium

- NWOT Shoulder Tote Bag synthetic Leather , Hot Pink , Unbranded

The task is to extract named aspects from the titles. Examples of aspect names are "Brand", "Color" (generally applied to color names), "Size" (applied to words providing information about a typical size such as “Large”), "Department" (applied to gender specifications), and so on.

The list of aspect names with definitions and examples can be found in the last section of this document.

Rules & Tips

- A tag is assigned to every token in the title.
- Context matters, for example “gold” might be tagged as “Material” in a handbag handle context, but as “Color” in an exterior context.
- Misspellings and abbreviations are tagged whenever possible.
- In general, tokens that are literal aspect names such as the words "Pattern", "Color" or "Size" should not be tagged. In rare cases, however, they have a semantic function (i.e. are part of the phrase being tagged) and should be tagged. For instance, the token “Color” should be tagged in “Multi Color” (as it is a valid color name) as being part of the two-token aspect value “Multi Color” for the aspect name “Color”. In “Color Pink” only “Pink” should be tagged with the aspect name “Color”, but the token “Color” should not be tagged.
- It is important that titles are not modified in any way during the tagging process, that is, do not correct spelling errors. If a word is a spelling variation of a word that falls under a specific tag then it is tagged that way. For instance, "sansumgg" (a misspelling of the brand name "Samsung") should be tagged as “Brand”.
- If an abbreviation stands for a word that belongs to any tag then it is tagged accordingly. Example: "LV" stands for "Louis Vuitton", so it is tagged as “Brand”.
- Words which belong to multiple semantic tags are tagged with only one tag by the annotator using their best judgment for the given context.

Special Tags

In the training dataset, you will observe two other special tags besides the aspect names listed in the final section.

No Tag

- “No Tag” is used for words and punctuation that do not add meaning to the title.

- The “&” in "blue & green" should be marked “No Tag” because it serves as punctuation only. However, special characters are tagged when they add meaning to the title:
    - The "&" in "Abercrombie & Fitch" should be tagged as "Brand" because it is part of the trademarked brand name.

- Certain words in English are just connectors between other words and not part of the meaning. This will frequently be the case for prepositions “with” and “for”, while the word “of” may or may not be part of the meaning (see next bullet). Connector words are tagged as “No Tag”.

- But in other cases the preposition will be an integral part of the meaning, especially the preposition “of”: In the title “Elvis Presley Army Messenger Bag The King of Rock and Roll” the tokens “The King of Rock and Roll” are a span (belonging to the aspect “Theme”) where in this case the preposition “of” is part of the meaning.

Obscure

- “Obscure” is used for words that could not be deciphered or tagged during the human annotation process.

- Words and terms not in the native language (English for this challenge) should be tagged as “Obscure” unless:
    - The word or term is commonly used in the native language. For instance: "attaché" and "Art Nouveau" are commonly used in English.
    - The word is part of a brand name or a product name. For instance, the French brand name "Petit Bateau".

- If the majority of words in a term are not in the native language then all words should be marked as “Obscure”, unless an exception applies as listed above.

- Improperly tokenized words such as “Anniversary1998” or “Quilted12x17” are tagged as “Obscure”.

Additional Notes

Product Line vs Model

Product Line and Model can look similar, but they are effectively a hierarchy (where Product Line is higher than the Model). For example, the brand Louis Vuitton has the product line “Neverfull” with several models such as “Neverfull MM”, among others. In this case, the general Product Line is the “Neverfull” line and the Model is “Neverfull MM”. The human annotators were not always consistent in applying this distinction, and you will find “Neverfull” both as a Product Line aspect value and as a Model aspect value. No effort has been made to clean up such inconsistencies; they are part of real-world data.

Tagging Accuracy

The train / quiz / test data have been tagged by human annotators, and as such are subject to human errors, besides different annotators making different judgements for related listing titles. The resulting inconsistencies are a key part of real-world data.

Tokenization

The listing titles have been tokenized from their raw form. During tokenization a certain amount of text cleaning and transformation was performed. In particular the provided (tokenized) titles do not contain any tab / newline / linefeed characters. The provided titles are to be split on whitespace into tokens without any additional transformation, and the resulting tokens are what should be tagged. One example of note is the following. The raw title “Women’s handbag” would be tokenized as **Women**, **'s**, and **handbag**, that is, the tokenized title would contain three tokens. The first two of these should be tagged and combined as an aspect with the name “Department” with one combined value “Women 's”. Notice that there is a space in the resulting aspect value, it should not be removed in submission files.

Data Format

For all provided data files the following applies:

- Gzip compressed
- UTF-8 encoded
- Windows End-Of-Line characters: \r\n
- TAB-separated (all values are free of TAB characters)
- No CSV-style quoting (all text is presented as it is)

- Text may contain non-ASCII characters (for example ❤)

Data Layout

Listing Data

Two columns: Record Number, Title

The record numbers start at 1.

The dataset contains 19,999,921 data records and 1 header record.

Train Data

The tagged Train Data is provided in a separate file from the Listing Data.

It has four columns: Record Number, Title, Token, Tag.

The Train Data matches records 1 to 5000 of the listing data, inclusively.

The tagged Train Data contains one or more records per listing because it contains one record per token in the title.

The dataset contains 55,121 data records and 1 header record.

Quiz Data

There is no separate dataset distributed for the Quiz Data to be used for submission to the leaderboard at eval.ai.

The Quiz Data consists of records 5001 to 30000 of the listing data, inclusively.

As described elsewhere only about 2500 of these listings are evaluated for the leaderboard score. The subset of which listings exactly make up the scored records of the Quiz Data is not disclosed.

Test Data

There is no separate dataset distributed for the Test Data to be used for submission by the high-ranking teams at the end of the competition.

The Test Data consists of 25000 records of the listing data, the precise record numbers will be disclosed to the leading teams at the end of the competition.

As described elsewhere only about 2500 of these listings are evaluated for the winning score. The subset of which listings exactly make up the scored records of the Test Data is not disclosed.

NER Tagged Data Format Specification

The data used for training / evaluating NER is human-annotated data. In an item title, the data consists of tokens (aspect values) with a tag (aspect name) assigned to each. Aspect values can be composed of several tokens belonging to the same semantic entity, labeled with the aspect name.

Record Number	Title	Token	Tag
1	LOUIS VUITTON M40096 Handbag Priscilla Multi-color canvas Multi-color canvas	LOUIS	Brand
1	LOUIS VUITTON M40096 Handbag Priscilla Multi-color canvas Multi-color canvas	VUITTON	
1	LOUIS VUITTON M40096 Handbag Priscilla Multi-color canvas Multi-color canvas	M40096	MPN
1	LOUIS VUITTON M40096 Handbag Priscilla Multi-color canvas Multi-color canvas	Handbag	Type
1	LOUIS VUITTON M40096 Handbag Priscilla Multi-color canvas Multi-color canvas	Priscilla	Model
1	LOUIS VUITTON M40096 Handbag Priscilla Multi-color canvas Multi-color canvas	Multi-color	Color
1	LOUIS VUITTON M40096 Handbag Priscilla Multi-color canvas Multi-color canvas	canvas	Fabric Type
1	LOUIS VUITTON M40096 Handbag Priscilla Multi-color canvas Multi-color canvas	Multi-color	Color
1	LOUIS VUITTON M40096 Handbag Priscilla Multi-color canvas Multi-color canvas	canvas	Fabric Type
2	LOUIS VUITTON Petit Noe Drawstring Shoulder Bag Monogram Leather M42226 39SD442	LOUIS	Brand
2	LOUIS VUITTON Petit Noe Drawstring Shoulder Bag Monogram Leather M42226 39SD442	VUITTON	
2	LOUIS VUITTON Petit Noe Drawstring Shoulder Bag Monogram Leather M42226 39SD442	Petit	Model
2	LOUIS VUITTON Petit Noe Drawstring Shoulder Bag Monogram Leather M42226 39SD442	Noe	
2	LOUIS VUITTON Petit Noe Drawstring Shoulder Bag Monogram Leather M42226 39SD442	Drawstring	Closure
2	LOUIS VUITTON Petit Noe Drawstring Shoulder Bag Monogram Leather M42226 39SD442	Shoulder	Type
2	LOUIS VUITTON Petit Noe Drawstring Shoulder Bag Monogram Leather M42226 39SD442	Bag	Type
2	LOUIS VUITTON Petit Noe Drawstring Shoulder Bag Monogram Leather M42226 39SD442	Monogram	Pattern
2	LOUIS VUITTON Petit Noe Drawstring Shoulder Bag Monogram Leather M42226 39SD442	Leather	Material
2	LOUIS VUITTON Petit Noe Drawstring Shoulder Bag Monogram Leather M42226 39SD442	M42226	MPN
2	LOUIS VUITTON Petit Noe Drawstring Shoulder Bag Monogram Leather M42226 39SD442	39SD442	No Tag

Each row in the above table contains the following fields:

Record Number: An ID which is unique to each title, and is synchronized with the Listing Data. The ID will be repeated for each row which has a token belonging to that title.

Title: The text of the title.

Token: A single token belonging to the title. Note: tokens will be in the order they appear in the title.

Tag: The annotation for each token. All tokens have a tag, which might be the empty tag. If the field is empty that indicates the token in that row belongs to the same semantic entity as the token before it, in other words, the title has a multi-token entity. In this case the tag of the previous row would apply to the current token and the tokens would need to be combined with a single whitespace to obtain the corresponding aspect value. If two (or more) consecutive rows have the same non-missing entry present in the Tag field, it means they have the same tag, but are different entities, and should not be combined.

Examples

Below are the annotations of the first two listings in the Train Data shown above.

The first title is made up of 9 tokens. There are 8 aspect values in the titles, and two of them occur twice.

Consider the first two tokens “Louis” and “Vuitton” of the first listing title.

“Louis” is tagged as Brand. Because the token “Vuitton” has an empty Tag field this indicates that “Louis” and “Vuitton” are part of the same entity, and the extracted aspect will have the aspect name “Brand” and the aspect value “Louis Vuitton”.

Notes:

1. There is no limit to the number of consecutive tokens allowed for a given Aspect Value.

2. If an aspect value consists of two or more tokens then the tokens should be concatenated with spaces in between, even if that space only happens to occur as a consequence of the tokenization. For example, **Women’s** gets tokenized into the two tokens **Women** and **'s** with the resulting extracted combined aspect value being the single combined value **Women 's** with a space.

Consider the last four tokens “Multi-color”, “canvas”, “Multi-color”, and “canvas” of the first listing title.

These are repeated aspects, and they all need to be extracted. As such the last four extracted aspects would contain two duplicate aspects. They should not be dropped.

The final set of records from the first listing title in submission format (for the Quiz Data for upload to eval.ai, and also for the Test Data for the leaders at the end of the challenge) is given below.

Record Number	Aspect Name	Aspect Value
1	Brand	LOUIS VUITTON
1	MPN	M40096
1	Type	Handbag
1	Model	Priscilla
1	Color	Multi-color
1	Fabric Type	canvas
1	Color	Multi-color
1	Fabric Type	canvas

Notes:

1. The submission files should contain three tab-separate fields, and should not have a header line. The above inclusion of a header is only to illustrate the meaning of the columns.

2. The order in which the records appear in the submission file does not matter.

3. If there are multiple extractions for a given aspect name then they all need to be included, even if the value is the same.

The second listing title consists of 11 tokens, of which two have a missing entry in the Tag field, meaning those two rows are to be combined with their respective preceding rows. The second title also has two consecutive rows with the same tag “Type”, these should not be concatenated into a single aspect, but rather should be parsed into two separate aspects with the same aspect name. (Whether this is or is not correctly tagged by the human annotators is not debated here, the annotation tags are provided “as is” in this real-world dataset.) Finally, this listing has a “No Tag” token, which should be excluded from any submission file.

The final set of records from the second listing title in submission format (to eval.ai for the Quiz Data or for the leaders for the Test Data) is given below.

Record Number	Aspect Name	Aspect Value
2	Brand	LOUIS VUITTON
2	Model	Petit Noe
2	Closure	Drawstring
2	Type	Shoulder
2	Type	Bag
2	Pattern	Monogram
2	Material	Leather
2	MPN	M42226

Notes:

1. The submission files should contain three tab-separate fields, and should not have a header line. The above inclusion of a header is only to illustrate the meaning of the columns.

2. The order in which the records appear in the submission file does not matter.

3. If there are multiple extractions for a given aspect name then they all need to be included, even if the value is the same.

List of Aspects Names

The table below gives the aspect names to be extracted, along with descriptions and example values. Note that the two tags "No Tag" and "Obscure" described previously in this document are not in this table, and should not be submitted.

Aspect Name	**Definition and Examples**
Accents	Designates non-functional extra parts of the products.

	Examples: Beaded, CC Logo, Embossed, Embroidered, Embroidery, Empreinte, Logo, Quilted, Studded, Tassel
Brand	The brand, designer, artist for a product (may be the same or different from the manufacturer). Examples: CHANEL, Chanel, Coach, GG, GUCCI, Gucci, LOUIS VUITTON, Louis Vuitton, LV, Michael Kors
Character	The recognized character that this product has on the packaging or on the product itself. Examples: Harry Potter, Hello Kitty, Mickey, Mickey Mouse, Minnie, Minnie Mouse, Minnie Witch, Sleeping Beauty, Snoopy, Winnie the Pooh
Character Family	The recognized character family that this product has on the packaging or on the product itself. Examples: Alice In Wonderland, Avatar, Beauty and the Beast, Hello Kitty, Lady and The Tramp, Little Mermaid, My Little Pony, Star Wars, The Nightmare Before Christmas, Tinkerbell, Toy Story
Closure	Describes the type of closure of a product. Examples: Buckle, Double Zip, Drawstring, Lock, Top Zip, Turnlock, ZIP, Zip, Zipper, Zippered
Color	The color of the product itself (not the packaging). The manufacturer-specific color name.

	Examples: Beige, Black, Blue, Brown, Gold, Green, Navy, Pink, Red, White
Country/Region of Manufacture	Designates where the product was manufactured, not the country where it is sold. Examples: France, ITALIAN, Italian, ITALY, Italy, Japan, Japanese, Paris, US, USA
Department	Describes the age, gender, etc. of the intended user of a product. Examples: Female, Girl, Girls, Ladies, Lady, Unisex, Woman, Women, Women 's, Womens
Fabric Type	Indicates the type of fabric by construction (not the material constituents or fiber contents). Examples: Braided, CANVAS, Canvas, canvas, Coated Canvas, Denim, Jacquard, Mesh, Tweed, Woven
Features	Design or stylistic features of the product. Examples: 2WAY, 2Way, 2way, Adjustable, Convertible, Half Moon, Quilted, Soft, Vegan, Waterproof
Handle Drop	Measures the length between a handbag handle and the top of the handbag.

	Examples: 11-inch, 24CM, Long Drop, SHORT, Strap 47 "
Handle Style	Designates the style of handle on a handbag. Examples: Handle, Handles, Shoulder Strap, Single, Sling, STRAP, Strap, strap, Top Handle, Wristlet
Handle/Strap Material	The material of the handle or strap. Examples: Bamboo, CHAIN, Chain, chain, LEATHER, Leather, leather, Metal, Rope, Wooden
Hardware Material	Designates the material of hardware on a product. Examples: 24k, Brass, Gold, Gold Plated, Gold-Plated, Gunmetal, Metal, metal, Palladium, Ruthenium
Lining Material	Designates the inner material of a product. Examples: lamb, Nylon, Satin, silky
MPN	Manufacturer Part Number: This may be the same as the model number or part number for a product. Examples: M51130, N51109

Material	Describes the material of the product. Use Material only when a more specific material tag (such as lining material) does not apply. Examples: Epi, Epi Leather, Faux Leather, Lambskin, LEATHER, Leather, leather, Nylon, PVC, PVC Leather
Measurement, dimension	Length, height, width, or other measurements of a handbag. Also weight and volume. Dimension should be explicit, not arbitrary (e.g. "large" is not a dimension). Often will be measured in some kind of standard units (inches, cm, km, g, miles, lbs, fl oz, etc.) but not required. Excludes: Sizing information without an (explicit or implied) unit of measure and non-numeric clothing sizes, and also words such as "size", "length", "width", etc. unless part of the full size specification. Examples: 11 ", 12 ", 18 ", 25L
Model	The brand's specific model name. Examples (from the brand Louis Vuitton): GM, MM, Neverfull MM, PM, Speedy 30 Examples (from the brand Coach): HAMPTONS, Kristin, Legacy, Peyton, Willow Examples (from the brand Michael Kors): Bedford, Hamilton, KENLY, Mercer, Selma
Occasion	Designates the occasion a product might be used. Please note that Season is a separate tag.

	Examples: Beach, Business, Casual, EVENING, Evening, Party, School, TRAVEL, Travel, Work
Pattern	The pattern design on the product. Construction processes that create some pattern-like visuals such as “quilted” should NOT be considered a pattern. Examples: Check, Floral, Flower, Intrecciato, Matelasse, MONOGRAM, Monogram, Pebbled, Signature, Zucca
Pocket Type	Designates the style and placement of a product pockets. Examples: Cargo, Concealed, CONCEALED CARRY, Concealed Carry, concealed carry, Concealment, Front Pocket, Multi Pocket, Pouch, Side Pockets
Product Line	The manufacturer collection or collaboration that the product belongs to. Examples (from the brand Louis Vuitton): DAMIER, Damier, Keepall, Pochette, Vernis Examples (from the brand Coach): East West, Legacy, MADISON, Madison, Soho Examples (from the brand Michael Kors): CHARLOTTE, Charlotte, East West, JET SET, Jet Set
Season	Designates the appropriate season for a product. Examples: Autumn, Spring, SUMMER, Summer, summer, Winter

Size	The size of the product, measured in diameter or a standard term that dictates an accepted measurement. This is for non-numeric measurements only, not for numeric sizes such as US size 6 (in shoe sizes), for example. Examples: Giant, LARGE, Large, Medium, MINI, Mini, mini, SMALL, Small, XL
Strap Drop	Measures the length between a handbag strap and the top of the handbag. Examples: 12CM, 18 ", 23 "
Style	Designates the style of the product. Examples: Belt, Boho, Dome, Hippie, Retro, Round, Slim, Vanity, Waist, Western
Theme	Designates a similar theme the product can be grouped with. Examples: 90s, Animal, Anime, Cartoon, Frida Kahlo, Indian, Lolita, Luxury, Novelty, Tribal
Trim Material	Material embellishments on the outer edges of the item. Examples: camel leather, Fur, LEATHER, Leather, leather, Lizard-Trimmed, Natural leather, Saffiano Leather, SHEARLING, Suede Leather

Type	Basic item type. Examples: Backpack, BAG, Bag, bag, Crossbody, Hand Bag, Handbag, Purse, Shoulder, Tote