Friedrich-Schiller-Universität Jena
Institute of Computer Science
Degree Programme Computational and Data Science,
M.Sc.

Verifying Query Logs from Unkown Sources

Master's Thesis

Benjamin Schneg

1. Referee: Prof. Dr. Matthias Hagen
2. Referee: Jan Heinrich Merker

Submission date: April 15, 2025

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Jena, April 15, 2025

. . . . . .
Benjamin Schneg

Abstract

This thesis aims to analyze the Archive Query Log (AQL), a query log of unknown source, and examine if it reflects realistic user behavior or if its query distribution is rather anomalous. To accomplish this, we compare it to other publicly available query logs, which are collected from real-world logs. We consider the AOL Log, the MS-MARCO Web Search Log and the ORCAS Log. The comparison is based on various metrics that aim to capture distinctive characteristics of a query log.

The analysis in this thesis is composed of three parts: structure-related, inference-based and temporal-based analysis. For each part we measure distributions of the query logs and compare them to each other. The numeric comparison is based on Wasserstein distances, a method to measure similarities of probability distributions. Besides the numeric comparison of distributions we provide concrete insights, displaying tables of top queries, top words and visualizations of temporal patterns.

The results show that the AQL exhibits similar distributions to the other query logs, but the similarity lacks magnitude to conclude that the AQL belongs to the comparison group. Furthermore, temporal pattens of the AQL are substantially different to the temporal patterns present in real-word query logs.

Contents

Chapter 1

Introduction

Search engine query logs are an important resource to promote research in information retrieval [Agosti et al., 2012]. They enable, for instance, analyzing user behavior and user experience, improving query suggestions and query reformulations or provide data to train retrieval models for re-ranking [Reimer et al., 2023]. Consequently, a vast access to search engine query logs would be highly beneficial for the research community. Additionally, public access to search engine query logs is valuable to create transparency and facilitate investigations on the fairness of major search engine's ranking algorithms [Reimer et al., 2023]. However, despite the high value of query logs, they remain publicly unavailable for the most part. This is due to multiple reasons. On the one hand, the publication of query logs brings up privacy concerns, as they contain sensitive information about the users [Reimer et al., 2023]. On the other hand, search engine operators' interests may not align with the previous mentioned motivations to publish query logs, as higher competition in the search engine market may arise or a comprehensive transparency of the search engines' behavior may simply not be desired in the first place by the operators [Reimer et al., 2023]. Nonetheless, a few public query logs exist, among which the AOL query log [Pass et al., 2006] is the most prominent and comprehensive one [Reimer et al., 2023]. However, publicly available query logs are not on par with private query logs collected by major search engine operators. This applies not only for the mere size of the log, specifically the number of queries, but also for the temporal span the queries stem from. Furthermore, publicly available query logs are outdated for the most part and lack an overall high quality in the aforementioned aspects [Reimer et al., 2023]. To fill this gap, Reimer et al. [2023] published a query log from a new source that had not been exploited before. A set of 356 million queries, stemming from the past 25 years, was collected from the *Internet Archive*. The new resource, called the *Archive Query Log* (AQL), is on par with private query logs in terms of

number of queries, time span and further quality aspects. The scope of the AQL supports many tasks in information retrieval research and can be used to advance research in this field. This is however only possible if the query log is trustworthy and reflects realistic user bahaviour. In this thesis, we provide a comprehensive analysis of the AQL and aim to examine if the AQL reflects realistic user behaviour or if its queries are rather statistically biased. To accomplish this, we compare it to other publicly available query logs, which are known to show realistic user behaviour. The comparison is based on various metrics that aim to capture distinctive characteristics of a query log. The analysis in this thesis is composed of the following parts:

- **Structure-related:** This analysis aims at capturing structural charasteristics of the involved query logs. Frequency distributions of the query length, word length, word count, character count, search operator count and other variables are measured. In addition, we examine if the query logs comply with liguistic laws like zipf's law.
- **Inference-based:** We infer characteristics from queries by applying various language models to them. The involved models either classify the queries on different taxonomies or extract meaningful information. The query logs are investigated with regard to query intent, personal identifiable information (PII) entities and questions.
- **Temporal-based:** We utilize google trends to examine the temporal aspect of the queries. We analyze the temporal distribution of the queries in the AQL and their relation to real-world queries. This involves a correlation study.

Chapter 2

Background

In this chapter we review some aspects of the literature that are relevant to this work. In particular, we describe prevalent analyses of search logs that have been conducted in the past and their motivations. This includes simple syntactical analyses of the queries as well as the classification into meaningful taxonomies. We discuss whether these analyses would contribute to the objective of this work. Furthermore, as an additional background to the temporal analysis in this work, we critically point out limitations of utilizing data from Google Trends[1] to investigate temporal patterns in query logs. Thereon, we elaborate on sentence embeddings and their potential usage in this work. Finally, we introduce Wasserstein distances as a method to compare distributions.

2.1 Traditional Query Log Analysis

With the increasing advent of the internet during the late 1990s, search engines became an essential tool for users to navigate the web. This led to a growing interest in understanding the behavior of users when searching the web and hence an interest to study search queries. A lot of publications from this time investigate query logs to understand user behavior and the differences between information retrieval on the web and traditional information retrieval.

Query Frequencies

A common analysis among the early works is the examination of query frequencies. Silverstein et al. [1999] consider individual and repeated queries and measure their frequency. They find out that most queries are individual and

[1] `https://trends.google.com/trends/`

conclude that the user's information need is quite diverse. In addition, they partition queries into groups that appear 1, 2, 3, or more than 3 times and measure the frequencies of these groups. Spink et al. [2001] also measure frequencies of queries with respect to the occurence in the log. They also find a small number with high frequency and a high number of unique queries. Generally, a lot of works analyze the frequency distribution of queries and discover highly skewed distributions.

Terms in Queries

Another common approach is the consideration of terms in query logs and their distribution. Silverstein et al. [1999] measure the frequency distribution of the number of terms in a query and related statistics like the mean and standard deviation of the number of terms. Jansen et al. [2000] consider mere frequency distributions of terms and create a descending-ranked distribution. They test for Zipf's Law in the ranked distribution of terms by displaying the frequency of terms in a log-log plot. Spink et al. [2001] also measure the length of queries with respect to the number of occuring terms and display this distribution. Besides testing for Zipf's Law in the term distribution, they prepare tables of the most frequent terms and their respective frequencies to provide concrete insights to the queries. Wolfram et al. [2001] investigate the diversity of terms in queries and conclude that more unique terms are present in query logs than in large text corpora.

Search Operators

Lastly, investigations concerning the usage of search operators in queries are covered by many works. Search operators are special characters or words that are used to refine the search. Silverstein et al. [1999] measure the fraction of queries that contain search operators and the number of search operators in a query. They find that most queries do not contain search operators. Jansen et al. [2000] also measure the frequency of search operators in queries. Spink et al. [2001] provide a list of selected search operators and their frequencies in the query logs.

Based on this overview, we deduce the following domains to analyze query logs:

- **Query Frequencies:** Generating frequency distributions of queries. From this, the creation of rankings of the queries and a subsequent test for Zipf's Law is feasible. Furthermore, a list of the most frequent queries and their frequencies might provide concrete insights.

- **Terms in queries:** Extracting terms and generating the distribution of terms in queries. Likewise, a test for Zipf's Law after creating a ranking is possible. In addition, we can measure a query's length in the number of its terms and generate a frequency distribution from this. For a concrete insight, we can create tables of the most frequent terms and their frequencies.
- **Search operators:** Creating a list of common search operators and measuring their frequencies in the query logs.

2.2 Query Classification

A common approach to refine search results is a preceding classification of the search query. Different ideas regarding what taxonomy of search queries would be useful to refine search results exist.

Query Intent

One of the most prominent approaches is the idea to understand the intent behind a query and classify it accordingly. The idea of Broder [2002] to focus on the intent and distinguish between navigational, informational, and transactional queries was widely adopted. According to Broder [2002], navigational queries are used to find a specific website, informational queries are used to find information about a topic, and transactional queries are used to perform a transaction, such as purchasing a product or download a file. Rose and Levinson [2004] and Kang and Kim [2003] as well emphasize the importance of understanding the user's intent behind a query and adopt the taxonomy of Broder [2002]. Those early approaches measure, among others, conditional probabilities of terms being present in a certain category and use them for classification. Alexander et al. [2022] further refine the taxonomy of Broder [2002] and introduce a more fine-grained classification of informational queries. They subdivde the informational category into instrumental (i.e., how to do sth.), factual (search of facts or pieces of information) and abstain (rest of informational queries).

Topic-related Categories

Similar attention was given to the idea of classifying queries into topic-related categories. Beitzel et al. [2005] state that classifying queries into topic-related categories would make topic-specific databases employable. Hence, a more efficient and effective search could be performed. They attempt to classify

queries into one of 18 topic-related categories and use a supervised learning approach to do so. Moreover, they develop an ensemble-classifier that involves a perceptron trained on labeled queries and a rule-based classifier which uses conditional probabilities of bigrams to classify queries. However, this method performs rather low and achieves a F1-Score of 0.12.

Challenges

Despite the potential benefits of classifying queries, there are several challenges associated with this task. One of the main challenges is the sparsity of information present in queries. Most queries are short and lack context, making it difficult to classify them accurately. Furthermore, query streams vary over time heavily, which even further complicates the classification task [Beitzel et al., 2005]. Due to their brevity, queries are often ambiguous and can have multiple meanings. Because of these challenges, first attempts to classify queries were not very accurate regardless the taxonomy [Beitzel et al., 2005, 2007, Kang and Kim, 2003]. Though, the method of Alexander et al. [2022] is more promising. They achieve an accuracy of 0.90 when classifying queries into the three categories informational, navigational and transactional. Because of this, we focus on the intent-based taxonomy of Broder [2002] and Alexander et al. [2022] in this work.

2.3 Named Entities in Queries

According to Guo et al. [2009] and Zhang et al. [2015], over 70% of queries contain named entities. They seem to be a substantial part of queries. Zhang et al. [2015] use query logs as a source to learn entity types and Guo et al. [2009] perform named entity recognition in queries and a subsequent classification into entity types. Lin et al. [2012] state that a lot of web search queries involve actions on entities and propose a method to automatically find queries bearing entities and suggest the most desired actions on such entities for the user. By actions the authors mean, for instance, reading reviews, watching demo videos etc. with regard to an entity that is also part of the query. Those early approaches involve manual labeling of training data and human engineering in designing domain-specific features and rules [Li et al., 2022]. Present Named Entity Recognizers (NER), however, are based on deep learning and are not dependent on designing domain-specific features by humans. Those approaches achieve state-of-the-art performance [Li et al., 2022] and are provided by open source libraries such as spaCy[2]. Considering the large presence of named

[2] `https://spacy.io/`

entities in queries and the availability of state-of-the-art NER, detecting named entities in the involved query logs seems feasible and insightful.

2.4 Questions in Queries

Over the years, an increasing amount of queries is formulated in the form of natural language instead of a set of keywords [White et al., 2015]. Accordingly, users also pose questions to search engines. White et al. [2015] study the behaviour of search engines when handling questions. They filter questions from search logs by a rule-based approach and find that 2-3% of queries are questions. Bondarenko et al. [2020] as well investigate the presence of questions in search queries. In their study, they aim at finding comparative questions and develop classifiers to do so. The work of Reimer [2023] also identifies questions in queries among other things. Here, a rule-based classifier for detecting questions in queries achieves a recall of 0.89 and an even higher precision of 0.99. Based on theses findings, a question classification of queries seems to be feasible and valuable.

Summary

Based on the previous findings we deduce further following domains to characterize query logs:

- **Query Classification:** Classifying queries into a intent-based taxonomy or a topic-related taxonomy. The resulting distributions deliver additional insights into the query log's properties. Since the classification into topic-related categories is not very accurate, we focus on the intent-based taxonomy.
- **Named Entity Recognition:** Identifying named entities in queries and classifying them into entity types. We obtain distributions of the entity types in the query log and can make another comparison.
- **Questions in queries:** Identifying questions in queries from rule-based approaches. This analysis provides another insight and additionally is simple to implement.

2.5 Temporal Patterns in Query Logs

The temporal popularity of queries is another phenomenon studied in the literature. Query frequencies change over time and reflect information about

general popularity of different topics and trends within society. Some queries even show periodic patterns, such as seasonal trends. Shokouhi [2011] state that seasonal queries are frequent and need to be detected in order to be met appropriately by search engines. According to Shokouhi [2011], seasonal queries demand the most recent web pages and not the most clicked which is why they have to be treated differently. They propose a method that determines the seasonality of queries with high accuracy by applying time series analysis on historical frequency distributions. Chien and Immorlica [2005] suppose that temporally correlated queries are semantically related and define a new measure of the temporal correlation of two queries. Their method successfully captures temporally correlated queries and shows that, indeed, temporal correlation is a good indicator of semantic similarity.

Google Trends and its Limitations

In order to investigate temporal patterns in the AQL, we must specify a subset of queries whose temporal popularity we want to determine and compare. To make comparisons, we need realistic and reliable time series of query popularity from an external source. We choose Google Trends for this. Google Trends is a free tool that provides insights into the temporal popularity of search queries from Google. It allows us to make the necessary comparisons and subsequent conclusions on the observed temporal query popularities from the AQL.

Despite its usefulness, the usage of Google Trends data entails some limitations that need to be taken into account. Behnen et al. [2020] conducted a study on the reliability of Google Trends data and conclude that the data is not always reliable. In their study, they review, among others, google's claim that inconsistencies are due to overall low search volume and should only appear for unpopular queries. Behnen et al. [2020] examine inconsistencies of three german queries for different time spans. They find that inconsistencies occur particularly for short time spans and queries with overall low search volume. According to this, data of google trends for english queries of large time spans should be robust. They even acknowledge that for time spans larger than eight months, the data seem to be reliable.

In summary, we can state that comparing the popularity of different queries with data from Google Trends is feasible and valuable since we can assess temporal patterns. However, we need to be careful which queries and what time spans we choose. Considering the findings of Behnen et al. [2020], a comparison of queries with high search volume and large time spans is reasonable.

2.6 Wasserstein Distances

To evaluate the resulting distributions of this works's experiments, we need methods to compare the distributions of the AQL with the distributions of AOL, MS-MARCO Web Search and ORCAS, our comparison group. *Wasserstein distances* quantify differences of probability distributions and are widely adopted in statistics [Panaretos and Zemel, 2019]. Let μ and ν be two probability measures on $\mathbb{R}^d$, the p-Wasserstein distance is defined as

$$W_p(\mu, \nu) = \inf_{\substack{X \sim \mu \\ Y \sim \nu}} \left(\mathbb{E}\|X - Y\|^p\right)^{1/p}, \quad p \geq 1 \tag{2.1}$$

where the infimum is taken over all pairs of d-dimensional random vectors X and Y marginally distributed as μ and ν [Panaretos and Zemel, 2019]. An intuitive idea of what Wasserstein distances measure would be the minimuim effort to make two distributions coincide. They originally stem from the theory of optimal transport where one tries to find the optimal way to transport a distribution of mass to another distribution of mass [Panaretos and Zemel, 2019]. Conveniently, Wasserstein distances satisfy all properties of a metric, i.e., non-negativity, symmetry and validity of the triangle inequality [Panaretos and Zemel, 2019]. This makes Wasserstein distances suitable for comparisons between distributions. Wasserstein distances even generalize to probability measures defined on much more general spaces: if $(\mathcal{X}, \rho)$ is any complete metric space, then the p-Wasserstein distance can be defined in the same way, with $\|X - Y\|$ replaced by the metric $\rho(X, Y)$ [Panaretos and Zemel, 2019]. That is, the distance is able to inherit the metric that is already defined in the space of the considered distributions which gives a natural interpretation of Wasserstein distances.

Chapter 3

Data Sets

In this chapter, we introduce the data sets that we use in our experiments. We provide an overview of the data sets, present their origin and describe possibly needed data cleaning processes that ensure a fair comparison.

3.1 Archive Query Log

The Archive Query Log (AQL) was published by Reimer et al. [2023] in order to expand the publically available query logs and promote research in the field of information retrieval. It was mined from the Internet Archive's *Wayback Machine* [1] by identifying search engine result pages (SERP) and parsing the queries from the corresponding URLs. Reimer et al. [2023] name two possible reasons why SERPs are archived in the Wayback Machine: First, SERPs can be linked to by web pages and may thus be included in automated web crawls of the Internet Archive. Secondly, any user of the Internet Archive can request archiving a specific URL, hence SERPs may be included, as well. Considering these two cases as the main sources, the AQL's queries would be a mixture of specific queries that users wanted to archive and queries which were linked to by web pages. This might bias the query distribution compared to a real-world search engine query log. The AQL contains around 356.6M queries from the years 1999 to 2022. Moreover, it covers a wide range of languages, namely 104 different languages, and a variety of domains. The most prominent domains in the AQL are Baidoo, Google, StackOverflow, Twitter and Youtube. The most frequent query languages in the AQL are English and Cantonese.

[1] `https://web.archive.org/`

Column Name	Description	Example	Data Type
serp_query_text_url	Query-string	weather	str
serp_timestamp	Query-timestamp	1680604622	int
search_provider_name	The query's search provider	google	str
serp_url	URL of a query's SERP	<SERP-URL>	str
Data Generation	SERPs from WayBack Machine		
Languages	Multilingual, 104 languages		
Time Span	1999-2022		
Num. Rows	356.6 M		

Table 3.1: Summary of the Archive Query Log.

3.1.1 AQL Cleaning

Some results of analysing the AQL have uncovered outliers and further abnormalities that should be removed to ensure a fair comparison between the query logs. First of all, the analysis of the query length distribution has shown that there are some exceptionally frequent lengths in the distribution. In Figure 3.1 we can see that the query length distribution has a peak at the lengths 14, 16 and 24. Therefore, we take a look at the most frequent queries of these lengths. For each length, we find a particularly frequent query whose frequency outnumbers the second most frequent query by several orders of magnitude and likely produces the outlier in the distribution. We remove these uncommonly frequent queries from the data set. Secondly, we noted that a subset of the AQL's queries is subject to a decode error and, as a result, consist of replacement characters. We remove these queries from the data set as well. Lastly, we remove all empty queries, i.e., queries with an empty string, from the data set.

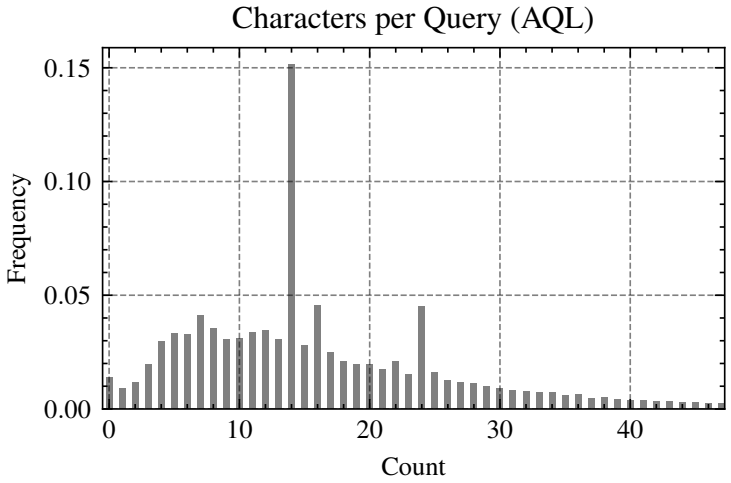

Figure 3.1: Query length distribution of the AQL before cleaning. Note the peaks at the lengths 14, 16 and 24.

3.2 AOL User Session Collection

The AOL User Session Collection (AOL) was published by Pass et al. [2006] to provide real query log data based on real users that was intended for research in personalization, query reformulation and other types of search research. It was a dircet publication from AOL to encourage research in the named fields. The dataset contains around 36 million queries from AOL's search engine entered by about 650,000 users. The queries were randomly sampled over a 3-month period from March to May 2006. There were no specific filters applied to the queries, hence the dataset depicts a real-world query distribution. Of the roughly 36 million queries, around 20 million are unique. The language distribution of the queries is highly skewed towards English, as the majority of the queries are in English. The queries are provided with a respective timestamp. Despite the controversial publication of this data set and its violation of user privacy, it remained a useful asset for research in the field of information retrieval [MacAvaney et al., 2022].

Column Name	Description	Example	Data Type
serp_query_text_url	Query-string	weather	str
serp_id	Query-ID	1733	int
serp_timestamp	Query-timestamp	(2006,3,2)	datetime
serp_offset	Rank of clicked URL	3	int
Data Generation	Random sample over a 3-month period		
Languages	Mostly english		
Time Span	May, 2006		
Num. Rows	36 M.		

Table 3.2: Summary of the AOL Log.

3.3 MS-MARCO Web Search

The MS-MARCO Web Search dataset, published by Chen et al. [2024], contains around 10 million queries from 93 different languages. The queries were mined from Microsoft's Bing search engine. According to the publishers, the dataset closely mimics real-world web document and query distributions and was created to serve as a critical data foundation for future research in downstream tasks of information retrieval. The dataset was constructed by filtering the Bing query log from a selected time span for queries that have a click connection to the ClueWeb22 document set. ClueWeb22 is a large dataset of web documents and is aligned with the document distributions in commercial web search [Overwijk et al., 2022]. Hence, the publishers expect the query distribution of the MS-MARCO Web Search dataset to mimic real-world distributions, as well. The query set was further filtered to remove queries that are rarely triggered, contain personally identifiable information, offensive content or adult content. By rarely triggered the authors mean that queries are removed if they were triggered by less than K users, where K is a high number. Moreover, the dataset contains only unique queries. The most frequent language in the dataset is English, followed by Chinese, Japanese, German, French and Spanish.

Column Name	Description	Example	Data Type
`serp_query_text_url`	Query-string	`JFK airport`	`str`
`serp_id`	Query-ID	`1733`	`int`
`language`	Query-language	`en-US`	`int`
Data Generation	Bing queries with a clicked connection to ClueWeb22		
Languages	Multilingual, 93 languages		
Time Span	-		
Num. Rows	10 M.		

Table 3.3: Summary of the MS-MARCO Web Search Log.

3.4 ORCAS

The Open Resource for Click Analysis in Search (ORCAS) dataset was provided by Craswell et al. [2020]. It is, like MS-MARCO Web Search, based on a subsample of Bing's query logs. In this case, the subsample stems from a 26-month period up to January 2020. Queries of this dataset were aggregated based on a click-connection to the TREC Deep Learning documents. The queries of TREC DL were selected in a way that favored natural language questions. Since ORCAS' queries were selected based on a connection to the TREC DL documents, they still might be biased towards questions but, according to Craswell et al. [2020], also have words in the top-10 that are more rare in TREC DL such as "www", "the", "best" and "free". The queries of ORCAS were filtered for potentially offensive queries, like queries related to hate speech or pornography, and for queries that had very negative post-click signals, such as a short dwell time. Additionally, only english queries which were typed by K different users from the United States for a high value of k were kept in the dataset. The dataset contains around 10 million unique queries and is intended for web mining, query autocompletion and ranking tasks.

Column Name	Description	Example	Data Type
`serp_query_text_url`	Query-string	`"weather"`	`str`
`serp_id`	Query-ID	`120133`	`int`
Data Generation	Bing queries with a clicked connection to TREC Deep Learning documents		
Languages	English		
Time Span	26-month period up to January, 2020		
Num. Rows	10 M.		

Table 3.4: Summary of ORCAS.

Chapter 4

Experiments

This chapter covers the analyses that we carry out on the involved query logs and related information regarding the used computational framework and general setup. For each analysis, we again provide a brief motivation and, if necessary, add information to complete the theoretical background. Furthermore, we describe in detail how the analysis is conducted. Eventually, we evaluate each analysis and discuss the obtained results. In this section, we provide detailed descriptions on how we carry out the analyses. Though, in order to precisely reprodcue the obtained results, we refer to this GitLab-Repository [1]. The repository contains the source code to run the analyses and instructions on how to set up the environment.

4.1 Computational Framework

Since we are processing large data sets (recall that the AQL contains around 356 million queries), it is favorable to employ a distributed computing framework which allows for parallel processing and distributed memory management. Moreover, many pretrained models and ML-related libraries are available in python. Therefore, we choose *Ray*[2], an open-source distributed computing framework for python applications, as our environment for implementing and executing the analyses. Ray provides a high-level API that enables parallelizing python code without much effort. The Ray API includes methods whose call starts parallel processing of a parsed function. By this, we can easily parallelize our implemented functions that perform desired transformations on the data. In addition, Ray provides API-calls for parallel reading and writing of data on the distributed system, thus providing a fully parallel pipeline.

[1] `https://git.webis.de/code-teaching/theses/thesis-schneg`
[2] `https://docs.ray.io/en/latest/index.html`

4.1.1 Ray API-Calls

There are different paradigms to transform data in Ray. In this work, we mainly make use of Ray's following API-calls:

- `map_batches()`: This API-call is used to perform transformations on the data set. In this case, transformations are carried out on batches of the data set, enabling vectorized transformations from, e.g., numpy operations. Map_batches() takes a user-defined function or a callable class as an argument and applies it to each batch of the data set. The execution is configurable: We can specifiy the the number of parallel workers, the amount of required memory for each worker or the number of CPUs or GPUs for each worker. Map_batches() is Ray's preferable API-call for performing offline batch inference.

- `flat_map()`: This API-call is mainly used to extract specific elements of the data. It is detached from the constraint of returning at most one result per row and can handle multiple results from each row. The flat_map()-call is applied to each row of the data set and returns a new data set with the extracted elements. The resulting data set can be of different size than the original one. As in map_batches(), we can specify the configuration of execution.

- `groupby()`: This API-call is used to group the data set by a specified key. It returns a new data set with the grouped elements. In addition, we can apply further functions to each group, e.g., we can apply a function to count the number of elements in each group. This API-call does not support specifiying the configuration of execution.

4.2 Structure-related Statistics

In this section, we generate a set of structure-related statistics from the query logs. The analyses in this section are based on the findings of Section 2.1. Additionally, aspects of the analysis of named entities are included in this section, too. The goal is to perform a comparison of the query logs' linguistic and structural composition, initially neglecting semantic characteristics. We collect a set of distributions from the cleaned AQL and the other involved query logs. Ultimately, we evaluate similarities of the distributions by computing distances between them. This allows for identifying syntactical differences or similarities.

For this analysis, we look at queries from different syntactic perspectives

and carry out measurements in the defined perspectives. We define the following categories as syntactic perspectives:

1. **Queries**
2. **Named Entities**
3. **Words**
4. **Characters**

Even though named entities are not considered a syntactic category primarily, we include them in this analysis since they are frequent enough to be regarded a structural element of queries. Also, their analysis might provide an additional valuable insight to the structure of query logs. To add on that, we recall from Section 2.3 that up to 70% of queries contain named entities.

For each of the aforementioned categories (queries, named entities, words and characters), we carry out two types of measurements:

1. **Frequencies of Linguistic Elements:** We extract all elements of a category from the query log and determine the frequency of each element. For instance, we extract all existing words from the query log and measure each word's frequency. We proceed accordingly for all categories.
2. **Length-related Frequencies:** We measure the lengths of all extracted items from a category in terms of all possible subcategories. The defined syntactic categories are subject to a hierachical order, e.g., queries can be described as a set of named entities, words or characters. Words, in turn, can not be described as a set of named entities. Accordingly, we measure lengths of queries in terms of named entities, words and characters. Named entities are measured by the count of words or characters. By continuing this procedure for all categories, we gain a thorough set of measurements for each query log.

4.2.1 Frequencies of Linguistic Elements

To obtain the frequency of liguistic elements, we first extract all elements of a category from the query log and subsequently measure each element's frequency. Besides reading and writing the data, this experiment consists of two major steps in the Ray environment:

1. **Extraction of Linguistic Elements:** We apply the `flat_map()` API-call to extract all elements of a linguistic category from the query log.

Max. Number of Workers	32
Max. Number of CPUs per Worker	1
Max. Memory per Worker	12 GB
Max. Duration	24h
Used Models	*spaCy-Tokenizer*, *spaCy-NER*

Table 4.1: The parameter values of this table are the extreme values of the configuration to run the extractions of linguistic elements. Both models, the spaCy-Tokenizer and the spaCy-NER are part of spaCy's `en_core_web_sm`-model, which was used for this analysis.[3]

For named entities and words, we parse a spaCy-model to the API-call that performs named entity recognition or tokenization into words. The model is applied to each query of the data set and returns the found elements which are appended to the result set. In the end, we obtain a data set that contains all extracted elements of a liguistic category.

2. **Frequency Measurement:** We apply the `groupby()` API-call to group the data set by the extracted linguistic elements. Thereon, we call a `count()` to count the items of each group, which provides us with each elements's frequency in the query log. Eventually, we obtain a data set with the linguistic elements and their frequencies.

In Table 4.1, we note some key parameters of the experiments to enhance reproducibility. We only list the parameters' extreme values, such as the maximum number of used workers, to indicate the maximum of required ressources.

Evaluation of Linguistic Elements

As stated in Section 2.1, a well-studied phenomenon is the frequency distribution of terms in query logs. Several studies conclude that ranked frequency distributions of terms in query logs resemble Zipf's law. Zipf's law originates from linguistics and is meaningful, among other things, to describe frequency distributions of words in natural language texts [Piantadosi, 2014]. The law states that the frequency f of an element is inversely proportional to its rank r in the frequency table with some scaling constant c and exponent $\alpha \approx 1$:

$$f \propto \frac{c}{r^{\alpha}} \tag{4.1}$$

We investigate the frequencies of all considered categories, namely queries, named entities, words and characters, and evaluate two things: the similarities of the resulting distributions and if they resemble Zipf's law. To achieve

[3] `https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.7.1`

this, we take two approaches: Firstly, we visualize the results in log-log-scaled graphs to assess the proximity to Zipf's law. A constant slope would suggest a Zipfian distribution. Albeit primarily studied for words, we attempt to retrieve Zipf's law also in the frequencies of queries, named entities and characters since they as well are linguistic categories and probably follow linguistic dynamics. Secondly, we numerically compare the distributions by measuring distances between them. In this case, we are primarily interested in the differences of two groups: the pairwise distances of the AQL to each of the other query logs and the pairwise distances within the comparison group, namely AOL, MS-MARCO Web Search and ORCAS.

For evaluation we consider the frequencies of the linguistic elements sorted in descending order, i.e., rank-size distributions. Accordingly, we initially only evaluate the relationship of rank and frequency.

Rank-size Distributions of Named Entities and Words

To assess similarities to Zipf's law, we display the frequencies sorted in descending order in a log-log-scaled graph. Figure 4.1 shows the ordered frequencies of named entities and words in the query logs. As for named entities, we can state that all distributions fairly resemble Zipf's law. The distributions show a relatively constant slope in the log-scaled dimensions. The slope of the word frequencies, in turn, is not as constant. We can observe a small deviation: The constant slope is interrupted by a small curvature in the central part of the distribution. However, this curvature is present in all distributions. Consequently, from a visual perspective, we can conclude that the distributions of the different data sets are quite similar. No striking outliers which indicate clear differences are visible. Though, the rank-size distribution of named entities is more similar to a Zipfian distribution than the distribution of words.

Rank-size Distributions of Queries and Characters

In Figure 4.2, the frequencies of queries and characters are displayed in the log-scaled graph. The query distributions as well comply reasonably well with Zipf's law. The distributions are similar and a constant slope is present in the log-scaled dimensions. In turn, the distributions of characters do not resemble Zipf's law. They are actually quite different among the query logs. Though, it is striking that AOL and ORCAS are similar as well as MS-MARCO Web Search and AQL. This is probably due to their language distribution. AQL and MS-MARCO Web Search are multilingual query logs, whereas AOL and ORCAS are English-only query logs. Hence, the character distributions of the English-only query logs are similar and the distributions of the multilingual query logs are also similar. Additionally we can observe a significant difference

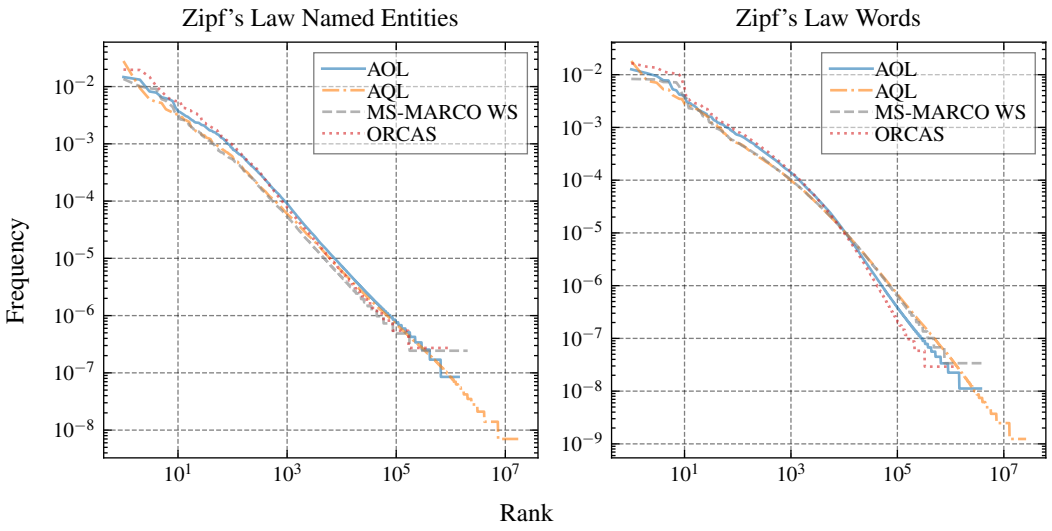

Figure 4.1: The relative frequencies of extracted named entities and words are displayed in a log-scaled graph for each query log. The frequencies were ordered in a descending order to create a ranking. A straight line in the log-scaled dimensions indicates similarities to Zipf's law.

in the number of characters. The multilingual query logs contain much more characters then the English-only logs which seems logical. Moreover, all query logs commonly show that there is a group of very frequent and rather uniformly distributed characters while the frequencies of less frequent characters are decreasing even more rapidly than Zipf's law would suggest. This might be the case because alphanumeric characters are probably significantly more frequent than special characters, indicating why the distribution is even more skewed.

Numeric Comparison: Linguistic Elements

The distributions of linguistic elements are now compared numerically to each other. To evaluate similarities between the resulting distributions, we apply the Wasserstein distance with $p = 1$. For each analysis, we compute the Wasserstein distance between the distributions of all pairs of query logs. In Table 4.2 the obtained Wasserstein distances are displayed with regard to the conducted analysis. To assess if the AQL's distributions are similar or rather anomalous, we first calculate the average Wasserstein distance of the comparison group, i.e., AOL, MS-MARCO Web Search and ORCAS. We denote the average Wasserstein distance of the group excluding the AQL by $W_\mu(\overline{AQL})$. Secondly, we determine the average Wasserstein distance between the AQL and the other query logs. For this we consider all possible pairs that contain

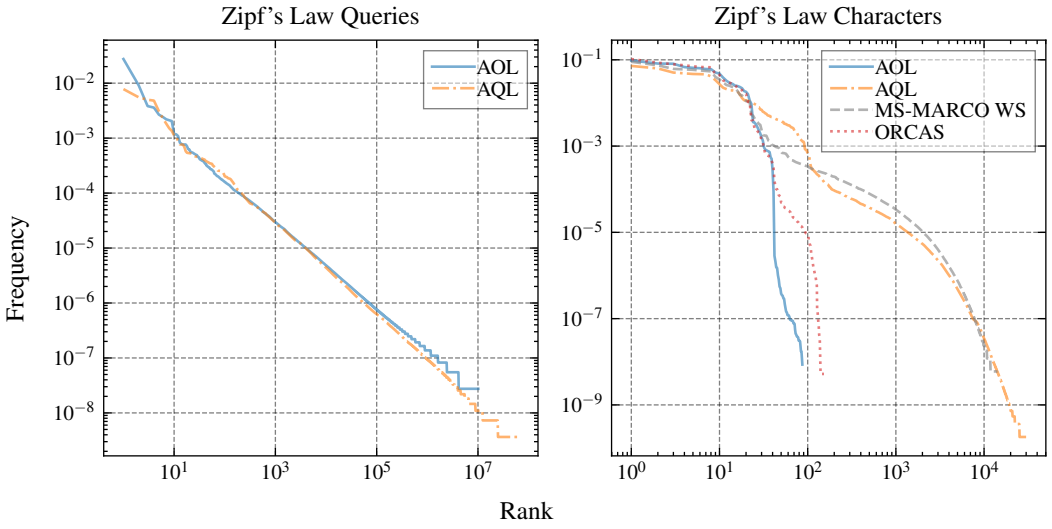

Figure 4.2: The relative frequencies of queries and extracted characters are displayed in a log-scaled graph for each query log. The frequencies were ordered in a descending order to create a ranking. A straight line in the log-scaled dimensions indicates similarities to Zipf's law.

the AQL and one of the other query logs. We call this average Wasserstein distance $W_\mu(AQL)$. If $W_\mu(AQL) \leq W_\mu(\overline{AQL})$ is true, we can argue that the AQL's distributions are similar enough to be considered a part of the comparison group. If ths is not the case, we can conclude that there are significant differences. For a further evaluation, we compute the standard deviation of the Wasserstein distances within the comparison group. We denote the standard deviation of distances within the group excluding the AQL by $\sigma_{\overline{AQL}}$. This allows us to put differences of $W_\mu(AQL)$ and $W_\mu(\overline{AQL})$ into the perspective of average deviations within the comparison group. In all three cases, named entities, words and characters, $W_\mu(AQL)$ is smaller than $W_\mu(\overline{AQL})$. This indicates that the AQL's distributions are similar to the comparison group, suggesting an non-anomalous behaviour of rank-size distributions of linguistic elements when compared to "true" query logs.

Top Queries

To further evaluate the linguistic elements, we take a look at the top 20 most frequent queries in the query logs. The most frequent queries are listed in Table 4.5. Since MS-MARCO Web Search and ORCAS only contain unique queries, we only have data from the AQL and AOL available to create this table. Moreover, only top queries with latin characters are considered for this

	Named Entities			
	AOL	AQL	MS WS	ORCAS
AOL	-	91895	391408	27423
AQL	91895	-	301466	97748
MS WS	391408	301466	-	397262
ORCAS	27423	97748	397262	-

	WS-Values	
$\mathbf{w}_\mu(\overline{\mathbf{AQL}})$		272K
$\mathbf{w}_\mu(\mathbf{AQL})$		163K
$\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})$		108K
$\sigma_{\overline{\mathbf{AQL}}}$		172K
$\frac{\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})}{\sigma_{\overline{\mathbf{AQL}}}}$		0.63

	Words			
	AOL	AQL	MS WS	ORCAS
AOL	-	55120	216678	77063
AQL	55120	-	161679	132184
MS WS	216678	161679	-	293741
ORCAS	77063	132184	293741	-

	WS-Values	
$\mathbf{w}_\mu(\overline{\mathbf{AQL}})$		195K
$\mathbf{w}_\mu(\mathbf{AQL})$		116K
$\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})$		79K
$\sigma_{\overline{\mathbf{AQL}}}$		90K
$\frac{\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})}{\sigma_{\overline{\mathbf{AQL}}}}$		0.89

	Characters			
	AOL	AQL	MS WS	ORCAS
AOL	-	57	92	1
AQL	57	-	44	57
MS WS	92	44	-	92
ORCAS	1	57	92	-

	WS-Values	
$\mathbf{w}_\mu(\overline{\mathbf{AQL}})$		61
$\mathbf{w}_\mu(\mathbf{AQL})$		53
$\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})$		8.9
$\sigma_{\overline{\mathbf{AQL}}}$		43
$\frac{\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})}{\sigma_{\overline{\mathbf{AQL}}}}$		0.21

Table 4.2: On the left: Wasserstein distances of rank-size distributions of linguistic elements. On the right: various values to evaluate the AQL's similarity to the comparison group, which were motivated in the preceding paragraph (see Section 4.2.1). The Wasserstein distances are computed with $p = 1$.

table. This is due to the fact that the AQL contains a lot of queries with non-latin characters, e.g., Chinese characters. Since we can not compare these queries with queries from AOL, we replace the six queries of the AQL's top 20 with the next most frequent queries of latin characters.

It is striking that the top queries of the AOL log are mostly navigational queries, either domain names like "www.yahoo.com" or "www.google.com" or the mere names of those websites, e.g., "yahoo" or "google". The AQL's top queries are more diverse and contain a lot of queries that are not navigational. They don't seem very coherent and quite random. A lot of them are even difficult to interpret, such as "place:86f203b1e5d", "{srch_str}" or "p2045576.m1710". Some top queries of the AQL even are parts of URLs, e.g., "http://extras.denverpost.com/media/maps /kml/co...". In summary, the top queries are substantially different and the AQL's top queries don't seem to

reflect realistic user behaviour.

Top Words

Besides the top 20 queries, we additionally take a look at the top 20 most frequent words in the query logs. The most frequent words are listed in Table 4.3 and Table 4.4. To evaluate similarities, we take a look at intersections of the top words. Particularly, we first determine the average intersection of top words within AOL, MS-MARCO Web Search and ORCAS. Considering a collection G of n sets of top words W_i, the average cardinality of the pairwise intersections is given by:

$$S_\mu(G) = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} |W_i \cap W_j| \tag{4.2}$$

We denote the average cardinality of the group excluding the AQL by $S_\mu(\overline{AQL})$. Secondly, we determine the average intersection of top words between the AQL and the other query logs. For this we consider all possible pairs that contain the AQL and one of the other query logs. We call this average cardinality $S_\mu(AQL)$. We obtain:

- $S_\mu(\overline{AQL}) = 11.66$
- $S_\mu(AQL) = 9.33$

To evaluate how similar the two cardinalities are, we compute the deviation present in the group $\overline{AQL}$ and assess if the deviation of the AQL's average cardinality matches the standard deviation of cardinalities within the comparison group $\overline{AQL}$. For the standard deviation of the group $\overline{AQL}$ we obtain:

- $\sigma_{\overline{AQL}} = 1.25$

Even though the cardinalities of intersections of top words seem to be in the same range, the deviation of the AQL is slightly higher than the standard deviation in the comparison group. We could conclude that we observe similarites but, despite this, the AQL's top words are still too different to the comparison group to be considered a part of this group. However, the queries of ORCAS and MS-MARCO Web Search are biased to more quetion-like queries, naturally exhibiting higher similarities. Taking this into account, we could also argue that the AQL's top words are not too different from the other query logs.

	AOL			AQL		
Rank	Word	Count	Ratio	Word	Count	Ratio
1	of	1,126,030	1.26%	=	29,301,944	1.81%
2	in	946,200	1.06%	of	12,284,222	0.76%
3	the	839,233	0.94%	and	11,385,504	0.71%
4	for	698,847	0.78%	the	8,532,900	0.53%
5	and	692,798	0.78%	site	8,466,226	0.52%
6	to	471,360	0.53%	in	6,584,034	0.41%
7	free	450,322	0.50%	for	6,084,390	0.38%
8	a	373,919	0.42%	-wikipedia	5,620,112	0.35%
9	google	366,059	0.41%	to	5,588,686	0.35%
10	's	359,448	0.40%	vector	4,544,168	0.28%
11	new	270,823	0.30%	finance	4,353,432	0.27%
12	http	263,056	0.29%	$	4,211,514	0.26%
13	on	254,673	0.29%	from	4,105,194	0.25%
14	pictures	236,860	0.27%	-site	3,983,136	0.25%
15	county	232,176	0.26%	on	3,794,202	0.24%
16	yahoo	219,822	0.25%	a	3,675,640	0.23%
17	how	209,175	0.23%	kak	3,574,962	0.22%
18	lyrics	190,043	0.21%	2	3,554,028	0.22%
19	my	188,189	0.21%	de	3,349,984	0.21%
20	school	183,790	0.21%	free	3,132,780	0.19%

Table 4.3: Ranking of the top 20 words in the AOL and AQL. The ratio is computed by dividing the count of a word by the total number of words in the query log.

	MS-MARCO Web Search			ORCAS		
Rank	Word	Count	Ratio	Word	Count	Ratio
1	the	245,317	0.83%	of	567,790	1.65%
2	de	242,823	0.82%	to	484,322	1.41%
3	in	239,674	0.81%	in	465,981	1.36%
4	to	214,375	0.73%	how	393,224	1.15%
5	for	212,892	0.72%	what	374,409	1.09%
6		212,760	0.72%	is	367,364	1.07%
7	of	211,873	0.72%	for	355,161	1.03%
8	a	175,659	0.59%	the	343,583	1.00%
9	and	127,821	0.43%	a	268,469	0.78%
10	is	107,218	0.36%	and	192,209	0.56%
11	how	104,567	0.35%	on	118,872	0.35%
12	on	68,471	0.23%	online	117,947	0.34%
13	2021	67,953	0.23%	free	106,799	0.31%
14	what	66,618	0.23%	does	104,246	0.30%
15	download	65,442	0.22%	definition	102,850	0.30%
16	sale	65,000	0.22%	best	98,291	0.29%
17	online	64,211	0.22%	do	97,787	0.28%
18	en	62,309	0.21%	login	94,548	0.27%
19	free	61,086	0.21%	's	93,077	0.27%
20	la	59,558	0.20%	county	91,291	0.26%

Table 4.4: Ranking of the top 20 words in MS-MARCO Web Search and ORCAS. The ratio is computed by dividing the count of a word by the total number of words in the query log.

	AOL			AQL		
Rank	Query	Count	Ratio	Query	Count	Ratio
1	-	1,000,375	2.75%	finance	2,135,183	0.78%
2	google	332,192	0.92%	#	1,515,866	0.55%
3	ebay	139,207	0.38%	query	1,368,346	0.50%
4	yahoo	130,538	0.35%	$	1,319,815	0.49%
5	yahoo.com	97,518	0.27%	speed force	850,884	0.31%
6	mapquest	88,279	0.24%	place:86f203b1e5dc4397	414,853	0.15%
7	google.com	79,991	0.22%	"Kurdish Referendum"	371,211	0.14%
8	myspace.com	77,211	0.21%	video	294,899	0.11%
9	myspace	74,365	0.20%	Latoya Cantrell	239,827	0.09%
10	myspace.com	43,036	0.12%	http://extras.denverpost.com/media/maps/kml/co...	208,918	0.08%
11	www.yahoo.com	42,597	0.12%	{srch_str}	208,725	0.08%
12	www.google.com	39,622	0.11%	la teachers strike	186,735	0.07%
13	internet	30,125	0.08%	rabble.ca	157,519	0.06%
14	http	28,516	0.08%	Sarah Stierch	146,697	0.05%
15	my space	27,887	0.08%	Aspects	143,905	0.05%
16	weather	27,845	0.08%	#communitywebs	139,392	0.05%
17	www.myspace.com	27,842	0.08%	http://dcist.com/2015/07/metros_-1000-series_ra...	130,715	0.05%
18	map quest	25,856	0.08%	prom electric	127,972	0.05%
19	ebay.com	22,893	0.07%	p2045576.m1710	124,971	0.05%
20	american idol	22,652	0.07%	GFW	123,257	0.04%

Table 4.5: Ranking of the top 20 queries in the respective query logs. Since MS-MARCO and ORCAS only contain unique queries, we only consider AQL and AOL for this table. We also only display queries with latin characters. Because of this, six queries of the AQL's top 20 are replaced with the next most frequent queries of latin characters.

4.2.2 Length-related Frequencies

Besides considering the frequency of linguistic elements, we also measure the lengths of the elements in terms of different subcategories. We describe the length of an element by the occuring counts of a possible subcategory, e.g., the length of a query by the count of characters the query contains. From the extraction of named entites and words, we have three sets of categories whose lengths we can measure: queries, named entities and words. To illustrate how we obtain the frequencies of query lengths, named entity lengths and word lengths, we provide a description of how the lengths of queries are measured. Accordingly, we proceed for all categories. The measurement of lengths is carried out in two steps:

1. **Computing Lengths:** We apply the `map_batches()` API-call to compute the lengths of queries. Since we are interested in the count of named entities, words and characters, we parse a function that performs multiple measurements on each query. We again apply spaCy's `en_core_web_sm`-model to perform the measurements (e.g., count words or named entities). For each type of measurement (i.e., entity count, word count or character count) the function appends a new column to the batch which contains the corresponding lengths. Eventually, a batch with the computed lengths is returned.

2. **Frequency Measurement:** We apply the `groupby()` API-call to group the resulting data set by the computed lengths. By this, we obtain a group for each length. Thereon, we call a `count()` in order to count each length's occurence in the query log. In the end, we obtain a data set with the lengths and their frequencies.

In Table 4.6, we again note some key parameters of the experiments that outline the maximum of required ressources and the applied models.

Max. Number of Workers	32
Max. Number of CPUs per Worker	1
Max. Memory per Worker	9 GB
Max. Duration	22h
Used Models	*spaCy-Tokenizer*, *spaCy-NER*

Table 4.6: The parameter values of this table are the extreme values of the configuration to run the frequency measurement of lengths. Both models, the spaCy-Tokenizer and the spaCy-NER are part of the spaCy-model `en_core_web_sm`[4], which was used for this analysis.

Distribution of Character Counts: Queries and Named Entities

In Figure 4.3, the distributions of query and named entity lengths in characters are displayed. As for queries, the distributions appear to be not very close but still similar to each other. Though, the range of occuring query lengths is quite similar among all logs. Regarding the shape, AOL and AQL are similar, whereas the AQL's distribution is more noisy. ORCAS stands out presenting a binomial-like distribution. MS-MARCO Web Search's distribution is also unique but unlike ORCAS' not symmetrical. Regarding the distributions of named entities, we can again observe a common scope of lengths among all logs and similar shapes. Especially ORCAS and AOL are very similar to one another. The AQL's distribution is similar to a poisson distribution whereas distribution of MS-MARCO Web Search is not exactly assignable to a popular distribution.

Distributions of Characters per Word and Entities per Queries

In Figure 4.4 the distributions of word and query lengths measured in characters and named entities are displayed. We can observe very similar frequencies of entity counts in queries. The distributions are visually almost equivalent among the involved query logs. As for the word lengths, the distributions are also similar but depict some clear differences, as well. While MS-MARCO Web Search and ORCAS show a very similar distribution, AOL's distribution is clearly different. We can observe an unusual peak of word lengths between 10 and 20 characters. Reviewing AOL's words of this lengths showed that the most frequent words are website addresses in this range. The extensions before and after the website's name (e.g., "www." or ".com") cause a shift of frequent words towards longer words. This has been confirmed by filtering out website addresses from the AOL and subsequently measure the word legth distribution. In that case, AOL's distribution of word lengths aligns with the other query logs' distributions. Added to this, the top queries of AOL (see Table 4.5) contain many web sites and domain names, which also confirms the peak in the distribution.

Distributions of Word Counts: Queries and Named Entities

In Figure 4.5 the distributions of query and named entity lengths measured in the number of words are displayed. Concerning the lengths of named entities, we can observe almost equivalent distributions of AOL and MS-MARCO Web Search. The distribution of ORCAS is still similar to AOL and MS-MARCO Web Search while the distribution of the AQL is slightly different. The AQL

[4] `https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.7.1`

contains significantly more named entities comprised of one word than the other logs. In contrast to the other logs, named entities comprised of one word a the most common named entities. A similarity among all logs is that named entities consisting of two words are the most frequent. As of the query lengths measured in words, we can observe more diverse distributions among the different logs. While the number of words is distributed similarily in AOL and AQL, the distributions of ORCAS and MS-MARCO are different. The distribution of ORCAS resembles a poisson distribution and the distribution of MS-MARCO Web Search is not clearly assignable to a popular distribution. However, the range of the most frequent number of words is similar among all logs.

Numeric Comparison of Lengths

The resulting distributions of the measurements are now compared numerically to each other. We again apply the 1-Wasserstein distance to evaluate similarities between the resulting distributions. In Table 4.7 the obtained Wasserstein distances are displayed with regard to the conducted analysis. The table is structured in a similar way as Table 4.7. We again compute the average Wasserstein distance within the comparison group, i.e., AOL, MS-MARCO Web Search and ORCAS, denoted by $W_\mu(\overline{AQL})$. We refer to the standard deviation of wasserstein distances within that group as $\sigma_{\overline{AQL}}$. Additionally, we determine the average Wasserstein distance between the AQL and the other query logs, denoted by $W_\mu(AQL)$. It is striking that we obtain $W_\mu(AQL) > W_\mu(\overline{AQL})$ for all length-related measurements. The closest distribution of the AQL to the comparison group is the distribution of characters per word. In this case, we measure a deviation from the AQL to the comparison group of 0.3 $\sigma_{\overline{AQL}}$. The most dissimilar distribution in turn is the query length measured in characters. The deviation of this distribution is 6.1 $\sigma_{\overline{AQL}}$. Since we measure substantial deviations of the AQL's distributions we conclude that the AQL is rather anomalous to the comparison group for length-related measurements of linguistic elements.

Characters per Query

	AOL	AQL	MS	WS	ORCAS
AOL	-	5.54	2.56	3.72	
AQL	5.54	-	4.75	7.75	
MS WS	2.56	4.75	-	3.00	
ORCAS	3.72	7.75	3.00	-	

$\mathbf{w}_\mu(\mathbf{AQL})$	6.0	$\mathbf{w}_\mu(\overline{\mathbf{AQL}})$	3.1
$\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})$	-2.9	$\sigma_{\overline{\mathbf{AQL}}}$	0.5
$\frac{\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})}{\sigma_{\overline{\mathbf{AQL}}}}$	-6.1		

Characters per Entity

	AOL	AQL	MS	WS	ORCAS
AOL	-	2.12	1.79	0.80	
AQL	2.12	-	2.86	1.98	
MS WS	1.79	2.86	-	1.03	
ORCAS	0.80	1.98	1.03	-	

$\mathbf{w}_\mu(\mathbf{AQL})$	2.3	$\mathbf{w}_\mu(\overline{\mathbf{AQL}})$	1.2
$\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})$	-1.1	$\sigma_{\overline{\mathbf{AQL}}}$	0.4
$\frac{\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})}{\sigma_{\overline{\mathbf{AQL}}}}$	-2.6		

Characters per Word

	AOL	AQL	MS	WS	ORCAS
AOL	-	3.69	6.31	4.20	
AQL	3.69	-	6.29	4.17	
MS WS	6.31	6.29	-	2.20	
ORCAS	4.20	4.17	2.20	-	

$\mathbf{w}_\mu(\mathbf{AQL})$	4.7	$\mathbf{w}_\mu(\overline{\mathbf{AQL}})$	4.2
$\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})$	-0.5	$\sigma_{\overline{\mathbf{AQL}}}$	1.7
$\frac{\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})}{\sigma_{\overline{\mathbf{AQL}}}}$	-0.3		

Words per Query

	AOL	AQL	MS	WS	ORCAS
AOL	-	0.58	0.69	0.89	
AQL	0.58	-	0.81	1.25	
MS WS	0.69	0.81	-	0.44	
ORCAS	0.89	1.25	0.44	-	

$\mathbf{w}_\mu(\mathbf{AQL})$	0.9	$\mathbf{w}_\mu(\overline{\mathbf{AQL}})$	0.7
$\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})$	-0.2	$\sigma_{\overline{\mathbf{AQL}}}$	0.2
$\frac{\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})}{\sigma_{\overline{\mathbf{AQL}}}}$	-1.1		

Words per Entity

	AOL	AQL	MS	WS	ORCAS
AOL	-	0.19	0.06	0.11	
AQL	0.19	-	0.18	0.30	
MS WS	0.06	0.18	-	0.17	
ORCAS	0.11	0.30	0.17	-	

$\mathbf{w}_\mu(\mathbf{AQL})$	0.2	$\mathbf{w}_\mu(\overline{\mathbf{AQL}})$	0.1
$\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})$	-0.1	$\sigma_{\overline{\mathbf{AQL}}}$	0.04
$\frac{\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})}{\sigma_{\overline{\mathbf{AQL}}}}$	-2.5		

Entities per Query

	AOL	AQL	MS	WS	ORCAS
AOL	-	0.18	0.12	0.03	
AQL	0.18	-	0.06	0.15	
MS WS	0.12	0.06	-	0.09	
ORCAS	0.03	0.15	0.09	-	

$\mathbf{w}_\mu(\mathbf{AQL})$	0.13	$\mathbf{w}_\mu(\overline{\mathbf{AQL}})$	0.08
$\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})$	-0.05	$\sigma_{\overline{\mathbf{AQL}}}$	0.04
$\frac{\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})}{\sigma_{\overline{\mathbf{AQL}}}}$	-1.3		

Table 4.7: Wasserstein distances between the distributions of the query logs. For each analysis, e.g. characters per query, distances of all pairs of query logs are computed. Besides the distances, the table contains values that are relevant for assessing the simlarity of the AQL to the comparison group.

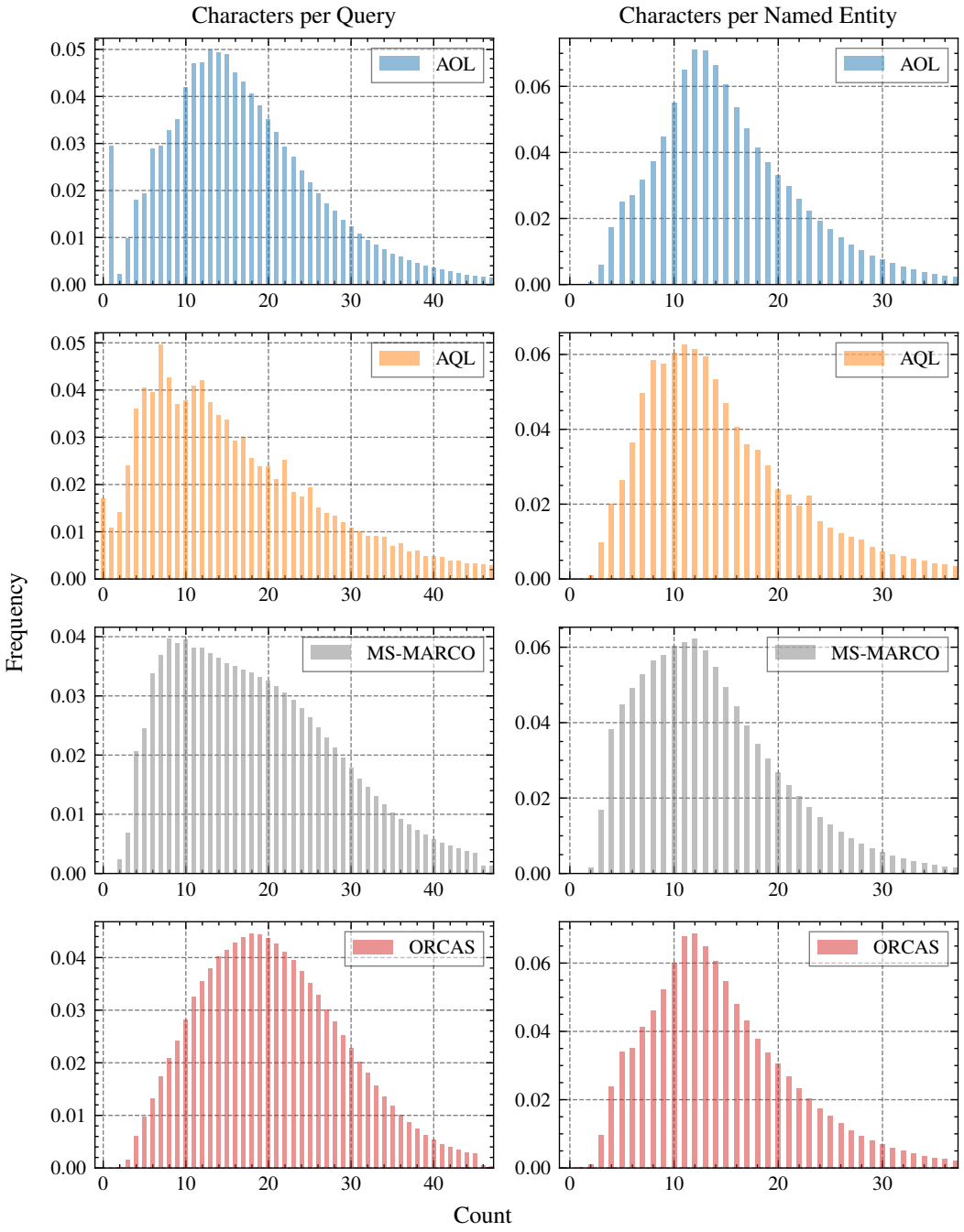

Figure 4.3: The relative frequencies of query and named entity lengths measured in the number of characters are displayed.

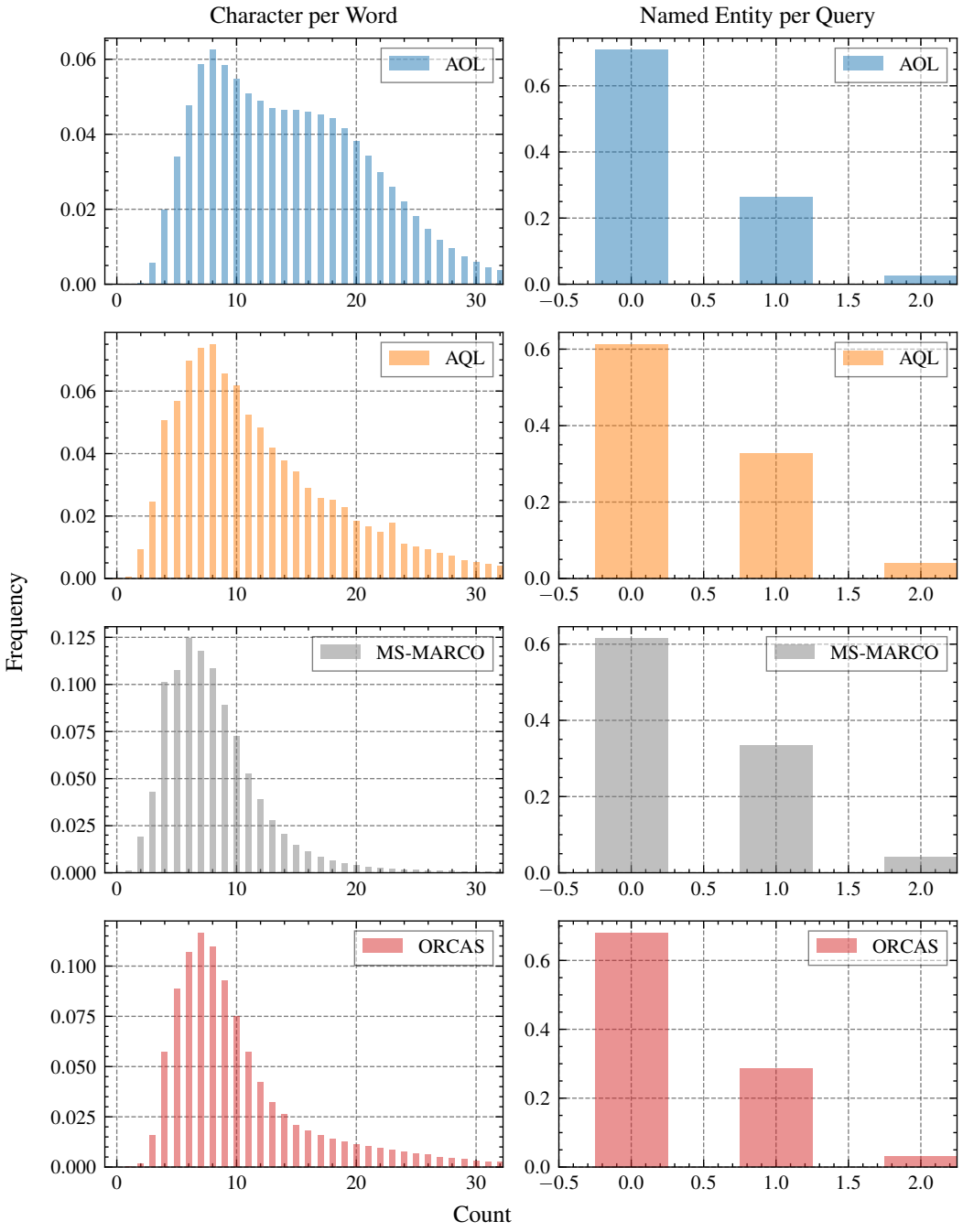

Figure 4.4: The relative frequencies of word and query lengths are displayed. In this collection, the word length is measured in number of characteres. The query length is measured in the number of occuring named entities.

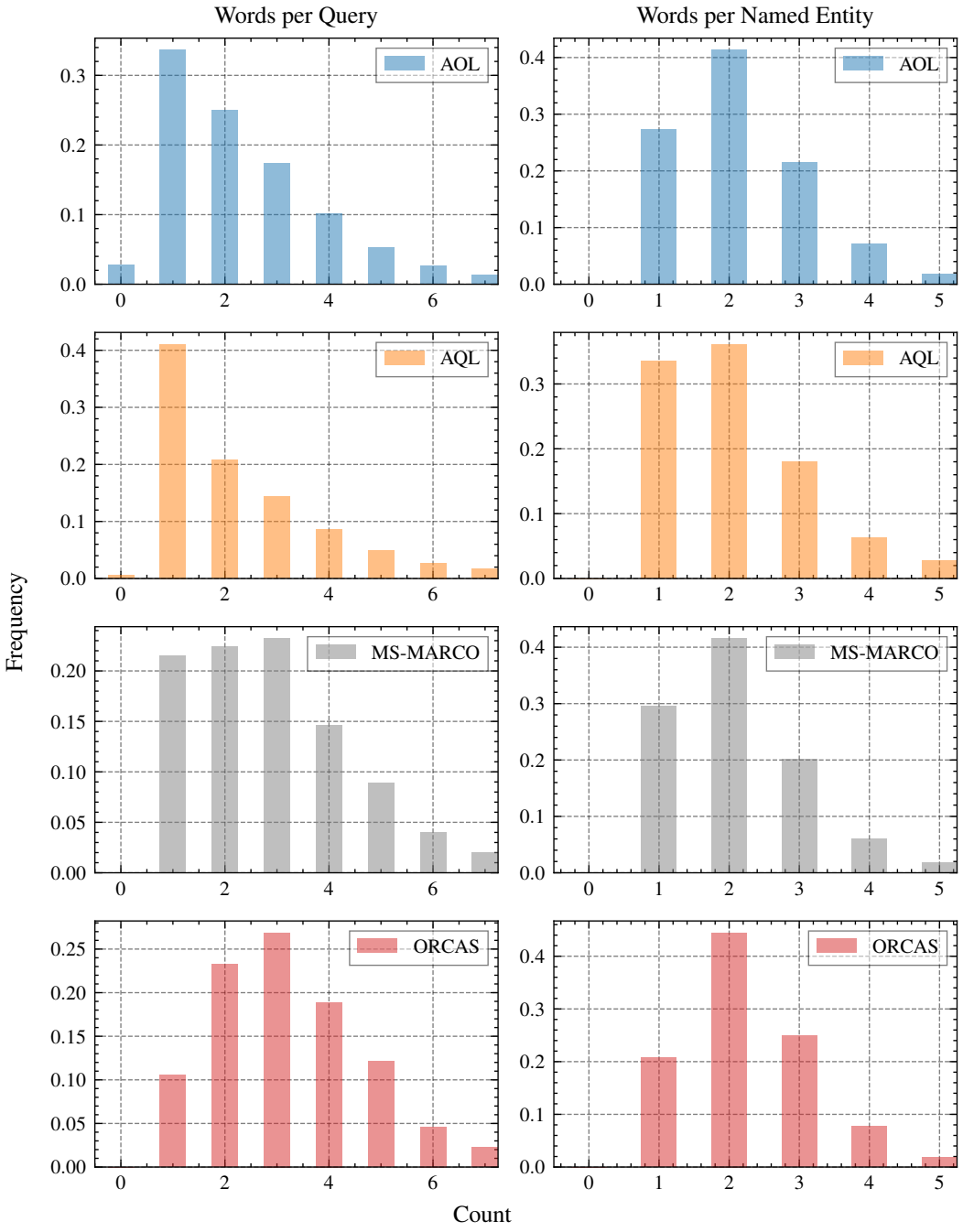

Figure 4.5: The relative frequencies of query and named entity lengths are displayed. Both, the query length and the named entity length are measured in number of occuring words.

4.2.3 Search Operators

As stated in Section 2.1, search operators are a common feature of search engines. They are used to filter the search results or specify specific requirements for the search results. In this section, we analyze the usage of search operators in the query logs. In particular, we determine the ratio of queries, that contain search operators, measure frequencies of the search-operator-count in queries and present the most prominent search operators per query log. The considered search operators of this analysis are:

- `AND`
- `OR`
- `around()`
- `site:`
- `filetype:`
- `intitle:`
- `allinurl:`
- `allintitle:`
- `intext:`
- `allintext:`
- `related:`
- `define:`
- `chache:`

Search Operator Frequencies

First, to obtain the frequencies of the considered search operators in a query log, we apply the `flat_map()` API-call to extract all search operators from the query log. We parse a function to the API-call that checks for each query if it contains one of the search operators. If so, the search operator is appended to the result set. The function simply checks for the presence of a search operator's string. Thereon, we apply the `groupby()` API-call to group the data set by the extracted search operators and call a subsequent `count()` to get the count of each search operator. Similarly we proceed to get the frequencies of search operator counts in queries. We apply the `map_batches()` API-call to compute the search operator counts of each query. The count is again obtained by a string matching. Then we apply the `groupby()` API-call to group the data set by the computed search operator counts and call a subsequent `count()` to get the frequency of each search operator count.

Evaluation Search Operators

In Table 4.8 the total frequencies of search operators, the search operator ratio and the fraction of queries with different numbers of search operators are displayed. Unfortunately, we couldn't detect any search operators in the AOL log. Since MS-MARCO Web Search and ORCAS only contain unique queries that were filtered by a certain minimum popularity threshold, only very few search operators were detected in these logs. This makes a comparison of the

Query Log	SO-count	SO-ratio	count = 0	count = 1	count = 2	count > 2
AOL	0	0%	100%	0%	0%	0%
AQL	7,663,010	2.21%	97.79%	1.83%	0.31%	0.07%
MS WS	6,727	0.07%	99.93%	0.07%	< 0.00%	< 0.00%
ORCAS	3,937	0.04%	99.96%	0.04%	< 0.00%	< 0.00%

Table 4.8: Ratios of queries that contain 0, 1, 2, 3, or more than 3 search operators.

	AQL			MS-MARCO WS			ORCAS		
Rank	SO	Count	Ratio	SO	Count	Ratio	SO	Count	Ratio
1	site:	5,507,994	1.59%	site:	6,564	0.07%	site:	3,307	0.03%
2	related:	1,471,351	0.42%	intitle:	80	0.00%	define:	624	0.01%
3	OR	391,561	0.11%	define:	24	0.00%	intitle:	4	0.00%
4	AND	188,382	0.05%	intext:	22	0.00%	related:	1	0.00%
5	intitle:	25,635	0.01%	filetype:	18	0.00%	intext:	1	0.00%
6	cache:	22,566	0.01%	cache:	12	0.00%	allintext:	0	0.00%
7	define:	21,676	0.01%	allintext:	3	0.00%	allinurl:	0	0.00%
8	allinurl:	19,451	0.01%	allintitle:	3	0.00%	allintitle:	0	0.00%
9	allinurl:	5,564	0.00%	related:	1	0.00%	filetype:	0	0.00%
10	intext:	4,950	0.00%	AND	0	0.00%	AND	0	0.00%
11	allintitle:	3,258	0.00%	OR	0	0.00%	OR	0	0.00%
12	allintext:	619	0.00%	around():	0	0.00%	around()	0	0.00%
12	around()	3	0.00%	allinurl:	0	0.00%	chache:	0	0.00%

Table 4.9: Ranking of the used search operators (SO) in the respective query logs. The table does not inlcude data from the AOL log because not a single occurence of the considered search operators was found in it.

AQL to the other logs difficult and less meaningful. In Table 4.9 a ranking of the considered search operators for each query log is displayed. Again, this comparison lacks meaningfulness since the search operators are not used frequently in the comparison group. Though, the search operator "site:" is the most frequent search operator in all query logs. Apart from this commonality, no other meaningful results are visible.

4.3 Inference-based Statistics

In this section, we classify queries or components of queries, like named entities, according to selected taxonomies. As described in Section 2.2, the intent behind a query is a meaningful taxonomy according which queries are classified. We also investigate the presence of personally identifiable information (PII) entities in the query logs and their distribution.

4.3.1 Query Intent

As a first step, we classify queries into the categories informational, navigational and transactional. Since the classifier from Alexander et al. [2022] performs fairly accurate, we apply it in this work. In their work, Alexander et al. [2022] train different models to perform intent classification. We apply the BERT-based model from their collection in this work which achieves an accuracy of 0.90. Since this classifier of was trained on english queries only, we apply it only to the english subset of the multiligual query logs (AQL and MS-MARCO Web Search). We produce the labels by passing the model to Ray's preferable API-call `map_batches()` for offline batch inference. After that, we apply the `groupby()` API-call to group the data set by the labels and get counts for each label. The classifier is not provided publicly, but was obtained by a personal request to Alexander et al. [2022]. In Table 4.10 we note some key parameters of the experiments to enhance reproducibility. We only list the parameters' extreme values, indicating the maximum of required ressources.

Query Intent Distributions

In Figure 4.6 the query intent distributions of the query logs are displayed. The labels were produced by the classifier of Alexander et al. [2022]. The figure shows an equal ranking of the labels. The most frequent label is "informational" followed by "navigational" and "transactional". Apart from the ratio of navigational queries, the distributions are very similar.

4.3.2 PII Entity Labels

Another aspect that we consider is the prevalence of personally identifiable information (PII) in the query logs. We employ Microsoft's Presidio-Model[5] to recognize PII entities in the query logs and store their labels. Microsoft provides an evaluation of the model, demonstrating that it achieves a precision of 0.94 and a recall of 0.55. Again, we produce the labels by applying Ray's preferable API-call `map_batches()` for offline batch inference. A subsequent `groupby()`-call provides the counts for each label. In Table 4.10, the needed resources and related information to run the classification are listed.

Label Distributions of PII Entities

In Figure 4.7 the distributions of PII entities in the query logs is displayed. The labels were produced by Microsoft's Presidio-Model. The figure shows that the

[5] `https://github.com/microsoft/presidio`

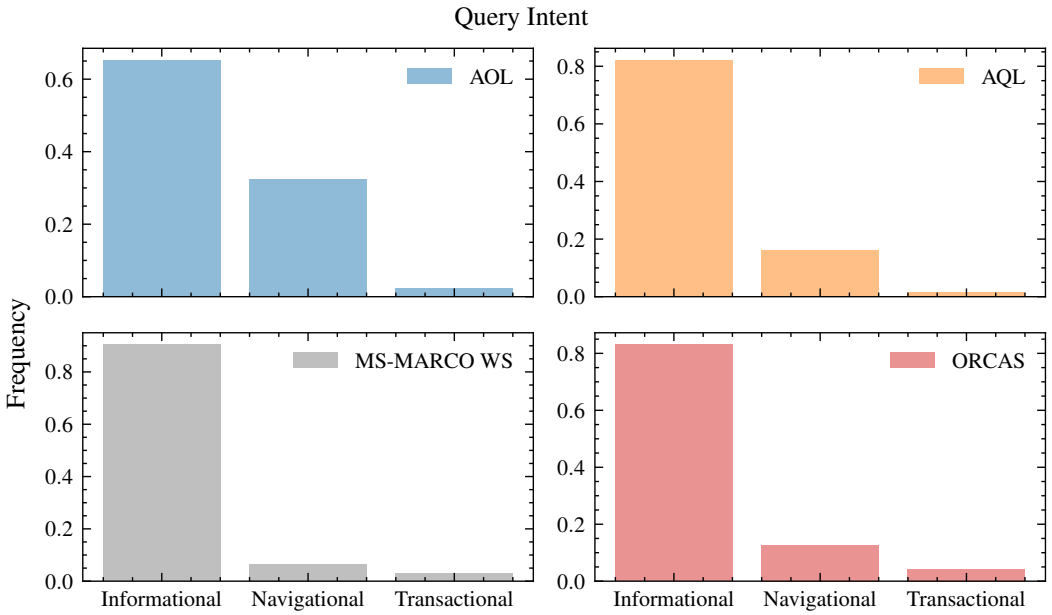

Figure 4.6: Distribution of query intents in the query logs. The labels were produced by the classifier of Alexander et al. [2022].

same four PII labels are the most frequent ones in all query logs: "Person", "Location", "NRP" (a person's Nationality, religious or political group) and "Datetime". Among these four labels however, the distribution varies across the query logs.

4.3.3 Question Classification

As a last taxonomy we consider the classification of queries into questions and non-questions. We apply a rule-based classifier provided by Reimer [2023] to classify queries. The classifier is based on a set of rules that are applied to the

	Intent Labels	**PII Labels**
Max. Number of Workers	32	320
Max. Number of CPUs per Worker	3	1
Max. Memory per Worker	7 GB	8GB
Max. Duration	1d 6h	12h
Used Models	*Intent-Classifier*	*Presidio-Model*

Table 4.10: The parameter values of this table indicate the most expensive configuration used to produce intent and PII labels

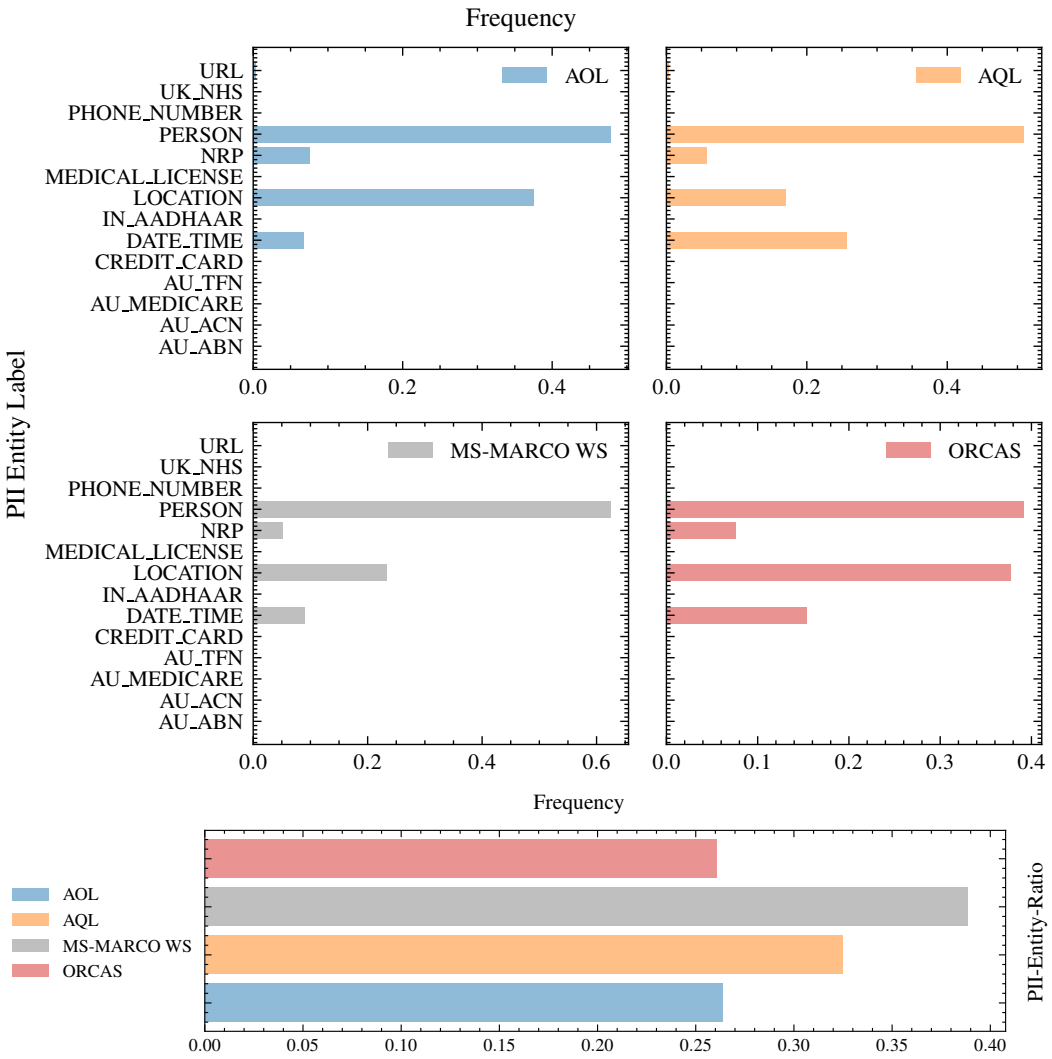

Figure 4.7: The figure displays the distribution of PII entities in the query logs. The labels were produced by Microsoft's Presidio-Model. The figure shows that the same four PII labels are the most frequent ones in all query logs: "Person", "Location", "NRP" (a person's Nationality, religious or political group) and "Datetime". Among these four labels however, the distribution varies across the query logs.

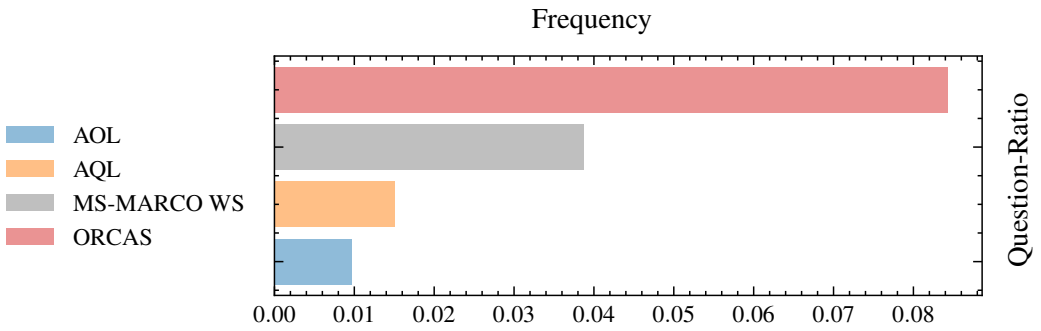

Figure 4.8: The figure displays the ratio of questions in the query logs. The labels were produced by a rule-based classifier provided by Reimer [2023].

queries. The rules are based on the presence of question words (e.g., "who", "what", "where", "when", "how") or the presence of a question mark at the end of the query. Since the classifier is designed for english queries only, we apply it only to the english subset of the multilingual query logs (AQL and MS-MARCO Web Search). In the study of Reimer [2023] the classifier achives a recall of 0.89 and a precision of 0.99. This classifier as well is not provided publicly, but was obtained by a personal request to Reimer [2023]. We produce the labels by passing the model to Ray's API-call `map_batches()` for offline batch inference and apply the `groupby()` API-call to get the counts for each label.

Distributions of Questions

In Figure 4.8 the ratio of questions in each query log is displayed. In general, the ratio of questions is low among all query logs. Not surprisingly, ORCAS contains the most questions as its generation is biased towards questions. The AQL and AOL log contain a similar ratio of questions, while the ratio of questions in the MS-MARCO Web Search log is between ORCAS and the two other logs.

Numeric Comparison: Inference-based Distributions

We compute Wasserstein distances for all pairs of distributions of the different characteristics. Again, we consider the average distance of the AQL to AOL, MS-MARCO Web Search and ORCAS and the average Wasserstein distance within AOL, MS-MARCO Web Search and ORCAS (see Section 4.2.1 for a detailed motivation). Regarding query intent and questions, we note that $W_\mu(AQL) < W_\mu(\overline{AQL})$ which indicates that the AQL is similar to the comparison group. In turn, considering the distribution of PII labels, we note

	Query Intent			
	AOL	AQL	MS WS	ORCAS
AOL	-	0.17	0.26	0.20
AQL	0.17	-	0.10	0.04
MS WS	0.26	0.10	-	0.07
ORCAS	0.20	0.04	0.07	-

WS-Values	
$\mathbf{w}_\mu(\overline{\mathbf{AQL}})$	0.18
$\mathbf{w}_\mu(\mathbf{AQL})$	0.10
$\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})$	0.08
$\sigma_{\overline{\mathbf{AQL}}}$	0.08
$\frac{\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})}{\sigma_{\overline{\mathbf{AQL}}}}$	1.00

	PII Labels			
	AOL	AQL	MS WS	ORCAS
AOL	-	0.22	0.17	0.09
AQL	0.22	-	0.17	0.22
MS WS	0.17	0.17	-	0.23
ORCAS	0.09	0.22	0.23	-

WS-Values	
$\mathbf{w}_\mu(\overline{\mathbf{AQL}})$	0.16
$\mathbf{w}_\mu(\mathbf{AQL})$	0.21
$\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})$	-0.04
$\sigma_{\overline{\mathbf{AQL}}}$	0.06
$\frac{\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})}{\sigma_{\overline{\mathbf{AQL}}}}$	-0.75

	Questions			
	AOL	AQL	MS WS	ORCAS
AOL	-	0.5e-2	2.9e-2	7.5e-2
AQL	0.5e-2	-	2.4e-2	6.9e-2
MS WS	2.9e-2	2.4e-2	-	4.6e-2
ORCAS	7.5e-2	6.9e-2	4.6e-2	-

WS-Values	
$\mathbf{w}_\mu(\overline{\mathbf{AQL}})$	0.05
$\mathbf{w}_\mu(\mathbf{AQL})$	0.03
$\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})$	0.02
$\sigma_{\overline{\mathbf{AQL}}}$	0.02
$\frac{\mathbf{w}_\mu(\overline{\mathbf{AQL}}) - \mathbf{w}_\mu(\mathbf{AQL})}{\sigma_{\overline{\mathbf{AQL}}}}$	0.90

Table 4.11: On the left: Wasserstein distances of inference-based distributions. On the right: various values to evaluate the AQL's similarity to the comparison group, which were motivated in Section 4.2.1. The Wasserstein distances are computed with $p = 1$.

that $W_\mu(AQL) > W_\mu(\overline{AQL})$ which indicates that the AQL is less similar to the comparison group. However, the deviation is equal to $0.75\sigma_{\overline{AQL}}$ which is not very significant. In summary, we can state that the AQL is reasonably similar for the two insights "query intent" and "questions" but less similar for the insight "PII labels".

4.3.4 Case Study: Probabilistic Approach

So far we have calculated mean and standard deviation of Wasserstein distances to assess whether the AQL is similar or anomalous to the comparison group of AOL, MS-MARCO Web Search and ORCAS. Following this logic, one could estimate a probability distribution from the Wasserstein distances and use it to model the "true" distribution for a considered characteristic.

From the pairwise calculation of Wasserstein distances, we obtain a distance matrix $D \in \mathbb{R}^{4\times4}$. Since the Wasserstein distance is a metric, a configuration of the distributions in three dimensions is naturally induced. We denote the space of this embedding the Wasserstein space $\mathcal{W}$. Provided that $\mathcal{W}$ is a Euclidean space in our use cases, mean and standard deviation can be calculated straightforward. Given that the ground truth W is normally distributed in the Wasserstein space,

$$W \sim \mathcal{N}(\mu, \sigma^2)$$

one could predict the probability of a distribution belonging to the ground truth by the distance to the mean of the ground truth. An estimation of the ground truth mean and standard deviation is obtained by the mean $W_\mu(\overline{AQL})$ and standard deviation $\sigma_{\overline{AQL}}$ of distances within the comparison group. According to this, the probability of a new sample belonging to ground truth can be estimated by its deviation from the ground truth mean. For simplification, we express the deviation of the AQL from the mean of the comparison group $W_\mu(\overline{AQL})$ in the number of sigma deviations $\sigma_{\overline{AQL}}$ which is also called z-score. The probabilities of samples with a certain z-score are straightforward:

$$\begin{aligned} P(|X-\mu| \geq \sigma) &= 0.32 \\ P(|X-\mu| \geq 2\sigma) &= 0.04 \\ P(|X-\mu| \geq 3\sigma) &= 0.003 \end{aligned}$$

From Table 4.2, Table 4.7 and Table 4.11 we can now obtain the calculated z-scores for the AQL. For simplification we round the z-scores to integers. For a z-score of 0 we assign a probability of 100% and a z-score that is higher than 4 we assign 0%. The resulting probabilites are displayed in Table 4.12. Averaging these probabilites would give us a probability of 58.66% that the AQL belongs to the comparison group. However, this approach is simplified and requires the satisfaction of many assumptions. In Figure 4.10 and 4.9 a visualization of the probabilistic approach is displayed. The original configuration in 3D was embedded into 2D by applying multidimensional scaling (MDS) [Kruskal, 1964].

Lengths	100%	100%	100%
Elements	0%	4%	100%
Elements	32%	4%	32%
Inference-based	100%	32%	100%

Table 4.12: Z-scores of the AQL to the comparison group of AOL, MS-MARCO Web Search and ORCAS for the different characteristics of the query logs

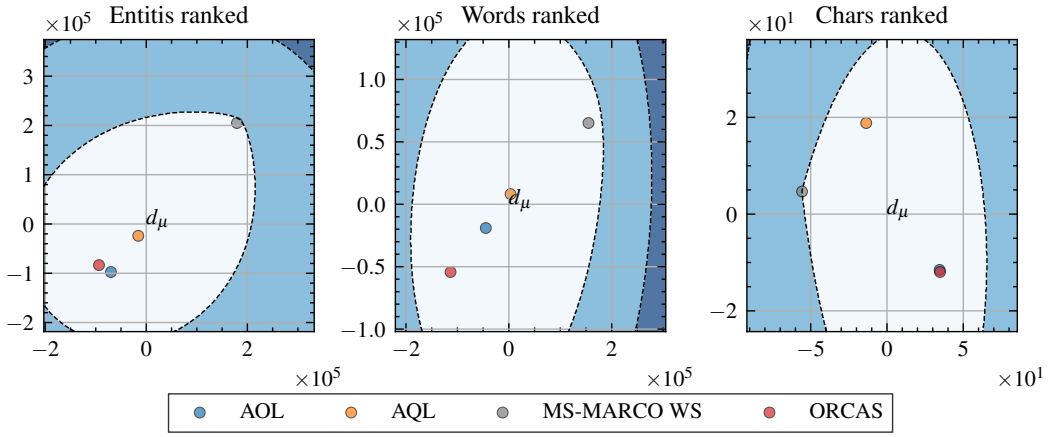

Figure 4.9: Wasserstein distances of rank-size distributions of linguistic elements. The distance matrix induces a configuration of the four distributions in $\mathbb{R}^3$. By applying multidimensional scaling (MDS) [Kruskal, 1964], we obtain the visualized embedding in $\mathbb{R}^2$. The white area corresponds to a point's average distance d_μ to the comparison group that satisfies $d_\mu < W_\mu(\overline{AQL})$. A point's location in one of the colored areas satisfies $W_\mu(\overline{AQL}) + n \cdot \sigma_{\overline{AQL}} \leq d_\mu \leq W_\mu(\overline{AQL}) + (n+1) \cdot \sigma_{\overline{AQL}}$ for $n \in \mathbb{N}_0$.

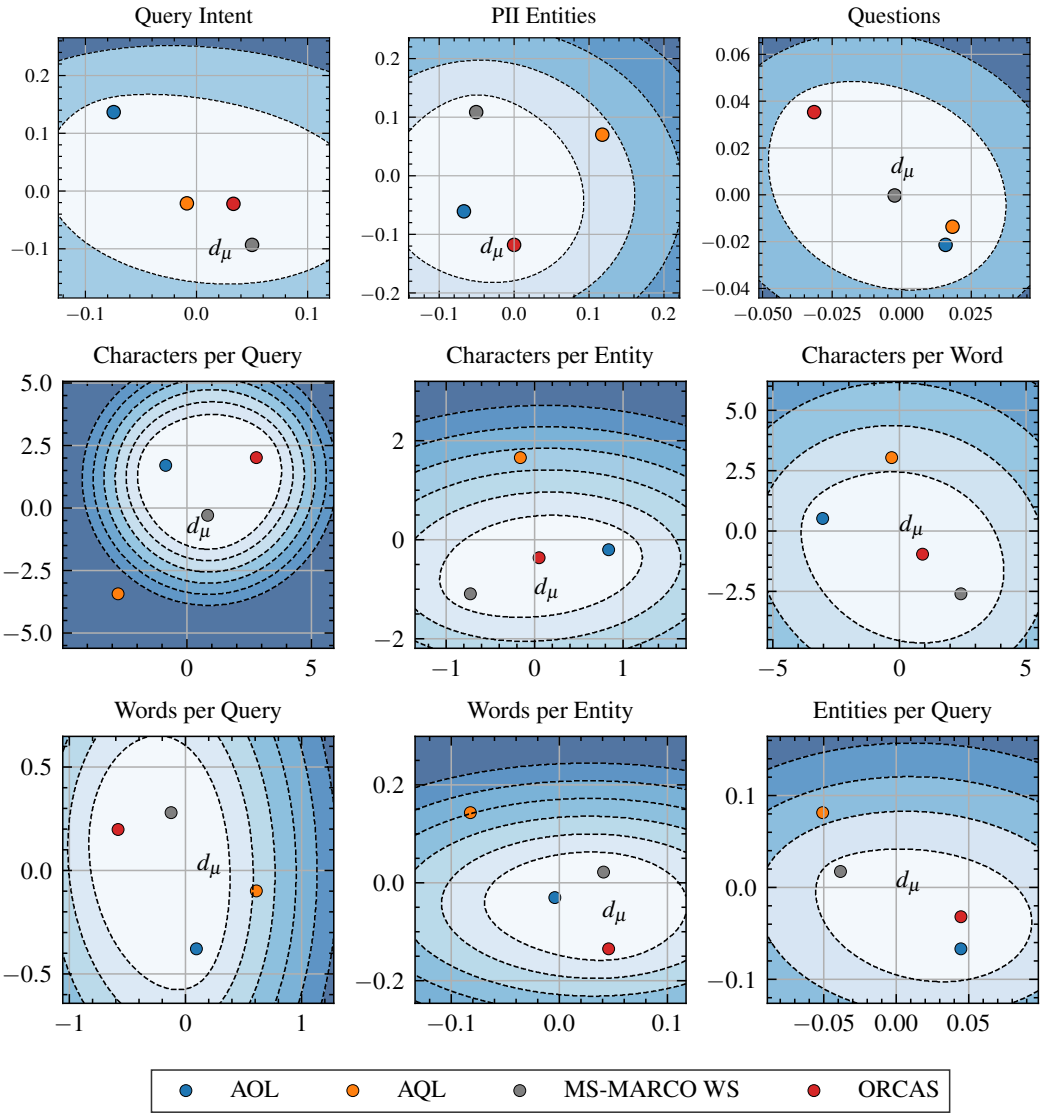

Figure 4.10: Wasserstein distances of distributions of lengths. The distance matrix induces a configuration of the four distributions in $\mathbb{R}^3$. By applying multidimensional scaling (MDS) [Kruskal, 1964], we obtain the visualized embedding in $\mathbb{R}^2$. The white area corresponds to a point's average distance d_μ to the comparison group that satisfies $d_\mu < W_\mu(\overline{AQL})$. A point's location in one of the colored areas satisfies $W_\mu(\overline{AQL}) + n \cdot \sigma_{\overline{AQL}} \leq d_\mu \leq W_\mu(\overline{AQL}) + (n+1) \cdot \sigma_{\overline{AQL}}$ for $n \in \mathbb{N}_0$.

4.4 Temporal-based Analysis

In this section, we aim at performing temporal-based comparisons of real-world queries and queries of the AQL. To realize this, we select a set of Google queries which we download from the tool "Google Trends". We select two sets, one with annual top queries and one with monthly top queries. As for the annual top queries, we simply compare the top queries of Google Trends with the AQL's top queries. For the monthly top queries, we compare temporal patterns of real-world queries and queries of the AQL. We expect monthly top queries to be more volatile that annual top queries which is why we prefer them for a comparison of temporal patterns. We carry out this comparison by computing the temporal correlation of the queries's popularity.

Google Trends

Before conducting the temporal comparison, we must select a set of queries that we want to utilize for the comparison. In order to meet the constraints of Google Trends (see Section 2.5), we select queries that are popular. Moreover, we consider a sufficiently long time span for the comparison. Google Trends offers to download the top 25 popular queries from a selected time span. We use this option to create two sets of queries for our comparison. First, we consider the annual top queries of Google and secondly the monthly top queries of Google. For the annual top queries we perform a simple comparison to the AQL's annual top queries. For the monthly top queries, we compute temporal correlations between the AQL and Google Trends.

4.4.1 Annual Top Queries

First, we consider annual top queries of Google and use them for comparisons. We create a list of the annual top 25 from 2004 until 2022. 2004 is the earliest year for which Google Trends provides data and 2022 is the latest year for which the AQL contains queries. Similarly, we create a list of the annual top 25 queries from the AQL for the same time span. For this, we first filter the AQL for queries that stem from Google and afterwards create the annual top 25. As a first comparison, we simply determine the intersection of the two lists with respect to the year. That is, for each year we check if the top 25 queries of Google Trends are also present in the AQL's top 25 queries. We can observe that the intersection is quite small. In fact, only 7 queries out of possible 475 queries are present in both respective top 25 annual queries. The queries are `facebook` (2 times), `google` (2 times), `youtube` (2 times) and `yahoo`. This indicates that the distribution of Google queries in the AQL is not very similar to the real distribution of Google queries.

4.4.2 Temporal Correlation

Secondly, we consider monthly top queries of Google. For this analysis we attempt to assess the similarity of temporal patterns between queries from Google and the AQL. We choose monthly top queries because we expect monthly top queries to be more volatile than annual top queries. Since there are lots of duplicates in the monthly top queries, we need to create a ranking of the monthly top queries. By this, we can compare the most popular queries of the monthly top queries. To do this, we first create a list of the monthly top 25 queries from Google from 2004 until 2022. For each month, we obtain a list of the top 25 queries. We then create a ranking of theses queries by employing reciprocal rank fusion [Cormack et al., 2009]. Reciprocal rank fusion is a method to combine multiple ranked lists into a single ranking. Given a set I of items to be ranked and a set R of rankings, the RRF-Score is computed by

$$RRF(i \in I) = \sum_{r \in R} \frac{1}{k + r(i)} \tag{4.3}$$

where k is a constant (usually set to 60) and $r(i)$ is the rank of item i in ranking r. The method assigns a score to each item in the list based on its positions in the original lists. The score is calculated as the reciprocal of the rank, so that higher-ranked items receive higher scores. The final score for each item is then computed by summing the scores from all lists. We feed all monthly top 25 Google queries from 2004 until 2022 into the RRF-Score function to obtain a final ranking of the top 25 monthly queries. From this, we obtain the following ranking:

1. `google`	8. `news`	15. `translate`	22. `map`
2. `yahoo`	9. `you`	16. `mp3`	
3. `weather`	10. `ebay`	17. `maps`	23. `face`
4. `youtube`	11. `amazon`	18. `msn`	
5. `hotmail`	12. `games`	19. `fb`	24. `video`
6. `facebook`	13. `free`	20. `mail`	
7. `gmail`	14. `twitter`	21. `instagram`	25. `juegos`

Measuring Query Frequencies in the AQL

We then search these queries in the AQL and measure their frequency over time. We achieve this by applying the `map_batches()` API-call to first filter the AQL's queries for the considered top queries and secondly to map

the timestamps of the queries into the format `(YYYY-MM)`. Then, we apply the `groupby()` API-call to group by queries and timestamp. Thereon, we call a `count()` to get the count of each query per month. The result is a data set with the counts of each query per month. In case a query is not present in the AQL during a specific month between 2004 and 2022, we set the count to 0. Eventually, we get a list of queries with their counts per month. In order to obtain the popularity of a query in a specific month, we divide the counts of each query by the sum of all counts of the present Google queries in the AQL in that month. This gives us a relative frequency of each query in that month which represents its popularity. Since the data of Google trends is projected to the scale [0,100], we also scale the relative frequencies of the AQL to the interval [0,100]. After this, we can proceed to compute the temporal correlation of the queries between the AQL and Google Trends.

Computing Temporal Correlations

To assess the similarity of the temporal patterns, we compute the temporal correlation of the queries between the AQL and Google Trends. Given two queries q and p, their respective frequency functions $X_{i,q}$ and $X_{i,p}$ with d time steps, their mean μ_q and μ_p and their standard deviation σ_q and σ_p, then, according to Chien and Immorlica [2005], their temporal correlation is computed by:

$$\rho_{q,p} = \frac{1}{d} \sum_i \left(\frac{X_{i,q} - \mu_q}{\sigma_q} \right) \cdot \left(\frac{X_{i,p} - \mu_p}{\sigma_p} \right) \tag{4.4}$$

The correlation coefficient $\rho_{q,p}$ indicates the strength and direction of the linear relationship between the two queries' frequency functions. The value of $\rho_{q,p}$ ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation and 1 indicates a perfect positive correlation. Since the time spans are aligned, we are looking only for positive correlations as an indicator of similarity. We compute the temporal correlation for each query in the AQL with respect to its counterpart in Google Trends. In Table 4.13 we display the resultig correlation coefficients. The table shows that most queries have a positive correlation with their counterpart in Google Trends, but the correlation coefficients are mostly very low. In addition, there are also some queries with a negative correlation. To get a visual impression of the similarity, we plot the two time series that refelct the highest correlation.

#	Query	$\rho_{q,p}$	#	Query	$\rho_{q,p}$	#	Query	$\rho_{q,p}$
1	google	0.22	2	yahoo	0.16	3	weather	0.24
4	youtube	0.17	5	hotmail	0.01	6	facebook	0.04
7	gmail	-0.02	8	news	0.01	9	you	0.09
10	ebay	0.07	11	amazon	0.20	12	games	-0.04
13	free	-0.11	14	twitter	0.27	15	translate	0.40
16	mp3	-0.01	17	maps	0.06	18	msn	0.11
19	fb	0.17	20	mail	0.19	21	instagram	0.28
22	map	-0.06	23	face	0.25	24	video	0.01
25	juegos	0.09						

Table 4.13: The table displays the resulting correlation coefficients of the temporal popularity. For a given query q in the AQL, the correlation coefficient $\rho_{q,p}$ is computed with respect to its counterpart p in Google Trends. The table shows that most queries have a positive correlation with their counterpart in Google Trends, but the correlation coefficients are mostly very low. In addition, there are also some queries with a negative correlation, indicating that the temporal popularity of queries is substantially different between the AQL and Google Trends.

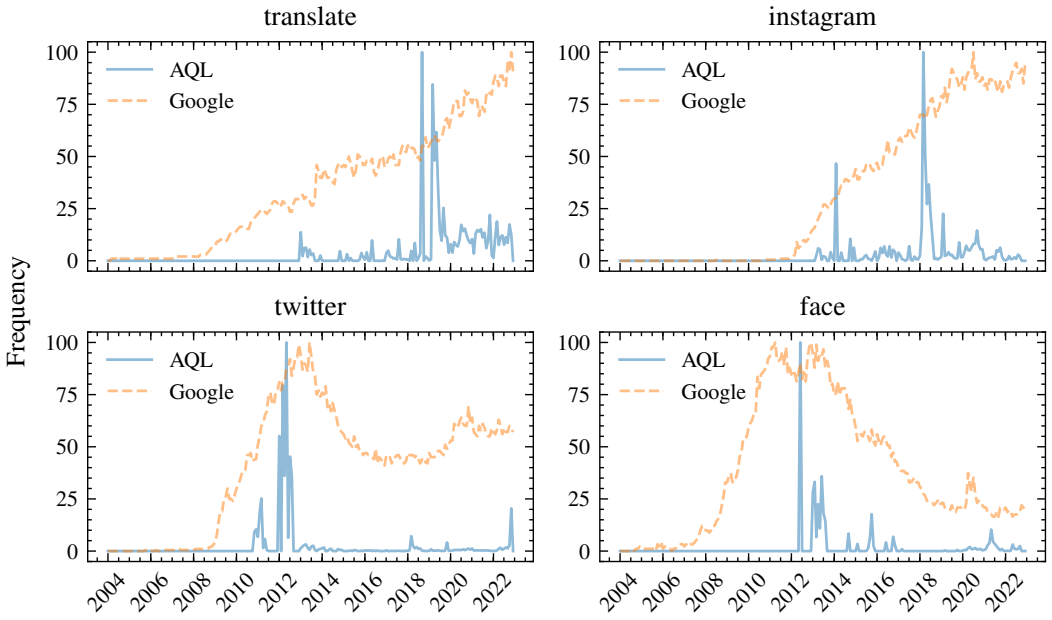

Figure 4.11: The figure shows temporal popularity of the queries "translate", "instagram", "twitter" and "face" in the AQL and Google. We consider these 4 queries because they have the highest correlation coefficients, showcasing that even the highest correlated time series are quite different.

Chapter 5

Discussion

In this chapter we briefly discuss the results of the three parts structure-related, inference-based and temporal-based analysis. We summarize the results and give an outlook on future work.

Structure-related Analysis

The structure-related analysis shows that the AQL exhibits similar distributions especially for the rank-size distribuitions of linguistic elements. In this regard, the AQL fits to the comparison group sinc its mean Wasserstein distance is smaller than the mean Wasserstein distance of the comparison group. Concerning length-related measurements of linguistic elements we can observe rather dissimilar distributions. For all considered length-related measurements the AQL exhibits a higher mean Wasserstein distance than the comparison group. Especially the distribution of characters per query seems anomalous. Moreover, the top queries of the AQL are quite different from the top queries of the AOL, which is our only reference point in this regard. It is striking that the most popular queries in the AQL seem quite random and are even hard to interpretate whereas the top queries of the AOL are easy to understand and make sense. For the top words in turn, the AQL is arguably similar to the comparison group. We can detect an average intersection of 9 words of the top 25 words between AQL and the comparison group. However, the average intersection of top words within the comparison group is even higher with 12 words. As a last aspect, we analyzed the presence of search operators in the query logs. However this comparison lacks meaningfulness as search operators were barely found in the comparison group. The absence of search operatos in the comparison group is due to preceding filtering processes during data generation.

In summary, one could argue that similrities were found in the structure-related analysis, but the AQL is rather not on par with the comparison group.

Inference-based Analysis

The inference-based analysis shows that the AQL exhibits similar distributions especially for its query intent and question distributions. In this case, a lower Wasserstein distance than the comparison group was found. Concerning the distribuition of PII entities, we can observe rather dissimilar distributions. The AQL exhibits a higher mean Wasserstein distance than the comparison group. Summarizing the results, we can argue that the AQL is similar to the comparison group since it matches the distributions of the comparison group in two of the three cases.

Temporal-based Analysis

The temporal-based analysis shows that the AQL exhibits a very different temporal query popularities compared to the comparison group. From the list of top 25 monthly google queries, we could find overall low correlations of temporal patterns between AQL queries and google queries. The highest found correlation is 0.4 which is not significant. the correlation of most query popularities are around 0, indicating dissimilar temporal patterns. Regarding the annual top queries, we could observe only a tiny intersection of AQL queries and google queries.

Future Work

In this work we have characterized queries by structural features, temporal correlations and some prevalent taxonomies. To capture even more information of queries, especially semantic information, one could employ sentence embedding models that are used for semantic search. These models have shown to create meaningful representations of text semantics and could be used to compare embedding distributions of the AQL with the comparison group. This may be done by measuring Wasserstein distances or one could perform clustering algorithms on the embeddings to detect present topics in the query logs.

Bibliography

Maristella Agosti, Franco Crivellari, and Giorgio Maria Di Nunzio. Web Log Analysis: A Review of a Decade of Studies about Information Acquisition, Inspection and Interpretation of User Interaction. *Data Mining and Knowledge Discovery*, 24(3):663–696, 2012. doi: 10.1007/S10618-011-0228-8. URL `https://doi.org/10.1007/s10618-011-0228-8`.

Daria Alexander, Wojciech Kusa, and Arjen P. de Vries. ORCAS-I: Queries Annotated with Intent using Weak Supervision. In Enrique Amigó et al., editor, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3057–3066. ACM, 2022. doi: 10.1145/3477495.3531737. URL `https://doi.org/10.1145/3477495.3531737`.

Philipp Behnen, René Keßler, Felix Kruse, Jorge Marx Gómez, Jan Schoenmakers, and Sergej Zerr. Experimental Evaluation of Scale, and Patterns of Systematic Inconsistencies in Google Trends Data. In Irena Koprinska et al., editor, *ECML PKDD 2020 Workshops - Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases ECML PKDD 2020: SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14-18, 2020, Proceedings*, volume 1323 of *Communications in Computer and Information Science*, pages 374–384. Springer, 2020. doi: 10.1007/978-3-030-65965-3_25. URL `https://doi.org/10.1007/978-3-030-65965-3_25`.

Steven M. Beitzel, Eric C. Jensen, Ophir Frieder, David A. Grossman, David D. Lewis, Abdur Chowdhury, and Aleksander Kolcz. Automatic Web Query Classification using Labeled and Unlabeled Training Data. In Ricardo A. Baeza-Yates et al., editor, *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005*, pages 581–582. ACM, 2005. doi: 10.1145/1076034.1076138. URL `https://doi.org/10.1145/1076034.1076138`.

Steven M. Beitzel, Eric C. Jensen, David D. Lewis, Abdur Chowdhury, and Ophir Frieder. Automatic Classification of Web queries using Very Large Unlabeled Query Logs. *ACM Transactions on Information Systems*, 25(2): 9, 2007. doi: 10.1145/1229179.1229183. URL `https://doi.org/10.1145/1229179.1229183`.

Alexander Bondarenko, Pavel Braslavski, Michael Völske, Rami Aly, Maik Fröbe, Alexander Panchenko, Chris Biemann, Benno Stein, and Matthias Hagen. Comparative Web Search Questions. In James Caverlee et al., editor, *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 52–60. ACM, 2020. doi: 10.1145/3336191.3371848. URL `https://doi.org/10.1145/3336191.3371848`.

Andrei Z. Broder. A Taxonomy of Web Search. *SIGIR Forum*, 36(2):3–10, 2002. doi: 10.1145/792550.792552. URL `https://doi.org/10.1145/792550.792552`.

Qi Chen, Xiubo Geng, Corby Rosset, Carolyn Buractaon, Jingwen Lu, Tao Shen, Kun Zhou, Chenyan Xiong, Yeyun Gong, Paul N. Bennett, Nick Craswell, Xing Xie, Fan Yang, Bryan Tower, Nikhil Rao, Anlei Dong, Wenqi Jiang, Zheng Liu, Mingqin Li, Chuanjie Liu, Zengzhong Li, Rangan Majumder, Jennifer Neville, Andy Oakley, Knut Magne Risvik, Harsha Vardhan Simhadri, Manik Varma, Yujing Wang, Linjun Yang, Mao Yang, and Ce Zhang. MS MARCO Web Search: A Large-scale Information-rich Web Dataset with Millions of Real Click Labels. In Tat-Seng Chua et al., editor, *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, pages 292–301. ACM, 2024. doi: 10.1145/3589335.3648327. URL `https://doi.org/10.1145/3589335.3648327`.

Steve Chien and Nicole Immorlica. Semantic Similarity between Search Engine Queries using Temporal Correlation. In Allan Ellis and Tatsuya Hagino, editors, *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005*, pages 2–11. ACM, 2005. doi: 10.1145/1060745.1060752. URL `https://doi.org/10.1145/1060745.1060752`.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal Rank Fusion outperforms Condorcet and Individual Rank Learning Methods. In James Allan et al., editor, *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23,*

2009, pages 758–759. ACM, 2009. doi: 10.1145/1571941.1572114. URL https://doi.org/10.1145/1571941.1572114.

Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. ORCAS: 18 Million Clicked Query-Document Pairs for Analyzing Search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2983–2989. Association for Computing Machinery, 2020. doi: 10.1145/3340531.3412779. URL https://doi.org/10.1145/3340531.3412779.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. Named Entity Recognition in Query. In James Allan et al., editor, *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 267–274. ACM, 2009. doi: 10.1145/1571941.1571989. URL https://doi.org/10.1145/1571941.1571989.

Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing and Management*, 36(2):207–227, 2000. doi: 10.1016/S0306-4573(99)00056-4. URL https://doi.org/10.1016/S0306-4573(99)00056-4.

In-Ho Kang and Gil-Chang Kim. Query Type Classification for Web Document Retrieval. In Charles L. A. Clarke et al., editor, *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada*, pages 64–71. ACM, 2003. doi: 10.1145/860435.860449. URL https://doi.org/10.1145/860435.860449.

Joseph B. Kruskal. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika*, 29(1):1–27, 1964. doi: 10.1007/BF02289565.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions On Knowledge Data Engineering*, 34(1):50–70, 2022. doi: 10.1109/TKDE.2020.2981314. URL https://doi.org/10.1109/TKDE.2020.2981314.

Thomas Lin, Patrick Pantel, Michael Gamon, Anitha Kannan, and Ariel Fuxman. Active Objects: Actions for Entity-Centric Search. In Alain Mille et al., editor, *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, pages 589–598. ACM, 2012.

doi: 10.1145/2187836.2187916. URL https://doi.org/10.1145/2187836.2187916.

Sean MacAvaney, Craig Macdonald, and Iadh Ounis. Reproducing Personalised Session Search Over the AOL Query Log. In Matthias Hagen et al., editor, *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I*, volume 13185 of *Lecture Notes in Computer Science*, pages 627–640. Springer, 2022. doi: 10.1007/978-3-030-99736-6_42. URL https://doi.org/10.1007/978-3-030-99736-6_42.

Arnold Overwijk, Chenyan Xiong, and Jamie Callan. ClueWeb22: 10 Billion Web Documents with Rich Information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3360–3362. Association for Computing Machinery, 2022. doi: 10.1145/3477495.3536321. URL https://doi.org/10.1145/3477495.3536321.

Victor M. Panaretos and Yoav Zemel. Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application*, 6:405–431, 2019. doi: https://doi.org/10.1146/annurev-statistics-030718-104938. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-030718-104938.

Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A Picture of Search. In Xiaohua Jia, editor, *Proceedings of the 1st International Conference on Scalable Information Systems, Infoscale 2006, Hong Kong, May 30-June 1, 2006*, volume 152 of *ACM International Conference Proceeding Series*, page 1. ACM, 2006. doi: 10.1145/1146847.1146848. URL https://doi.org/10.1145/1146847.1146848.

Steven T. Piantadosi. Zipf's Word Frequency Law in Natural Language: A critical Review and Future Directions. *Psychonomic Bulletin & Review*, 21: 1112–1130, 2014.

Jan Heinrich Reimer. Identifying and Answering Health-Related Questions. Master's thesis, Martin-Luther University Halle-Wittenberg, Institut für Informatik, Halle (Saale), Germany, April 2023. URL https://downloads.webis.de/theses/papers/reimer_2023.pdf.

Jan Heinrich Reimer, Sebastian Schmidt, Maik Fröbe, Lukas Gienapp, Harrisen Scells, Benno Stein, Matthias Hagen, and Martin Potthast. The Archive Query Log: Mining Millions of Search Result Pages of Hundreds

of Search Engines from 25 Years of Web Archives. In Hsin-Hsi Chen et al., editor, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2848–2860. ACM, 2023. doi: 10.1145/3539618.3591890. URL https://doi.org/10.1145/3539618.3591890.

Daniel E. Rose and Danny Levinson. Understanding User Goals in Web Search. In Stuart I. Feldman et al., editor, *Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004*, pages 13–19. ACM, 2004. doi: 10.1145/988672.988675. URL https://doi.org/10.1145/988672.988675.

Milad Shokouhi. Detecting Seasonal Queries by Time-Series Analysis. In Wei-Ying Ma et al., editor, *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 1171–1172. ACM, 2011. doi: 10.1145/2009916.2010104. URL https://doi.org/10.1145/2009916.2010104.

Craig Silverstein, Monika Henzinger, Hannes Marais, and Michael Moricz. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum*, 33 (1):6–12, 1999. doi: 10.1145/331403.331405. URL https://doi.org/10.1145/331403.331405.

Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, and Tefko Saracevic. Searching the Web: The Public and their Queries. *Journal of the American Society for Information Science and Technology*, 52(3): 226–234, 2001. doi: https://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.0.CO;2-R. URL https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/1097-4571%282000%299999%3A9999%3C%3A%3AAID-ASI1591%3E3.0.CO%3B2-R.

Ryen W. White, Matthew Richardson, and Wen-tau Yih. Questions vs. Queries in Informational Search Tasks. In Aldo Gangemi et al., editor, *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 135–136. ACM, 2015. doi: 10.1145/2740908.2742769. URL https://doi.org/10.1145/2740908.2742769.

Dietmar Wolfram, Amanda Spink, Bernard J. Jansen, and Tefko Saracevic. Vox Populi: The Public Searching of the Web. *Journal of the Association for Information Science and Technology*, 52(12):1073–1074, 2001. doi: 10.1002/ASI.1157. URL https://doi.org/10.1002/asi.1157.

Jingyuan Zhang, Luo Jie, Altaf Rahman, Sihong Xie, Yi Chang, and Philip S. Yu. Learning Entity Types from Query Logs via Graph-Based Modeling. In James Bailey et al., editor, *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 603–612. ACM, 2015. doi: 10.1145/2806416.2806498. URL `https://doi.org/10.1145/2806416.2806498`.