CAHL

Computational Approaches to Human Learning

Design and deployment of a better university course search: Inferring latent keywords from enrollment networks

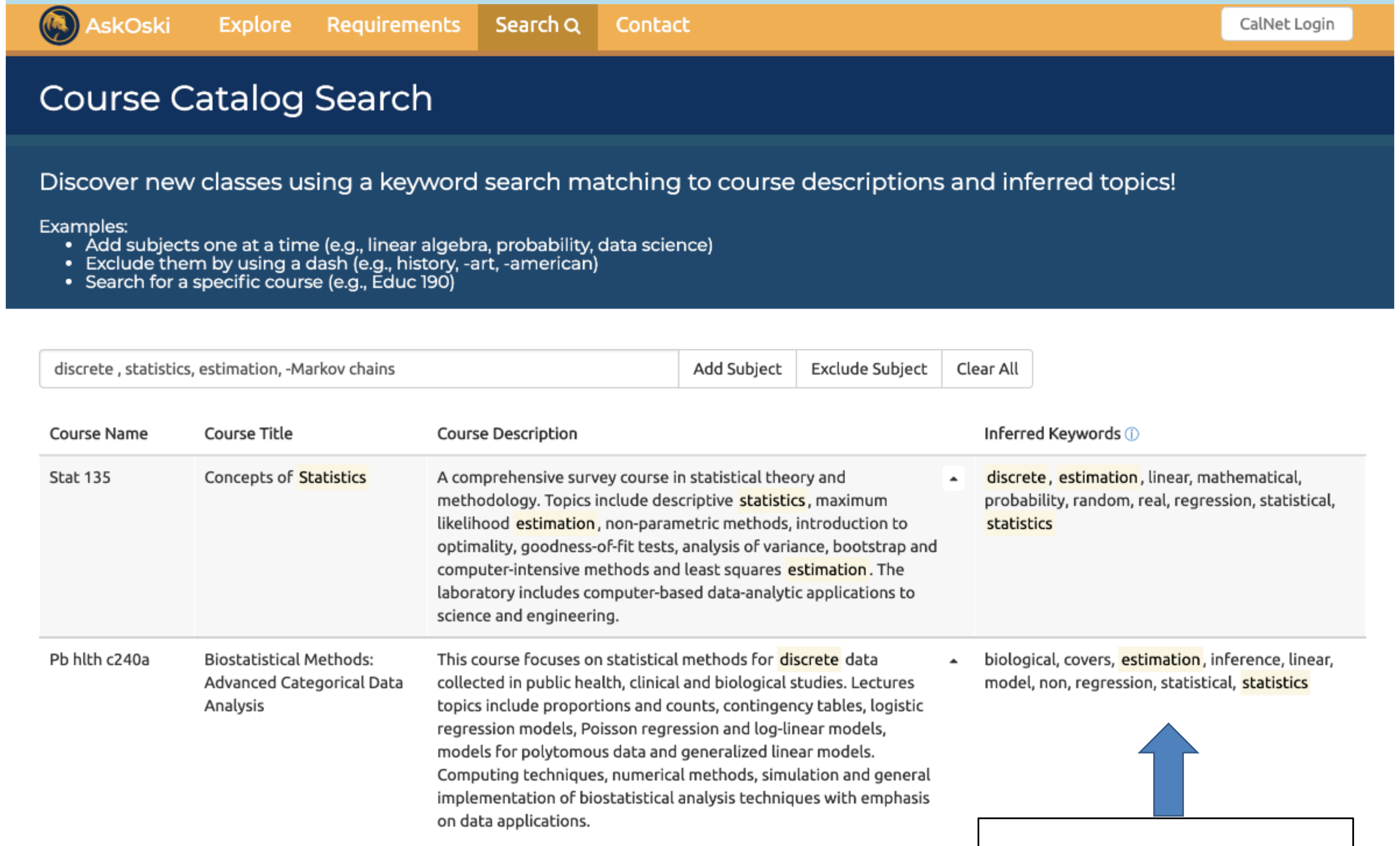
Matthew Dong[1]; Run Yu[2]; Zachary Pardos[1]

[1]University of California, Berkeley, [2]Wuhan University

Website: github.com/CAHLR

Contact: askoski@berkeley.edu

Live demo: askoski.berkeley.edu/search



Course Catalog Search

Discover new classes using a keyword search matching to course descriptions and inferred topics!

Examples:
- Add subjects one at a time (e.g., linear algebra, probability, data science)
- Exclude them by using a dash (e.g., history, -art, -american)
- Search for a specific course (e.g., Educ 190)

discrete , statistics, estimation, -Markov chains

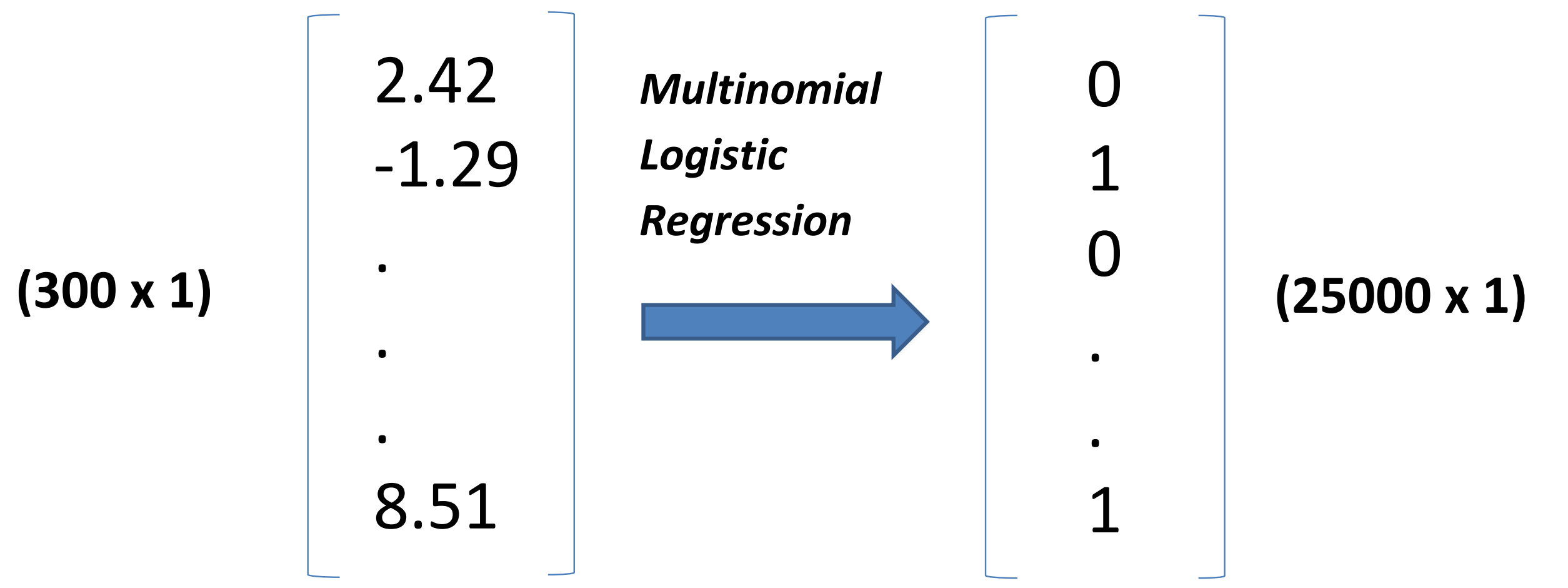
Course Name	Course Title	Course Description	Inferred Keywords
Stat 135	Concepts of Statistics	A comprehensive survey course in statistical theory and methodology. Topics include descriptive statistics, maximum likelihood estimation, non-parametric methods, introduction to optimality, goodness-of-fit tests, analysis of variance, bootstrap and computer-intensive methods and least squares estimation. The laboratory includes computer-based data-analytic applications to science and engineering.	discrete, estimation, linear, mathematical, probability, random, real, regression, statistical, statistics
Pb hlth c240a	Biostatistical Methods: Advanced Categorical Data Analysis	This course focuses on statistical methods for discrete data collected in public health, clinical and biological studies. Lectures topics include proportions and counts, contingency tables, logistic regression models, Poisson regression and log-linear models, models for polytomous data and generalized linear models. Computing techniques, numerical methods, simulation and general implementation of biostatistical analysis techniques with emphasis on data applications.	biological, covers, estimation, inference, linear, model, non, regression, statistical, statistics

Figure 1. User study relevancy ratings by keyword group

These inferred keywords provide an additional vetting field to improve search relevancy, distinguishing it from other course information services.

Methodology

Our approach to generating inferred keywords for a particular course comprises of training a model from its course2vec vector to its respective natural language course description (bag-of-words) vector.



$$(300 \times 1) \quad \begin{bmatrix} 2.42 \\ -1.29 \\ . \\ . \\ . \\ 8.51 \end{bmatrix} \xrightarrow{\textit{Multinomial Logistic Regression}} \begin{bmatrix} 0 \\ 1 \\ 0 \\ . \\ . \\ 1 \end{bmatrix} \quad (25000 \times 1)$$

This is essentially a machine translation, not between languages [1], but between a course representation space and a semantic space constructed from catalog descriptions.

Datasets

Course Information

- Class titles and descriptions sourced from the official University of California at Berkley (UCB) course catalog API.

Course Vectors (Course2Vec)

- Distributed representation of courses learned from individual UCB student course selection histories [2].
- The anonymized data consisted of per-semester course enrollment information for 108,033 undergraduates with a total of 2.2M course enrollment records from Fall 2008 through Fall 2016.

User Study Validation

To gauge how the model results aligned with students' perception of relevance, undergraduates were recruited to participate in a personalized online survey based on their course history in exchange for a $10 Amazon gift card. Students were asked to rate the relevancy of keywords to courses they had taken on a scale of 1-5. These keywords were generated from the embedding model and other sources (e.g., random within and outside of description words) which served as baselines to evaluate the utility of the semantics generated.

Model Sorted (All): Top five overall keywords from the entire corpus as predicted by the model.

Model Sorted (Description): Top five words in the description in order of likelihood as predicted by the model.

Model Sorted (Non-Description): Top five words not in the description in order of likelihood as predicted by the model.

Random (Description): Five random words from within the description.

Random (All): Five random words across all collective descriptions.

Course Title	STAT 135 – Concepts of Statistics
Description	A comprehensive survey course in statistical theory and methodology. Topics include descriptive statistics, maximum likelihood estimation, non-parametric methods, introduction to optimality, goodness-of-fit tests, analysis of variance, bootstrap and computer-intensive methods and least squares estimation. The laboratory includes computer-based data-analytic applications to science and engineering.
Model Sorted (All)	regression, statistics, random, statistical, estimation
Model Sorted (Description)	statistics, statistical, estimation, variance, tests
Model Sorted (Non-Description)	regression, random, real, linear, discrete
Random (Description)	course, engineering, includes, methods, computer-based
Random (All)	diverse collection, topics problems, year credit, planning research, user interfaces

Table 1. Example of course keyword groups

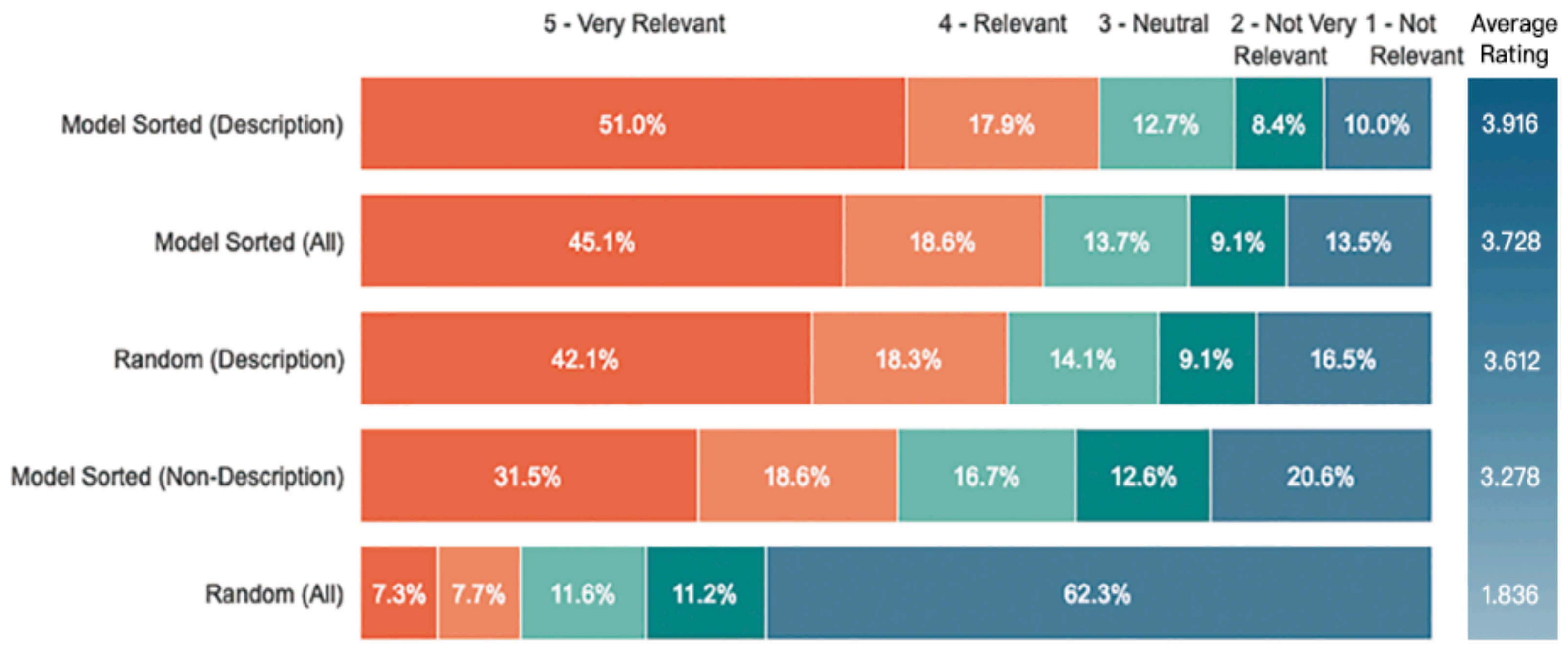


Figure 3. User study relevancy ratings by keyword group

Results

The benefit of the model-based approach in terms of improving displayed keyword relevance is shown in Fig 3. It can be quantified by the difference in average ratings and proportion of *very relevant* scores between the random within-description selection group, and the model-based within-description group. Additionally, model predicted non-description keywords perform statistically on par with the random description group and also performs considerably higher than the random (all) relevancy baseline.

References

[1] T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. CoRR, abs/1309.4168, 2013.

[2] Z.A. Pardos, and A.J.H. Nam. A map of knowledge. CoRR, abs/1811.07974, 2018.