MOTIVATION.

[question1] For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
[answer1] The dataset in its current form was created with the purpose of helping solve an industrial challenge of optimal catalogue structure prediction.
[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
[answer2] Original raw data collection was performed as part of the day-to-day operations of the company Tjek A/S, which aggregates product catalogues for viewing in a digital format. The curation and preprocessing was performed by the authors of this paper.
[question3] Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
[answer3] The research is funded through an Innovation Fund Denmark research grant that Tjek A/S is a beneficiary of (grant number 9065-00017B).
[question4] Any other comments?
[answer4] No

COMPOSITION.

[question5] What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
[answer5] The instances represent 3 types of entities. The most atomic entity is an offer, which represents a specific product with a text heading and description, which often includes its on-offer price. Individual product offers are then grouped into sections, which represent pages in a physical catalogue brochure. Finally, an ordered list of sections comprise a single catalogue, for which a prediction about its optimal structure is made. This takes the form of permuting the input set of offers into an ordered list, with section breaks marking the start and end of a section.
[question6] How many instances are there in total (of each type, if appropriate)?
[answer6] The dataset consists of just over 10 thousand catalogs (11063), almost a quarter of a million sections (238256) and over 1.5 million offers (1613686). These are further grouped into a suggested 80/20 train and test split, with 8850 catalogs in the train set and 2212 in the test set.
[question7] Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).
[answer7] The dataset is not a sample, it contains all catalogue instances from the years 2015 -

2019 available for viewing in the Tjek A/S app. No other selection filter was used.

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

[answer8] Each instance consists of both raw data and pre-processed features.
Each offer instance consists of its unique id, its related section and catalogue ids, a text heading and description in both raw form and as word tokens using the nltk tokenizer [Bird, 2006], the total token count, and finally the full offer text as a vector referencing a vocabulary of 300 thousand word tokens. Additionally, each offer is categorized into a priority class, representing how visually prominent it was in the original catalogue in terms of relative image size (on a 1-3 integer scale).
Each catalogue instance consists of its unique id, an ordered list of associated section ids, and an ordered list of offer ids that comprise the catalogue in question, including section break markers. Additionally, each catalogue instance also includes information in the form of ordered lists of offers as vectors, grouped into sections, their corresponding priority class and the catalogue's total number of offers. Finally a shuffled x of offer vectors (with section breaks) is provided for each catalogue, along with the target y representing the permutation required to restore the original order.

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9] Yes, each catalogue instance is pre-processed into a shuffled x of offer vectors and section break markers, along with the target y representing the permutation required to restore the human-designed structure of the original catalogue.

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

[answer10] No data is missing.

[question11] Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

[answer11] Yes, every offer instance is tied to its section and catalogue via their ids in the appropriate columns of the provided comma-separated files.

[question12] Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

[answer12] Yes, the entire catalogue set is grouped into a suggested 80/20 train and test split, with 8850 catalogs in the train set and 2212 in the test set. Catalogues were assigned to each group randomly. A validation set can be extracted from the train set based on each researcher's individual preference.

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13] There are no known errors, sources of noise or redundancies in the dataset, however there is a possibility of some degree of overlap between individual offers in terms of the underlying product.

[question14] Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or

relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
[answer14] The dataset is self-contained.
[question15] Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
[answer15] The dataset does not contain data that might be considered confidential.
[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
[answer16] The dataset does not contain data that the authors would consider offensive, insulting, threatening or causing anxiety.
If the dataset does not relate to people, you may skip the remaining questions in this section.
[SkipA] YES
[question17] Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
[answer17]
[question18] Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.
[answer18]
[question19] Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
[answer19]
[question20] Any other comments?
[answer20]

COLLECTION PROCESS.

[question21] How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

[answer21] The data was acquired through a combination of feed readers and custom scraping scripts developed by Tjek A/S. For further details, see the answer to the next question.

[question22] What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

[answer22] The scripts read the feeds and scrape a list of stores and PDF catalogs associated with said stores. This provides the basic tooling and processing of the data and communicates this to the company's core API, running the scrapers on a defined schedule as well as on-demand. Following that, a human curation step is performed by the operations department to make sure the obtained data is correct. The data is directly observable.

[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?

[answer23] The dataset is not a sample.

[question24] Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

[answer24] The data collection process was done as part of the day-to-day operations of Tjek A/S, by properly compensated full-time employees.

[question25] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

[answer25] The data was collected within the full 4 year period between 2015 and 2019.

[question26] Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

[answer26] No

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipB] YES

[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?

[answer27]

[question28] Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how

notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
[answer28]
[question29] Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
[answer29]
[question30] If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
[answer30]
[question31] Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
[answer31]
[question32] Any other comments?
[answer32]

PREPROCESSING/CLEANING/LABELING.

[question33] Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.
[answer33] Yes, the raw text features of each offer instance were tokenized using the nltk tokenizer [Bird, 2006], a vocabulary of word tokens was limited to 300 thousand words and used to obtain offer vectors. Each offer instance was truncated or padded to 30 word tokens, with over 75% of offers consisting of fewer than 24 tokens. Each catalogue instance was truncated or padded to 200 offer instances, with over 75% of catalogues consisting of fewer than 163 offers.
Additionally, to obtain the prominence class per offer per section, signifying the relative size of the offer's image on the page, a proprietary algorithm was used.
[question34] Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
[answer34] Yes, raw data is also provided.
[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.
[answer35] Yes, the nltk library is available under the Apache License 2.0.

[question36] Any other comments?
[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.
[answer37] The dataset is actively being used to help predict the optimal structure of product catalogues given a provided set of offers, based on their textual description and to recommend complementary offers. It has not been used in prior research.
[question38] Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
[answer38] The repository containing the scripts for repeated experiments will include links to any and all papers using this dataset. For more information, see the appendix subsection A.1.
[question39] What (other) tasks could the dataset be used for?
[answer39] The dataset can be used for representation learning through the co-occurrence of offers within the same section, leading to a complementariness-based recommendation system. It can also be used for learning to cluster a set of offers into a variable number of sections, which is an implicit step in the main task of predicting the entire structure of a catalogue through permutation learning (as it includes the section break markers).
[question40] Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
[answer40] It is important to remember that the provided catalogues represent the Danish market between 2015-2019, and thus might not represent patterns that will hold in other societies. This, however, has no bearing on demonstrating a machine learning model's ability to learn structure through joint clustering and permutation learning, which is the intended use of the dataset.
[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.
[answer41] The dataset is not meant to be used as a representation of the market for any form of trend prediction.
[question42] Any other comments?
[answer42] No

DISTRIBUTION.

[question43] Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

[answer43] The dataset will be made publicly available under the chosen license to any and all parties. For more information see the appendix subsection A.1.

[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

[answer44] The dataset is distributed through a dataset hosting service and has a DOI, for details see the appendix subsection A.1.

[question45] When will the dataset be distributed?

[answer45] The dataset will be distributed by the time of the paper's submission.

[question46] Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

[answer46] The dataset will be distributed under the Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0). The dataset should not be used for commercial purposes

[question47] Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

[answer47] No

[question48] Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

[answer48] No

[question49] Any other comments?

[answer49] No

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?

[answer50] The dataset is hosted by figshare, an open access repository where researchers can preserve and share their research outputs, including figures, datasets, images and videos.
It is supported by Digital Science & Research Solutions Ltd.
It is maintained by the authors of this paper.

[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?

[answer51] Via the emails provided in the contact information above the abstract, repeated here for convenience: maju@itu.dk; leod@itu.dk.

[question52] Is there an erratum? If so, please provide a link or other access point.
[answer52] There is currently no erratum, it will be added to both the main sharing link and the github repository containing the code for repeated experiments should the need to create an erratum occur.
[question53] Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?
[answer53] If labeling errors are found, they will be corrected. The dataset may be expanded with further instances, depending on the academic interest and number of downloads.
[question54] If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
[answer54] The dataset does not relate to people.
[question55] Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
[answer55] Yes, all previous versions of the dataset will continue to be available.
[question56] If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.
[answer56] Others are encouraged to extend the dataset and can choose to either do so in cooperation with the authors of this paper after contacting them via the provided email addresses or individually in accordance with the chosen license.
[question57] Any other comments?
[answer57] No