Amazon KDD Cup '22

Shopping Queries Data Set

ESCI Challenge for Improving Product Search

$21,000 Cash + **$10,500 AWS Credit** Prize Pool

ACM SIGKDD 2022 Workshop

Chandan Reddy and Nurendra Choudhary

Amazon Team

https://amazonkddcup.github.io

ORGANIZERS – AMAZON TEAM

Chandan Reddy

Nurendra Choudhary

Lluís Marquez

Fran Valero

Nikhil Rao

Hugo Zaragoza

Sambaran Bandyopadhyay

Arnab Biswas

Introduction

- Improving the relevance of search results can significantly improve the customer experience and their engagement with search.

- Despite the recent advancements in the field of machine learning, correctly classifying items for a particular user search query for shopping is challenging.

- Reasons that contribute to the complexity of this problem.
  - Presence of noisy information in the results
  - Difficulty of understanding the query intent
  - Diversity of the available items

- A small number of unwanted items in the results can break the user experience and hence extremely high accuracy is required for this problem.

Search Relevance

- The notion of binary relevance limits the customer experience. Hence, we break down relevance into four classes (ESCI) which are used to measure the relevance of the items in the search results:

  - **Exact (E):** the item is relevant for the query, and satisfies all the query specifications (e.g., water bottle matching all attributes of a query “plastic water bottle 24oz”, such as material and size)

  - **Substitute (S):** the item is somewhat relevant: it fails to fulfill some aspects of the query but the item can be used as a functional substitute (e.g., fleece for a “sweater” query)

  - **Complement (C):** the item does not fulfill the query, but could be used in combination with an exact item (e.g., track pants for “running shoe” query)

  - **Irrelevant (I):** the item is irrelevant, or it fails to fulfill a central aspect of the query (e.g. socks for a “pant” query)

Shopping Queries Dataset

- We introduced the **“Shopping Queries Data Set”,** a large dataset of difficult search queries, published with aim of **fostering research in the area of query-product semantic matching.**
- **Massive dataset with 2.6 million human annotations (ImageNet for Product search).**

- Some important Characteristics of this dataset:
  - It is derived from **real customers searching for real products online**. Products are linked to online catalog.
  - For each query, the dataset provides a **list of up to 40 potentially relevant results**, together with ESCI relevance judgements.
  - The dataset is **Multilingual**, as it contains queries in English, Japanese, and Spanish. It provides both **breadth** (a large number of queries) **and depth** (≈20 results per query), unlike other public datasets.
  - All results have been **manually labeled with multi-valued relevance labels** in the context of e-shopping.
  - **Queries are not randomly sampled**, but rather, subsets of the queries have been sampled specifically to provide a variety of challenging problems (such as negation, attribute parsing, etc.).
  - Each query-product pair is **accompanied by some additional public catalog information** (including title, product description, and additional product related bullet points). This has both categorical and textual metadata, and multiple levels of representation (from a short title to a long description of the product).

Challenge Tasks

- The primary objective of this competition is to build new ranking strategies and simultaneously identify interesting categories of results (i.e., substitutes) that can be used to improve the customer experience when searching for products. **The three tasks for this KDDCup** competition using the Shopping Queries Dataset.

- **TASK 1: Query-Product Ranking:** Given a user specified query and a list of matched products, the goal of this task is to rank the products so that the relevant products are ranked above the non-relevant ones. (measured by nDCG metric)

- **TASK 2: Multi-class Product Classification:** Given a query and a result list of products retrieved for this query, the goal of this task is to classify each product as being an Exact, Substitute, Complement, or Irrelevant match for the query. (measured by micro-F1 metric)

- **TASK 3: Product Substitute Identification:** This task will measure the ability of the systems to identify substitute products in the list of results for a given query. (measured by micro-F1 metric)

Crowdsourcing AI to Solve Real-World Problems

AIcrowd enables data science experts and enthusiasts to collaboratively solve real-world problems, through challenges.

246+
Completed Challenges

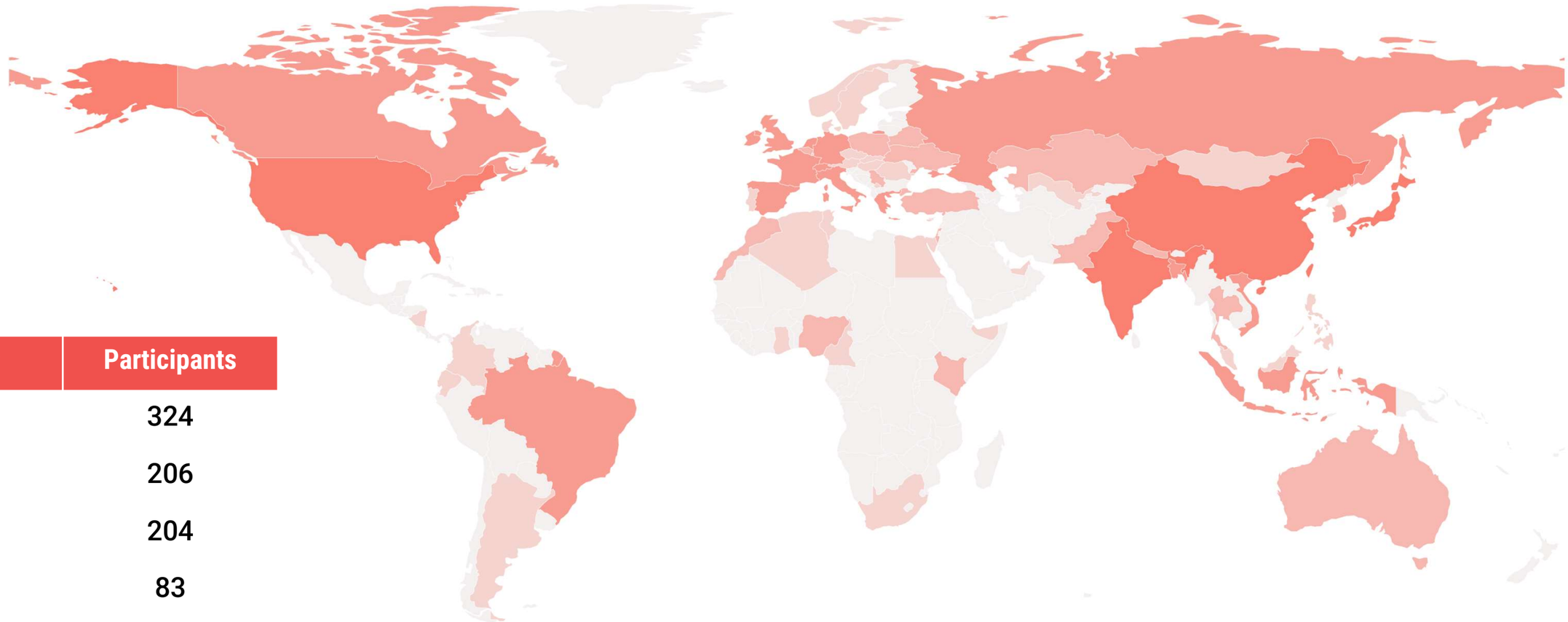
59k+
Community Members

$823k+
Awarded in Prizes

60+
Research Papers Published

13 TB+
Codes, Models & Datasets Hosted

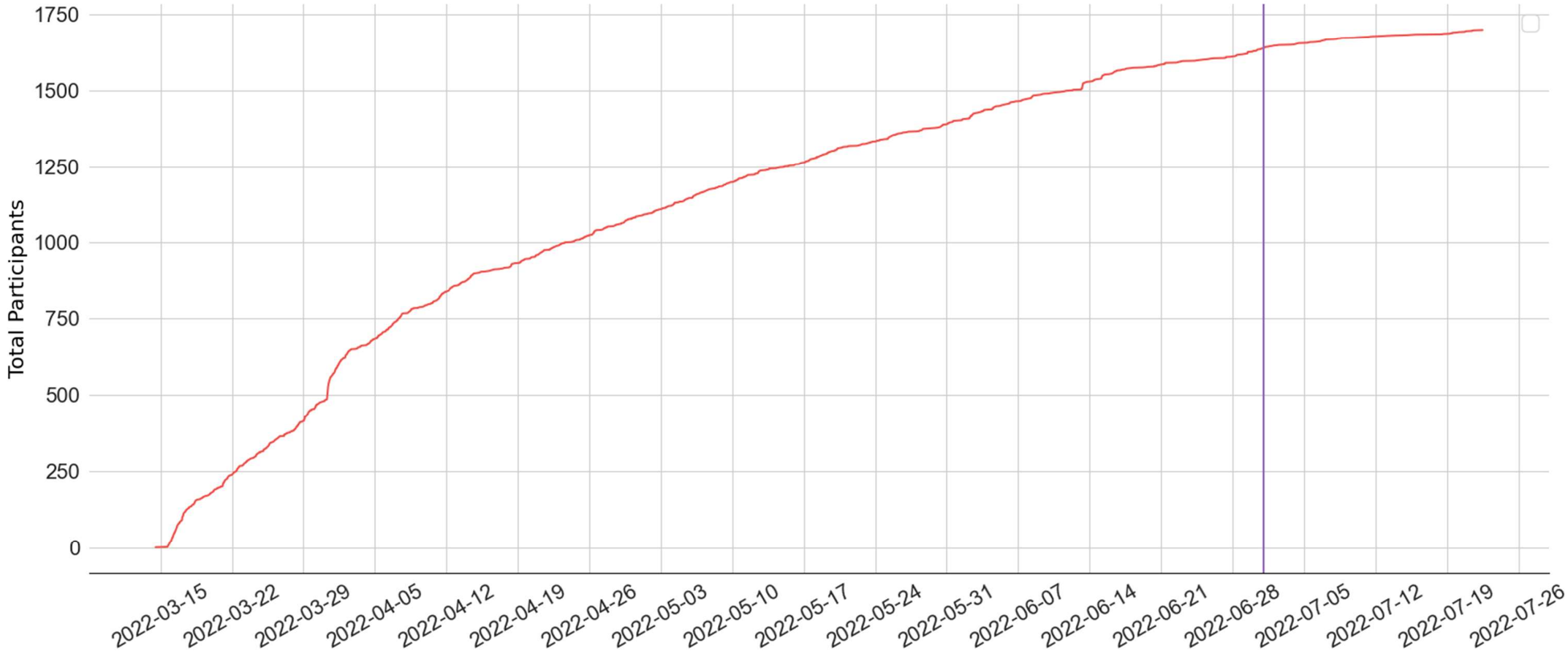
Participants (by Country)



Country	Participants
China	324
USA	206
India	204
Japan	83
Taiwan	80

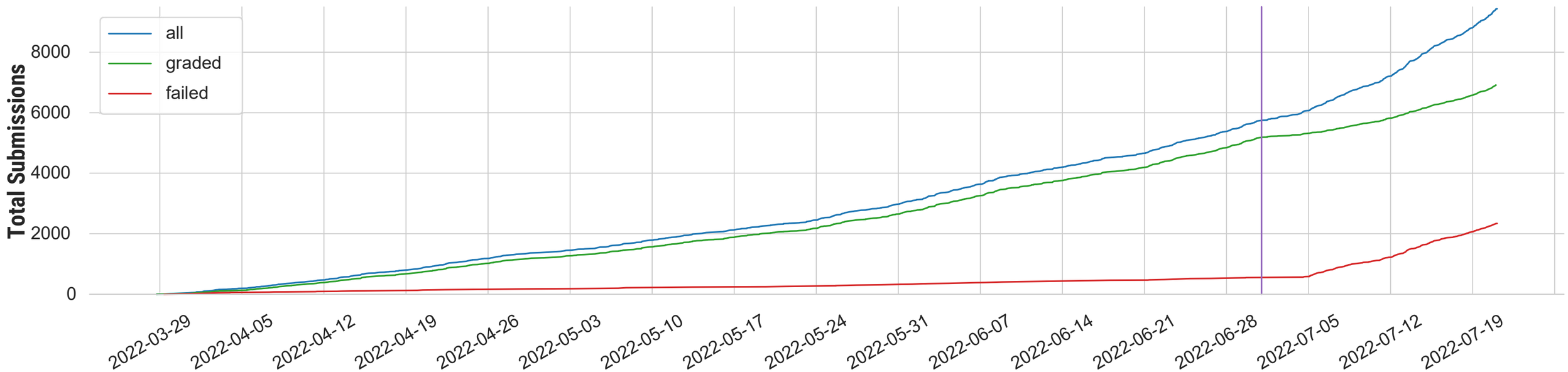
Total Countries Represented : 65

Participants (over time)



Total Participants : **1699**

Submissions (over Time)



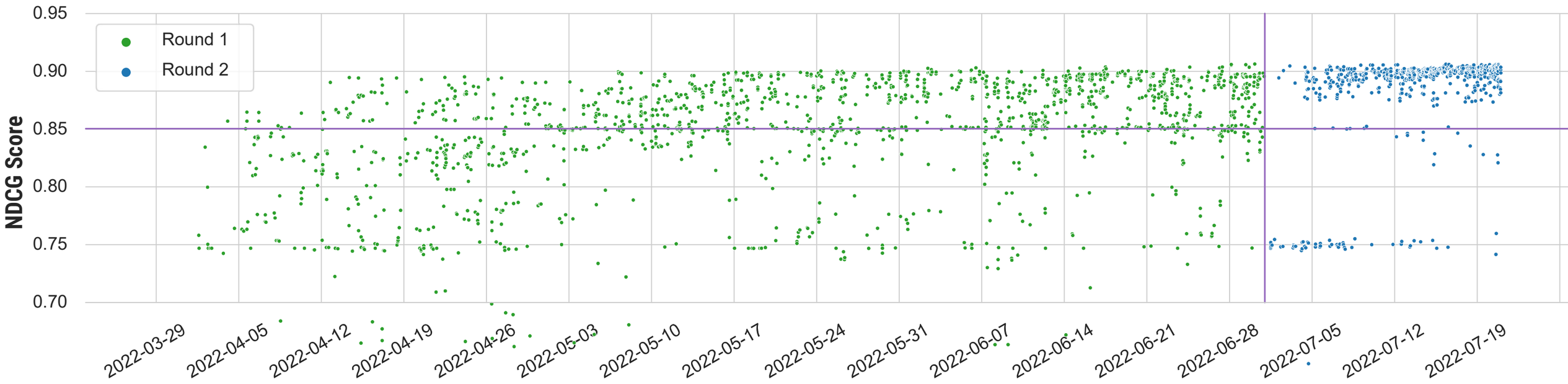
Total Submissions : 7464

Total Successful Submissions : 6911

Total Failed Submissions : 553

Total Size of Code & Models Submitted : 2.5TB

Submissions (by Score) Task - 1



Total Submissions : 2865

Total Successful Submissions : 2689

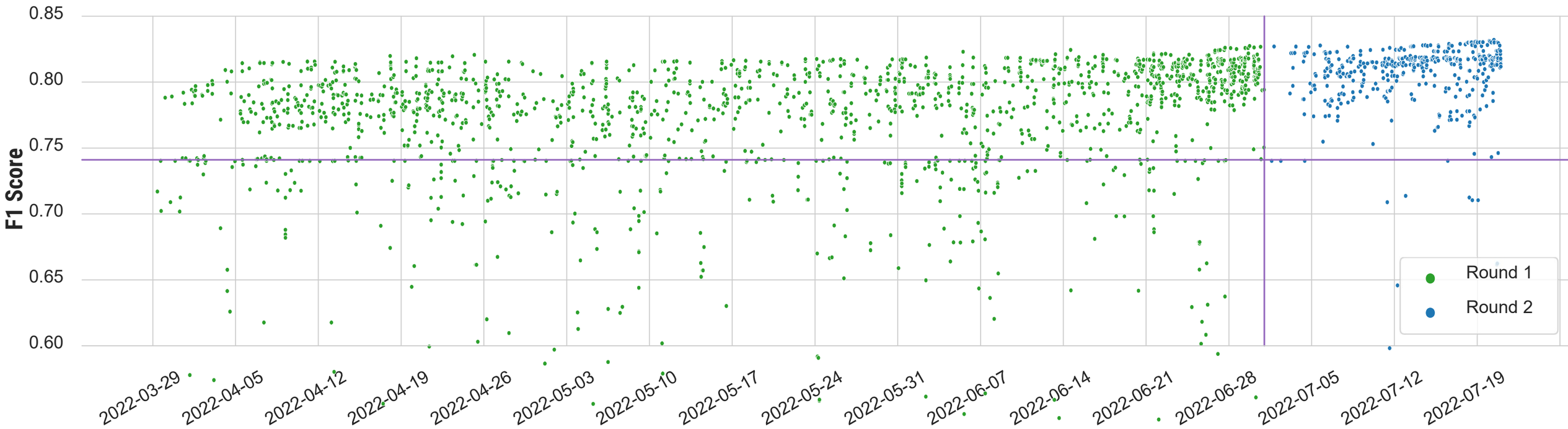
Total Failed Submissions : 176

Baseline Score : 0.8503

Submissions Above Baseline (Round 1): 1153

Submissions Above Baseline (Round 2): 712

Submissions (by Score) Task - 2



Total Submissions : 2849

Total Successful Submissions : 2645

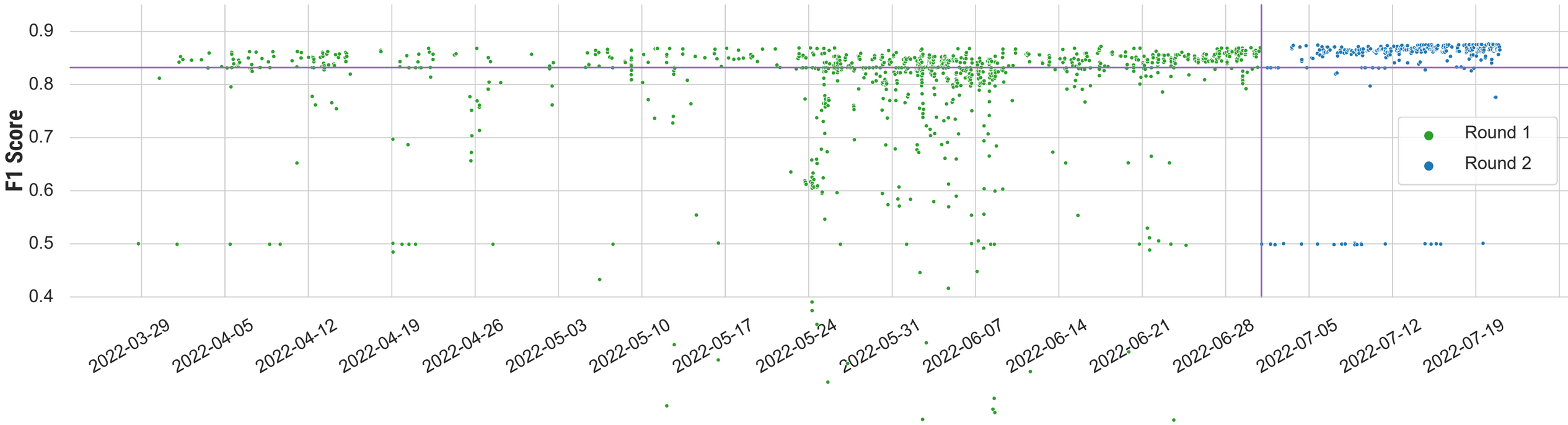
Total Failed Submissions : 204

Baseline Score : 0.7410

Submissions Above Baseline (Round 1): 1635

Submissions Above Baseline (Round 2): 470

Submissions (by Score) Task - 3



Total Submissions : 1750

Total Successful Submissions : 1577

Total Failed Submissions : 173

Baseline Score : 0.8319

Submissions Above Baseline (Round 1): 680

Submissions Above Baseline (Round 2): 340

Aggregated Submission Statistics

	Task-1			Task-2			Task-3		
	Round-1	Round-2	Overall	Round-1	Round-2	Overall	Round-1	Round-2	Overall
Total Submissions	*2069*	*1433*	***3502***	*2307*	*1301*	***3608***	*1360*	*778*	***2138***
Total Successful Submissions	*1893*	*797*	***2690***	*2103*	*542*	***2645***	*1187*	*390*	***1577***
Total Failed Submissions	*176*	*636*	***812***	*204*	*759*	***963***	*173*	*388*	***561***
Total Submissions (score > baseline_score)	*1153*	*712*	***1865***	*1635*	*470*	***2105***	*680*	*340*	***1020***

Final Ranks - Task - 1

Rank	Team Name	Private Test Set Score (NDCG)	Prizes
#1 👑	www	0.9043	$4000
#2 👑	qinpersevere	0.9036	$2000
#3 👑	day-day-up	0.9035	$1000
#4	GraphMIRAcles	0.9028	$500 (in AWS credits)
#5	ZhichunRoad	0.9025	$500 (in AWS credits)
#6	ETS-Lab	0.9014	$500 (in AWS credits)
#7	ALONG	0.9014	$500 (in AWS credits)
#8	ljr333	0.9008	$500 (in AWS credits)
#9	NeuralMind	0.9007	$500 (in AWS credits)
#10	zackchen	0.8998	$500 (in AWS credits)

Final Winners - Task - 1

First place: Team www - Interactive Entertainment Group of Netease Inc., Guangzhou, China

Qi Zhang, Zijian Yang, Yilun Huang, Zijian Cai, Kangxu Wang.

Second place: Team qinpersevere - Netease Games AI Lab, Hangzhou, China

Xiaolei Qin, Nan Liang, Hongbo Zhang, Wuhe Zou, and Weidong Zhang.

Third place: Team day-day-up - Ant Group, Hangzhou, Zhejiang, China

Jinzhen Lin, Lanqing Xue, Zhenzhe Ying, Changhua Meng, Weiqiang Wang, Haotian Wang, and Xiaofeng Wu.

Final Ranks - Task - 2

Rank	Team Name	Private Test Set Score (F1)	Prizes
#1 👑	day-day-up	0.8326	$4000
#2 👑	ETS-Lab	0.8325	$2000
#3 👑	Uni	0.8273	$1000
#4	hahaha	0.8251	$500 (in AWS credits)
#5	MetaSoul	0.8207	$500 (in AWS credits)
#6	www	0.8204	$500 (in AWS credits)
#7	ZhichunRoad	0.8194	$500 (in AWS credits)
#8	qinpersevere	0.8191	$500 (in AWS credits)
#9	zackchen	0.8189	$500 (in AWS credits)
#10	LYZD-fintech	0.8183	$500 (in AWS credits)

Final Winners - Task - 2

First place: Team day-day-up - Ant Group, Hangzhou, Zhejiang, China

Jinzhen Lin, Lanqing Xue, Zhenzhe Ying, Changhua Meng, Weiqiang Wang, Haotian Wang, and Xiaofeng Wu.

Second place: Team ETS-Lab - Purdue University & Tsinghua University

Fanyou Wu (Purdue University), Yang Liu (Tsinghua University), and Xiaobo Qu (Tsinghua University)

Third place: Team Uni -

Ruiqing Yan (Chinese Academy of Sciences and Beijing Institute Of Petrochemical Technology), Peng Zhang (Zhejiang University), Linghan Zheng (Ant Group), Changyu Li (University of Electronic Science and Technology of China), and Rui Hu (Zhejiang University)

Final Ranks - Task - 3

Rank	team_name	Private Test Set Score (F1)	Prizes
#1 👑	day-day-up	0.8790	$4000
#2 👑	ETS-Lab	0.8771	$2000
#3 👑	Uni	0.8754	$1000
#4	hahaha	0.8734	$500 (in AWS credits)
#5	LYZD-fintech	0.8708	$500 (in AWS credits)
#6	qinpersevere	0.8701	$500 (in AWS credits)
#7	wookiebort	0.8687	$500 (in AWS credits)
#8	ZhichunRoad	0.8686	$500 (in AWS credits)
#9	NTT-DOCOMO-LABS-GREEN	0.8677	$500 (in AWS credits)
#10	rein20	0.8668	$500 (in AWS credits)

Final Winners - Task - 3

First place: Team day-day-up - Ant Group, Hangzhou, Zhejiang, China

Jinzhen Lin, Lanqing Xue, Zhenzhe Ying, Changhua Meng, Weiqiang Wang, Haotian Wang, and Xiaofeng Wu.

Second place: Team ETS-Lab - Purdue University & Tsinghua University

Yang Liu (Tsinghua University), Fanyou Wu (Purdue University), and Xiaobo Qu (Tsinghua University)

Third place: Team Uni -

Ruiqing Yan (Chinese Academy of Sciences and Beijing Institute Of Petrochemical Technology), Peng Zhang (Zhejiang University), Linghan Zheng (Ant Group), Changyu Li (University of Electronic Science and Technology of China), and Rui Hu (Zhejiang University)

Key Issues Faced (mostly in Code Submissions)

- We did not release the private test set on which the competition was evaluated upon. We asked the participants to submit their code so that we can run their models on our end and report the results.

- **Organizers**
  - Dataset Sanity Checks
  - Resource Constraints: Compute & Time limits
- **Participants**
  - Packaging of Software Runtime
  - Packaging of Large models
  - Use of Large Ensembles
  - Lack of Debug Output in Private-Test Phase
  - Submission Timeouts
  - Large evaluation Queues

Data is Publicly Available !!

The data is made publicly available at this website
https://github.com/amazon-research/esci-data

If you plan to use this dataset for your own research, please cite this paper.

@article{reddy2022shopping,
title={Shopping Queries Dataset: A Large-Scale {ESCI} Benchmark for Improving Product Search},
author={Chandan K. Reddy and Lluís Màrquez and Fran Valero and Nikhil Rao and Hugo Zaragoza and Sambaran Bandyopadhyay and Arnab Biswas and Anlu Xing and Karthik Subbian},
year={2022},
eprint={2206.06588},
archivePrefix={arXiv}
}

Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search

Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, Karthik Subbian
Amazon, USA
{ckreddy,lluismv,fvalero,nikhilsr,hugzarag,sambarab,abisway,anluxing,ksubbian}@amazon.com

ABSTRACT

Improving the quality of search results can significantly enhance users experience and engagement with search engines. In spite of several recent advancements in the fields of machine learning and data mining, correctly classifying items for a particular user search query has been a long standing challenge, which still has a large room for improvement. This paper introduces the "Shopping Queries Dataset", a large dataset of difficult Amazon search queries and results, publicly released with the aim of fostering research in improving the quality of search results. The dataset contains around 130 thousand unique queries and 2.6 million manually labeled (query,product) relevance judgements. The dataset is multilingual with queries in English, Japanese, and Spanish. The Shopping Queries Dataset is being used in one of the KDDCup'22 challenges. In this paper, we describe the dataset and present three evaluation tasks along with baseline results: (i) ranking the results list, (ii) classifying product results into relevance categories, and (iii) identifying substitute products for a given query. We anticipate that this data will become the gold standard for future research in the topic of product search.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**; *Query representation*; • **Applied computing → Online shopping**.

KEYWORDS

search relevance, querying, e-commerce, semantic matching

1 INTRODUCTION

Improving the relevance of search results can significantly improve "iPhone", would an iPhone charger be relevant, irrelevant, or somewhere in between? In fact, many users issue the query "iPhone" to find and purchase a charger for the iPhone. They simply expect the search engine to understand their need. For this reason, we break down relevance into the following four classes which are used to measure the relevance of items in the search results:

- **Exact (E):** the item is relevant for the query, and satisfies all the query specifications (e.g., a water bottle matching all attributes of a query "plastic water bottle 24oz", such as material and size)
- **Substitute (S):** the item is somewhat relevant, i.e., it fails to fulfill some aspects of the query but the item can be used as a functional substitute (e.g., fleece for a "sweater" query)
- **Complement (C):** the item does not fulfill the query, but could be used in combination with an exact item (e.g., track pants for "running shoes" query)
- **Irrelevant (I):** the item is irrelevant, or it fails to fulfill a central aspect of the query (e.g., socks for a "telescope" query, or a wheat flour bread for a "gluten–free bread" query)

In this paper, we introduce the "Shopping Queries Dataset", a large dataset of difficult search queries published with the aim of fostering research in the area of semantic matching of queries and products. For each query, the dataset provides a list of up to 40 results, together with their ESCI relevance judgements (Exact, Substitute, Complement, or Irrelevant) indicating the relevance of the product to the query [11]. Each query-product pair is accompanied by additional information from the Amazon catalog, including: product title, product description, and additional product related bullet points. This information is public, as it is displayed at the Amazon website when searching for those products. The Shopping Queries Dataset is multilingual, as it contains queries in English,

Final Dataset Statistics

	Total	Total	Total	Train	Train	Train	Test	Test	Test
Language	# Queries	# Judgements	Avg. Depth	# Queries	# Judgements	Avg. Depth	# Queries	# Judgements	Avg. Depth
English	29,844	601,354	20.15	20,888	419,653	20.09	8,956	181,701	20.29
Spanish	8,049	218,774	27.18	5,632	152,891	27.15	2,417	65,883	27.26
Japanese	10,407	297,883	28.62	7,284	209,094	28.71	3,123	88,789	28.43
Overall	**48,300**	**1,118,011**	**23.15**	**33,804**	**781,638**	**23.12**	**14,496**	**336,373**	**23.20**

Table 1: Summary of the Shopping queries data set for task 1 (reduced version) - the number of unique queries, the number of judgements, and the average number of judgements per query.

	Total	Total	Total	Train	Train	Train	Test	Test	Test
Language	# Queries	# Judgements	Avg. Depth	# Queries	# Judgements	Avg. Depth	# Queries	# Judgements	Avg. Depth
English	97,345	1,818,825	18.68	74,888	1,393,063	18.60	22,458	425,762	18.96
Spanish	15,180	356,410	23.48	11,336	263,063	23.21	3,844	93,347	24.28
Japanese	18,127	446,053	24.61	13,460	327,146	24.31	4,667	118,907	25.48
Overall	**130,652**	**2,621,288**	**20.06**	**99,684**	**1,983,272**	**19.90**	**30,969**	**638,016**	**20.60**

Table 2: Summary of the Shopping queries data set for tasks 2 and 3 (larger version) - the number of unique queries, the number of judgements, and the average number of judgements per query.

Acknowledgements

- **Amazon Search Leadership**
- **KDDCup Chairs**
- **AICrowd Platform**
- **AWS Partners**
- **Participants**

Thank You for Your Participation !!

Workshop:

https://amazonkddcup.github.io/

Data:

https://github.com/amazon-research/esci-data

KDDCup Challenge:

https://www.aicrowd.com/challenges/esci-challenge-for-improving-product-search