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:

- Damen Sneaker Wedges Schnürer Keilabsatz Plateau Stoff 836273 Schuhe
- ARKK Copenhagen Damensneaker - 60 % unter UVP

- PUMA Tsugi Jun Cubism EU 42,5 US 9,5 Laufschuhe 365490-02 Schwarz Low Tops
- Puma Streetballer Mid Winter Leder burnt olive Schuhe grün 358798 03

The task is to extract named aspects from the titles. Examples of aspect names are "Marke" (English: brand), "Farbe" (color), generally applied to color names, "EU-Schuhgröße" (EU Shoe Size), "Abteilung" (Department), generally applied to gender specifications, and so on. Note that the aspect names are in German.

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 "New" might be tagged as "No Tag" in "New shoes", but as part of "Marke" (brand) in "New Balance".
- Misspellings and abbreviations are tagged whenever possible.
- In general, tokens that are literal aspect names such as the words "Muster" (pattern), "Farbe" (color) or "Größe" (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. In "Farbe Weiß" (color white) only "Weiß" should be tagged with the aspect name "Farbe" (color), but the token "Farbe" 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, "Herrenshuhe" (a misspelling of "Herrenschuhe" (men's shoes)) should be tagged as "Abteilung" (department).
- 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 "Marke" (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 "Schwarz & Weiß " (black & white) 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 German (as in English) are just connectors between other words and not part of the meaning. This will frequently be the case for prepositions "mit" (with) and "für" (for), while the word "der" (the) or "die" (the) 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 article "der" (the) or "die" (the): In the title "Sneaker für die ganz Kleinen" (sneaker for the very small ones [children]) the token "die" is a connector and should be tagged as "No Tag", while in the span "Zurück In Die Zukunft" (Back To The Future) the article "die" 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 (German for this challenge) should be tagged as "Obscure" unless:
    - The word or term is commonly used in the native language. This in particular applies to English words which are nowadays very frequently used by German speakers.
    - The word is part of a brand name or a product or model name. For instance, the Sketchers model name "Summits".

- Improperly tokenized words such as "All-Star-Schuhe" or "LaufschuheT642N" are tagged as "Obscure".

Additional Notes

Product Line vs Model

"Produktlinie" (product line) and "Modell" (model) can look similar, and while they are effectively a hierarchy (where product line is higher than the model), it isn't always clear what is what. For example, the brand Reebok has a "Royal" product line, a "Classic" product line, and also has "Royal Classic" shoes. Given a listing for "Reebok Royal Classic Jogger 2.0" the question arises whether "Royal Classic" is the product line and "Jogger 2.0" is the model, or "Royal" is the product line and "Classic Jogger 2.0" is the model. And what about "Reebok Classic Royal Glide LX", is "Classic" the product line and "Royal Glide LX" the model, or is "Classic Royal" the product line and "Glide LX" the model? The human annotators were not always consistent in applying the product line / model distinction, and you will find the token "Classic" both as (part of) product line aspect values and as (part of) model aspect values. 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 Sneakers" would be tokenized as **Women**, **'s**, and **Sneakers**, that is, the tokenized title would contain three tokens. The first two of these should be tagged and combined as a single aspect with the name "Abteilung" (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 and not at 0.

The dataset contains 10,000,000 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 55,184 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 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 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.

Each row in the above tagged data file 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. Note that record numbers start at 1 and not at 0.

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 several listings from the Train Data.

Example 1

Record Number	Title	Token	Tag
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Supreme	Modell
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Nike	Marke
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	SB	Produktlinie
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Dunk	
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	High	Schuhschaft-Typ
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	By	Modell
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	any	
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Means	
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Red	Farbe
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	US10	US-Schuhgröße
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	EU44	EU-Schuhgröße
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Supreme	No Tag
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Box	No Tag
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Logo	Akzente
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Air	Produktlinie
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Force	

The title is made up of 16 tokens. There are 9 meaningful aspect names in the tag column for this listing (plus "No Tag"), and one of them occurs twice. There are also tokens with missing tags (NaN if pandas is used to read the data, shown empty in the screenshot above to better illustrate the point).

Consider the first third and the fourth tokens "SB" and "Dunk". "SB" is tagged as "Produktlinie" (product line). Because the token "Dunk" has an empty tag field this indicates that "SB" and "Dunk" are part of the same entity, and the extracted aspect will have the aspect name "Produktlinie" (product line) and the combined aspect value "SB Dunk".

For this listing the first token "Supreme" and the three consecutive tokens "By any Means" are tagged as "Modell" (model). These are repeated aspects, and they all need to be extracted. They should not be dropped. The same applies to the second product line value in this listing, "Air Force".

There is also a second occurrence of the token "Supreme" which is labeled as “No Tag”, and therefore should not be extracted.

Whether any of these are or are not correctly tagged by the human annotators is not debated here, the annotation tags are provided "as is" in this real-world dataset.

Notes:

1. There is no limit to the number of consecutive tokens allowed for a given Aspect Value. For example the listing above has a "Modell" (model) aspect with the value consisting of the three consecutive tokens "By any Means".

2. If an aspect value consists of two or more tokens then the tokens should be concatenated with spaces in between to form the aspect value, 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.

The final set of records for this 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	Akzente	Logo
1	EU-Schuhgröße	EU44
1	Farbe	Red
1	Marke	Nike
1	Modell	Supreme
1	Modell	By any Means
1	Produktlinie	SB Dunk
1	Produktlinie	Air Force
1	Schuhschaft-Typ	High
1	US-Schuhgröße	US10

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 (an example of this is shown below).

Example 2

Record Number	Title	Token	Tag
2	New Balance 530 Männer und Frauen Laufschuhe mit Buchstaben N bequeme Laufschuhe	New	Marke
2	New Balance 530 Männer und Frauen Laufschuhe mit Buchstaben N bequeme Laufschuhe	Balance	
2	New Balance 530 Männer und Frauen Laufschuhe mit Buchstaben N bequeme Laufschuhe	530	Modell
2	New Balance 530 Männer und Frauen Laufschuhe mit Buchstaben N bequeme Laufschuhe	Männer	Abteilung
2	New Balance 530 Männer und Frauen Laufschuhe mit Buchstaben N bequeme Laufschuhe	und	No Tag
2	New Balance 530 Männer und Frauen Laufschuhe mit Buchstaben N bequeme Laufschuhe	Frauen	Abteilung
2	New Balance 530 Männer und Frauen Laufschuhe mit Buchstaben N bequeme Laufschuhe	Laufschuhe	Produktart
2	New Balance 530 Männer und Frauen Laufschuhe mit Buchstaben N bequeme Laufschuhe	mit	No Tag
2	New Balance 530 Männer und Frauen Laufschuhe mit Buchstaben N bequeme Laufschuhe	Buchstaben	Muster
2	New Balance 530 Männer und Frauen Laufschuhe mit Buchstaben N bequeme Laufschuhe	N	
2	New Balance 530 Männer und Frauen Laufschuhe mit Buchstaben N bequeme Laufschuhe	bequeme	No Tag
2	New Balance 530 Männer und Frauen Laufschuhe mit Buchstaben N bequeme Laufschuhe	Laufschuhe	Produktart

This listing title has two rows with the same token "Laufschuhe" (running shoes) and the same tag "Produktart" (product type), these should be parsed into two separate aspects with the same aspect name and same aspect value. The duplicate should not be dropped.

The final set of records from this 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	Abteilung	Männer
2	Abteilung	Frauen
2	Marke	New Balance
2	Modell	530
2	Muster	Buchstaben N
2	Produktart	Laufschuhe
2	Produktart	Laufschuhe

Example 3

Record Number	Title	Token	Tag
24	Herren Damenschuhe Laufschuhe Atmungsaktiv Mesh Running Shoes Sneaker Gr.35-45	Herren	Abteilung
24	Herren Damenschuhe Laufschuhe Atmungsaktiv Mesh Running Shoes Sneaker Gr.35-45	Damenschuhe	Abteilung
24	Herren Damenschuhe Laufschuhe Atmungsaktiv Mesh Running Shoes Sneaker Gr.35-45	Laufschuhe	Produktart
24	Herren Damenschuhe Laufschuhe Atmungsaktiv Mesh Running Shoes Sneaker Gr.35-45	Atmungsaktiv	Besonderheiten
24	Herren Damenschuhe Laufschuhe Atmungsaktiv Mesh Running Shoes Sneaker Gr.35-45	Mesh	Gewebeart
24	Herren Damenschuhe Laufschuhe Atmungsaktiv Mesh Running Shoes Sneaker Gr.35-45	Running	Aktivität
24	Herren Damenschuhe Laufschuhe Atmungsaktiv Mesh Running Shoes Sneaker Gr.35-45	Shoes	Produktart
24	Herren Damenschuhe Laufschuhe Atmungsaktiv Mesh Running Shoes Sneaker Gr.35-45	Sneaker	Stil
24	Herren Damenschuhe Laufschuhe Atmungsaktiv Mesh Running Shoes Sneaker Gr.35-45	Gr.35-45	EU-Schuhgröße

This listing title also has two consecutive rows with the same tag "Abteilung" (department).These should not be concatenated into a single aspect, but rather should be parsed into two separate aspects with the same aspect name because they are both separately tagged (and are not a tag followed by an empty tag).

The final set of records from this 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
24	Abteilung	Herren
24	Abteilung	Damenschuhe
24	Aktivität	Running
24	Besonderheiten	Atmungsaktiv
24	EU-Schuhgröße	Gr.35-45
24	Gewebeart	Mesh
24	Produktart	Laufschuhe
24	Produktart	Shoes
24	Stil	Sneaker

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
Abteilung (Department)	Gender and/or age grouping characterized by the category. Examples: DAMEN, Damen, Damenschuhe, HERREN, Herren, Herren Damen, Herrenschuhe, Unisex, W, M
Aktivität (Performance/Activity)	Type of activity the product is best suited for. Examples: Basketball, FITNESS, Fitness, Laufen, Running, Skate, Skater, Tennis, Trail, Trekking
Akzente (Accents)	Attributes that give the product its distinctive look. Examples: Cut Out, Glitter, Glitzer, Logo, Nieten, Pailletten, Print, Prints, Spitze, Strass
Anlass (Occasion)	Occasion or celebration the product is affiliated with. Examples: CASUAL, Freizeit, Freizeit Sport, Gym, Outdoor, Sport, Sport Freizeit, Sportliche
Besonderheiten (Features)	Secondary attributes or functions that are not essential to the product’s main function. Examples: Atmungsaktiv, gefüttert, Leicht, Profil-Sohle, Profilsohle, Ultraleicht, Warm Gefütterte
Charakter (Character)	Recognized character that the product has on itself or on the packaging. Examples: Beauty And The Beast, Bruce Lee, Han Solo, Hello Kitty, Pepe Le Pew
Charakter Familie (Character Family)	Recognized character family that the product has on itself or on the packaging. Examples: DISNEY, Looney Tunes, Pokemon, Sanrio, STAR WARS, Star Wars
Dämpfungsgrad (Cushioning Level)	Level of shock absorption of the product. Examples: Airsoft, Barefoot, Barfuß, cushion, Luftpolster, Luftpolster Airsoft, Luftpolstersohle
Erscheinungsjahr (Release Year)	Year the product was released by the manufacturer. Examples: 2020, 2018, 2007
EU-Schuhgröße (EU Shoe Size)	Size of the shoes, using European standard sizes. Examples: 36, 38, 39, 40, 41, 42
Farbe (Color)	Main color of the product itself and other prominent colors. This doesn't include the product's packaging. Examples: beige, Black, black, Blau, Braun, braun, Grau, Grün, Neon
Futtermaterial (Lining Material)	Main material of the product's lining. Examples: Fell, Fleece Fellfutter, Fur, Kunstfell, Kunstpelz, Lammfell, Textilfutter
Gewebeart (Fabric Type)	Type of fabric by construction, not the material constituents or fiber contents. Examples: canvas, Denim, Grob, Lack, Mesh, Netz, Optik, Strick

Herstellernummer (Style Code)	Style Code (may also be called "MPN" or "Manufacturer Part Number") is a product identifier given by the manufacturer, can be the same as the model number or part number. Characterized by a combination of numbers, letters, and/or symbols. Examples: 1339-14, 365208, CT8527-114, M7652C, ML574EAG, V94M
Herstellungsland und -region (Country/Region of Manufacture)	Geographic location where the product is manufactured. Examples: Germany, DE, West Germany, DDR, Italien, Italy, Portugal, England, USA
Innensohlenmaterial (Insole Material)	Main material of the product's insole. Examples: EVA-Sohle, Gel, Laufsohle, Lederfußbett, Luftkissen, Memory, Memosoft, OrthoLite
Jahreszeit (Season)	Time of year the product is intended to be worn, characterized by season name. Examples: Herbst, Herbst Winter, Sommer, Summer, Winter, Winterschuhe
Laufsohlenmaterial (Outsole Material)	Main material of the product's outer sole. Examples: Cupsole, Gummi, Gummisohl, Gummisohle, Gummschalensohle
Marke (Brand)	Name of the brand, designer, or artist that produces the product. This may be the same or different from the manufacturer. Examples: Adidas, adidas, Asics, Converse, New Balance, NIKE, Nike, Puma
Maßeinheit (Unit of Measure)	Units of measure, such as a length in inches/cm, a weight (pounds, grams, kg, etc.), or other measurements. Note that the German language uses a comma "," where English uses a decimal point ".", and conversely uses a period "." instead of a comma "," to separate thousands, for example 1,234.56 in English is 1.234,56 in German. Examples: 26 cm, 27,5 cm, 28 cm, 28,0 cm
Modell (Model)	Brand or manufacturer’s specific name used for the product. Examples: 70, Retro, Smash, Smash v2, ST, ZX 8000, 1 Retro, Classic, III, Mexico 66, Plus, Quantum
Muster (Pattern)	Pattern on the product. Examples: Camo, Camouflage, Graffiti, Leopard, Snake, Zebra
Obermaterial (Upper Material)	Main material of the product's upper component. Examples: Echtleder, Knit, Leather, Leder, Stoff, Suede, Synthetik
Produktart (Type)	Specific type of product that is being sold in the product listing. Examples: Sneaker, Freizeitschuhe, Halbschuhe, Laufschuhe, Sportschuh, Sportschuhe, Trainers
Produktlinie (Product Line)	Manufacturer collection or collaboration that the product belongs to. Examples: Air Force 1, Air Jordan, Air Max, Chuck Taylor All Star, Classic, Flex, Gel, Yeezy

Schuhschaft-Typ (Shoe Shaft Style)	Distinct design appearance of the product characterized by height. Examples: Hi, High, High Top, low, Low Top, Low-Top, Mid
Schuhweite (Shoe Width)	Measured horizontal distance from side to side of the shoe. Often but not always a single capital letter. Examples: B, D, G, G-Weite, H, K, WIDE
Stil (Style)	Distinct design appearance (shape) of the product. Examples: Ballerina, Ballerinas, Keilabsatz, Sneaker, Sneakers
Stollentyp (Cleat Type)	Type of cleats. Examples: Spikes, Nokken, Schraubstollen
Thema (Theme)	Type of visual style or design subject of the product, but NOT the shape of the product: see "Stil (Style)". Examples: 90er, Retro, Retro Vintage, Sportlich, Vintage
UK-Schuhgröße (UK Shoe Size)	Size of the shoes, using UK standard sizes. Note that the German language uses a comma "," where English uses a decimal point ".", for example the size UK 11,5 in German is UK 11.5 in English. Examples: UK 10, UK 11, UK 11,5, UK 7
US-Schuhgröße (US Shoe Size)	Size of the shoes, using US standard sizes. Note that the German language uses a comma "," where English uses a decimal point ".", for example the size US 11,5 in German is US 11.5 in English. Examples: US 10, US 11, US 11,5, US 12, US 8
Verschluss (Closure)	Type of closing mechanism the product uses. Examples: Klett, Klettverschluss, Lace Up, Lace-Up, Reißverschluss, Schnür, Schnüren
Zwischensohlen-Typ (Midsole Type)	Type of supportive structure that is layered between the shoe's insole and outsole Examples: Air, Cloudfoam, Croslite, Dämpfung, Federsohle, Foam, Memory Foam, Soft Foam