

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Mutability of Web Search Engines Results — Data Collection and Brief Analysis

Bruno Miguel Faustino Moreno



Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Carla Teixeira Lopes

September 28, 2021

Mutability of Web Search Engines Results — Data Collection and Brief Analysis

Bruno Miguel Faustino Moreno

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Prof. João Correia Lopes

External Examiner: Prof. Ricardo Campos

Supervisor: Prof. Carla Teixeira Lopes

September 28, 2021

Abstract

Web search is an essential tool that has changed the way we access information. The dynamic characteristics of the web and the evolution of search engine algorithms may cause search volatility, meaning that the results retrieved for a query might change based on several factors, like time or location. Additionally, web search personalization has been evolving, as now different results for the same query are shown to different users based on their individual characteristics.

It is our objective to gather web search results data for the variance of several search factors, such as time, location, safe search, privacy mode, user agent, presence of cookies, and authentication status, for each search engine examined, and describe the mutability in the search results caused by each search factor. We also compare how much the results vary between the search engines and explore the influence of query topics. The final contribution of this study consists in making the obtained data available publicly so it can be used for further researches. Another goal is to perform a superficial analysis of the results obtained. There isn't enough information about results variability outside of the business scope of each search engine's company. The relevance of this study comes from the necessity to obtain more knowledge about each factor's influence on the results retrieved from web searches, and how much the results change according to user personalization. It is necessary to better understand this volatility, as it can be useful for research projects that use search engine results in their studies. Past works have studied the volatility of web search results but have focused mostly on changes caused by time or different search engines.

To perform this study, we designed and conducted seven data collection experiments for the variance of the following search factors — time, location, safe search, privacy mode, user agent, cookies, and authentication status — using the Google and Bing search engines, in order to evaluate the impact of each factor on the volatility of search results for each search engine. A query set of 298 queries was used for the experiments, and the top 10 results were retrieved for each query. The experiments were divided into three groups with distinct methods of data collection: retrievals via APIs, manual retrievals followed by web scraping, and retrievals by real users followed by web scraping. Then, the results were analyzed and described. The datasets generated are available to the community. Finally, a superficial analysis of the results obtained was conducted, demonstrating that the datasets generated can be useful for future deeper research. Our brief results analysis shows that the location and time factors have the largest impact on the volatility of search results. The use of different search engines and query topics also impacts the results' volatility. Additionally, higher ranked results often maintain their ranking positions.

This research is innovative as it analyses the changes in search results caused by the many search factors mentioned, instead of only exploring the changes caused by time or different search engines. The datasets produced in this study may aid future information retrieval investigations in obtaining further insights about web search results volatility.

Keywords: information retrieval, web, search results, ranking volatility, search engines

Resumo

A pesquisa na Web é uma ferramenta essencial que mudou a maneira como acedemos a informação. As características dinâmicas da web e a evolução dos algoritmos dos motores de busca podem causar volatilidade na pesquisa, o que significa que os resultados retornados para uma pesquisa podem mudar com base em vários fatores, como o tempo ou a localização. Para além disso, a personalização da pesquisa na web tem vindo a evoluir, uma vez que agora resultados diferentes para a mesma pesquisa são mostrados a utilizadores diferentes, com base nas suas características individuais.

O nosso objectivo é recolher dados de resultados de pesquisa na web para a variação de vários fatores de pesquisa, tais como tempo, localização geográfica, pesquisa segura, modo de privacidade, *user agent*, presença de cookies, e estado de autenticação, para cada motor de busca examinado, e descrever a mutabilidade nos resultados da pesquisa causada por cada fator de pesquisa. Também comparamos o quanto os resultados variam entre os motores de busca e exploramos a influência dos tópicos de consulta. A contribuição final deste estudo consiste na disponibilização dos dados obtidos publicamente de modo a que estes possam ser usados em futuras investigações. Outro objetivo é realizar uma análise superficial dos resultados obtidos. Não há informação suficiente sobre a variabilidade dos resultados fora do contexto empresarial das empresas de cada motor de busca. A relevância deste estudo reside no facto de ser necessário obter mais conhecimento sobre o quanto cada fator influencia os resultados obtidos a partir das pesquisas na web, e o quanto os resultados mudam de acordo com a personalização do utilizador. É necessário perceber melhor esta volatilidade, pois pode ser útil para projetos de investigação que utilizem resultados de motores de busca nos seus estudos. Trabalhos passados estudaram a volatilidade dos resultados das pesquisas na web, mas concentraram-se principalmente nas mudanças causadas pelo tempo ou por diferentes motores de busca.

Para realizar este estudo, foram planeadas e conduzidas sete experiências de recolha de dados para a variância dos seguintes fatores de pesquisa — tempo, localização geográfica, pesquisa segura, modo de privacidade, *user agent*, cookies, e estado de autenticação — utilizando os motores de busca Google e Bing, de modo a avaliar o impacto de cada fator na volatilidade dos resultados de pesquisa para cada motor de busca. Um conjunto de 298 expressões de pesquisa foi usado nas experiências, e os top 10 resultados foram recolhidos para cada expressão de pesquisa. As experiências foram divididas em três grupos com métodos distintos de recolha de dados: recolhas via *APIs*, recolhas manuais seguidas de *web scraping*, e recolhas por utilizadores reais seguidas de *web scraping*. Em seguida, os resultados foram analisados e descritos. Os ficheiros de dados gerados estão disponíveis para a comunidade. Por fim, foi conduzida uma análise superficial dos resultados obtidos, demonstrando que os ficheiros de dados gerados podem ser úteis para futuras investigações mais profundas. A nossa breve análise de resultados mostra que a localização geográfica e o tempo são os fatores que têm o maior impacto na volatilidade dos resultados de pesquisa. A utilização de diferentes motores de busca e tópicos de consulta também influencia a volatilidade dos resultados. Além disso, os resultados com posições de ranking mais altas costumam

nam manter as suas posições.

Esta investigação é inovadora, pois analisa as alterações nos resultados de pesquisa causadas pelos diversos fatores de pesquisa mencionados, em vez de apenas explorar as alterações causadas pelo tempo ou por diferentes motores de busca. Os ficheiros de dados produzidos neste estudo podem auxiliar futuras investigações sobre recuperação de informação na obtenção de mais conhecimento acerca da volatilidade dos resultados de pesquisa na web.

Keywords: recuperação de informação, web, resultados de pesquisa, volatilidade de classificação, motores de busca

Acknowledgements

I would like to thank my supervisor, professor Carla Teixeira Lopes, for the guidance provided during the development of this project. I would also like to thank professors Gonçalo Gonçalves and José Ornelas for the assistance provided during the execution of the study's experiments, as well as the students of the 2nd year of the MIEIC course of 2020/2021 that participated in the experiments.

Thank you to my friends and family as well.

Bruno Miguel

*“The alchemists in their search for gold
discovered many other things of greater value.”*

Arthur Schopenhauer

Contents

1	Introduction	1
1.1	Context	1
1.2	Objectives	2
1.3	Motivation	2
1.4	Document Structure	3
2	State of the Art	5
2.1	Background	5
2.1.1	Web Information Dynamics	5
2.1.2	Web Information Retrieval	7
2.1.3	Search Engines and their Evolution	9
2.1.4	Search Results Evolution Over Time	11
2.2	Related Work	12
2.2.1	Mutability of Search Results Over Time	12
2.2.2	The Impact of Search Engines on the Volatility of Search Results	14
2.2.3	Mutability of Search Results by Context Features	15
3	Objectives and Methodology	19
3.1	Objectives	19
3.2	Methodology	20
3.2.1	Search Engine Selection	20
3.2.2	Search Factors Examined	22
3.2.3	Search Queries Selection	23
3.2.4	Collecting Data for the Analysis of the <i>time, location, and safe search</i> Factors	24
3.2.5	Collecting Data for the Analysis of the <i>privacy and user agent</i> Factors	30
3.2.6	Collecting Data for the Analysis of the <i>cookies and authentication status</i> Factors	35
3.2.7	Metrics Used in Results Analysis	37
4	Data Extraction	41
4.1	Implementation Details	41
4.1.1	Execution of the <i>time, location, and safe search</i> Experiments	41
4.1.2	Execution of the <i>privacy and user agent</i> Experiments	42
4.1.3	Execution of the <i>cookies and authentication status</i> Experiments	42
4.2	Datasets Obtained	43

5	Brief Analysis	45
5.1	Analysis Strategy	45
5.2	Impact of Time	46
5.3	Impact of Location	51
5.4	Impact of Safe Search	56
5.5	Impact of Privacy Mode	61
5.6	Impact of User Agent	67
5.7	Impact of Cookies	72
5.8	Impact of Authentication Status	74
5.9	Results Discussion	76
6	Conclusions and Future Work	79
6.1	Conclusions	79
6.2	Future Work	80
A	User Survey	81
B	Real User Experiments Scripts	83
B.1	<i>Cookies</i> Experiment Script:	84
B.2	<i>Authentication status</i> Experiment Script:	100
	References	121

List of Figures

2.1	WebIR System Components	8
3.1	General Methodology Diagram	21
3.2	Timeline Chart Showing All the Experiments Performed	21
3.3	Methodology Diagram for the <i>time</i> Experiment	25
3.4	Methodology Diagram for the <i>location</i> and <i>safe search</i> Experiments	26
3.5	Example of a Web Search Result Including One Main URL and Six Secondary URLs	32
3.6	Methodology Diagram for Experiment Group <i>Exp-Manual</i>	32
3.7	Methodology Diagram for Experiment Group <i>Exp-RealUsers</i>	36
5.1	Graph Showing the % <i>New URL</i> Metric Results Between the First Retrieval and Each Subsequent Retrieval, for the Top 10, 5, 3 and 1, for Google (<i>time</i> Exp.)	47
5.2	Graph Showing the % <i>New URL</i> Metric Results for Each Pair of Consecutive Retrievals, for the Top 10, 5, 3 and 1, for Google (<i>time</i> Exp.)	47
5.3	Graph Showing the % <i>New URL</i> Metric Results Between the First Retrieval and Each Subsequent Retrieval, for the Top 10, 5, 3 and 1, for Bing (<i>time</i> Exp.)	48
5.4	Graph Showing the % <i>New URL</i> Metric Results for Each Pair of Consecutive Retrievals, for the Top 10, 5, 3 and 1, for Bing (<i>time</i> Exp.)	48
5.5	Graph Showing the % <i>New URL</i> Metric Results Between the First Retrieval and Each Subsequent Retrieval, for the Top 10, for Both Google and Bing (<i>time</i> Exp.)	49
5.6	Graph Showing the % <i>New URL</i> Metric Results for Each Pair of Consecutive Retrievals, for the Top 10, for Both Google and Bing (<i>time</i> Exp.)	49
5.7	Graph Showing the <i>Rank Movements Per Rank Position</i> Metric Results Between the First Retrieval and Each Subsequent Retrieval, for the Top 10, 5, and 3, for Google (<i>time</i> Exp.)	50
5.8	Graph Showing the <i>Rank Movements Per Rank Position</i> Metric Results for Each Pair of Consecutive Retrievals, for the Top 10, 5, and 3, for Google (<i>time</i> Exp.)	50
5.9	Graph Showing the <i>Rank Movements Per Rank Position</i> Metric Results Between the First Retrieval and Each Subsequent Retrieval, for the Top 10, 5 and 3, for Bing (<i>time</i> Exp.)	50
5.10	Graph Showing the <i>Rank Movements Per Rank Position</i> Metric Results for Each Pair of Consecutive Retrievals, for the Top 10, 5, and 3, for Bing (<i>time</i> Exp.)	50
5.11	Graph Showing the <i>Rank Movements Per Rank Position</i> Metric Results Between the First Retrieval and Each Subsequent Retrieval, for the Top 10, for Both Google and Bing (<i>time</i> Exp.)	51
5.12	Graph Showing the <i>Rank Movements Per Rank Position</i> Metric Results for Each Pair of Consecutive Retrievals, for the Top 10, for Both Google and Bing (<i>time</i> Exp.)	51

5.13	Graph Showing the % New URL Metric Results for Google, for the Top 10, 5, 3 and 1 (<i>location</i> Exp.)	52
5.14	Graph Showing the % New URL Metric Results for Bing, for the Top 10, 5, 3 and 1 (<i>location</i> Exp.)	53
5.15	Graph Showing the % New URL Metric Results, for the Top 10, for Google and Bing (<i>location</i> Exp.)	53
5.16	Graph Showing the % New URL Metric Results for Each Query Topic, for the Top 10, for Google and Bing (<i>location</i> Exp.)	54
5.17	Graph Showing the Rank Movements Per Rank Position Metric Results for Google, for the Top 10, 5, and 3 (<i>location</i> Exp.)	55
5.18	Graph Showing the Rank Movements Per Rank Position Metric Results for Bing, for the Top 10, 5, and 3 (<i>location</i> Exp.)	55
5.19	Graph Showing the Rank Movements Per Rank Position Metric Results, for the Top 10, for Google and Bing (<i>location</i> Exp.)	56
5.20	Graph Showing the Rank Movements Per Rank Position Metric Results for Each Query Topic, for the Top 10, for Google and Bing (<i>location</i> Exp.)	57
5.21	Graph Showing the % New URL Metric Results for Google, for the Top 10, 5, 3 and 1 (<i>safe search</i> Exp.)	57
5.22	Graph Showing the % New URL Metric Results for Bing, for the Top 10, 5, 3 and 1 (<i>safe search</i> Exp.)	58
5.23	Graph Showing the % New URL Metric Results, for the Top 10, for Google and Bing (<i>safe search</i> Exp.)	58
5.24	Graph Showing the % New URL Metric Results for Each Query Topic, for the Top 10, for Google and Bing (<i>safe search</i> Exp.)	59
5.25	Graph Showing the Rank Movements Per Rank Position Metric Results for Google, for the Top 10, 5, and 3 (<i>safe search</i> Exp.)	60
5.26	Graph Showing the Rank Movements Per Rank Position Metric Results for Bing, for the Top 10, 5, and 3 (<i>safe search</i> Exp.)	60
5.27	Graph Showing the Rank Movements Per Rank Position Metric Results, for the Top 10, for Google and Bing (<i>safe search</i> Exp.)	61
5.28	Graph Showing the Rank Movements Per Rank Position Metric Results for Each Query Topic, for the Top 10, for Google and Bing (<i>safe search</i> Exp.)	62
5.29	Graph Showing the % New URL Metric Results for Google, for the Top 10, 5, 3 and 1 (<i>privacy</i> Exp.)	62
5.30	Graph Showing the % New URL Metric Results for Bing, for the Top 10, 5, 3 and 1 (<i>privacy</i> Exp.)	63
5.31	Graph Showing the % New URL Metric Results, for the Top 10, for Google and Bing (<i>privacy</i> Exp.)	63
5.32	Graph Showing the % New URL Metric Results for Each Query Topic, for the Top 10, for Google and Bing (<i>privacy</i> Exp.)	64
5.33	Graph Showing the Rank Movements Per Rank Position Metric Results for Google, for the Top 10, 5, and 3 (<i>privacy</i> Exp.)	65
5.34	Graph Showing the Rank Movements Per Rank Position Metric Results for Bing, for the Top 10, 5, and 3 (<i>privacy</i> Exp.)	65
5.35	Graph Showing the Rank Movements Per Rank Position Metric Results, for the Top 10, for Google and Bing (<i>privacy</i> Exp.)	66
5.36	Graph Showing the Rank Movements Per Rank Position Metric Results for Each Query Topic, for the Top 10, for Google and Bing (<i>privacy</i> Exp.)	66

5.37	Graph Showing the % New URL Metric Results for Google, for the Top 10, 5, 3 and 1 (<i>user agent</i> Exp.)	67
5.38	Graph Showing the % New URL Metric Results for Bing, for the Top 10, 5, 3 and 1 (<i>user agent</i> Exp.)	68
5.39	Graph Showing the % New URL Metric Results, for the Top 10, for Google and Bing (<i>user agent</i> Exp.)	68
5.40	Graph Showing the % New URL Metric Results for Each Query Topic, for the Top 10, for Google and Bing (<i>user agent</i> Exp.)	69
5.41	Graph Showing the Rank Movements Per Rank Position Metric Results for Google, for the Top 10, 5, and 3 (<i>user agent</i> Exp.)	70
5.42	Graph Showing the Rank Movements Per Rank Position Metric Results for Bing, for the Top 10, 5, and 3 (<i>user agent</i> Exp.)	70
5.43	Graph Showing the Rank Movements Per Rank Position Metric Results, for the Top 10, for Google and Bing (<i>user agent</i> Exp.)	71
5.44	Graph Showing the Rank Movements Per Rank Position Metric Results for Each Query Topic, for the Top 10, for Google and Bing (<i>user agent</i> Exp.)	71
5.45	Graph Showing the % New URL Metric Results for Google, for the Top 10, 5, 3 and 1 (<i>cookies</i> Exp.)	72
5.46	Graph Showing the % New URL Metric Results for Each Query Topic, for the Top 10, for Google (<i>cookies</i> Exp.)	73
5.47	Graph Showing the Rank Movements Per Rank Position Metric Results for Google, for the Top 10, 5, and 3 (<i>cookies</i> Exp.)	73
5.48	Graph Showing the Rank Movements Per Rank Position Metric Results for Each Query Topic, for the Top 10, for Google (<i>cookies</i> Exp.)	74
5.49	Graph Showing the % New URL Metric Results for Google, for the Top 10, 5, 3 and 1 (<i>authentication status</i> Exp.)	75
5.50	Graph Showing the % New URL Metric Results for Each Query Topic, for the Top 10, for Google (<i>authentication status</i> Exp.)	75
5.51	Graph Showing the Rank Movements Per Rank Position Metric Results for Google, for the Top 10, 5, and 3 (<i>authentication status</i> Exp.)	76
5.52	Graph Showing the Rank Movements Per Rank Position Metric Results for Each Query Topic, for the Top 10, for Google (<i>authentication status</i> Exp.)	77
A.1	User Survey Conducted During the Real User Experiments	81

List of Tables

2.1	Summary of Related Work on the Mutability of Search Results	16
3.1	User Agents Used	23
3.2	Summary of Query Topics Used	24
3.3	Queries Used in Each Retrieval, When Varying the <i>time</i> Factor, for Google	27
3.4	Queries Used in Each Retrieval, When Varying the <i>time</i> Factor, for Bing	28
3.5	Queries Used in Each Retrieval, When Varying the <i>location</i> Factor, for Google . .	29
3.6	Queries Used in Each Retrieval, When Varying the <i>location</i> Factor, for Bing . . .	29
3.7	Queries Used in Each Retrieval, When Varying the <i>safe search</i> Factor, for Google	30
3.8	Queries Used in Each Retrieval, When Varying the <i>safe search</i> Factor, for Bing .	31
3.9	Queries Used in Each Retrieval, When Varying the <i>privacy</i> Factor, for Both Google and Bing	33
3.10	Queries Used in Each Retrieval, When Varying the <i>user agent</i> Factor, for Google	34
3.11	Queries Used in Each Retrieval, When Varying the <i>user agent</i> Factor, for Bing . .	34
3.12	Queries Used in Each Session, When Varying the <i>cookies</i> Factor, for Google . . .	37
3.13	Queries Used in Each Session, When Varying the <i>authentication status</i> Factor, for Google	38
3.14	Simple Example of New URLs Introduced from One Retrieval to Another	38
4.1	API Call Parameters Used in Group <i>Exp-APIs</i>	42
5.1	Results of Both Metrics for the Experiments Performed, for the Top 10 Results, for Both Search Engines	77

Abbreviations

IR	Information Retrieval
WebIR	Web Information Retrieval
WWW	<i>World Wide Web</i>
API	Application Programming Interface
HTTP	Hypertext Transfer Protocol
URL	Uniform Resource Locator
IP	Internet Protocol
HTML	Hypertext Markup Language
AI	Artificial Intelligence
CSV	Comma-Separated Values

Chapter 1

Introduction

This chapter presents the initial setting of this project and summarizes the contents and purpose of this research. Section 1.1 describes the surrounding environment where the work is inserted, Section 1.2 explains what this project sets out to achieve and the course of action to do it, Section 1.3 outlines the importance of this study in the scope of information retrieval, and Section 1.4 summarizes the structure of this document and presents an overview of each chapter's contents.

1.1 Context

Web searching has become an essential tool for finding information. It is used by the majority of people for its ability to disseminate knowledge and share ideas. Search engines are the main tool used to search the web, allowing the user to find content and other websites.

The web's size has grown immensely over the years [10, 38, 27]. Data is constantly being added, changed, or deleted at an extremely fast rate, and it has only gotten faster as the years go by [43, 35]. The dynamic aspect of the web can lead to changes in the information that can be retrieved by search engines [44]. Furthermore, search engine technology and algorithms keep evolving with time, and personalization has been one of the main focus in recent years. This means that the search results retrieved by web search systems for a particular query can vary according to several factors. While the most apparent cause of search results volatility is time, many other search factors, such as location or information related to the user behind the search, can impact the set of retrieved results.

Search results mutability means that web pages in the results list fluctuate over time or disappear from it altogether. This means that previously discovered documents may become harder or impossible to find, which can lead to an inconsistent user experience. Additionally, information about the search and ranking algorithms in search engines is not always disclosed outside companies associated with commercial search engines, which leads to a lack of data about this subject.

The volatility of search results is an area of interest in information retrieval research, as well as in the area of search engine optimization for companies that want to control where their company's results appear in the ranking.

1.2 Objectives

The aim of this project is to collect web search results data while varying several search conditions of the query submission, and to describe the volatility of the search results according to the different search conditions. While time is the most apparent search factor worth analyzing, there are many other relevant search factors that may impact the variance of search results, such as location, authentication status, presence of cookies, safe search, privacy mode, and user agent. It is also our objective to explore the influence of each search factor mentioned above in different search engines and be able to compare the differences between them. The same analysis can be conducted regarding different query topics. In essence, the focal goal of this study is to collect and analyze search results data for the variance of each search factor, and to store the obtained search results data and its analysis in organized datasets, which are made available publicly in order to aid future researches in this topic. Additionally, another goal is to provide a superficial analysis of the results obtained to raise some hypotheses about the impact of the studied search factors, search engines, and query topics on the mutability of web search results, showing that the data gathered can be useful for further research.

This project should allow us to attain more knowledge about how search engines' results change through the manipulation of different factors, as well as raising hypotheses about how this influence varies based on the search engine or search topic used.

1.3 Motivation

Web search is a tool used by everyone, both the general public and researchers. Even though many investigations have been carried out in the area of information retrieval, the evolution of the web and web technologies continually instigates the need for further developments.

Research in the area of information retrieval has grown a lot with the expansion of the web and web search. Some of the research is carried out by companies linked to commercial search engines, but most of it is done by academic researchers aiming to find ways to improve and evaluate how commercial search engines work. Furthermore, it is important to note that research works on information retrieval are done using results returned by commercial search engines [51, 55, 46, 39, 2]; therefore, a better understanding of how web search results are retrieved can influence decisions when it comes to researches' methodological planning.

The importance of this investigation lies in the fact that it is necessary to obtain more knowledge about how much each search factor impacts the results returned by web search systems, as well as how much user personalization affects the results delivered to the users. The insights and data obtained from this study about web search results are important, both for the general public

and for advancements in the field of information retrieval research, as they can provide more tools and knowledge regarding how search engines customize search results based on different users and different search factors.

1.4 Document Structure

The rest of this document is organized as follows: Chapter 2 presents the relevant background information and past work related to this project, Chapter 3 describes the goals of this work and the proposed method to achieve them, Chapter 4 describes how the data retrieval was executed and how the results of this study were organized and made available, Chapter 5 presents a brief analysis of the collected search results data, and Chapter 6 presents the conclusions for this project.

Chapter 2

State of the Art

This chapter will present background information and past work relevant to the current project. Section 2.1 will present background information on the matter of web information evolution, web information retrieval, search engines, and search results evolution over time. Section 2.2 will examine past projects developed on the subjects of search results mutability and user personalization when retrieving search results.

2.1 Background

In this section we will present information about the context of this project. Section 2.1.1 presents details about the web and its evolution, Section 2.1.2 describes web information retrieval systems, their components and characteristics, Section 2.1.3 presents information about search engines, their components and algorithms, and Section 2.1.4 describes the instability of search results over time.

2.1.1 Web Information Dynamics

‘The web is becoming a universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before’ [5, p. 2].

The World Wide Web was first conceptualized by Tim Berners-Lee et al. [9], in a project that presented hypertext as a way to link and access information. The web has now evolved to be the largest service in the Internet, and has become an essential tool for millions of people all over the world.

In 1999, Lawrence et al. [38] analyzed the disposition of data in the web and concluded that the static web contained approximately 800 million pages with 6 terabytes of text data. It also stood out that 83% of pages were used for commercial purposes. The replication of content on the

web has also been studied, and Bharat et al. [11] pointed out that 10% of the web's information was duplicated. Other studies have found the page duplication rate to be close to 30% [53].

Furthermore, two studies, one performed in 1998 [10] and the other in 2005 [27], followed similar methodologies to estimate the size of the web. The first submitted queries in five different search engines, and then used statistical analysis to conclude that the static public web contained, at least, 200 million documents. Some years later and resorting to a similar methodology, the second study estimated the indexable web to contain 11.5 billion pages, at least. These studies, along with the one from Lawrence et al. [38], can give us some understanding about the growth of the web's content over time. It's relevant to realize that these conclusions about the web's size do not account for the size of the Deep Web, which is estimated to be 100 times larger than the static web [28]. The Deep Web is a concept that refers to contents that are "hidden" in the web, as they can't be accessed by search engines.

It is also important to understand that information available on the web is highly mutable. New information is constantly being created in the world and added to the web. In 2004, it was estimated that 320 million web pages were created each week [43]. In 2005, through the monitoring of 34,000 web sites for 100 days, it was determined that, on a weekly basis, about 5% of URLs detected were new [35]. It was also determined that 62% of the content of newly created pages is actually new. This data, along with the increase in overall page sizes, allows us to get a quick look at how much new information is made available in the web as time progresses.

At the same time, data in the web can also be changed, and the databases or web sites hosting it can become inactive or be taken down. To put it into perspective, in the span of 8 months, between August 2002 and March 2003, nearly 20% of addresses in scholarly articles became inactive [20]. In 2004, it was also estimated that 50% of all web sites go offline every year, just to be replaced by new ones, which can "borrow" most of the information of the old ones [43]. Web site mutability is also common, as pages can undergo changes very often; approximately 15% of pages go through changes at least once a week. The evolution of the content and link structure of web pages was analyzed by Ntoulas et al. [43] and they found that 80% of existing pages were estimated to not be accessible in their original form after just one year. This is enough to demonstrate just how much the content of the web can be altered or lost in such a small amount of time. It was also found that link structure is more dynamic than page content and that the creation of new pages is much more frequent than updating existing ones.

Although there is a large amount of pages that change over time, in May 2003 [23], the mutability of web pages was analyzed and some relevant characteristics were found: most pages that undergo changes, usually only change in trivial ways; there is a relationship between the top-level domain of a web site and its content's frequency of change (for example, in the *.com* domain, pages change more frequently than in the *.gov* and *.edu* domains); the size of the pages is also an important variable to determine how much they can change, as larger documents change more often and go through larger changes than smaller ones.

Web information has been evolving towards a state where relevant and extremely recent data is often available to the users, but also where a lot of data is lost over time as well, either due to

web pages being altered or becoming inaccessible, which may create some knowledge gaps about the present time for future generations [17].

2.1.2 Web Information Retrieval

The searchable web and the search engines that analyse its contents have become essential tools for obtaining information.

Web information retrieval deals with the recovery, representation, storage and organization of data on the web [5]. It is important to distinguish between data retrieval and information retrieval. Data retrieval aims at searching a collection of documents and finding the ones that contain the keywords the user searched for, and it usually deals with data that is semantically sound and well structured. On the other hand, the focus of Information Retrieval (IR) is on content that is unstructured, usually expressed in natural language, and that can, at times, be semantically ambiguous. Therefore, an IR system must be able to analyse the contents of a document — its syntactic and semantic information — and decide if it suits the user's needs. A good IR system should be very effective at deciding whether a document's contents are relevant to the user or not [5].

In order to achieve its goal, there are several methodologies that can be followed when performing IR, which can be mathematically classified into three distinct model categories [5]:

1. Algebraic Models, where documents are represented as vectors or matrices. In order to calculate similarity, documents are transformed into one-dimensional entities using algebraic operations. The Vector Model [49] is a known example of this model category, and the most used one, because of its simplicity and efficiency.
2. Set-theoretic Models, where a document's content is characterized as a set that contains terms, and set-theoretic operations are performed to obtain a similarity value. The Standard Boolean Model and the Fuzzy Model are some examples of this model category.
3. Probabilistic Models, where probabilities are used to represent document's relevance. The similarities between documents and a query are viewed as probabilities. An example of a probabilistic theorem used in these models is the Bayes' theorem [37].

Web Information Retrieval (WebIR) consists in applying the methodologies of information retrieval to the World Wide Web, and its goal is to find relevant information on the web as a response to a user's query, as well as sorting it by order of predicted relevance. The first web search systems resorted to content analysis to reach their goals, using a simple collection of words to rank documents [26]. In order to accomplish this, term weighting methods, such as *TF.IDF* measures, were deployed [50, 48]. Over time, systems evolved towards the analysis of pages' HTML structure in order to rank them [19]; this way, a term in a heading or title HTML tag was given more importance than the same term in a paragraph, for example.

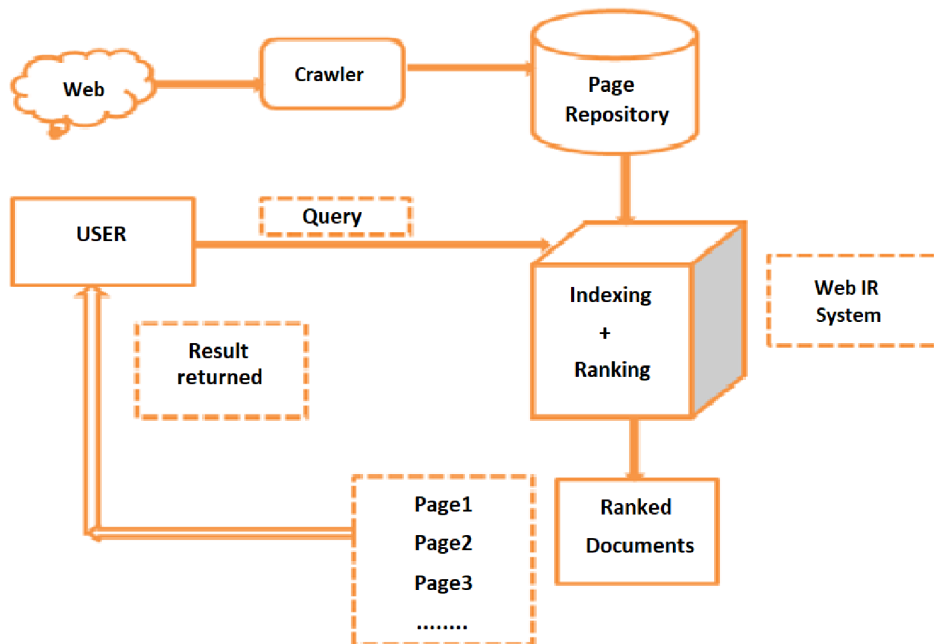


Figure 2.1: WebIR System Components [57]

However, the task of retrieving information in the web has always been challenging, because of its unique characteristics, such as: the large size of its contents, the unique link structure between documents, the volatility of its contents, its heterogeneous properties (such as different formats and languages), and the existence of a large amount of duplicated pages [44]. Also, a study by Craig et al. [18] estimated that over 85% of queries written by users contain less than three terms. All these factors, along with the usually poorly written user queries, can cause search results to be not relevant or redundant [58].

Web search systems need to develop strategies that allow them to rank their search results. These systems have several components: **the crawler**, which constantly scans and harvests web pages, and stores them in a central repository; **the page repository** which is a specialized database that holds the web documents fetched by the crawler, as full HTML text, and allows fast reading operations and high concurrent access; **the indexing module**, which takes each new complete page and extracts only the essential descriptors, generating a compressed description of the page that is stored in several indexes optimized for fast reads; **the query module**, which converts the user's search input in natural language to a format that can be interpreted by the system, and consults the existing indexes to answer the query, which results in a set of relevant pages that are forwarded to the ranking module; **the ranking module**, which ranks each page of the set of relevant pages received from the query module according to specific criteria, **the presentation**, which sorts and presents the ranked documents to the user, usually including the title of the document, its URL, and a short snippet of its contents [44, 57]. Figure 2.1 presents a simple diagram of the components of WebIR systems, and how they interact with each other.

It is worth mentioning that web crawlers scan and store pages in a methodical and automated

way [37]. A web crawling process usually begins with a specific collection of URLs and harvests new pages iteratively, while storing new URLs for later. The behaviour of a crawler is defined by several processes: a selection algorithm, that specifies which pages to harvest; a politeness policy, which helps not to overload web sites; and a parallelization approach, as multiple web crawlers can be coordinated to work in parallel [13]. Web crawlers must also deal with other issues, such as the deep web's content and URL identification.

Even so, web crawlers do not get a proper snapshot of the web, because the information they gather doesn't represent the web at any given time [5]. As crawlers take a long time to harvest the web pages, and because of the dynamic nature of the web, by the time a crawler has finished its crawl, many changes could have happened to the pages. A crawler must strive to minimize the fraction of time pages remain outdated by deploying a re-visit policy, which states when to re-scan pages for changes [15].

The hyperlink structure of the web is also an important component that has been analyzed and has enabled significant improvements in web information retrieval. Hyperlink analysis can be used for deciding which web pages to crawl and how to rank search results, as well as finding pages that are related to each other or duplicated [30, 31]. Approaches based on link structure analysis view the web as a directed graph, containing nodes and edges. Each node represents a web document and each edge represents a link between two documents [44]. Many algorithms used for ranking search results examine and utilize the linkage between web pages in order to obtain better results. One of the most relevant algorithms that use this concept is the PageRank algorithm [12], used by the Google¹ search engine. More details about this algorithm are discussed in Section 2.1.3.

Furthermore, instead of analyzing the link structure of a static snapshot of the web, it's also relevant to use multiple snapshots of the web in order to try to keep pace with its dynamic characteristics. A study conducted in 2004 [4] used the HTTP header field "last modified" to estimate the age of pages' contents in order to establish a timestamp for web documents. The notion is demonstrated by adapting link analysis ranking algorithms to adjust nodes' weights based on a page's timestamp. This subject has been addressed as temporal link analysis.

2.1.3 Search Engines and their Evolution

Search engines are an essential part of the web and have grown in importance over time.

Search engine websites are some of the most accessed pages, as they are the gateway for people to find other websites. Data suggests that, in 2020, Google is the most visited website on the planet, with over one billion active monthly users and an astonishing 6.9 billion searches performed every day, on average [47]. Following Google, some of the most relevant search engines at the time of this work include Bing² and Yahoo!³.

Search engines crawl web documents in advance to build a local index of the pages, which is then used to identify relevant pages and answer user's queries quickly. Since web sites change

¹<https://www.google.com>

²<https://www.bing.com>

³<https://www.yahoo.com>

very often, search engines must update their index regularly in order to always provide the most accurate results possible [43].

Search engine technology has evolved drastically to keep up with the immense expansion of the web (see Section 2.1.1), and that becomes even clearer when we consider the size of search engines' indexes through the years. In 1994, one of the first web search engines, the World Wide Web Worm, had 110,000 pages in its index. In 1997, the top search engines claimed to have an index of 2 million to 100 million documents [12]. As of today, Google claims to have hundreds of billions of web pages in its index [25].

Search engines have also grown when it comes to efficiency. With the purpose of providing information to the user faster, they have lately evolved into producing summaries, instant answers, and engaging navigational aids [21].

Search systems have also developed features that include semantics analysis. Semantics is the study of the logic and meaning of words. When applied to search systems, the analysis of a query's contextual meaning can help the system predict the user's intent and increase search accuracy. To enhance the semantic value of contents and improve search results, Google developed the Knowledge Graph, which consists of an extensive graph containing entities and the relationships between them, encapsulating over 70 billion facts [41]. Google uses this graph to identify terms and associated content based on the user's query, in order to quickly present to the user additional relevant information related to the query, and display it in knowledge panels on the right of the page.

To tackle the growing challenges of WebIR, search engines' architectures have become very specialized over the years, as well as the algorithms used to compute search results. One example is the PageRank algorithm, which was developed in 1998 by Brin et al. [12] and is one of the most relevant algorithms based on link analysis. It is used by Google to rank their search results. There are several variations of this algorithm, but they all follow the same general concept [22, 14], described as follows [44]. This algorithm ranks web pages based on the number of existing references to them. The more references to a page, the higher its rank, and the higher its position in the search results list. Also, each page's PageRank value (PR) is distributed uniformly to its outlinks. This means that the higher a page's PR, the more impact it will have on the outlink's PR. Viewing the web as a directed graph $G(V, E)$, V is the set of nodes (or vertices) representing web pages, and E is the set of edges representing the links between pages; for each vertex V_i , there's a set of vertices that point to it (inlinks), $In(V_i)$, and a set of vertices that it points to (outlinks), $Out(V_i)$. Equation 2.1 demonstrates how to calculate the PageRank value of a page.

$$PR(V_i) = (1 - d) + d * \sum_{V_j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|} \quad (2.1)$$

The sum of all PageRank values must equal 1; therefore, each page has an initial PR equal to $\frac{1}{N}$, where N refers to the total number of nodes (pages). The variable d is a damping factor which can have values between 0 and 1. It prevents the occurrence of null PageRank values by establishing a minimum value for each node. This algorithm is recursive, and its execution iterates

until a given threshold is reached. Additionally, this algorithm is run offline, not at query time. Also, the PageRank algorithm is biased against new web pages [6], as these are the most likely to have fewer pages referencing it, due to webmasters or search engine crawlers not having found them yet. The same can be said about very old pages, as they are more likely to be forgotten and not referenced by other pages.

When it comes to developer tools, out of the referenced search engines, only Google and Bing have web search APIs available for use. Yahoo! had a web search API in the past, but it is now inaccessible, as it was discontinued. Both the Google Custom Search API⁴ and the Bing Web Search API⁵ allow calls with several arguments, including relevant factors for this research such as location or safe search. However, none of them allow for controlling variables such as the presence of cookies, private/non-private window, or user authentication status. In addition, both APIs return 10 search results by default, and this amount can be extended. When it comes to pricing, the free version of the Google API provides 100 search queries per day, and the free version of the Bing API allows 1,000 queries per month. The number of queries available may be extended for a fee.

2.1.4 Search Results Evolution Over Time

It's common for search results to change with time, as web documents are frequently updated, and search engines try to keep up with these changes. But the issue is that changes in web search results are much larger than the variation of the web itself [52]. This means that, for a certain query, searched documents that are in the top rankings may fluctuate over time, even when there are no actual content changes in those documents. This denotes an instability of web search results over time [34].

This instability can become frustrating for users, especially if they wish to find a document that a previous search had returned. As it happens, it has been suggested that 40% of web queries are re-finding queries [56], and rank changes in the results — or the absence of the results altogether — can make it hard for users to get back to previously discovered documents.

We can distinguish between several types of search results instability [34]. Firstly, we can categorize it as structural and non-structural. Structural instability derives from search engines deploying new features or indexing techniques, which can cause changes to the results rankings for a wide range of queries. When there are no such events that explain the changes in the search results, we refer to it as non-structural instability. This work focuses on this type of instability. We can also characterize the instability according to indexing issues and ranking issues. Ranking issues cause variation in the ranking positions of documents, meaning that documents can swap positions in the ranking, which can be caused by document content changes, link structure changes, among other factors. The indexing issues refer to the availability of a document in the search engine index. If a document isn't present in the index, it can't be found in a search. These issues can be caused by new crawling techniques, architecture and capacity limitations, spam detection policies,

⁴<https://developers.google.com/custom-search/v1/overview>

⁵<https://www.microsoft.com/en-us/bing/apis/bing-web-search-api>

and others. From a top 10 results page perspective, indexing issues cause documents to be inserted or removed from the top 10, having no effect on the documents' ranking positions. Furthermore, changes in web search results can be short-term or long-term changes. The duration of a change measures how much time the documents keep their positions after going through the changes.

2.2 Related Work

In this section we will provide an overview of past works related to our study. Section 2.2.1 describes past works that analyzed how time impacts the changes in search results, Section 2.2.2 presents related works that studied how search results vary according to the search engine used, and Section 2.2.3 describes past works that analyzed how context features and user personalization affects the results retrieved in web searches.

2.2.1 Mutability of Search Results Over Time

Many investigations have analyzed the volatility web search results have over time. In 2011, Altingovde et al. [3] analyzed query results retrieved by the Yahoo! search engine public API in order to draw conclusions about the changes in the web and query results over time. This study used a large set of 630,000 queries sampled from the AOL Query Log [45], and the search results were gathered twice at two distant points in time, set apart by 3.5 years (from 2007 to 2010), and the top 1, top 10, top 20 and top 100 results were analyzed from each search. Many conclusions were drawn from the study. First, the number of retrieved results per query increased from 16.5 million to 52.3 million, which can be explained by the overall growth of the web between 2007 and 2010. Additionally, in 2010, the returned results' titles decreased in size, and their snippets increased in size. The main findings of the experiment indicated that 20% of the URLs returned in the top 1 result in 2010 were at the same position in 2007, and that value is approximately 11% for URLs returned in the top 10 results, which leads us to believe that relevant documents don't necessarily change with the growth of the web. Moreover, 37% of the unique domain names returned in the top 1 results in 2010 were also at the same position in 2007, and that value becomes approximately 27% for the top 10 results. These numbers suggest that domains retrieved by the queries in 2007 could successfully answer queries in 2010 but, even though there is a substantial amount of results that appear in the query results of both 2007 and 2010, most of the query results changed in the span of 3 years.

A study conducted in 2000 [52], measured the stability of search engines, studying the difference between search results returned for the same queries, submitted at two different times. A set of 25 queries was used in nine search engines (AltaVista, Excite, Lycos, HotBot, InfoSeek, Northern Light, PlanetSearch, WebCrawler and Yahoo!), and the top 10 and top 200 results for the queries were gathered 25 times over a period of a month (from 12/1998 to 01/1999). The time window between issuing the query set was increased along the experiment, starting at 15 minutes and moving up from there. The changes in the retrieved results were compared for each search engine individually. The study concluded that there is a very high rate of change in search results

over time, with an average of 54% for the top 10 results, and going as high as 64% in some search engines. This experiment demonstrates that there is a large change in web search results, which isn't fully explained by the web's dynamic nature. It is hypothesised that the high rate of change in search results is due to the fact that search engines would trade off search quality for speed, as search engines limit the resources available for each query. Even so, this study may prove to be less relevant, because of the very small set of queries used, and the age of the study itself, as the web and search engines have evolved considerably since this study was performed.

In 2018, Jimmy et al. [32] analyzed the volatility of search engines, by studying the differences in the results obtained for queries in several different points in time. This study also aimed at assessing the relevance of the top search results and analysing search engine effectiveness over time, in order to determine the impact of search results volatility on information retrieval research. In order to assess the differences in search engine results, the TREC 2013, 2014 Web Track and 2016 CLEF eHealth IR query collections were used, resulting in a total of nearly 400 queries. The Bing Search API was used to perform retrievals of the top 10 results nearly every two days for a period of just over 2 months (from 11/2017 to 01/2018). The number of new URLs was counted between different retrieval dates. Differences in URL ranking were also investigated. The conclusions were the following: on average, each day had 24.43% new URLs when compared to the initial sampling date, and a 10.72% new URLs when compared to the previous sampling date (usually two days prior); on average, each day, the URLs moved 11.36% up or down the ranking when compared to the initial sampling date, and 6.29% when compared to the previous sampling date. Furthermore, in order to assess the impact of search results volatility on information retrieval research, the relevance of the top 10 results of every sample date was evaluated using a simplified TREC 2013 [16] judgement scale, and the results were analysed according to specific TREC 2013 metrics. It was found that, on average, the graded relevance metric (nDCG@10) varied by nearly 20% for each query, with its biggest variation for a query being 143.51%. This denotes that search effectiveness is conditioned by search engine volatility, and that information retrieval research that uses commercial search engines as part of their study could be affected by the mutability of results retrieved by search engines. This means that the knowledge obtained from the present work's results could also have an impact on information retrieval research initiatives.

In a study performed by Kim et al. in 2011 [34], with the objective of analysing the instability of web search results over time, the daily variability in the top 10 returned results for the same query was measured over a period of a month (from 06/2010 to 07/2010). The Bing search engine was used to gather the results, along with a query set of 12,600 random queries, selected from the Bing search engine query logs. Both the changes in the positions (ranking) of the documents retrieved and the changes regarding the addition and removal of documents from the top 10 were analyzed. The conclusions drawn were that approximately 90% of queries experienced some change at top 10 (either because the documents were replaced in the top 10, or because of changes in their ranking positions), over the last two weeks of the experiment. The duration of changes was also analyzed during the last two weeks of the study, and it was concluded that at least 50% of inserted documents are removed from the top 10 in a span of 5 days, and at least 50% of rank

swaps in top 10 documents are revoked within 5 days. This denotes that most of the changes that the top results undergo don't last longer than a few days. It was also concluded that longer queries have higher instability than shorter ones.

While most of the past work conducted on this subject use a large set of queries to perform their investigation, a work published in 2008 [7] studied the number of search results obtained only for the query "infometrics", over the span of 8 years (from 1998 to 2006). In each year, the largest search engines at the time were used to gather the data. The study found there was a very large growth in the number of search results over the years, as it grew from less than 900 to over 24,000 different URLs in the 8-year period. It is also relevant to mention that 80% of the total unique URLs identified over the 8-year period were located at the last data gathering point (in 2006). Another pertinent observation is the fact that, after 8 years, only 18% of the original URLs from 1998 still existed.

2.2.2 The Impact of Search Engines on the Volatility of Search Results

Instead of focusing their efforts on analyzing how much search results change over time, some works have studied how other elements affect the volatility of the results, such as search engines. The study by Kim et al. [34] mentioned in the previous section also included another experiment to compare the instability of search results across different commercial search engines. Using a query set of 1,000 queries (randomly selected from the Bing query logs), the top 10 results were collected daily, over a period of 2 weeks (from 02/08/2010 to 16/08/2010), using the Bing, Yahoo! and Google APIs. Upon concluding the experiment, they observed that all three search engines revealed a similar level of instability when it comes to the search results, with approximately 20% of documents being replaced in the top 10, and around 40% of documents changing positions in the top 10 ranking, over the two-week period. The data showed that nearly 30% of queries experienced some change at top 10 every day.

In 2015, a study [1] was carried out about the differences in search results between the Google and Bing search engines. The results were gathered for two sets of queries: a set of 33 queries, taken from Google Trends in April 2014, and a set of 35 hand-picked queries. The Google custom search API and Bing search API were used, and the data was harvested daily, for a period of 21 days for the first set of queries, and for a period of 17 days for the second set (from 06/2014 to 07/2014). Only the top 10 results were analyzed in most of the tests. The study created two tools: use tensor analysis to obtain a representation of search results and study their differences between search engines; use machine learning to predict the similarities between the results returned from the two search engines. Several data analysis were performed and, contradicting previous and later works, the main conclusion was that the search results from Google and Bing are vastly similar and that that similarity remains consistent over time.

Under the same subject, in 2020, Dritsa et al. [21] conducted a study to analyze the similarity of the search results between the Google, Bing, and DuckDuckGo search engines. This similarity analysis used a metric that took into consideration, for each query, the number of common search results in the top 10, and the distance separating the common results' positions in the list. The

similarity between the results' titles and snippets was also taken into account. This analysis was performed for 300 queries, including around 200 queries from the U.S. version of Google Trends (from May 2016) and around 100 hand-picked queries. The data was gathered using the Bing Web Search API and the Google Custom Search API, as well as web scraping for retrieving the results from the DuckDuckGo website. The queries were performed daily and at the same time, in two separate phases: for a month in 2016 (July-August) and for 2 months in 2019 (May-July). The study showed that Bing and DuckDuckGo output very similar results, while Google's results are largely different from the first two. Also, while all engines produced almost identical titles, the snippets changed a large amount in all engine comparison pairs. Additionally, upon analyzing the results from the two phases, it was concluded that the amount of results similarity between the search engines remained consistent over time, both for the duration of the second experiment and for the three year period. Furthermore, it was also possible to gather data about the effect of time on the search results of each search engine: the average similarity of the search results for the 300 queries, between 2016 and 2019, was 37% for DuckDuckGo, 43% for Bing, and 48% for Google. These numbers denote a large amount of changes in the search results over time.

Table 2.1 presents a summary of the past work about the mutability of search results mentioned in the current section and in Section 2.2.1. It displays the main common methodology elements of the studies. It's clear that much research has been performed, making use of different approaches. Most of these studies focus on analyzing the influence time has on the mutability of search results, both for short and long periods of time. Some works also research how search results change according to the search engine used. The main conclusions that can be drawn from previous works are that the experiments often indicate a significant change in search results over time, which may be explained by the vast evolution of the web's contents over time. Also, it is worth noting that the experiments suggest that results can undergo changes very quickly, within a year, a week, or even daily. Furthermore, it is suggested that the volatility of search results can influence search effectiveness since results can become less relevant after undergoing changes. When it comes to the impact of search engines on the volatility of search results, it is suggested that this factor incurs variance in the results. We've also seen that past works reach some contrasting conclusions regarding how search engines' results differ from each other, which can be conditioned by aspects such as the date at which the study was conducted or the deployed methodology.

2.2.3 Mutability of Search Results by Context Features

Search engines have been deploying strategies to provide customized results for each specific user. This means that particular user characteristics, such as location, previous searches, personal data, among others, can be used by search engines to assess what results might be more appropriate for each individual user. Even though this approach has the potential to produce better search results, the personalization of web search can sometimes lead to the scenario where some users aren't able to receive certain search results because they are deemed as not relevant by the search engine algorithm.

Table 2.1: Summary of Related Work on the Mutability of Search Results

Authors	Search Engines	Query Set	Duration	Time Between Searches	N. Results
Selberg et al. (2000)	AltaVista, Excite, Lycos, InfoSeek, WebCrawler, Northern Light, HotBot, PlanetSearch and Yahoo!	25 queries	1 month (12/1998 to 01/1999)	Incremental	Top 10, Top 200
Bar-Ilan et al. (2008)	AltaVista, Excite, Hotbot, InfoSeek, Lycos, Northern Light, AllTheWeb, Google, Teoma, Wisenut, Gigablast, Exalead, MSN, and Yahoo!	1 query	8 years (1998-2006)	1 year	All Results
Altingovde et al. (2011)	Yahoo! API	630,000 queries sampled from the AOL Query Log	3.5 years (2007-2010)	3.5 years	Top 1, Top 10, Top 20, Top 100
Kim et al. (2011)	Bing, Yahoo! and Google APIs	12,600 queries + 1,000 queries from Bing query logs	4 weeks (June-July 2010), and 2 weeks (August 2010)	1 day	Top 10
Agrawal et al. (2015)	Bing API and Google API	33 queries (Google Trends April 2014) + 35 handpicked queries	21 days (June-July 2014), and 17 days (June-July 2014)	1 day	Top 10
Jimmy et al. (2018)	Bing API	400 queries (TREC 2013 + 2014 Web Track + 2016 CLEF eHealth IR)	2 months (11/2017 to 01/2018)	2 days	Top 10
Dritsa et al. (2020)	Bing API, Google API, DuckDuckGo (Web scraping)	300 queries (U.S. Google Trends May 2016 + handpicked)	1 month (July-August 2016), and 2 months (May-July 2019)	1 day	Top 10

There are also studies that analyze how search engines deliver results tailored to each user based on their specific characteristics. In 2013, Hannak et al. [29] performed an investigation where they tried to answer the following questions: "What user features influence Google's personalized results?" and "How does user personalization affect search results?". The Google search engine was used to perform the experiments. The investigation's methodology consisted of creating several Google accounts, each varying by one specific user feature. Each account submitted 120 queries (from 12 different topics) daily for 7 days, and the top 10 results were collected. The queries were chosen from the 2011 Google Zeitgeist (now Google Trends⁶) and WebMD⁷ query sets. Then, the returned results were compared to evaluate the amount of variability. Several experiments were conducted. For each experiment, several Google accounts were created, each varying by one specific feature; after performing daily identical queries from each account, the results for each query were analyzed. Some of the experiments also performed searches while not being authenticated in any account. The main experiment analyzed how the results changed based on many user features, such as cookies and authentication status, browser and OS (Operating System), IP Address (with varying geolocation), gender, age, and activity history. When it comes to the cookie tracking and authentication status features, the test revealed that the results were similar for all users (logged in, logged out, and with cookies disabled); still, there were some differences in the ranking order of the retrieved results (for example, logged in users received results that were reordered in two places, on average, when compared to users with no cookies enabled). When it comes to the geolocation feature (IP address), there was, on average, a 9% variation in the results, which means that the output of queries from different locations usually differs by one result. When it comes to the activity history factor, three items were analyzed for their effect on the retrieved results: search history, search history and result clicks, and browsing history; the test showed that, for all of these three items, accounts with different activity history retrieved similar results for the same queries. Also, no differences in the results were observed for browser and OS variation, or for varying demographic factors (age and gender). One of the tests performed compared the results retrieved by Google for both the newly created accounts and real user accounts (for which a lot of user data had already been collected). This test revealed that search results from real user accounts showed an 11.7% higher likelihood of differing from the results of the newly created accounts than the newly created accounts' results differing from each other. The overall results of this investigation show that Google personalizes results based on the user's location and based on authentication status and the presence of cookies. They also indicate that user personalization has relevant weight on the received results. Furthermore, the research also concluded that some query topic's results are more volatile than others. For example, the topics "politics" and "news" have more new results per day and the highest daily rate of results reordering, while the results for the topics "what is" and "green" vary the least. Finally, it was also observed that after the user performs one search, Google personalizes search results for consecutive searches based on the first search, for the following 10 minutes.

⁶<https://trends.google.com/trends>

⁷<https://www.webmd.com/news/year-in-health/default.htm>

Kliman-Silver et al. [36] performed a study focused on analyzing the impact of user location on Google search results personalization. In order to conduct the investigation, 240 queries from 3 different search topics were submitted to the Google search engine for 30 days. The search results were gathered using 59 different GPS coordinates in the United States at three granularities (county, state, and national locations). For each granularity, the search results were collected daily, and only the first page of results was gathered. This project concluded that the results retrieved by Google are influenced by the user's location and suggests that the differences in the search results due to this personalization grow as the physical distance between users' locations increases. Another relevant conclusion was that the variance of the results due to location-based personalization depends heavily on the query topic. While results for queries from the topics "politicians" and "controversial" fluctuate minimally, results for queries from the topic "places" (like "bank" or "Chipotle") varied a lot based on user location (between 18% and 34% of results changed, and 6-10 URLs had their ranking order shifted). It was also noted that personalization based on location is consistent over time.

The analysis of past works indicates that user personalization has a significant impact on the volatility of search results, with Google being the most investigated search engine in this field. It was also noted that factors such as location, authentication status, and presence of cookies can influence the search results retrieved, with location being the most analyzed factor. To the best of our knowledge, there haven't been any works that focus on studying the effect of other factors, such as private/non-private window or safe search, on the changes in search results.

Chapter 3

Objectives and Methodology

This chapter presents the focus of this research, and how the research was carried out. Section [3.1](#) outlines the goals of this project and Section [3.2](#) describes the methodology used to perform the research.

3.1 Objectives

The aim of this study is to gather web search results data and describe the volatility of the search results according to different search conditions when submitting the same query. This work examines the influence of the following search factors on the mutability of search results: time, location, safe search, privacy (private/non-private window), user agent, cookies, and authentication status. Two different search engines — Google and Bing — were used to evaluate the impact of each factor, which allowed us to make some comparisons between them after concluding the analysis of the results. We also explored the influence of query topics on the volatility of search results caused by each factor.

Concretely, the main objectives of this work are: to gather search results data for the variance of each search factor, for each search engine; to analyze the amount of changes that search results undergo; and, most importantly, to store all the data obtained in well organized datasets to be made available to the public, so they can be used for further researches regarding this topic. Furthermore, another goal of this study is to present a brief analysis of the results obtained in order to raise hypotheses about the influence of the several search factors, search engines, and query topics on the volatility of search results, serving as a proof of concept that the data collected can be used for future deeper analyses.

3.2 Methodology

To perform this investigation, seven experiments of search results retrieval were carried out, one for each search factor. The search factors to study were determined based on past works reviewed as well as other factors that could prove to be relevant. The search results for all experiments were gathered using the Google and Bing search engines, and the experiments used queries from the same set of 298 queries. Section 3.2.1 describes the search engine selection process, Section 3.2.2 describes the search factors used and their possible values, and Section 3.2.3 describes the selection process of the query set used. We also need to establish some terminology that will be used throughout the work. The term *retrieval* refers to a gathering of search results for a certain query/set of queries, using a certain value of the varying search factor. For example, when it comes to the safe search experiment, there was one *retrieval* for the used query set with safe search enabled and another *retrieval* for the query set with safe search disabled. In each experiment performed, the top 10 results were retrieved for each value of the varying search factor. Three different methods were used to retrieve the results for different groups of experiments: for group **Exp-APIs** (*time*, *location*, and *safe search* variables) results were gathered via API calls; for group **Exp-Manual** (*privacy* and *user agent* variables) results were gathered manually followed by web scraping; and for group **Exp-RealUsers** (*cookies* and *authentication status* variables) results were gathered by real users followed by web scraping. Sections 3.2.4 thru 3.2.6 outline the specific methods used for each of the 3 groups of experiments. All experiments were performed using the same IP network. After the results retrieval phase of each experiment, the results were all stored in the same format, in a CSV file. Section 3.2.7 presents the metrics used for the results analysis. Fig 3.1 shows the overall methodology followed for this work, and Fig 3.2 shows a timeline chart of all the experiments carried out.

3.2.1 Search Engine Selection

When it comes to the selection of search engines to study in the project, we considered Google, Bing, and Yahoo!, as these are the most used search engines worldwide, as of January 2021, with Google holding an overwhelming 91.86% worldwide market share, Bing having 2.71%, and Yahoo! having 1.46% [54]. Upon analyzing the developer tools made available by each of the considered search engines, we found that only Google and Bing have web search APIs available for use. Yahoo! used to have a web search API in the past, but it has been discontinued. In the end, the fact that Yahoo! doesn't have an API makes it ineligible for our experiments, so we chose the Google and Bing search engines to conduct our experiments.

Both the Google Custom Search API¹ and the Bing Web Search API² allow calls with several arguments, including relevant factors for this research such as location and safe search. When it comes to the pricing and conditions of each API, the free plan of the Google API has a limit of 100 search queries per day, and the free plan of the Bing API has a limit of 1,000 queries per month.

¹<https://developers.google.com/custom-search/v1/overview>

²<https://www.microsoft.com/en-us/bing/apis/bing-web-search-api>

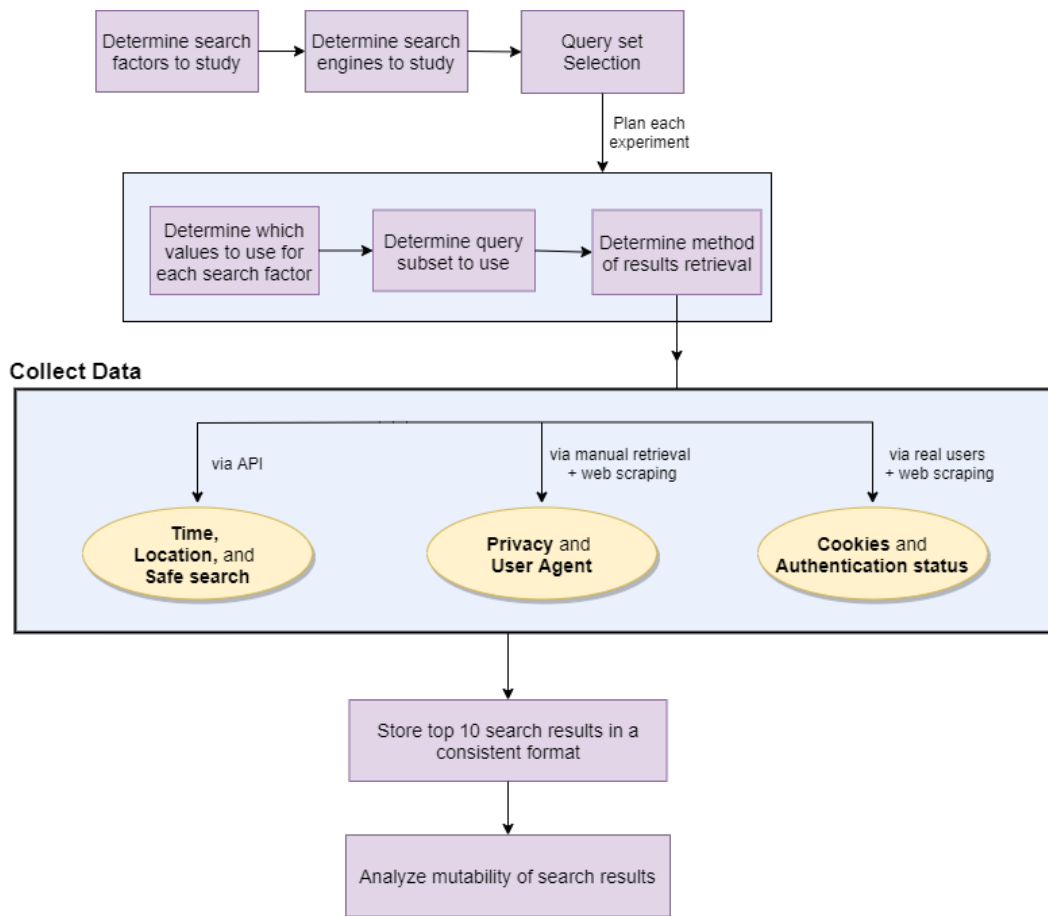


Figure 3.1: General Methodology Diagram

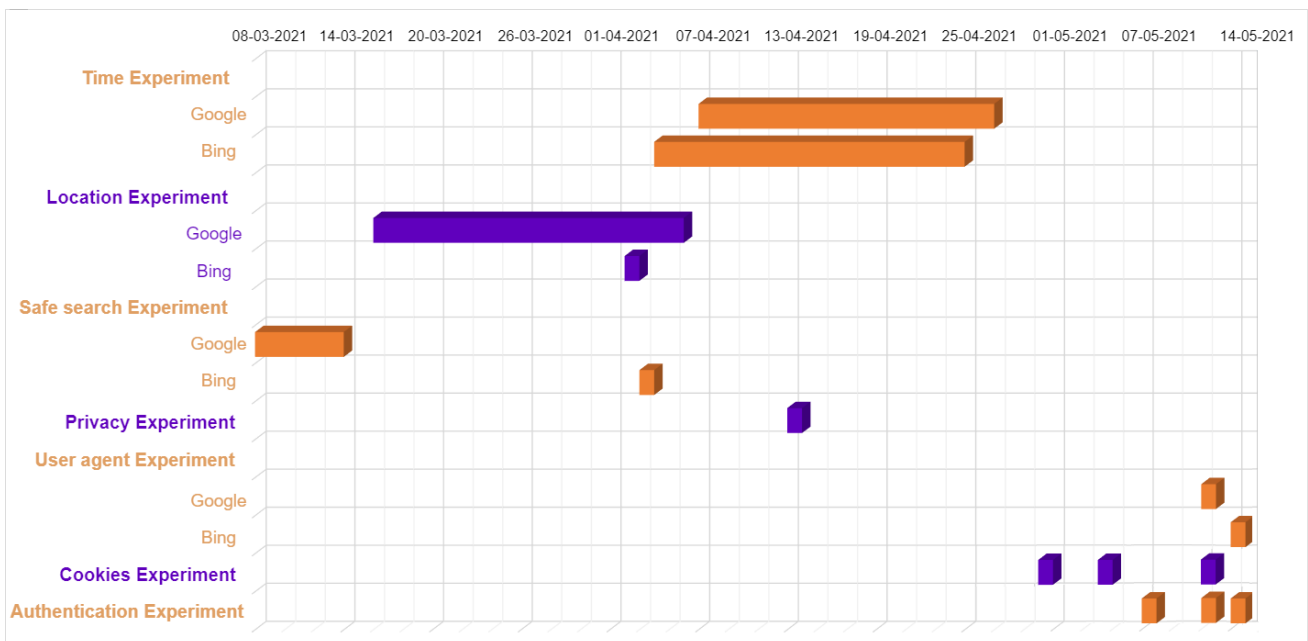


Figure 3.2: Timeline Chart Showing All the Experiments Performed

3.2.2 Search Factors Examined

The search factors studied in this work were *time*, *location*, *safe search*, *privacy*, *user agent*, *cookies*, and *authentication status*.

The *time* factor refers to the date in which a search was performed. For this factor, the top 10 results of each query were gathered every 3 days over approximately 3 weeks (see Section 3.2.4.1 for further details). Thus, for the *time* factor, the values used were all the dates in which retrievals took place.

The *location* factor refers to the geolocation where a search is performed, which may influence the results obtained. The search results of each query were gathered using seven different locations. The seven locations were picked based on the following method: for each of the following regions of the world — Asia, Europe, North America, Latin America, Africa, Oceania, and Middle East — the country with the largest population (as of January 2021) [60], belonging to a single continent, was chosen. Along with the continental regions, the Middle East was added to the list of regions so as to obtain a larger diversity of locations. Using the mentioned method, the countries obtained were: China, Germany, United States, Brazil, Nigeria, Australia, and Egypt. However, when it comes to the Bing search engine, some of the chosen locations (Nigeria and Egypt) weren't available to be used as arguments in the Bing API call parameters. This way, the next country of the corresponding region of the world with the largest population, available as an argument in the Bing API call parameters, was chosen. Therefore, Nigeria and Egypt were replaced with South Africa and Turkey (all other locations remained the same).

As for the *safe search* factor, safe search is a setting used to hide results that contain explicit content, which can be turned on and off. For this factor, the results were gathered with safe search enabled and disabled.

When it comes to the *privacy* factor, most web browsers available provide two possible privacy modes that can be used when searching the web: private window and non-private window. The private window allows the user to search the web without storing browsing history, cookies, and site data, and third party cookies will usually be blocked. For this factor, the results were gathered using a private window and using a non-private window.

As for the *user agent* factor, the user agent is an HTTP header used in HTTP requests that allows servers and network peers to identify the application, operating system, vendor, and/or version of the requesting browser [42]. Four different user agents were used to gather the search results. The 4 user agents chosen consisted of the user agents containing the latest versions of the 4 most used web browsers (as of April 2021): Chrome, Safari, Edge, and Firefox [59]. A summary of the used user agents can be seen in Table 3.1.

As for the *cookies* factor, cookies are small blocks of data created by a web server while a user is browsing a website. They track, personalize, and save information about each user's session (such as their logins, shopping carts, and more) [33]. The personal cookies accumulated in each user's browser's directory may impact the search results obtained. For this factor there will be as many values as real users carrying out the study (each user has a unique set of personal cookies).

Table 3.1: User Agents Used

ID	User Agent
Chrome-Win	Chrome 90.0.4430.85 for Windows 10
Firefox-Win	Firefox 88.0 for Windows 10
Safari-Mac	Safari 14.0.3 for Mac OS X 11.2.3
Edge-Win	Edge 90.0.818.46 for Windows 10

The *authentication status* factor refers to the personal search engine account a user is authenticated in while performing a web search. The personal data of each user's account may influence the search results obtained. Again, for this factor there will be as many values as real users carrying out the study.

3.2.3 Search Queries Selection

As for the search queries used in the study, two existing query sets were considered to be used to supply the necessary queries for the project: The AOL Query Logs 2004 query set [8] and the Google Trends worldwide (2020) [24]. The Google Trends worldwide (2020) query set contains queries obtained by using artificial intelligence to identify the most searched queries in a certain time period (in this case, in the year of 2020) in Google's web search service, for each topic considered. This query set includes queries from 25 search topics, and each topic has a maximum of 100 queries, 50 described as "rising" queries and 50 described as "top" queries. The AOL Query Logs 2004 query set contains 23,779 total queries randomly sampled from AOL Search in 2004, which were subsequently classified into 20 query topics by a team of human assessors. Each topic has an average of 1,189 queries. While the Google Trends query set had more recent and popular queries, it had many disadvantages: only a maximum of 100 queries for each search topic was available (which offers less diversity of queries to choose from); many of the queries were not representative of the current topic due to the fact that they are obtained through an AI system (for example, the topic "Autos & Vehicles" had queries like "whatsapp web", "microsoft teams", "Minecraft", and "weather tomorrow", which are clearly not related to the current topic); and there was a large amount of repeated queries inside several search topics (for example, the topic "Computers & Electronics" had 3 queries pertaining to the newly released PlayStation 5 console³, and the "Health" topic had 29 queries pertaining to COVID-19⁴). As for the AOL query set, even though its queries weren't as recent, it contained more queries for each topic and the queries were more representative of each search topic, as they were analyzed and classified by humans. In the end, the AOL query set was chosen as it offered more benefits than the Google Trends one.

³<https://www.playstation.com/en-us/ps5/>

⁴https://www.who.int/health-topics/coronavirus#tab=tab_1

When it comes to the selection of the query topics to use in our study, we chose topics that were present in both the AOL query set and the Google Trends query set, that appeared to be the most relevant to our research, and that had over 3% of the total amount of queries in the AOL query set. The queries used in this study’s experiments cover the 10 following search topics: *Arts & Entertainment, Autos & Vehicles, Business & Industrial, Computers & Electronics, Health, Home & Garden, Jobs & Education, News, Sports, and Travel*. Table 3.2 summarizes the query topics used.

Table 3.2: Summary of Query Topics Used

Topic ID	Topic Name	N°. Queries
T1	Arts & Entertainment	30
T2	Autos & Vehicles	30
T3	Business & Industrial	30
T4	Computers & Electronic	28
T5	Health	30
T6	Home & Garden	30
T7	Jobs & Education	30
T8	News	30
T9	Sports	30
T10	Travel	30

When it comes to the set of queries used for this study, we generated a query set of 298 queries, randomly selected from the AOL Query Logs 2004 query set, spanning the 10 chosen search topics. The selection criteria for the queries was as follows: for each of the 10 chosen search topics, thirty random numbers (IDs), between 0 and the total number of queries for that search topic, were generated, and the queries corresponding to those IDs were selected. Repeated numbers were ignored and more random numbers were picked until 30 queries had been obtained. This process was repeated for all 10 search topics. Therefore, the obtained query set contains 30 queries for each of the 10 topics. The chosen query set initially had 300 queries, but two of them needed to be removed due to not producing enough search results (didn’t return a minimum of 10 search results). All experiments used either the entire generated query set or a specific subset of it. A dataset detailing the complete set of search queries used for this project was made available in the repository associated with this study [40].

3.2.4 Collecting Data for the Analysis of the *time, location, and safe search* Factors

Experiments from group *Exp-APIs*, which cover the search factors *time, location, and safe search*, were performed for the Google and Bing search engines, using the Google Custom Search API⁵ and the Bing Web Search API⁶. The whole query set of 298 queries was used for the Google

⁵<https://developers.google.com/custom-search/v1/overview>

⁶<https://www.microsoft.com/en-us/bing/apis/bing-web-search-api>

experiments, whereas for Bing only a subset of 120 queries was used (containing the first 12 queries of each topic). The difference between the number of queries used for Google and Bing is due to the Bing API's monthly restriction of 1,000 query calls; this way, the amount of queries per topic had to be reduced, in order to include queries of every topic in the study, as well as to make sure there were enough query calls available to perform all the API dependent experiments. Fig 3.3 and Fig 3.4 show how the experiments were conducted for the Google and Bing search engines. The next sections will explain each experiment in more detail.

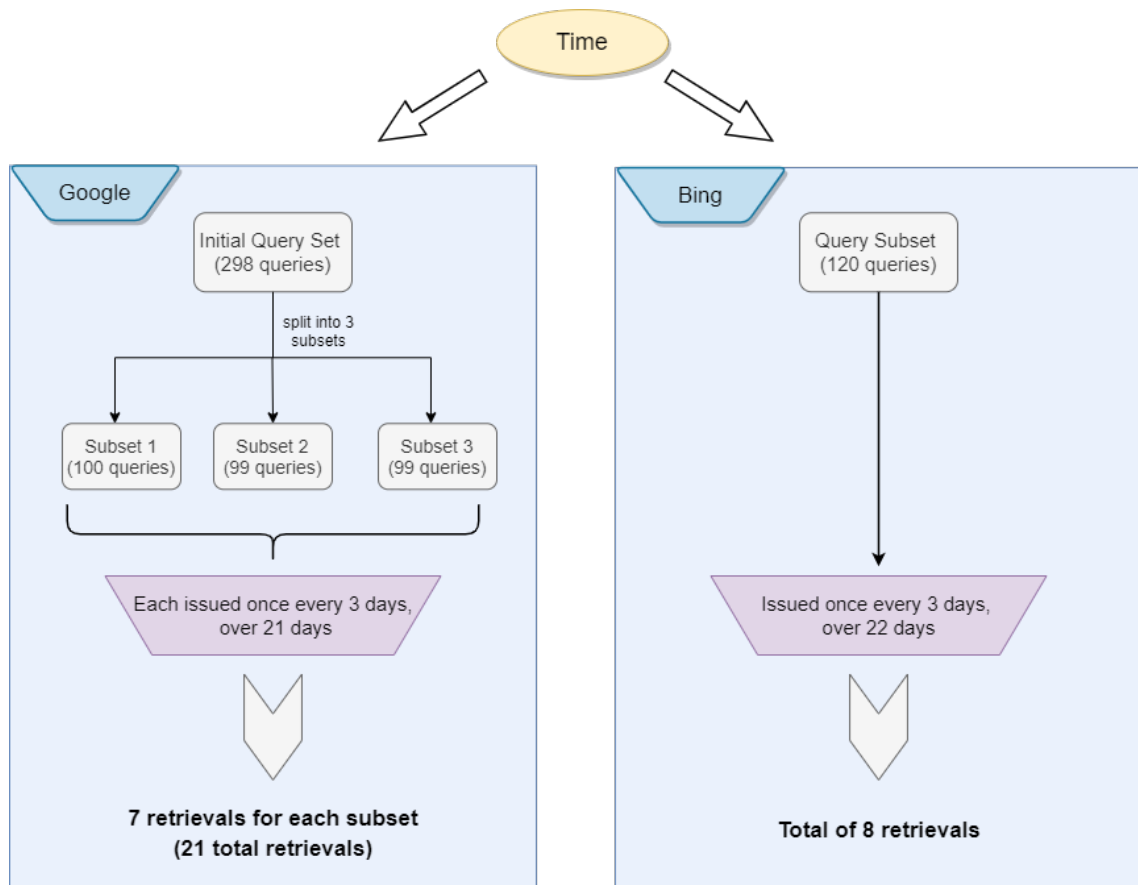


Figure 3.3: Methodology Diagram for the *time* Experiment

3.2.4.1 Time

This experiment evaluated the effect of time on the mutability of search results. The experiment was executed for Google and Bing using the previously mentioned APIs. Other search factors remained similar throughout the whole experiment since the objective is to evaluate only the time factor. For both search engines, the top 10 results of each query were gathered every 3 days over approximately 3 weeks. For Google, the search results were gathered from 07-04-2021 to 27-04-2021, and for Bing, the search results were gathered from 04-04-2021 to 25-04-2021.

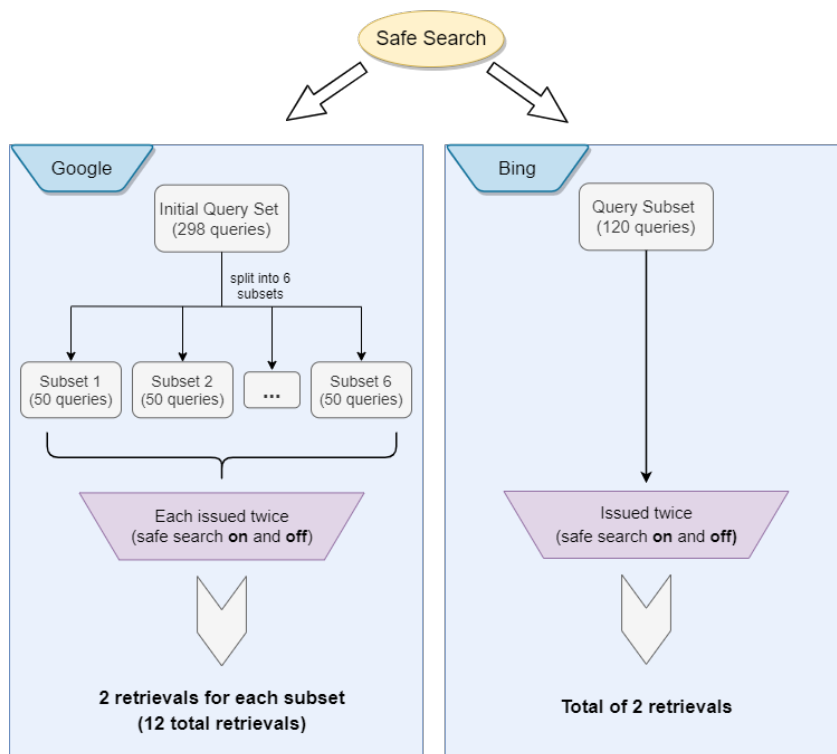
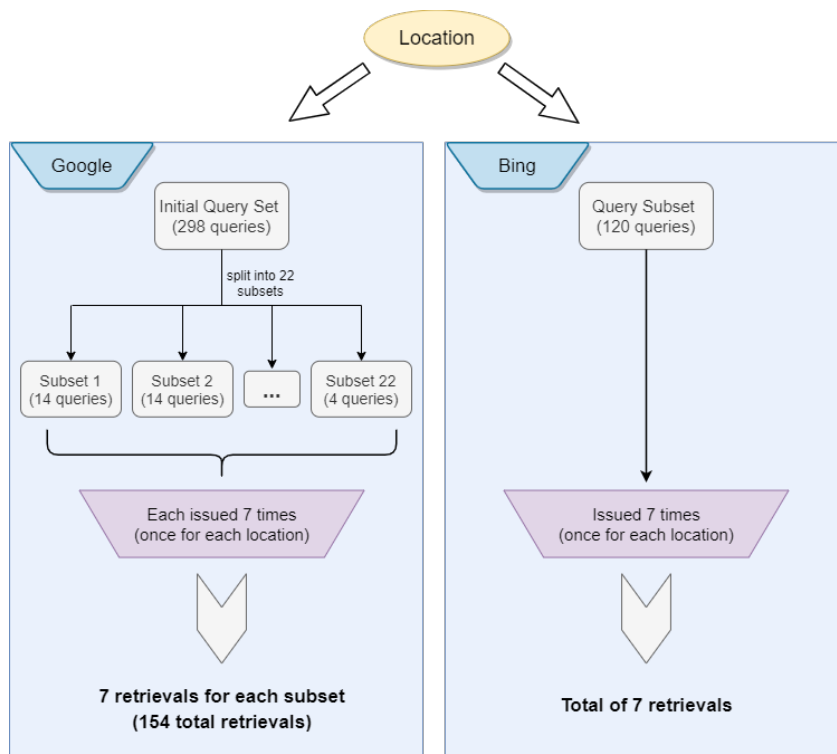


Figure 3.4: Methodology Diagram for the *location* and *safe search* Experiments

When it comes to the Google search engine, because of the Google API’s limit of 100 query calls per day, the initial query set (298 queries) was divided into 3 equally sized subsets, and each subset of 100 distinct queries (10 of each topic) was issued once every 3 days, for the duration of 21 days. Therefore, 100 queries were issued in day 1, 99 different queries were issued in day 2, and another 99 different queries were issued in day 3; in day 4, the same queries as day 1 were issued, and so on. So, for each query, the top 10 results were gathered on 7 different sampling dates. Table 3.3 shows the queries used in each retrieval performed, for Google. The variables *10a*, *10b*, and *10c* refer to the 1st-10th, 11th-20th, and 21st-30th queries of each search topic, respectively. We can see that there is a window of 3 days between the retrieval for each query subset. We can also see that starting on day 4, the pattern repeats itself until the end of the experiment.

Table 3.3: Queries Used in Each Retrieval, When Varying the *time* Factor, for Google

Topic \ Day	1	2	3	4	...
T1	10a	10b	10c	10a	
T2	10a	10b	10c	10a	
T3	10a	10b	10c	10a	
T4	10a	10b	10c	10a	
T5	10a	10b	10c	10a	
T6	10a	10b	10c	10a	...
T7	10a	10b	10c	10a	
T8	10a	10b	10c	10a	
T9	10a	10b	10c	10a	
T10	10a	10b	10c	10a	
Total	100	99	99	100	...
Date	07-04-2021	08-04-2021	09-04-2021	10-04-2021	...

As for the Bing search engine, the query set of 120 queries was issued once every 3 days, for the duration of 22 days. Thus, for each query, the top 10 results were gathered on 8 different sampling dates. Table 3.4 shows the queries used in each retrieval performed, for Bing. The variable *12a* refers to the 1st-12th queries of each search topic. We can see that there’s a window of 3 days between each retrieval, until the end of the experiment.

The implementation details regarding how the retrieval of search results was executed and how the data was stored are explained in Section 4.1.1.

3.2.4.2 Location

This experiment analyzed the effect of geolocation on the mutability of search results. The experiment was executed for Google and Bing using the previously mentioned APIs and other search factors remained similar throughout the whole experiment. The top 10 results of each query were gathered using the seven different locations mentioned earlier. The location of each query call was controlled via an API call parameter. The API parameters used, along with further implementation

Table 3.4: Queries Used in Each Retrieval, When Varying the *time* Factor, for Bing

Topic \ Day	Day				
	1	2	3	4	...
T1	12a	-	-	12a	
T2	12a	-	-	12a	
T3	12a	-	-	12a	
T4	12a	-	-	12a	
T5	12a	-	-	12a	
T6	12a	-	-	12a	...
T7	12a	-	-	12a	
T8	12a	-	-	12a	
T9	12a	-	-	12a	
T10	12a	-	-	12a	
Total	120	-	-	120	...
Date	04-04-2021	05-04-2021	06-04-2021	07-04-2021	...

details regarding how the retrieval of search results was executed and how the data was stored are explained in Section 4.1.1.

As for the Google search engine, due to the API's limit of 100 queries per day, the initial query set (298 queries) was divided into 22 subsets, and one subset was issued each day, using all locations, for the duration of 22 days (from 16-03-2021 to 06-04-2021). The query subsets were created in such a way that the amount of queries issued per day was maximized. This way, most of the subsets have 14 queries; as each query needs to be issued 7 times (once for each location), 14 queries * 7 equals 98 query calls per subset, out of the 100 daily query calls available. The 22nd subset is the only one that breaks the rule, having only the 4 remaining queries. The results for each distinct query, using all seven locations, are gathered in the same day, thus mitigating the effect of time on the collection. Table 3.5 shows the queries used in each retrieval performed, for Google. The variables *14a* and *14b* refer to the 1st-14th and 15th-28th queries of each search topic, respectively, and the variable *2c* refers to the last two queries of each topic (29th and 30th). It's clear that after the first two query subsets the pattern repeats itself 9 more times for the remaining query topics. The final two query subsets are structured differently in order to include the all the remaining queries.

As for Bing, the subset of 120 queries was completely issued, using all locations, on the same day (02-04-2021). Table 3.6 shows the queries used in each retrieval performed, for Bing. The variable *12a* refers to the 1st-12th queries of each search topic.

3.2.4.3 Safe Search

This experiment evaluated the effect of the safe search setting on the mutability of search results. The experiment was also executed using both Google and Bing's APIs. As always, other search factors remained similar throughout the whole experiment. The top 10 results of each query were

gathered with safe search enabled and disabled. This setting was also controlled via a parameter in the API call. The API parameters used, along with further implementation details regarding how the retrieval of search results was executed and how the data was stored are explained in Section 4.1.1.

When it comes to the Google search engine, once again, the API's limit of 100 queries per day led us to split the initial query set (298 queries) into 6 subsets of 50 queries each, and one subset was submitted each day, once with safe search enabled and once with safe search disabled, for the duration of 6 days (from 08-03-2021 to 14-03-2021). Table 3.7 shows the queries used in each retrieval performed, for Google. The variable *30a* refers to the 1st-30th queries of each search topic, the variables *20a* and *20b* refer to the 1st-20th and 11th-30th queries of each topic, respectively, and the variables *10a* and *10b* refer to the 1st-10th and 21st-30th queries of each topic, respectively. It can be seen that the date of 10-03-2021 was skipped; this is because we experienced difficulties during the results retrieval on this date, so the retrieval for the query subset needed to be performed again on 12-03-2021.

Table 3.7: Queries Used in Each Retrieval, When Varying the *safe search* Factor, for Google

Topic	Safe Search		On		Off		On		Off		On		Off	
	On	Off	On	Off	On	Off	On	Off	On	Off	On	Off	On	Off
T1	30a	30a	-	-	-	-	-	-	-	-	-	-	-	-
T2	20a	20a	10b	10b	-	-	-	-	-	-	-	-	-	-
T3	-	-	30a	30a	-	-	-	-	-	-	-	-	-	-
T4	-	-	10a	10a	20b	20b	-	-	-	-	-	-	-	-
T5	-	-	-	-	30a	30a	-	-	-	-	-	-	-	-
T6	-	-	-	-	-	-	30a	30a	-	-	-	-	-	-
T7	-	-	-	-	-	-	20a	20a	10b	10b	-	-	-	-
T8	-	-	-	-	-	-	-	-	30a	30a	-	-	-	-
T9	-	-	-	-	-	-	-	-	10a	10a	20b	20b	-	-
T10	-	-	-	-	-	-	-	-	-	-	30a	30a	-	-
Total	100		100		100		100		100		100		100	
Date	08-03-2021		09-03-2021		12-03-2021		11-03-2021		13-03-2021		14-03-2021			

As for Bing, the query subset of 120 queries was fully issued, once with the safe search setting enabled and again with the setting disabled, in a single day (03-04-2021). Table 3.8 shows the queries used in each retrieval performed, for Bing. The variable *12a* refers to the 1st-12th queries of each search topic.

3.2.5 Collecting Data for the Analysis of the *privacy* and *user agent* Factors

Experiments from group *Exp-Manual*, which cover the search factors *privacy* and *user agent*, were performed by manually retrieving the search results from the web and using web scraping to convert the search results data from HTML format into a concise CSV format. While manually retrieving the results, we ensured all other search factors remained similar throughout the whole

Table 3.8: Queries Used in Each Retrieval, When Varying the *safe search* Factor, for Bing

Topic	Safe Search	
	On	Off
T1	12a	12a
T2	12a	12a
T3	12a	12a
T4	12a	12a
T5	12a	12a
T6	12a	12a
T7	12a	12a
T8	12a	12a
T9	12a	12a
T10	12a	12a
Total	120	120
Date	03-04-2021	

experiments. Further implementation details regarding how the manual process of results retrieval was executed and how the data was stored are explained in Section 4.1.2. Only the first page of search results was retrieved for each query issued, and only the first 10 results, at most, were considered for the study. Because the number of results on the first page can vary from 6 to 12, when comparing the results of two distinct retrievals they may have a different number of results. In these cases, only the lower number of results between the two retrievals is taken into consideration when calculating the metrics. So, if *retrieval 1* has 9 results and *retrieval 2* has 8 results, only the top 8 results will be taken into consideration when calculating the metrics for this pair of retrievals. Thus, the metrics calculated for the top 10 results may take into consideration less than the first 10 results (it could be between 6 and 10 results).

When performing the manual collection of results, we considered as a search result all principal URLs obtained in the first page of the search. Therefore, in results that contain a main URL followed by several secondary URLs (as can be seen in Fig 3.5), only the main URL was considered. Also, when it comes to results that led to pages of the own search engine (such as Google Images, Bing Videos, etc), these were also considered as results as long as there was a clickable URL (usually a button displaying "View All") that led to a page showing all elements of that result (all images/videos/etc).

Both experiments from group *Exp-Manual* used a subset of the initial query set containing 100 queries (the first 10 queries of each topic) and both experiments were carried out for both Google and Bing. This smaller amount of queries used was due to time restrictions, since all the queries had to be issued manually by the investigators. Fig 3.6 shows an overview of how the experiments were conducted. The following sections explain each experiment in more detail.

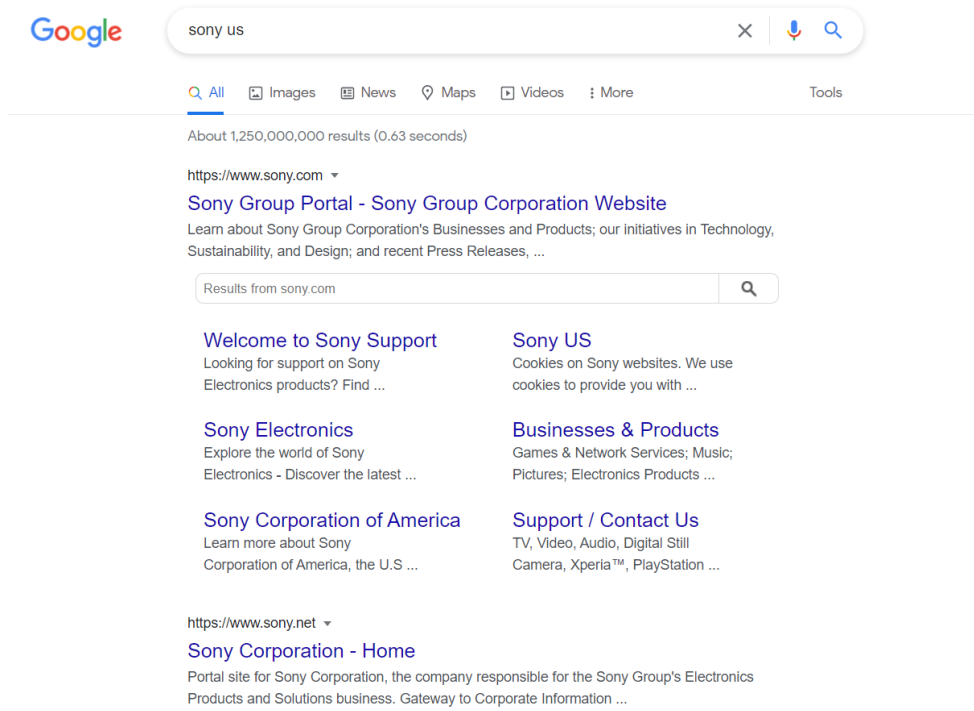


Figure 3.5: Example of a Web Search Result Including One Main URL and Six Secondary URLs

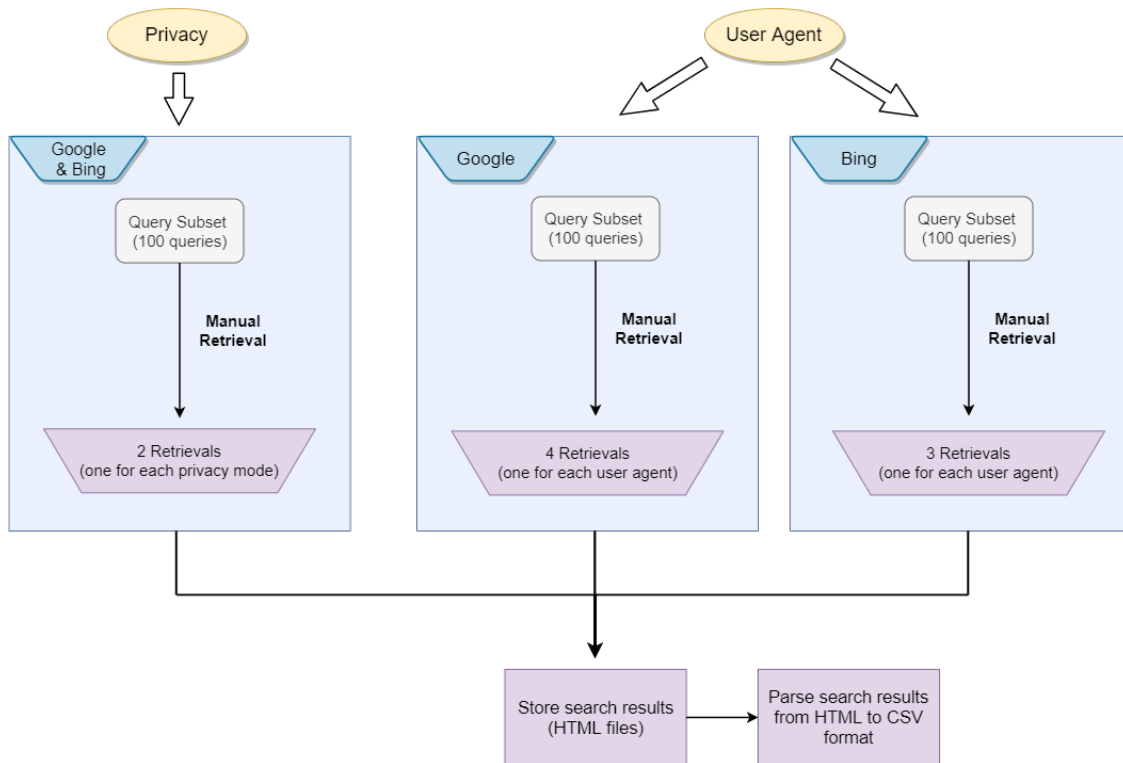


Figure 3.6: Methodology Diagram for Experiment Group *Exp-Manual*

3.2.5.1 Privacy

This experiment studied the effect of the privacy mode on the mutability of search results. Like all previous experiments, this one was also executed for both Google and Bing and all other search factors remained similar throughout the whole experiment. The top 10 results of each query were gathered using a private window and using a non-private window.

The method of results retrieval was similar for Google and Bing. The query subset of 100 queries was completely issued, once using a private window and then again using a non-private window, in a single day (13-04-2021), for each search engine. Table 3.9 shows the queries used in each retrieval performed, for Google and Bing. The variable *10a* refers to the 1st-10th queries of each search topic.

Table 3.9: Queries Used in Each Retrieval, When Varying the *privacy* Factor, for Both Google and Bing

Topic \ Privacy	Non-Private	Private
T1	10a	10a
T2	10a	10a
T3	10a	10a
T4	10a	10a
T5	10a	10a
T6	10a	10a
T7	10a	10a
T8	10a	10a
T9	10a	10a
T10	10a	10a
Total	100	100
Date	13-04-2021	

3.2.5.2 User Agent

This experiment evaluated the impact of the user agent on the mutability of search results. This experiment was also performed for both Google and Bing and all other search factors remained similar throughout the whole experiment. The top 10 results of each query were gathered using the four different user agents mentioned earlier.

As for Google, the query subset of 100 queries was fully issued 4 times, once for each user agent used, in a single day (11-05-2021). Table 3.10 shows the queries used in each retrieval performed, for Google. The variable *10a* refers to the 1st-10th queries of each search topic.

When it comes to Bing, the same query subset was used, but it was only issued 3 times (on 13-05-2021). This happened because the safari user agent (*Safari-Mac*) didn't allow Bing to produce search results, so only the other 3 user agents were used. Upon submitting a query while using the *Safari-Mac* user agent, Bing would output the message "There are no results for <query>. Check

Table 3.10: Queries Used in Each Retrieval, When Varying the *user agent* Factor, for Google

Topic \ User Agent	Chrome-Win	Firefox-Win	Safari-Mac	Edge-Win
T1	10a	10a	10a	10a
T2	10a	10a	10a	10a
T3	10a	10a	10a	10a
T4	10a	10a	10a	10a
T5	10a	10a	10a	10a
T6	10a	10a	10a	10a
T7	10a	10a	10a	10a
T8	10a	10a	10a	10a
T9	10a	10a	10a	10a
T10	10a	10a	10a	10a
Total	100	100	100	100
Date	11-05-2021			

your spelling or try different keywords". We can speculate that Bing didn't return search results because the user agent used had a different operating system than the one running on the local machine (Windows 10). Table 3.11 shows the queries used in each retrieval performed, for Bing. The variable *10a* refers to the 1st-10th queries of each search topic.

Table 3.11: Queries Used in Each Retrieval, When Varying the *user agent* Factor, for Bing

Topic \ User Agent	Chrome-Win	Firefox-Win	Edge-Win
T1	10a	10a	10a
T2	10a	10a	10a
T3	10a	10a	10a
T4	10a	10a	10a
T5	10a	10a	10a
T6	10a	10a	10a
T7	10a	10a	10a
T8	10a	10a	10a
T9	10a	10a	10a
T10	10a	10a	10a
Total	100	100	100
Date	13-05-2021		

3.2.6 Collecting Data for the Analysis of the *cookies* and *authentication status* Factors

The experiments covering the search factors *cookies* and *authentication status* (group *Exp-RealUsers*) were performed similarly to the experiments of the previous group (*privacy* and *user agent* experiments), with the difference being that the search results were manually retrieved by real users instead of by the investigators. The same subset of 100 queries was also used for the experiments of group *Exp-RealUsers*, but only the Google search engine was used in the experiments. The experiments were all carried out in the Faculty of Engineering of the University of Porto, and the recruitment of participants occurred during ongoing physical classes. For the *cookies* experiment, the participants recruited were required to have a browser which they regularly used in non-private mode, and which had been storing their cookies over time. For the *authentication status* experiment, the requirements consisted of the ownership of a Google account which the participant used regularly. The experiments lasted no more than 40 minutes, and a script was given out to each participant which contained the details of the experiment and the steps that should be followed to perform it. The scripts used can be seen in Annex B and also in the repository associated with this work. The experiment consisted of an initial setup to make sure only the search factor being studied varied between all participants (and all other factors remained similar throughout the whole experiment), followed by the submission of a specific set of queries using the Google search engine, the storage of the obtained search results on their local machines as HTML files, and, finally, the delivery of those files to the investigators. Additionally, the investigators were also present to supply further guidance and to validate all the executed steps of the experiments. Moreover, at the end of the experiments, every participant answered a small survey about some of their personal characteristics, such as gender, age, and how frequently they use Google for web searching. The survey conducted is presented in Annex A. A dataset was created which contains the answers of all users to this survey, which was made available for consultation.

The search results gathered during the experiments were stored as HTML files, and web scraping was used once again to convert the data from HTML format into CSV format. Like it happened in the previous experiment group, only the first page of search results was retrieved for each query issued, and only the first 10 results, at most, were taken into consideration. Fig 3.7 shows an overview of how the experiments were carried out. The *cookies* experiment was executed along 5 sessions, each having between 8 and 12 real users issuing the queries. The sessions were carried out on the following dates: 30-04-2021, 04-05-2021, and 11-05-2021. Similarly, the *authentication status* experiment was also executed along 5 sessions, each having between 7 and 11 real users. The sessions were carried out on the following dates: 07-05-2021, 11-05-2021, and 14-05-2021. The following sections explain each experiment in more detail. Further implementation details regarding how the retrieval of search results was executed and how the data was stored are explained in Section 4.1.3.

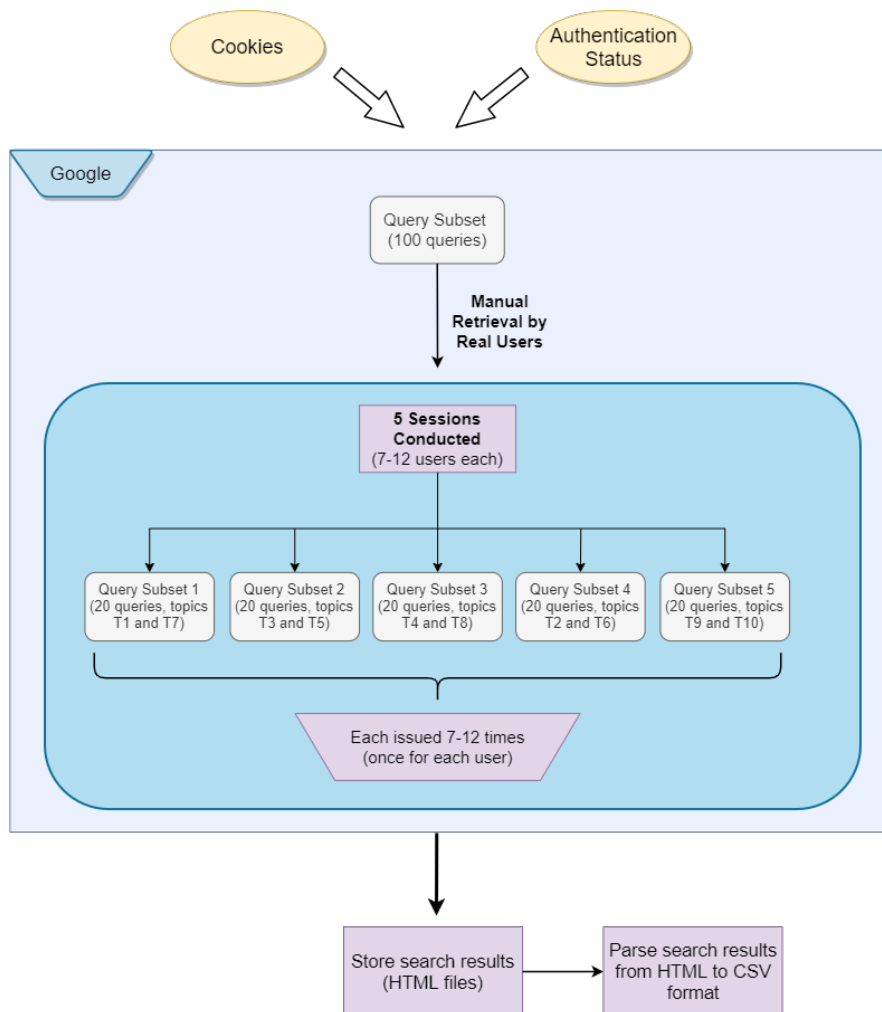


Figure 3.7: Methodology Diagram for Experiment Group *Exp-RealUsers*

3.2.6.1 Cookies

This experiment studied the effect of cookies on the mutability of search results. All other search factors remained similar throughout the whole experiment. The top 10 results of each query were gathered by 8-12 real users, each having a substantial sample of personal accumulated cookies in their browser's directory.

The experiment was carried out across 5 sessions. The query subset of 100 queries was also split into 5 subsets of 20 queries each (the first 10 queries of two specific topics), and one subset was used in each session. So, each subset of 20 queries was issued 8-12 times, once by each user. Table 3.12 shows the queries used, the number of users, and the dates of each session conducted. The variable *10a* refers to the 1st-10th queries of each search topic.

Table 3.12: Queries Used in Each Session, When Varying the *cookies* Factor, for Google

Topic \ Session	S1	S2	S3	S4	S5
T1	10a	-	-	-	-
T2	-	-	-	10a	-
T3	-	10a	-	-	-
T4	-	-	10a	-	-
T5	-	10a	-	-	-
T6	-	-	-	10a	-
T7	10a	-	-	-	-
T8	-	-	10a	-	-
T9	-	-	-	-	10a
T10	-	-	-	-	10a
Total	20	20	20	20	20
N°. Users	9	12	8	10	8
Date	30-04-2021	30-04-2021	04-05-2021	04-05-2021	11-05-2021

3.2.6.2 Authentication Status

This experiment analyzed the effect of authentication status on the mutability of search results. All other search factors remained similar throughout the whole experiment. The top 10 results of each query were gathered by 7-11 real users, while they were authenticated on their personal Google accounts.

The experiment's methodology was the same as the one used on the previous experiment. The experiment was carried out across 5 sessions. The same 5 subsets of 20 queries each were used, and one subset was used in each session. Table 3.13 shows the queries used, the number of users, and the dates of each session conducted. The variable *10a* refers to the 1st-10th queries of each search topic.

3.2.7 Metrics Used in Results Analysis

To express the mutability in the search results gathered, two metrics were created and calculated for every experiment performed. The first metric is called **% New URL** and it refers to the percentage of new URLs that are introduced from one retrieval to another. In essence, it evaluates the number of results/URLs that are not common between the *top k* results of two retrievals, divided by the number of total URLs (for example, when analyzing this metric for the top 10 results, we would divide by 10). A simple example of two retrievals of the top 10 results for the same query (using different values of the varying search factor) can be seen in the Table 3.14 below. We can see that there are two new URLs introduced in retrieval 2, which replace two of the URLs from retrieval 1 (both marked with an "*"). Because we are analyzing the top 10 results in this example, the value of the metric for this pair of retrievals is equal to 2 divided by 10, which

Table 3.13: Queries Used in Each Session, When Varying the *authentication status* Factor, for Google

Topic \ Session	S1	S2	S3	S4	S5
T1	10a	-	-	-	-
T2	-	-	-	10a	-
T3	-	10a	-	-	-
T4	-	-	10a	-	-
T5	-	10a	-	-	-
T6	-	-	-	10a	-
T7	10a	-	-	-	-
T8	-	-	10a	-	-
T9	-	-	-	-	10a
T10	-	-	-	-	10a
Total	20	20	20	20	20
N°. Users	11	8	8	8	7
Date	07-05-2021	07-05-2021	11-05-2021	14-05-2021	14-05-2021

equals 20%, which indicates there is a 20% difference between the URLs of retrieval 1 and the URLs of retrieval 2 (for the top 10 results). This metric only studies changes in search results caused by insertion and removal of results from the *top k* results of two retrievals; it doesn't take into consideration changes caused by shifts in the ranking positions of common results of two retrievals. The latter will be analyzed by the second metric.

Table 3.14: Simple Example of New URLs Introduced from One Retrieval to Another

Retrieval 1		Retrieval 2	
Rank	URL	Rank	URL
1	Link a	1	Link a
2	Link b	2	Link b
3	Link c	3	Link c
4	Link d	4	Link d
5	Link e	5	Link k*
6	Link f	6	Link f
7	Link g	7	Link e
8	Link h*	8	Link g
9	Link i*	9	Link m*
10	Link j	10	Link j

The second metric is called **Rank Movements Per Rank Position** and it refers to how many rank positions each common URL moves, on average, from one retrieval to another. Essentially, it calculates the sum of rank movements of each common URL inside the *top k* results and divides

that sum by the number of total URLs. If we take the example shown in Table 3.14, we can see that, out of the 8 common links, there are two, *link e* and *link g*, whose rankings moved positions from one retrieval to the other (all other links didn't experience any rank movements). The ranking of *link e* moved 2 positions and the ranking of *link g* moved 1 position. Because we are analyzing the top 10 results in this example, the value of the metric for this pair of retrievals is equal to 2+1 divided by 10, which equals 0.3, which indicates there are 0.3 rank movements per rank position (for the top 10 results).

Further details about the results analysis will be presented in Chapter 5.

Chapter 4

Data Extraction

This chapter presents how the data that resulted from the present study was collected, organized and made available. Section 4.1 describes the implementation details for each of the experiment groups. Section 4.2 describes the datasets generated using the data gathered in all the experiments performed and its analysis.

4.1 Implementation Details

This section explains the implementation details for each experiment group (*Exp-APIs*, *Exp-Manual*, and *Exp-RealUsers*).

4.1.1 Execution of the *time*, *location*, and *safe search* Experiments

When it comes to the implementation details for experiment group *Exp-APIs*, in order to perform the retrieval of the results via API, the Postman tool¹ was used. This tool allowed us to make calls to the Google and Bing APIs, using the parameter arguments to change the location and safe search factors. Table 4.1 shows the API call parameters used for the API based experiments. Furthermore, the user agent used in experiment group *Exp-APIs* was "PostmanRuntime/7.28.0", which was automatically set by Postman when executing all query calls.

Moreover, in order to store the data returned by the API calls, a script was continually running in the background which would store the retrieved search results in a specific CSV file. Two CSV files were created for each experiment, one for each search engine used. The data was stored in the CSV files always in the same format; each line in the CSV file contained: a query ID, query name, and its query topic, the search factor variables used in the retrieval (date, safe search, geolocation, privacy mode, user agent, cookies and authentication status), and the search result link and its ranking position. So, because the top 10 results were retrieved for each query call, the search results for each query generated 10 lines in the CSV file, one for each of the results/links retrieved.

¹<https://www.postman.com/product/api-client/>

Table 4.1: API Call Parameters Used in Group *Exp-APIs*

Search factor	Google API Params.		Bing API Params.	
	Geolocation	Safe Search	Geolocation	Safe Search
Time	United States (us)	off	United States (us)	off
Location	us, br, de, cn, au, eg, ng	off	us, br, de, cn, au, tr, za	off
Safe Search	United States (us)	off, on	Unites States (us)	off, on

The two scripts used to store the data returned by the API calls (one for Google and one for Bing) were made available in the institutional repository.

4.1.2 Execution of the *privacy* and *user agent* Experiments

When it comes to the experiments of group *Exp-Manual*, in order to perform the retrieval of the results manually, the Chrome browser was used, along with the Google and Bing search engine's web pages. To retrieve the results, the following process was used, for each experiment. First, we made sure every other search factor not being studied in the experiment remained the same throughout the retrieval of all the search results. This includes the safe search setting used, disabled cookies and no account authentication; also, all queries were issued at around the same time, from the same geolocation, and using the same IP network. Moreover, the user agent used in the *privacy* experiment was "Chrome 89.0.4389.90 for Windows 10" (due to it being the one containing the latest version for the Chrome browser at the time the experiment was carried out); and the privacy mode used in the *user agent* experiment was "non-private window". After this step, we manually searched the 100 queries and saved the first results page (in HTML format) returned by the search engine web page. We repeated this step for each different value of the varying search factor, and for both search engines. After storing the HTML files corresponding to each query's results, a script was created to parse the search results information in the HTML files into a CSV file, using the same format used in the experiments of group *Exp-APIs*, in order to ensure consistency during the data analysis stage of the project. Two CSV files were created for each experiment, one for each search engine used. The two scripts used (one for Google and one for Bing) to parse the HTML data into a CSV format were also made available in the repository.

4.1.3 Execution of the *cookies* and *authentication status* Experiments

When it comes to experiment group *Exp-RealUsers*, the search results were gathered by real users using a web browser along with the Google search engine's web page. The *cookies* experiment required the users to use the web browser they daily use, while the *authentication status* experiment didn't require the use of any specific browser. In order for the users to perform the retrieval of the results manually, the following process was used. First, we made sure every other search

factor not being studied in the experiment remained the same for all users throughout the retrieval of the search results. This includes the safe search setting used, the user agent used (which was "Chrome 90.0.4430.85 for Windows 10", due to it being the one containing the latest version for the Chrome browser at the time the experiment was carried out), and the privacy mode used (non-private window); also, all users issued the queries at the same time, all users were located in the same area when performing the queries, and all were using the same IP network. Moreover, in the *cookies* experiment, the users were signed out of their Google accounts during the whole process; and in the *authentication status* experiment, each user removed their existing cookies file from their browser's directory, thus disabling the influence their personal cookies could have on the results obtained. After this step, each user manually searched the 20 queries subset and saved the first results page (in HTML format) returned by the Google web page. This process was repeated in each session of both the experiments. Similarly to experiment group *Exp-Manual*, after saving the HTML files corresponding to each query's results and delivering them to the investigators, a similar script was run to parse the search results information in the HTML files into a CSV file, using the same format used in the experiments of groups *Exp-APIs* and *Exp-Manual*, in order to ensure consistency during the data analysis stage of the project. One CSV file was created for each experiment.

4.2 Datasets Obtained

After collecting all the experiments data, a script was created using the R software to perform analysis to the data and to create datasets (in CSV format) containing all the data and its analysis results in an organized and consistent format. The datasets generated are described below.

Seven dataset files were created, detailing the results of the analysis performed for each one of the seven experiments. Each experiment's dataset contains, for each query, one line for each distinct pair of retrievals analyzed (the two values of the varying search factor compared) and the results of both metrics for that pair, for each search engine used, and for each of the top results analyzed (top 10, 5, 3, or 1), along with the query's name and topic.

Two auxiliary datasets were also created for elucidating about some details regarding the user agents used and the characteristics of each user that participated in the *cookies* and *authentication* experiments. The first dataset describes all the user agents used in the *user agent* experiment, stating their corresponding operating system, browser, browser version, and complete description. The second dataset describes each real user who took part in the study, containing each user's user ID, gender, age, and whether or not he frequently uses the google search engine for web searching.

Furthermore, datasets containing all the search results information gathered in each experiment were also made available. Each experiment's dataset contains, for each query submitted, the query's ID, name and topic, the values of the search factors used in the search (date, geolocation, safe search setting, privacy, user agent, cookies and authentication status), and the URL of each search result obtained, along with its ranking position.

All the mentioned datasets are available to the community in the repository associated with this study [\[40\]](#).

Chapter 5

Brief Analysis

This chapter presents a brief analysis of the results obtained from the present study. The analysis presented not only allows for some conclusions to be drawn about the impact of the several search factors studied on the mutability of search results, but also provides a proof of concept that the data gathered can be used for further and deeper analyses. Section 5.1 explains the strategy used to analyze the results of the data gathered, Sections 5.2 thru 5.8 present the results of each experiment's gathered data, as well as some discussion about the results, and Section 5.9 shows an overall discussion of the main findings of the analysis.

5.1 Analysis Strategy

The strategy used to analyze the volatility of the search results in each experiment will be the same, as will be explained next. Considering every retrieval of the query set performed in the experiment (one for each value of the varying search factor), the differences in the search results (which are expressed by the two metrics defined earlier — *% New URL* and *Rank Movements Per Rank Position*) were calculated for each query, for each distinct pair of retrievals. Afterwards, the final value of each metric equals the average of all values of that metric calculated for all queries, for all distinct pairs of retrievals. So, for example, considering the safe search experiment, as there were only two retrievals of results (one with safe search enabled and one with safe search disabled), there was only one distinct pair of retrievals to analyze; hence, the values of the metrics for each query correspond to the values of the metrics calculated for this single pair of retrievals. On the other hand, when it comes to the location experiment, as there were 7 retrievals of results (one using *location 1*, one using *location 2*, ..., and one using *location 7*), there were a total of 21 distinct pairs of retrievals to analyze (the pair of retrievals using *location 1* and *location 2*, the pair of retrievals using *location 1* and *location 3*, ..., and the pair of retrievals using *location 6* and *location 7*); hence, for each query, the final value of each metric corresponds to the average of all values of that metric calculated for all 21 pairs of retrievals. All the analyses were executed using

the R software. The R script developed to execute this project's analyses was also made available to the community.

The results of this analysis allowed us to draw some conclusions about the impact of each search factor on the mutability of search results. We performed the analysis for the top 10, top 5, top 3 and top 1 search results. We also compared the analysis' results for Google and Bing, in order to understand how different the amount of variation in search results is, for each search factor, between both search engines. We also explore the differences in the results obtained between different query topics.

In order to present the obtained results, graphs detailing the values of the *% New Url* and *Rank Movements Per Rank Position* metrics will be used. Some graphs will present the differences between the values of these metrics obtained for top 10, top 5, top 3 and top 1 results, while some other graphs will present the differences that exist between the values obtained for each search engine. To present the results regarding query topics, graphs will also be used to compare the values of each metric across different query topics.

5.2 Impact of Time

This section presents the analysis of the *time* experiment's data, expressed by each metric used. The data was analyzed in two different ways. First, we explored how much the results of each retrieval changed when compared to the first retrieval, by calculating the *% New URL* and *Rank Movements Per Rank Position* metrics between the first retrieval and each retrieval after the first one. Secondly, we explored how much the results of each retrieval changed when compared to the previous retrieval, by calculating the two metrics for each pair of consecutive retrievals. These analyses were performed for the top 10, top 5, top 3 and top 1 search results, for both Google and Bing.

% New URL

5.2.0.1 General Analysis for Google

The plot in Fig 5.1 shows the values of the *% New URL* metric calculated between the first retrieval and each subsequent retrieval, for the top 10, top 5, top 3 and top 1 search results, for the Google search engine. We can see that the percentage of new URLs increases as the retrievals compared become further apart in time, regardless of examining top 10, 5, 3 or 1. This means that time causes changes in the search results obtained, in such a way that, as the window of time between retrievals becomes longer, the volatility of the results becomes larger. We can also see that the percentage of new URLs is lower when analyzing the top 5, top 3 and top 1 results. This indicates that there are less insertions/deletions of URLs in the higher ranking positions.

Fig 5.2 shows the values of the metric calculated for each pair of consecutive retrievals, for the top 10, 5, 3 and 1, for Google. We can see that that, despite some fluctuation between the several pairs of retrievals, the percentage of new URLs doesn't show an overall growth or decline.

Furthermore, we can see once more that the metric's values are overall lower when analyzing the top 5, top 3 and top 1 results, consolidating the notion that search results vary less in the higher ranking positions.

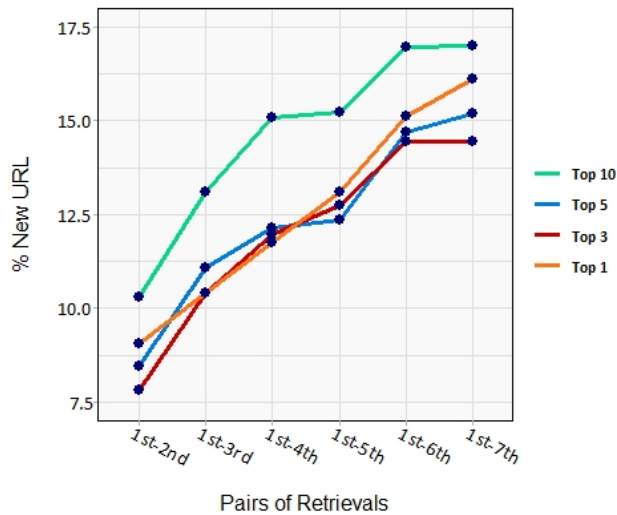


Figure 5.1: Graph Showing the % *New URL* Metric Results Between the First Retrieval and Each Subsequent Retrieval, for the Top 10, 5, 3 and 1, for Google (*time Exp.*)

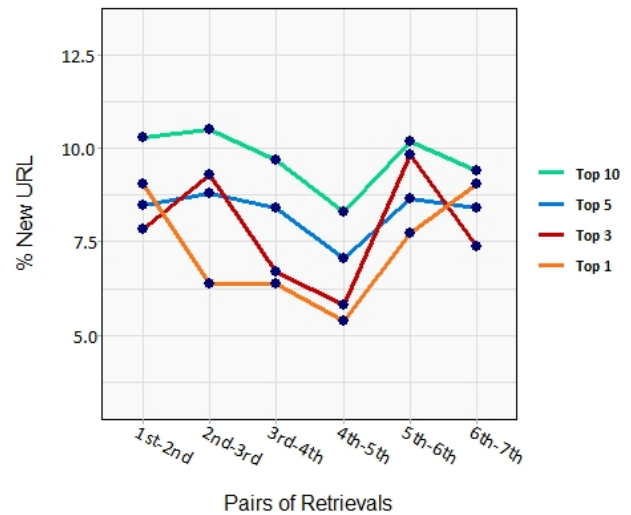


Figure 5.2: Graph Showing the % *New URL* Metric Results for Each Pair of Consecutive Retrievals, for the Top 10, 5, 3 and 1, for Google (*time Exp.*)

5.2.0.2 General Analysis for Bing

Fig 5.3 presents the values of the % *New URL* metric calculated between the first retrieval and each subsequent retrieval, for the top 10, top 5, top 3 and top 1 search results, for the Bing search engine. Similarly to Google's analysis, the percentage of new URLs generally increases as the retrievals compared become further apart in time, for all tops of results. It is also very apparent that the percentage of new URLs becomes lower as the top search results analyzed become higher, further supporting the conclusion that there are less insertions/deletions of URLs in the higher ranking positions.

In Fig 5.4 is shown the values of the metric calculated for each pair of consecutive retrievals, for the top 10, 5, 3 and 1, for Bing. Similarly to Google's analysis, we can see that there's some oscillation between the values of some pairs of retrievals, but the percentage of new URLs doesn't show an overall growth or decline. The fact that the values of the metric are lower when analyzing the higher top results remains apparent.

5.2.0.3 Search Engine Comparison

The plot in Fig. 5.5 shows the values of the metric calculated between the first and each subsequent retrieval, for both Google and Bing, for the top 10 results. There's clearly a noticeable difference between Google and Bing, with the latter presenting larger metric values than those of Google. The

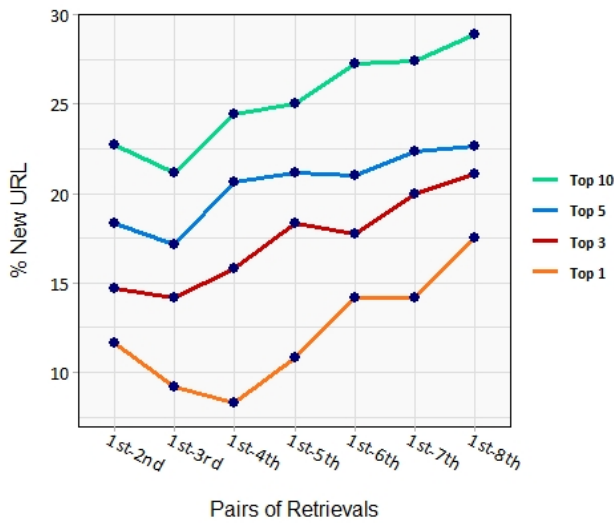


Figure 5.3: Graph Showing the % New URL Metric Results Between the First Retrieval and Each Subsequent Retrieval, for the Top 10, 5, 3 and 1, for Bing (time Exp.)

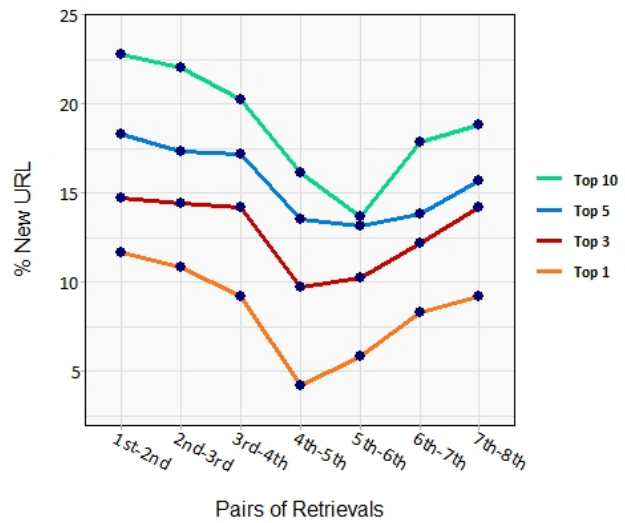


Figure 5.4: Graph Showing the % New URL Metric Results for Each Pair of Consecutive Retrievals, for the Top 10, 5, 3 and 1, for Bing (time Exp.)

percentage of new URLs peaked at 17% for Google and at nearly 29% for Bing, when calculated between the first and last retrievals.

Fig. 5.6 presents the metric’s values calculated for each pair of consecutive retrievals, for both Google and Bing, for the top 10 results. It’s clear there is still some perceptible difference between Google and Bing. Bing’s values are larger once again, peaking at 22.8%, while Google’s results peaked at 10.5%.

Rank Movements Per Rank Position

5.2.0.4 General Analysis for Google

The plot in Fig 5.7 shows the values of the Rank Movements Per Rank Position metric calculated between the first retrieval and each subsequent retrieval, for the top 10, top 5, and top 3 search results, for Google. Similarly to the previous metric’s results, the amount of rank movements per rank position also increases as the retrievals compared become further apart in time (for all tops of search results). The values of the metric are also lower when analyzing the top 5 results, and even lower for the top 3 results. This indicates that there are less rank shifts in the results with higher ranking positions. The value of this metric for the top 1 search results isn’t shown for simplicity purposes, as it will always be 0 (it isn’t possible for a common URL to move rankings and remain in the top 1 results; if a common URL shifts a single position in its ranking, say, from 1st to 2nd, it will no longer be part of the top 1 results, thus being counted as an insertion/deletion rather than a rank movement). This metric’s results for the top 1 will be hidden in the remaining analyses of this metric, in all experiments.

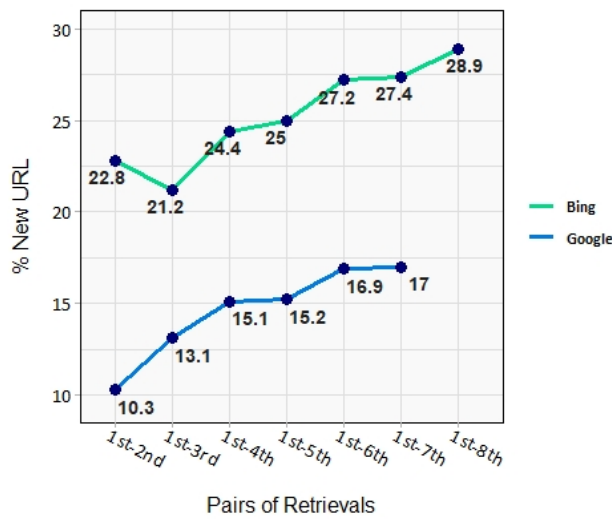


Figure 5.5: Graph Showing the % New URL Metric Results Between the First Retrieval and Each Subsequent Retrieval, for the Top 10, for Both Google and Bing (time Exp.)

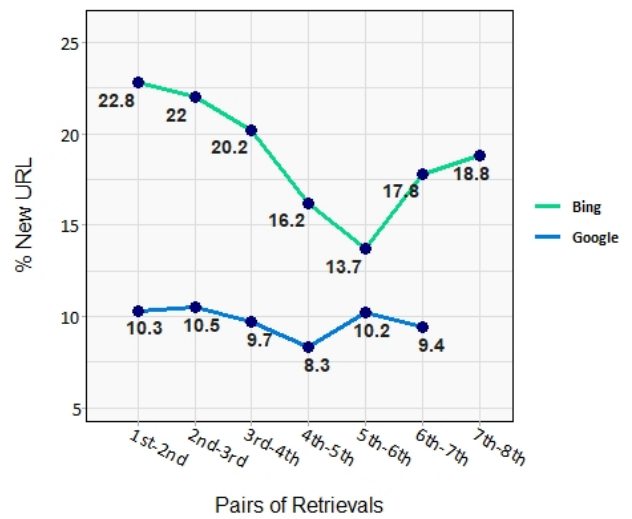


Figure 5.6: Graph Showing the % New URL Metric Results for Each Pair of Consecutive Retrievals, for the Top 10, for Both Google and Bing (time Exp.)

Fig 5.8 shows the values of the metric calculated for each pair of consecutive retrievals, for the top 10, 5, and 3, for Google. This analysis' conclusions are once again similar to those of the previous metric, as the amount of rank movements per rank position doesn't show an overall growth or decline, despite there being some fluctuations. The values of the metric are also lower for the top 5 and top 3 results.

5.2.0.5 General Analysis for Bing

Fig 5.9 presents the values of the Rank Movements Per Rank Position metric calculated between the first retrieval and each subsequent retrieval, for the top 10, top 5, and top 3 search results, for the Bing search engine. The analysis' conclusions are similar to Google's.

In Fig 5.10 is shown the values of the metric calculated for each pair of consecutive retrievals, for the top 10, 5, 3 and 1, for Bing. Similarly to Google's analysis, the amount of rank movements per rank position doesn't show an overall growth or decline, even though there is some noticeable oscillation between the values of some pairs of retrievals. The values of the metric are again lower for the top 5 and top 3 results, consolidating the notion that there are less rank shifts in the results with higher ranking positions.

5.2.0.6 Search Engine Comparison

The plot in Fig. 5.11 shows the values of the metric calculated between the first and each subsequent retrieval, for both Google and Bing, for the top 10 results. We can see that Google presents an overall larger amount of rank movements per rank position when compared to Bing, which seems to increase more rapidly with time than Bing's. The amount of rank movements per rank

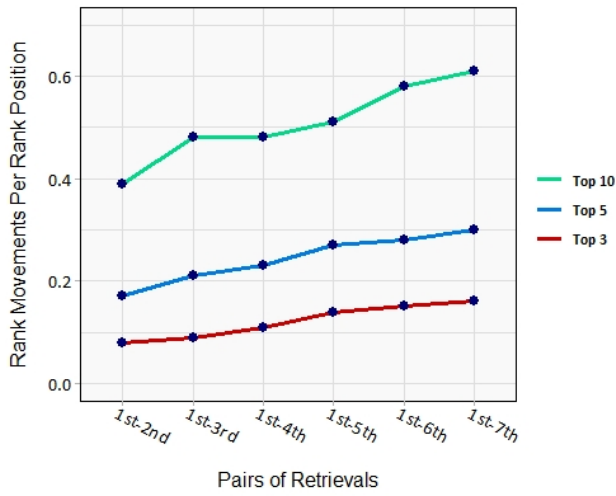


Figure 5.7: Graph Showing the *Rank Movements Per Rank Position* Metric Results Between the First Retrieval and Each Subsequent Retrieval, for the Top 10, 5, and 3, for Google (*time Exp.*)

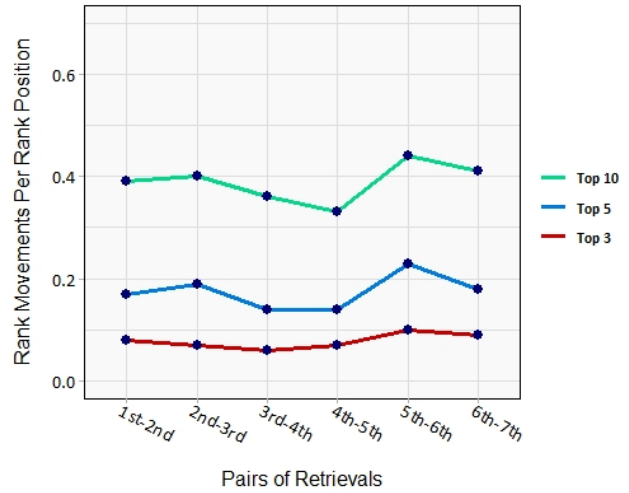


Figure 5.8: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Each Pair of Consecutive Retrievals, for the Top 10, 5, and 3, for Google (*time Exp.*)

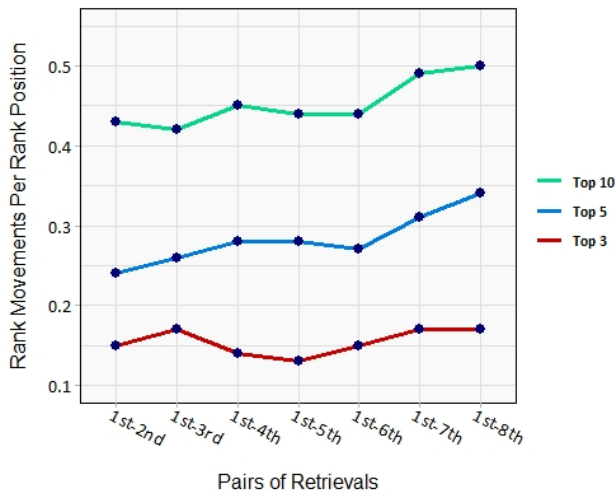


Figure 5.9: Graph Showing the *Rank Movements Per Rank Position* Metric Results Between the First Retrieval and Each Subsequent Retrieval, for the Top 10, 5 and 3, for Bing (*time Exp.*)

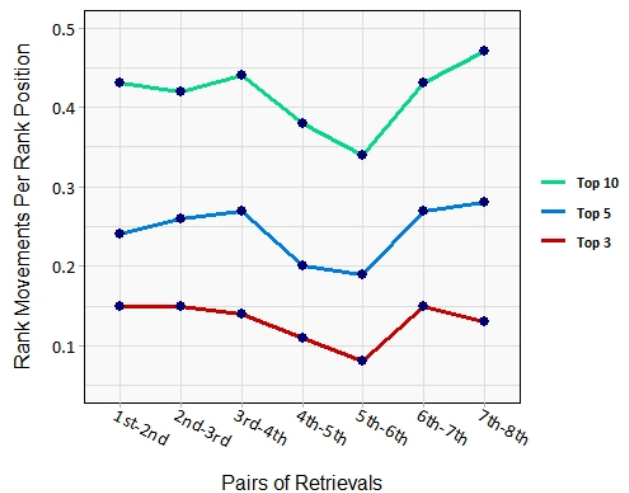


Figure 5.10: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Each Pair of Consecutive Retrievals, for the Top 10, 5, and 3, for Bing (*time Exp.*)

position peaked at 0.61 for Google and at 0.5 for Bing, when calculated between the first and last retrievals.

Fig. 5.12 presents the metric's values calculated for each pair of consecutive retrievals, for both Google and Bing, for the top 10 results. Bing's values are, for the most part, larger than Google's, even though the difference between both search engine's metric values isn't very large; Bing's results peaked at 0.47, while Google's results peaked at 0.44.

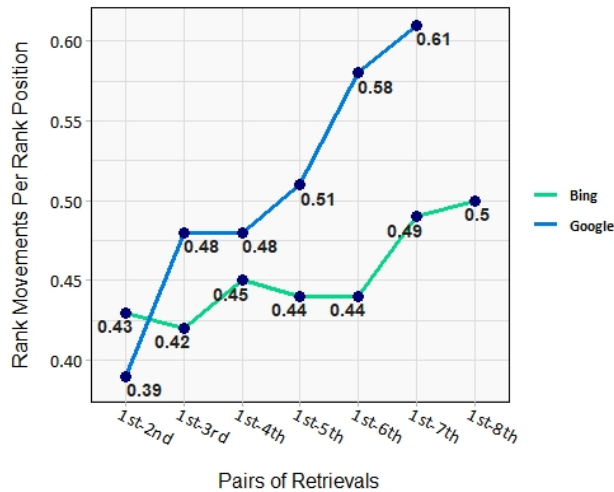


Figure 5.11: Graph Showing the *Rank Movements Per Rank Position* Metric Results Between the First Retrieval and Each Subsequent Retrieval, for the Top 10, for Both Google and Bing (*time Exp.*)

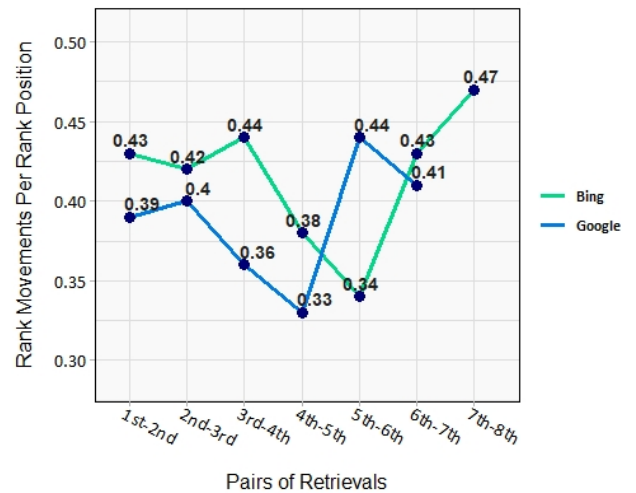


Figure 5.12: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Each Pair of Consecutive Retrievals, for the Top 10, for Both Google and Bing (*time Exp.*)

5.3 Impact of Location

This section presents the analysis of the *location* experiment's data, expressed by each metric used. First, we present the results obtained for each search engine, for the top 10, top 5, top 3 and top 1 search results. Afterwards, we compare the results obtained between Google and Bing, for the top 10 search results. Finally, we calculate both metrics for each set of queries belonging to the same query topic, for the top 10 search results, in order to attempt to evaluate if the search results' mutability is influenced by the query topic.

% New URL

5.3.0.1 General Analysis for Google

In Fig. 5.13 we can see the results of the *% New URL* metric for the top 10, top 5, top 3, and top 1 search results, using the Google search engine. Each box plot represents the aggregation of the metric values for all queries, and each query's metric value corresponds to the average of the

metric values calculated for all 21 distinct pairs of retrievals (pair using *location 1* and *location 2*, pair using *location 1* and *location 3*, etc...). The red square inside each box plot indicates the mean value of the metric. We can see that there is close to no variation between the results of top 10, top 5, top 3 and even top 1 (even though the latter presents a larger spread in the results, which is explained by the fact that the possible values of this metric for each pair of retrievals, for the top 1 results, are either 0% or 100%). This means that the percentage of new URLs that are inserted in the top 5, top 3 and top 1 is similar to the percentage of new URLs inserted in the top 10, when varying the location search factor.

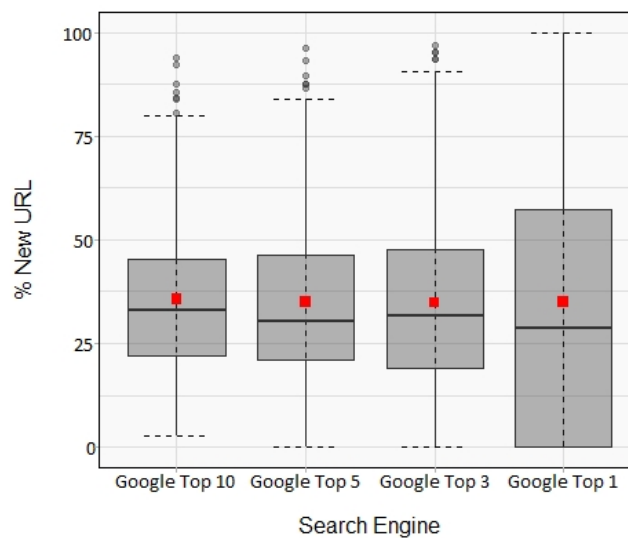


Figure 5.13: Graph Showing the % *New URL* Metric Results for Google, for the Top 10, 5, 3 and 1 (*location Exp.*)

5.3.0.2 General Analysis for Bing

In Fig. 5.14 we can see the results of the % *New URL* metric for the top 10, top 5, top 3 and top 1 search results, using the Bing search engine. We can see that there is some slight variation between the results of top 10, top 5, top 3 and top 1 (with the latter presenting a much larger spread again), as the mean value of the metric decreases as the top rankings analyzed become higher.

5.3.0.3 Search Engine Comparison

Fig. 5.15 shows the results of the metric, for the top 10 results, for both search engines. We can see that Bing's results (53.7% mean value) are larger than Google's (35.7% mean value). This means there is a difference of 18% between both search engine's percentage of URLs inserted/deleted from the top 10.

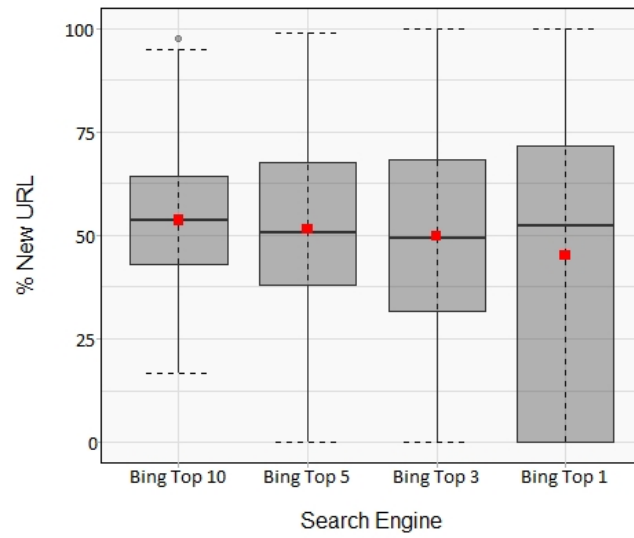


Figure 5.14: Graph Showing the % *New URL* Metric Results for Bing, for the Top 10, 5, 3 and 1 (*location Exp.*)

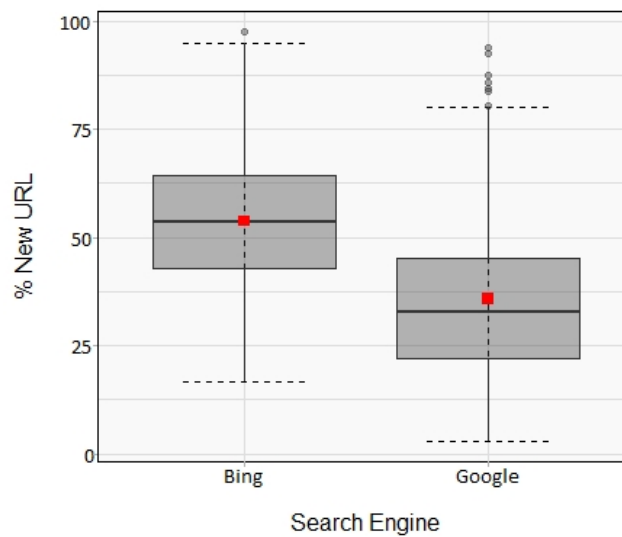


Figure 5.15: Graph Showing the % *New URL* Metric Results, for the Top 10, for Google and Bing (*location Exp.*)

5.3.0.4 Search Topic Comparison

Looking at Fig. 5.16, we can see the value of the *% New URL* metric for each query topic analyzed, for the top 10 search results, for both Google and Bing. It's clear that there are differences between the values of different query topics, for both search engines. We can see that, when it comes to Google, the difference between the values for "Computers & Electronics" and "News" (the largest and lowest values, respectively) is nearly **15%**. As for Bing, approximately the same difference can be seen, but the query topics with the largest and lowest values are different ("Travel" and "Business & Industrial", respectively).

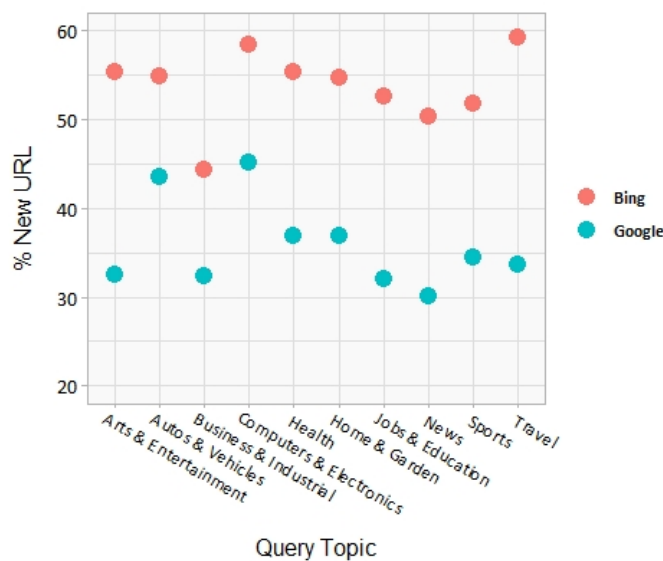


Figure 5.16: Graph Showing the *% New URL* Metric Results for Each Query Topic, for the Top 10, for Google and Bing (*location Exp.*)

Rank Movements Per Rank Position

5.3.0.5 General Analysis for Google

In Fig. 5.17 we can see the results of the *Rank Movements Per Rank Position* metric for the top 10, top 5 and top 3 search results, using the Google search engine. We can see that the amount of rank movements per rank position decreases for the analysis of the top 5, and decreases even more in the analysis of the top 3 results. This indicates that there are less rank shifts in the results with higher ranking positions.

5.3.0.6 General Analysis for Bing

In Fig. 5.18 we can see the results of the *Rank Movements Per Rank Position* metric for the top 10, top 5 and top 3 search results, using the Bing search engine. The results are similar to Google's, with the amount of rank movements per rank position decreasing for the analysis of the top 5, and decreasing even further for the analysis of the top 3 results.

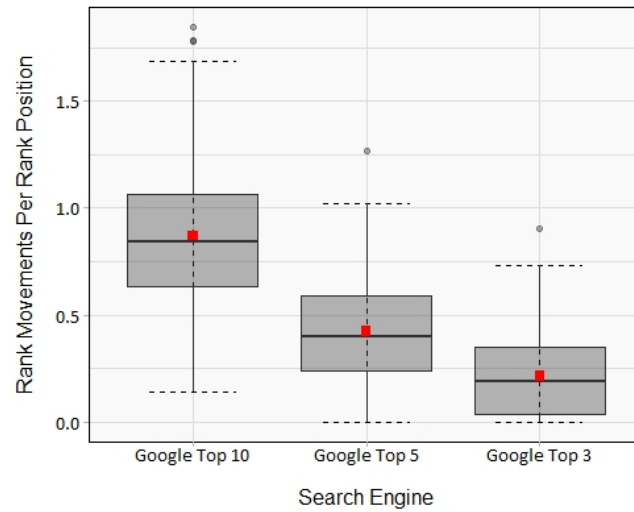


Figure 5.17: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Google, for the Top 10, 5, and 3 (*location Exp.*)

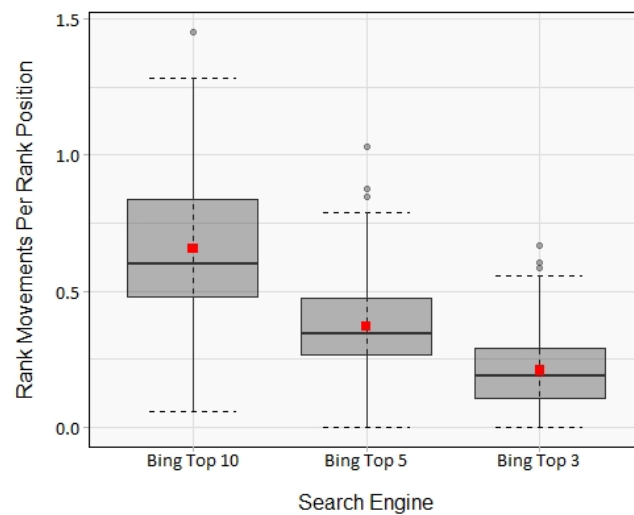


Figure 5.18: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Bing, for the Top 10, 5, and 3 (*location Exp.*)

5.3.0.7 Search Engine Comparison

Fig. 5.19 shows the obtained results of the metric, for the top 10 results, for both Google and Bing. It is clear that Google's results (**0.87** mean value) are larger than Bing's (**0.66** mean value). This could be explained by the fact that Google presents a lower value for the % New URL metric for the top 10 results (which indicates there are less insertions/deletions of URLs), meaning there is a larger amount of common results in the top 10 that have a chance to shift ranking positions.

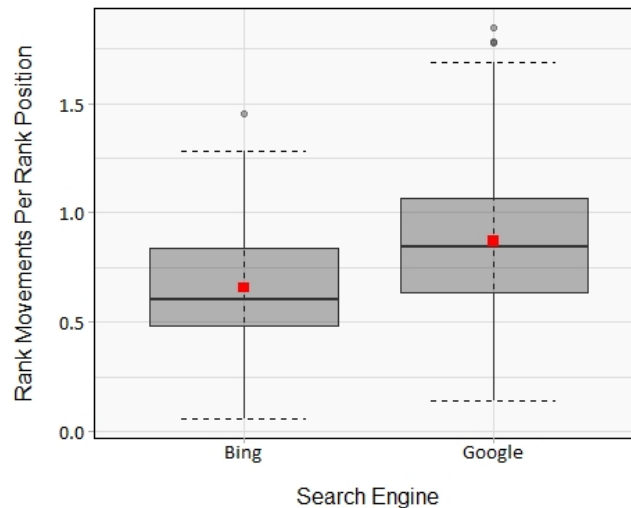


Figure 5.19: Graph Showing the *Rank Movements Per Rank Position* Metric Results, for the Top 10, for Google and Bing (*location Exp.*)

5.3.0.8 Search Topic Comparison

Fig. 5.20 presents the value of the *Rank Movements Per Rank Position* metric for each query topic analyzed, for the top 10 search results, for both Google and Bing. We can see that, for Google, the difference between the largest and lowest values ("Home & Garden" and "Business & Industrial", respectively) is nearly **0.17**, while for Bing the difference is larger (**0.24**) and the query topics with the largest and lowest values are different ("Sports" and "Travel", respectively)

5.4 Impact of Safe Search

The analysis of the *safe search* experiment's data is presented in this section. For each metric, the analysis' results will be presented using the same structure as the previous experiment.

% New URL

5.4.0.1 General Analysis for Google

In Fig. 5.21 we can see the results of the % *New URL* metric for the top 10, top 5, top 3, and top 1 search results, using the Google search engine. We can see that the results of the metric for all

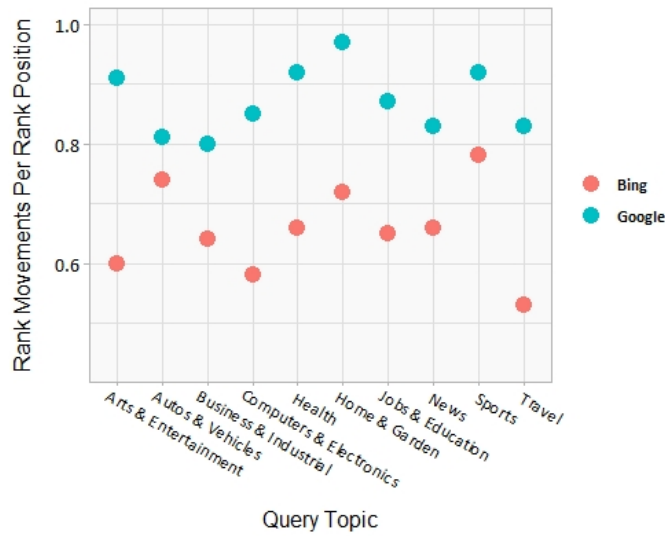


Figure 5.20: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Each Query Topic, for the Top 10, for Google and Bing (*location Exp.*)

tops of results are very low and present close to no variation between them. This indicates that varying the safe search factor doesn't cause nearly any insertion/deletion of new URLs in the top 10, 5, 3 and 1 results, for Google.

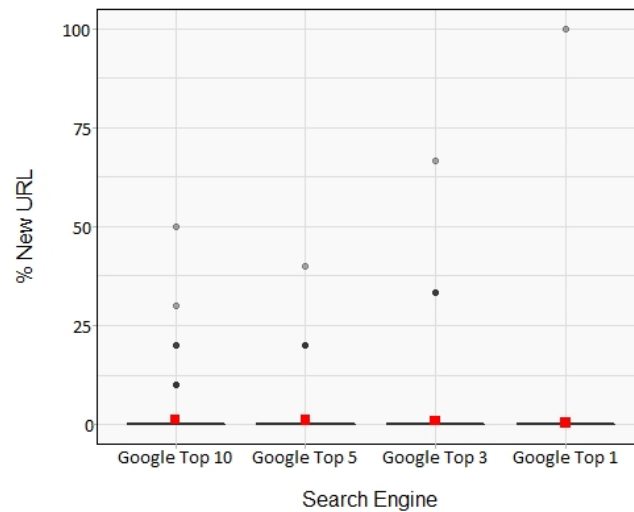


Figure 5.21: Graph Showing the *% New URL* Metric Results for Google, for the Top 10, 5, 3 and 1 (*safe search Exp.*)

5.4.0.2 General Analysis for Bing

In Fig. 5.22 we can see the results of the *% New URL* metric for the top 10, top 5, top 3 and top 1 search results, using the Bing search engine. We can see that there is some variation between

the results of top 10, top 5, top 3 and top 1, as the mean value of the metric decreases as the top rankings analyzed become higher.

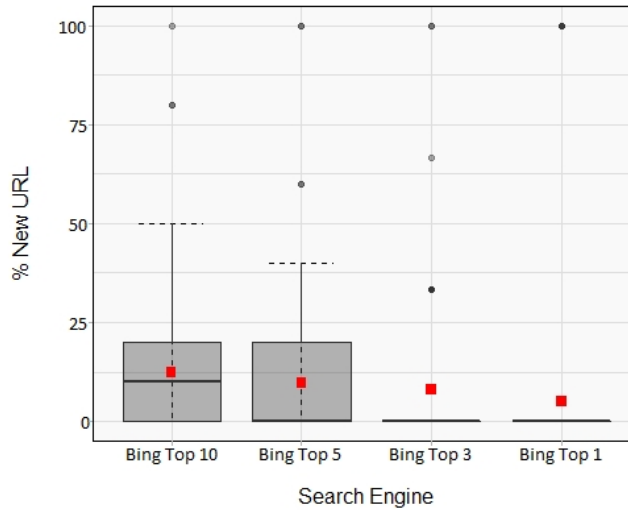


Figure 5.22: Graph Showing the % *New URL* Metric Results for Bing, for the Top 10, 5, 3 and 1 (*safe search Exp.*)

5.4.0.3 Search Engine Comparison

Fig. 5.23 shows the results of the metric, for the top 10 results, for both search engines. It's clear that Bing's results (12.4% mean value) are larger than Google's (1.2% mean value), with a difference of over 11% between both search engine's percentage of URLs inserted/deleted from the top 10.

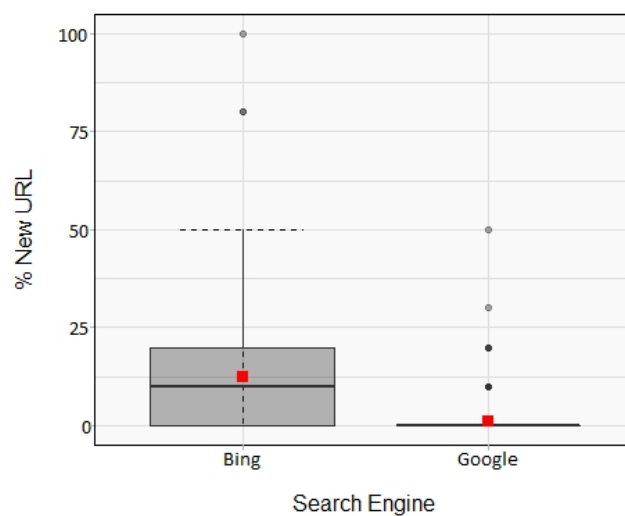


Figure 5.23: Graph Showing the % *New URL* Metric Results, for the Top 10, for Google and Bing (*safe search Exp.*)

5.4.0.4 Search Topic Comparison

Looking at Fig. 5.24, we can see the value of the *% New URL* metric for each query topic analyzed, for the top 10 search results, for both Google and Bing. We can see Bing presents larger differences between the metric's values for different query topics than Google. For Bing, the difference between the largest ("Health") and lowest ("Business & Industrial" and "News") values is nearly **17.5%**, while for Google this difference is just **2.9%**.

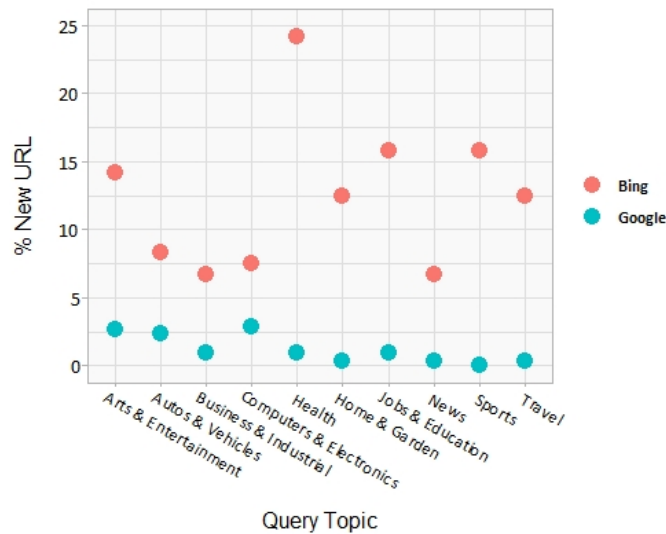


Figure 5.24: Graph Showing the *% New URL* Metric Results for Each Query Topic, for the Top 10, for Google and Bing (*safe search* Exp.)

Rank Movements Per Rank Position

5.4.0.5 General Analysis for Google

In Fig. 5.25 we can see the results of the *Rank Movements Per Rank Position* metric for the top 10, top 5 and top 3 search results, using the Google search engine. Similarly to the previous metric, the results of this metric for all tops of results are very low and present close to no variation between them. The conclusion can be drawn that the safe search factor has very little impact on the mutability of search results, when it comes to Google.

5.4.0.6 General Analysis for Bing

In Fig. 5.26 we can see the results of the *Rank Movements Per Rank Position* metric for the top 10, top 5 and top 3 search results, using the Bing search engine. Similarly to the previous metric, the mean value of the metric decreases as the top rankings analyzed become higher.

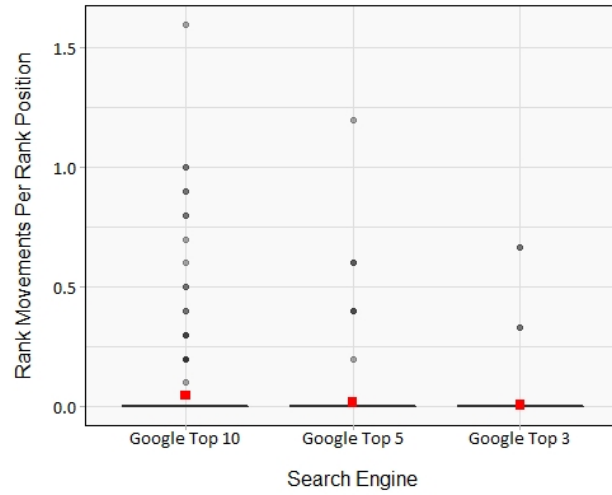


Figure 5.25: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Google, for the Top 10, 5, and 3 (*safe search Exp.*)

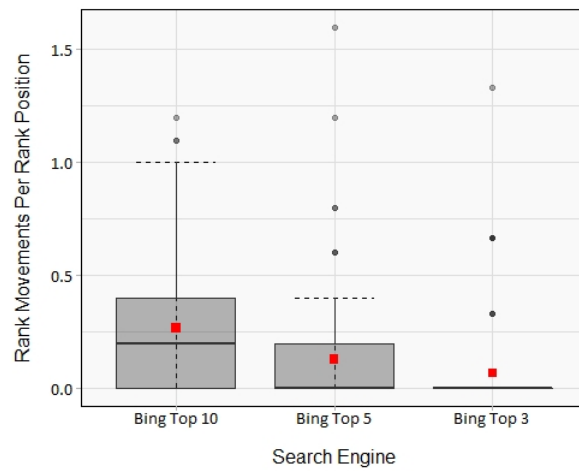


Figure 5.26: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Bing, for the Top 10, 5, and 3 (*safe search Exp.*)

5.4.0.7 Search Engine Comparison

Fig. 5.27 shows the obtained results of the metric, for the top 10 results, for both Google and Bing. Bing's results (**0.27** mean value) are once again larger than Google's (**0.05** mean value).

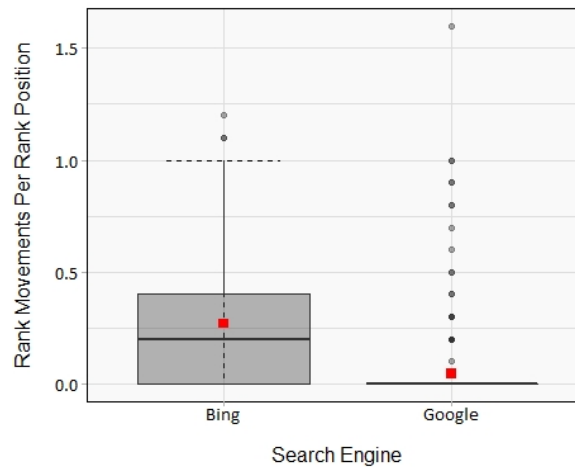


Figure 5.27: Graph Showing the *Rank Movements Per Rank Position* Metric Results, for the Top 10, for Google and Bing (*safe search Exp.*)

5.4.0.8 Search Topic Comparison

Fig. 5.28 presents the value of the *Rank Movements Per Rank Position* metric for each query topic analyzed, for the top 10 search results, for both Google and Bing. Similar to the observations made for the previous metric's results, Bing shows a larger difference between the metric's values for different query topics than Google. For Bing, the difference between the largest ("News") and lowest ("Business & Industrial") values is **0.33**, while for Google this difference is **0.09**.

5.5 Impact of Privacy Mode

This section presents the analysis of the *privacy* experiment's data for each metric, using the same structure as the previous experiment.

% New URL

5.5.0.1 General Analysis for Google

In Fig. 5.29 we can see the results of the *% New URL* metric for the top 10, top 5, top 3, and top 1 search results, using the Google search engine. We can see that the mean value of the metric increases as the top rankings analyzed become higher, indicating that varying the privacy factor causes a larger percentage of new URLs to be inserted in the results with higher ranking positions (for Google).

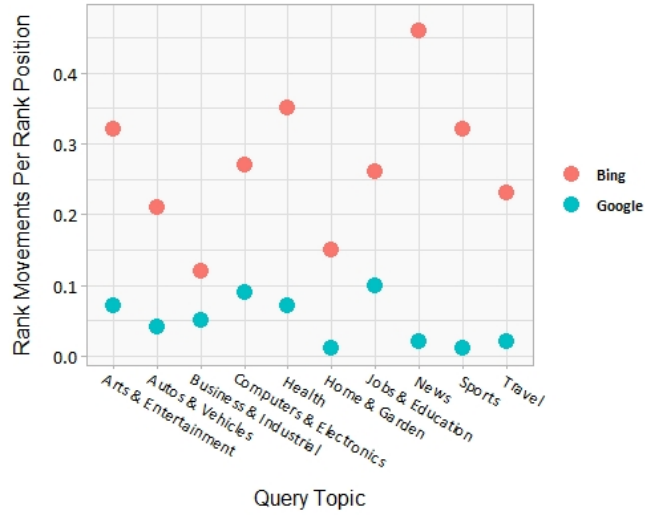


Figure 5.28: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Each Query Topic, for the Top 10, for Google and Bing (*safe search Exp.*)

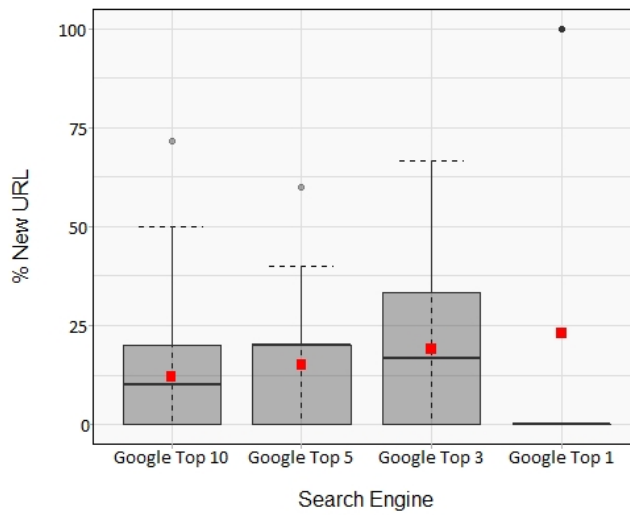


Figure 5.29: Graph Showing the *% New URL* Metric Results for Google, for the Top 10, 5, 3 and 1 (*privacy Exp.*)

5.5.0.2 General Analysis for Bing

In Fig. 5.30 we can see the results of the *% New URL* metric for the top 10, top 5, top 3 and top 1 search results, using the Bing search engine. We can see that the variation between the metric's results for top 10, 5, 3, and 1 is very slight.

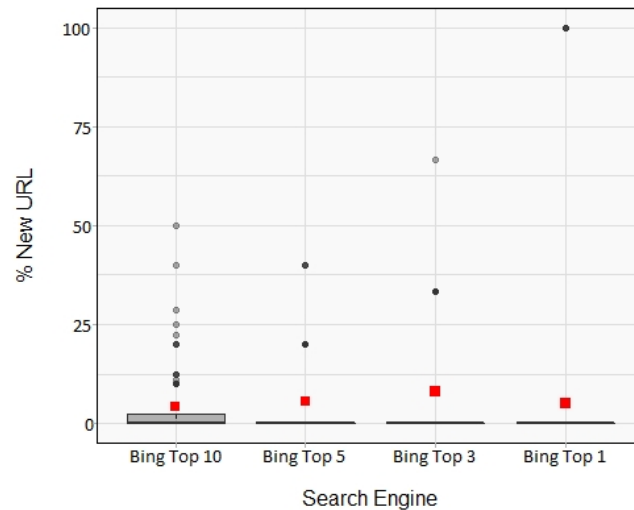


Figure 5.30: Graph Showing the *% New URL* Metric Results for Bing, for the Top 10, 5, 3 and 1 (*privacy Exp.*)

5.5.0.3 Search Engine Comparison

Fig. 5.31 shows the results of the metric, for the top 10 results, for both search engines. It's clear that Google's results (12% mean value) are larger than Bing's (4.2% mean value).

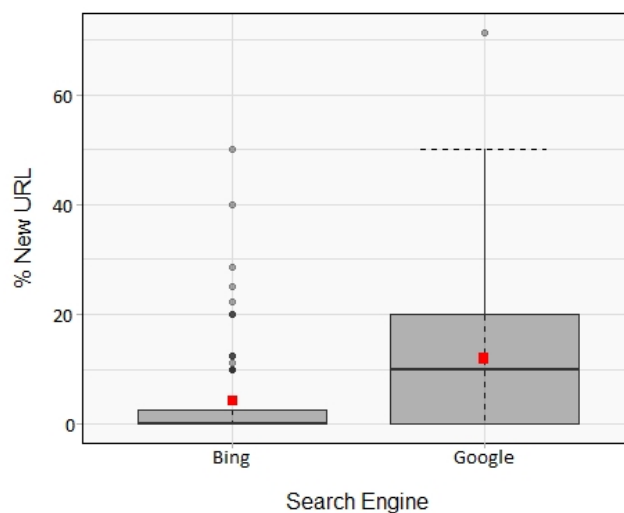


Figure 5.31: Graph Showing the *% New URL* Metric Results, for the Top 10, for Google and Bing (*privacy Exp.*)

5.5.0.4 Search Topic Comparison

Looking at Fig. 5.32, we can see the value of the *% New URL* metric for each query topic analyzed, for the top 10 search results, for both Google and Bing. We can see there are some differences between the metric's values for different query topics, for both search engines. For Google, the difference between the largest ("Home & Garden") and lowest ("Health") values is nearly **15.4%**, while for Bing this difference is **11.6%**.

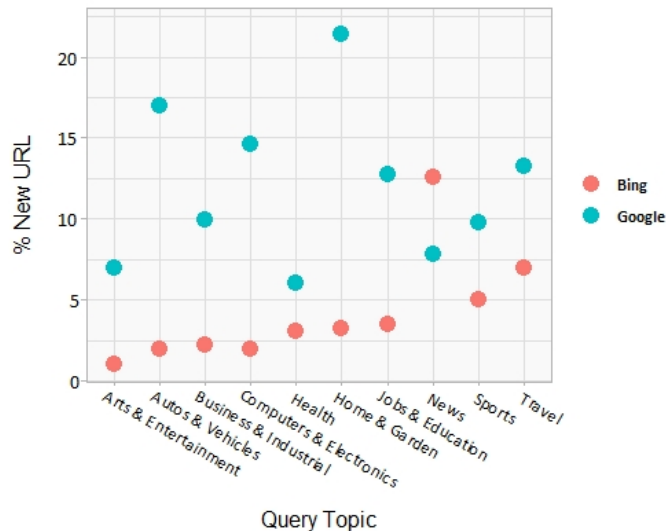


Figure 5.32: Graph Showing the *% New URL* Metric Results for Each Query Topic, for the Top 10, for Google and Bing (*privacy Exp.*)

Rank Movements Per Rank Position

5.5.0.5 General Analysis for Google

In Fig. 5.33 we can see the results of the *Rank Movements Per Rank Position* metric for the top 10, top 5 and top 3 search results, using the Google search engine. The amount of rank movements per rank position decreases as the top rankings analyzed become higher, showing that less rank shifts occur in the search results with higher ranking positions.

5.5.0.6 General Analysis for Bing

In Fig. 5.34 we can see the results of the *Rank Movements Per Rank Position* metric for the top 10, top 5 and top 3 search results, using the Bing search engine. Similarly to Google's results, the mean value of the metric decreases as the top rankings analyzed become higher.

5.5.0.7 Search Engine Comparison

Fig. 5.35 shows the obtained results of the metric, for the top 10 results, for both Google and Bing. Bing's results (**0.22** mean value) are larger than Google's (**0.08** mean value).

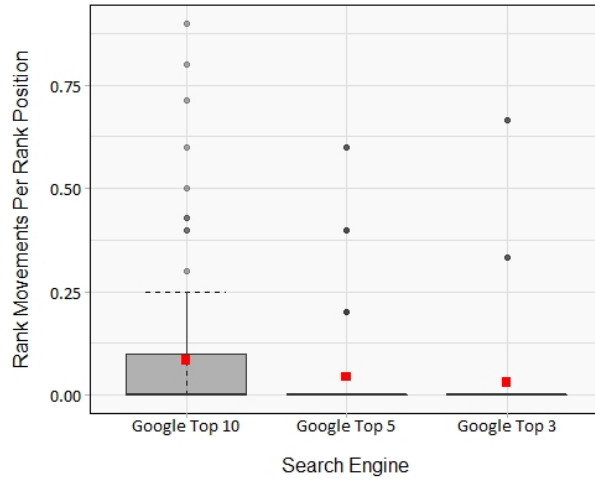


Figure 5.33: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Google, for the Top 10, 5, and 3 (*privacy Exp.*)

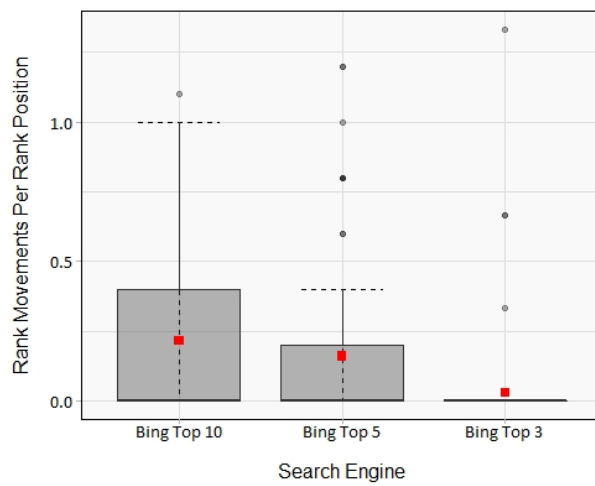


Figure 5.34: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Bing, for the Top 10, 5, and 3 (*privacy Exp.*)

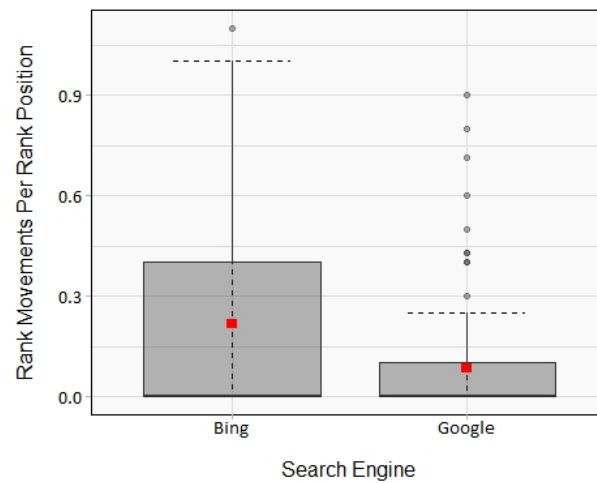


Figure 5.35: Graph Showing the *Rank Movements Per Rank Position* Metric Results, for the Top 10, for Google and Bing (*privacy Exp.*)

5.5.0.8 Search Topic Comparison

Fig. 5.36 presents the value of the *Rank Movements Per Rank Position* metric for each query topic analyzed, for the top 10 search results, for both Google and Bing. Bing shows a larger difference between the metric's values for different query topics than Google. For Bing, the difference between the largest ("Home & Garden") and lowest ("Autos & Vehicles") values is **0.51**, while for Google this difference is **0.27**.

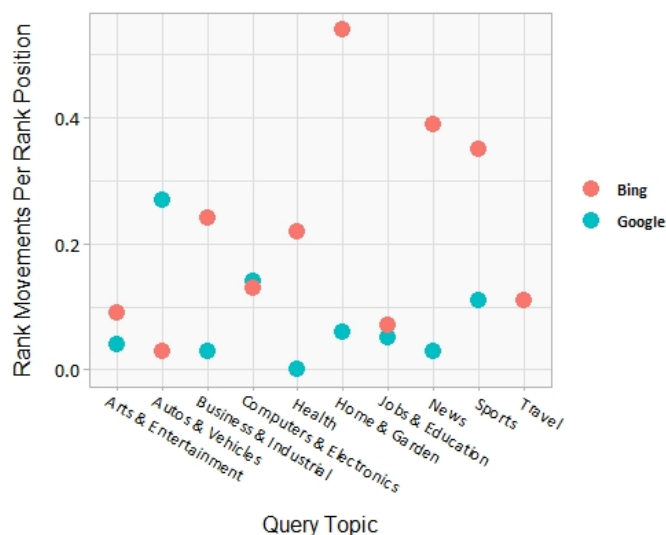


Figure 5.36: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Each Query Topic, for the Top 10, for Google and Bing (*privacy Exp.*)

5.6 Impact of User Agent

This section presents the analysis of the *user agent* experiment's data for each metric, using the same structure as the previous experiment.

% New URL

5.6.0.1 General Analysis for Google

In Fig. 5.37 we can see the results of the *% New URL* metric for the top 10, top 5, top 3, and top 1 search results, using the Google search engine. The metric's results increase as the top rankings analyzed become higher, which indicates that there's a larger percentage of new URLs inserted in the results with higher ranking positions (for Google), when varying the user agent search factor.

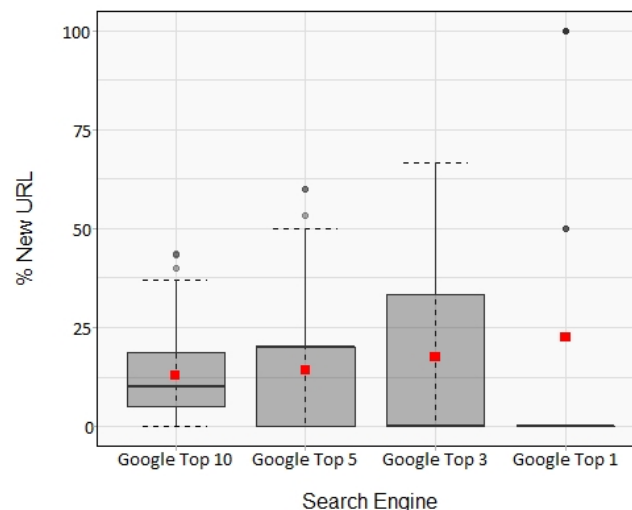


Figure 5.37: Graph Showing the *% New URL* Metric Results for Google, for the Top 10, 5, 3 and 1 (*user agent* Exp.)

5.6.0.2 General Analysis for Bing

In Fig. 5.38 we can see the results of the *% New URL* metric for the top 10, top 5, top 3 and top 1 search results, using the Bing search engine. We can see that the variation between the metric's results for top 10, 5, 3, and 1 is very slight and doesn't show an overall growth or decline.

5.6.0.3 Search Engine Comparison

Fig. 5.39 shows the results of the metric, for the top 10 results, for both search engines. Google's results (12.8% mean value) are slightly larger than Bing's results (10.1% mean value).

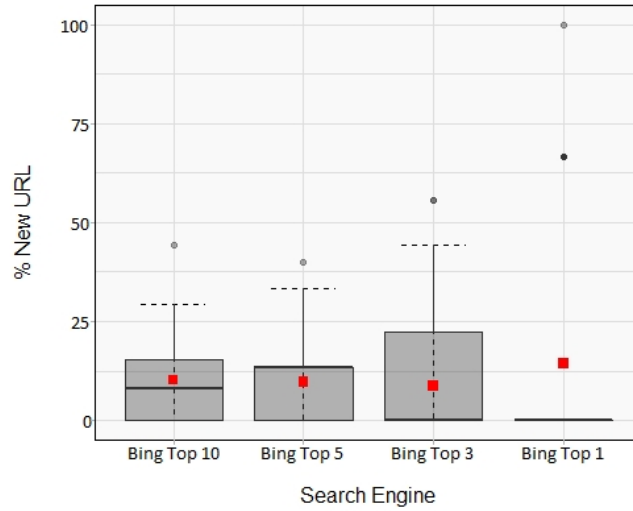


Figure 5.38: Graph Showing the % *New URL* Metric Results for Bing, for the Top 10, 5, 3 and 1 (*user agent Exp.*)

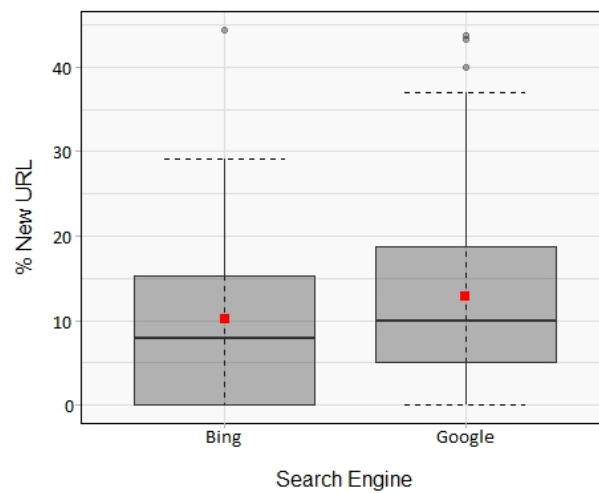


Figure 5.39: Graph Showing the % *New URL* Metric Results, for the Top 10, for Google and Bing (*user agent Exp.*)

5.6.0.4 Search Topic Comparison

Looking at Fig. 5.40, we can see the value of the *% New URL* metric for each query topic analyzed, for the top 10 search results, for both Google and Bing. We can see that the metric's values vary between different query topics, for both Google and Bing. For Google, the difference between the largest ("Home & Garden") and lowest ("Arts & Entertainment") values is nearly **13%**, while for Bing this difference is smaller (**5.4%**).

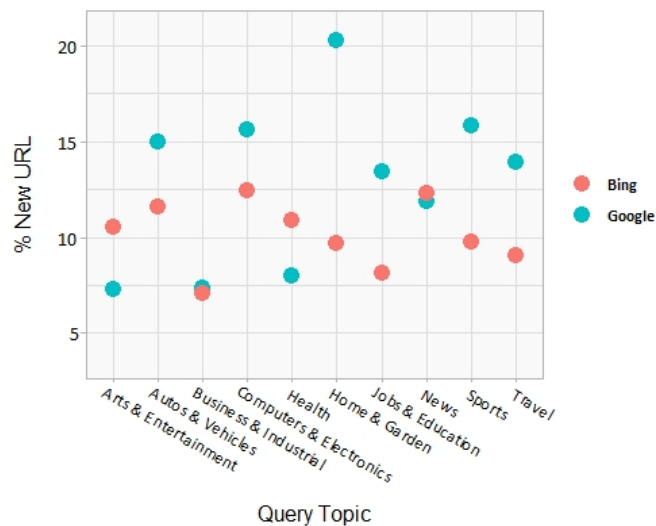


Figure 5.40: Graph Showing the *% New URL* Metric Results for Each Query Topic, for the Top 10, for Google and Bing (*user agent Exp.*)

Rank Movements Per Rank Position

5.6.0.5 General Analysis for Google

In Fig. 5.41 we can see the results of the *Rank Movements Per Rank Position* metric for the top 10, top 5 and top 3 search results, using the Google search engine. The metric's results decrease as the top rankings analyzed become higher, showing that less rank shifts occur in the search results with higher ranking positions.

5.6.0.6 General Analysis for Bing

In Fig. 5.42 we can see the results of the *Rank Movements Per Rank Position* metric for the top 10, top 5 and top 3 search results, using the Bing search engine. Similarly to Google's results, the mean value of the metric decreases as the top rankings analyzed become higher.

5.6.0.7 Search Engine Comparison

Fig. 5.43 shows the obtained results of the metric, for the top 10 results, for both Google and Bing. We can see that Bing's results (**0.35** mean value) are larger than Google's (**0.1** mean value).

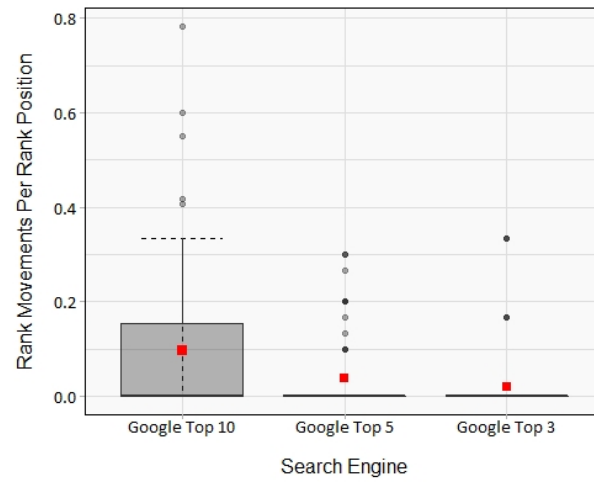


Figure 5.41: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Google, for the Top 10, 5, and 3 (*user agent Exp.*)

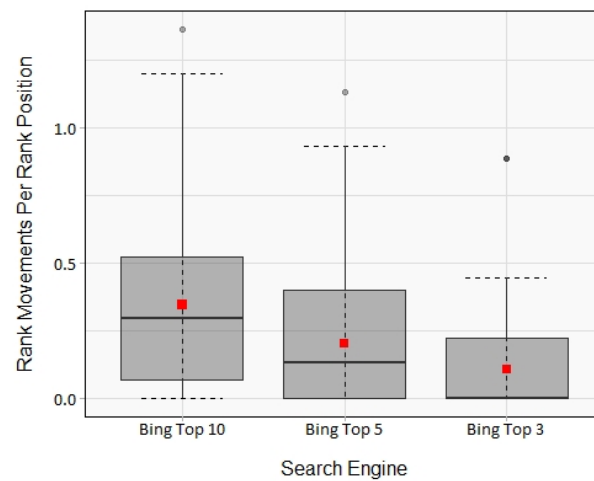


Figure 5.42: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Bing, for the Top 10, 5, and 3 (*user agent Exp.*)

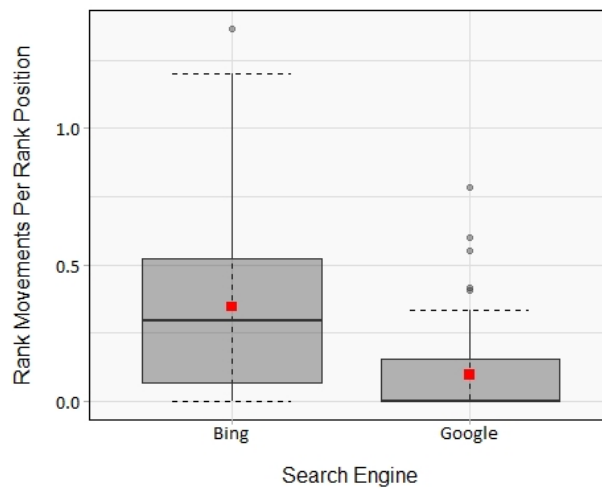


Figure 5.43: Graph Showing the *Rank Movements Per Rank Position* Metric Results, for the Top 10, for Google and Bing (*user agent Exp.*)

5.6.0.8 Search Topic Comparison

Fig. 5.44 shows the value of the *Rank Movements Per Rank Position* metric for each query topic analyzed, for the top 10 search results, for both Google and Bing. Both search engines present some differences between the metric's values for different query topics. For Bing, the difference between the largest and lowest values ("News" and "Business & Industrial", respectively) is **0.35**, while for Google this difference is **0.16**.

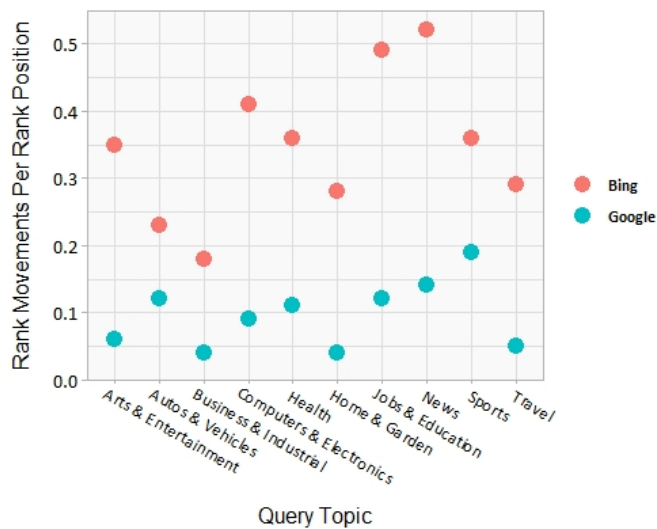


Figure 5.44: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Each Query Topic, for the Top 10, for Google and Bing (*user agent Exp.*)

5.7 Impact of Cookies

The analysis of the *cookies* experiment's data is presented in this section, expressed by each metric used. First, we present the results obtained for the Google search engine (the only one used in this experiment), for the top 10, top 5, top 3 and top 1 search results. Then, we calculate both metrics for each set of queries belonging to the same query topic, for the top 10 search results.

% New URL

5.7.0.1 General Analysis for Google

In Fig. 5.45 we can see the results of the *% New URL* metric for the top 10, top 5, top 3, and top 1 search results, using the Google search engine. There's some variation between the results of the top 10, 5, 3 and 1, as the mean value of the metric generally increases as the top rankings analyzed become higher. This indicates that there's a larger percentage of new URLs inserted in the results with higher ranking positions, when varying the cookies search factor.

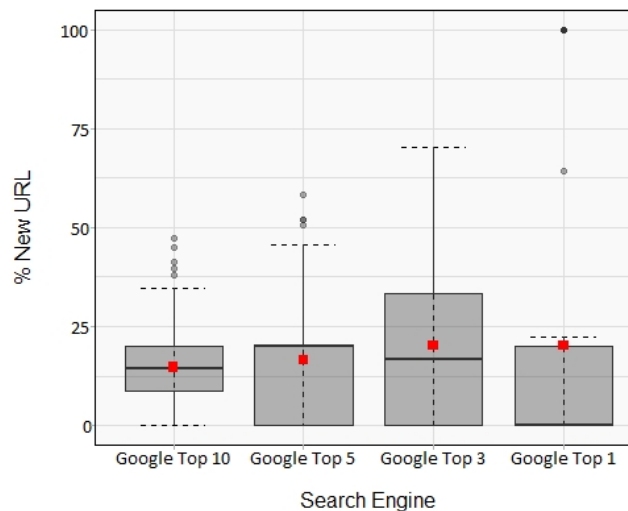


Figure 5.45: Graph Showing the *% New URL* Metric Results for Google, for the Top 10, 5, 3 and 1 (*cookies* Exp.)

5.7.0.2 Search Topic Comparison

Looking at Fig. 5.46, we can see the value of the *% New URL* metric for each query topic analyzed, for the top 10 search results, for Google. There's a large difference between the metric's values for different topics. The difference between the largest ("Home & Garden") and lowest ("Business & Industrial") values is nearly **14.6%**.

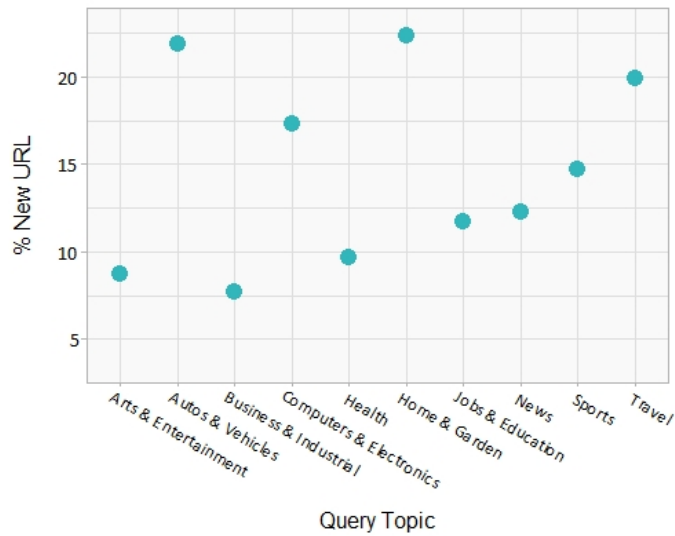


Figure 5.46: Graph Showing the % New URL Metric Results for Each Query Topic, for the Top 10, for Google (cookies Exp.)

Rank Movements Per Rank Position

5.7.0.3 General Analysis for Google

In Fig. 5.47 we can see the results of the Rank Movements Per Rank Position metric for the top 10, top 5 and top 3 search results, using the Google search engine. The amount of rank movements per rank position decreases for the analysis of the top 5, and decreases even more in the analysis of the top 3 results, indicating there are less rank shifts in the search results with higher ranking positions.

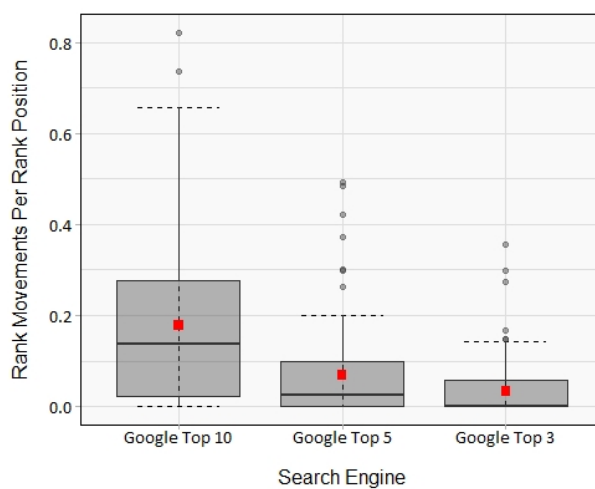


Figure 5.47: Graph Showing the Rank Movements Per Rank Position Metric Results for Google, for the Top 10, 5, and 3 (cookies Exp.)

5.7.0.4 Search Topic Comparison

Fig. 5.48 presents the value of the *Rank Movements Per Rank Position* metric for each query topic analyzed, for the top 10 search results, for Google. We can see there's some variation between the metric's values for different topics. The difference between the largest ("Autos & Vehicles") and lowest ("Jobs & Education") values is nearly **0.33**.

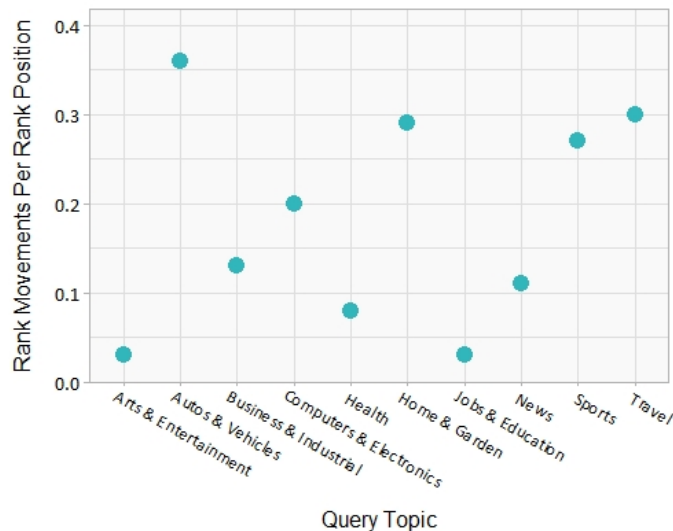


Figure 5.48: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Each Query Topic, for the Top 10, for Google (*cookies Exp.*)

5.8 Impact of Authentication Status

The analysis of the *authentication status* experiment's data is presented in this section, expressed by each metric used. Similarly to the analysis of the previous experiment, we present the results obtained for the Google search engine, for the top 10, top 5, top 3 and top 1 search results, followed by the the results obtained for each set of queries belonging to the same query topic, for the top 10 search results.

% New URL

5.8.0.1 General Analysis for Google

In Fig. 5.49 we can see the results of the *% New URL* metric for the top 10, top 5, top 3, and top 1 search results, using the Google search engine. We can see there's some variation between the results of the top 10, 5, 3 and 1, as the mean value of the metric increases as the top rankings analyzed become higher. This indicates that, when varying the authentication status search factor, there's a larger percentage of new URLs inserted in the results with higher ranking positions.

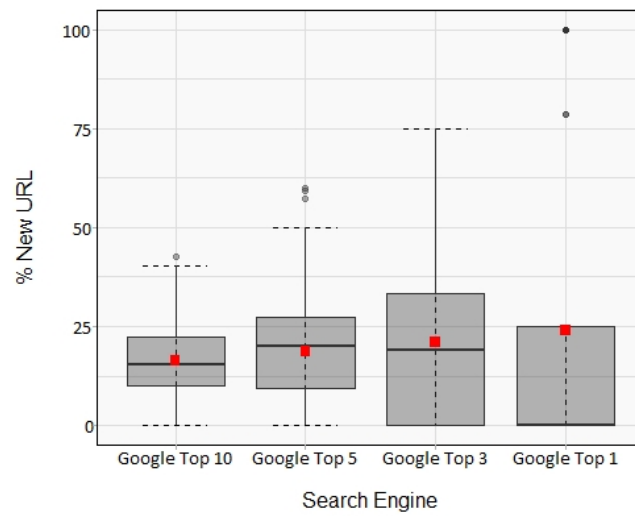


Figure 5.49: Graph Showing the % New URL Metric Results for Google, for the Top 10, 5, 3 and 1 (authentication status Exp.)

5.8.0.2 Search Topic Comparison

Looking at Fig. 5.50, we can see the value of the % New URL metric for each query topic analyzed, for the top 10 search results, for Google. The results show a perceptible difference between the metric's values for different topics. The difference between the largest ("Jobs & Education") and lowest ("Arts & Entertainment") values is nearly 10%.

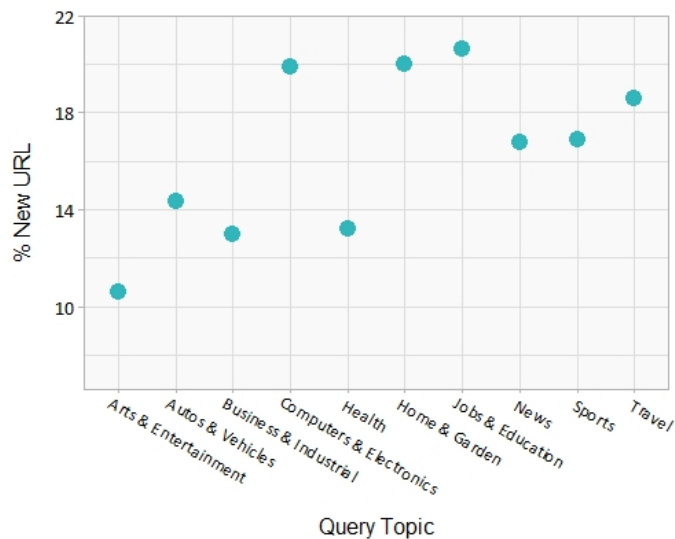


Figure 5.50: Graph Showing the % New URL Metric Results for Each Query Topic, for the Top 10, for Google (authentication status Exp.)

Rank Movements Per Rank Position

5.8.0.3 General Analysis for Google

In Fig. 5.51 we can see the results of the *Rank Movements Per Rank Position* metric for the top 10, top 5 and top 3 search results, using the Google search engine. It is clear that the amount of rank movements per rank position decreases for the analysis of the top 5, and decreases even more in the analysis of the top 3 results, indicating there are less rank shifts in the search results with higher ranking positions.

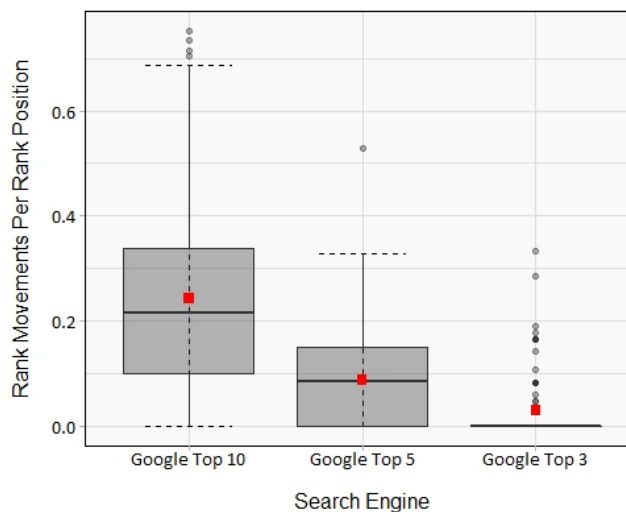


Figure 5.51: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Google, for the Top 10, 5, and 3 (*authentication status* Exp.)

5.8.0.4 Search Topic Comparison

Fig. 5.52 presents the value of the *Rank Movements Per Rank Position* metric for each query topic analyzed, for the top 10 search results, for Google. We can see there's some variation between the metric's values for different topics. The difference between the largest ("Computers & Electronics") and lowest ("Home & Garden") values is nearly **0.19**.

5.9 Results Discussion

This section presents the overall conclusions drawn from the brief results analysis carried out.

Firstly, the analyses' results show that the *location* search factor has the largest impact on the mutability of search results (as expressed by the values of both metrics calculated), followed by the *time*, *authentication status*, and *cookies* factors. The *safe search* and *privacy* factors seem to be the ones that have the lowest impact on search results volatility. To aid visualization, we present in Table 5.1 a concise summary of the results of both metrics for the experiments performed, for the top 10 results, for both Google and Bing. As for the results' comparison between the Google and

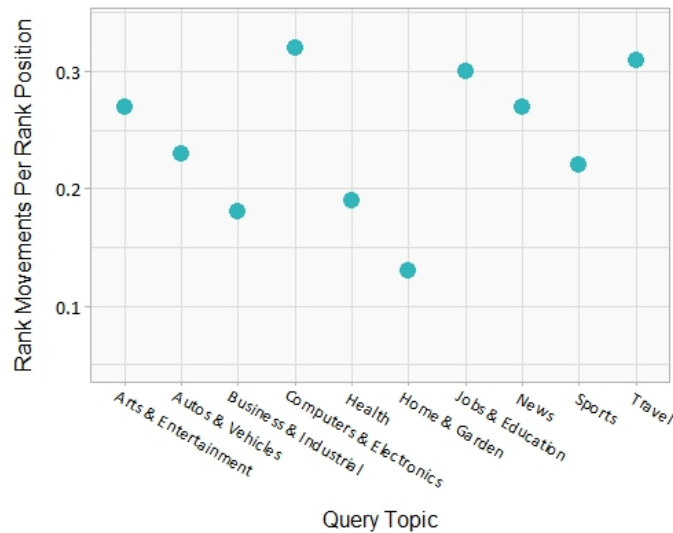


Figure 5.52: Graph Showing the *Rank Movements Per Rank Position* Metric Results for Each Query Topic, for the Top 10, for Google (*authentication status Exp.*)

Bing search engines, we observe that for some search factors the values of the metrics are larger for Bing, while for other factors the values of the metrics are larger for Google. This data indicates that the search engines influence the amount of search results' mutability differently depending on the varying search factor.

Table 5.1: Results of Both Metrics for the Experiments Performed, for the Top 10 Results, for Both Search Engines

	% New URL		Rank Movements Per Rank Position	
	Google	Bing	Google	Bing
Location	35.7%	53.7%	0.87	0.66
Safe search	1.2%	12.4%	0.05	0.27
Privacy	12%	4.2%	0.08	0.22
User Agent	12.8%	10.1%	0.1	0.35
Cookies	14.6%	-	0.18	-
Authentication Status	16.4%	-	0.24	-

When it comes to the differences between the metric's results obtained for the top 10, top 5, top 3 and top 1 search results, several conclusions can be drawn. First, when it comes to the *% New URL*, in some experiments (*time, location* and *safe search*) the results of the metric generally decrease as the top rankings analyzed become higher, while in some other experiments (*privacy, user agent, cookies, and authentication status*) the results of the metric generally increase as the top rankings analyzed become higher. This indicates that the percentage of new URLs increasing

or decreasing for higher top rankings is dependant on the varying search factor. Secondly, for all experiments analyzed, the amount of rank movements per rank position decreases as the top rankings analyzed become higher. This shows that there are less rank shifts in the search results with higher ranking positions, regardless of the varying search factor, which means higher top results often maintain their ranking positions. Also, since in the *time*, *location*, and *safe search* experiments both metrics' values decrease for higher top rankings, we can conclude that, when varying these search factors, search results in higher ranking positions are less volatile.

Furthermore, the search results' mutability seems to be influenced by the query topics, as many of the experiments' results show some variation between the values of the metrics for different query topics. The differences in the value of the *% New URL* metric for different query topics went as high as 15.4% for Google and 17.5% for Bing. As for the *Rank Movements Per Rank Position* metric, the largest difference was 0.33 for Google and 0.51 for Bing. We also notice that these differences have larger peaks for Bing than for Google, which may indicate that query topics have a larger influence on the mutability of search results when using Bing. Additionally, even though query topics seem to influence search result's mutability, the analyses' results don't supply enough information to conclude if there are specific topics that always have more or less impact on results volatility.

One additional interesting fact is that, often, when a search engines' results for one metric are larger than those of the other search engine, the inverse scenario can be seen when it comes to the other metric. This happens for the results of the *time*, *location*, *privacy*, and *user agent* experiments. One explanation for this could be the fact that having a lower value for the *% New URL* metric indicates there are less insertions/deletions of URLs in the *top k* results, which means there is a larger amount of common results in the *top k* that have a chance to shift ranking positions (contributing for the increase of the *Rank Movements Per Rank Position* metric). By the same logic, having a larger value for the *% New URL* metric leads to a smaller amount of common results in the *top k* and thus to less possible ranking shifts.

Chapter 6

Conclusions and Future Work

This chapter presents the final conclusions obtained from this study, as well as some features that could be improved in future works. Section 6.1 summarizes the results of this project and Section 6.2 explores possible enhancements to this work, as well as some next steps that could be taken by future projects.

6.1 Conclusions

Information on the web is unstable, due to the web's dynamic characteristics. Web search engines reflect this instability, as the search results returned can change based on several factors, such as time or the search engine used. The personalization of search results (based on factors such as location or cookies), has also evolved significantly recently.

This work allowed us to obtain insights about how much several search factors — *time*, *location*, *safe search*, *privacy*, *user agent*, *cookies*, and *authentication status* — can impact the mutability of the search results obtained in web searches, in each of the search engines examined, for each of the top results analyzed (top 10, 5, 3, and 1), as well as the influence of query topics on results volatility.

The main contribution of this work was the production of structured datasets detailing the analysis of the volatility of the obtained search results, for each search factor studied, and making these datasets available publicly so they may aid future investigations. Furthermore, a brief analysis of the data obtained was performed, serving as a proof of concept that the data created by this study could be used for further research, and revealing some conclusions. Firstly, the analysis performed reveals the *location* search factor to be the one with the largest impact on the mutability of search results, followed by the *time* factor. The factors *safe search* and *privacy* present the lowest impact. Furthermore, the use of different search engines also influences the volatility of the search results. Additionally, when varying some factors (*time*, *location* and *safe search*) the higher top results are less volatile (experiencing a lower percentage of insertions/deletions of URLs), while the inverse

scenario applies to the *privacy*, *user agent*, *cookies*, and *authentication status* factors. Still, when varying any of the factors studied, higher top results experience less rank shifts, meaning higher top results often maintain their ranking positions. Finally, the mutability of search results is also influenced by the query topics, even though the amount of impact each query topic has may not be consistent when varying different search factors.

The obtained insights about web search and search engines are relevant, not only to the general public, but also for developments in the area of information retrieval research. The results of this study may help shed more light on the subjects of search results volatility and personalization, by presenting an analysis of the variance of search results according to several search factors and making the obtained results available to the public.

6.2 Future Work

After finishing this study, we acknowledge there are a variety of aspects that could be improved or could be object of further research. First, the main aspect that could be focused on future work is to perform a deep analysis of the data present in the datasets obtained in this study, in order to obtain a deeper understanding of how different search factors influence the volatility of search results. Another relevant improvement of the current work would be to use a significantly larger query set when performing the queries; the use of a larger sample of queries could make the analysis' results obtained more reliable. Also, when it comes to the experiment involving the *time* factor, performing retrievals of results over a longer period of time could also help increase the reliability of the analysis' results. Furthermore, when it comes to the experiments involving real users, the use of a larger number of users could also enhance the reliability of the results. Additionally, in some of the experiments performed in this study (the *time*, *location*, and *safe search* experiments), the retrievals for Bing had to use only a subset of the queries used in the retrievals for Google, since the amount of monthly API calls was limited by Bing's API. This, allied with the fact that the query set used isn't very large to begin with, may cause some of the comparisons performed between both search engines to be less reliable. In order to perform a more coherent comparison between both search engines in the future, the exact same query set could be used for the data retrievals of both search engines.

Appendix A

User Survey

The image shows a screenshot of a web-based survey form. The form is titled "Impact of Cookies and Authentication Status on Search Results" and includes a subtitle "Formulário para submissão dos ficheiros recolhidos". It contains five sections, each with a required field indicator (* Required):

- ID do participante (deve colocar o ID que lhe foi atribuído) ***: A text input field with the placeholder "Your answer".
- Idade ***: A radio button selection with six options: "Abaixo de 18", "18-24", "25-34", "35-44", "45-54", and "Acima de 54".
- Género (Escolha aquele com que mais se identifica) ***: A radio button selection with three options: "Masculino", "Feminino", and "Prefiro não dizer".
- Costuma usar o Google regularmente para fazer pesquisas na Web? ***: A radio button selection with three options: "Sim", "Não", and "Às vezes".

Figure A.1: User Survey Conducted During the Real User Experiments

Appendix B

Real User Experiments Scripts

Here we present the scripts used to guide the users during the *cookies* and *authentication status* experiments. Several support pictures don't show the animation. The animated versions of the scripts are open for consultation in the community repository associated with this work.

B.1 Cookies Experiment Script:

Influência de cookies nos resultados de pesquisa

Esta experiência enquadra-se no trabalho associado a uma dissertação do Mestrado Integrado em Engenharia Informática e Computação em que se pretende estudar a mutabilidade de resultados em pesquisas web. Esta experiência foca-se na **influência de cookies e estado de autenticação** nestes resultados.

Durante a experiência, os participantes utilizarão o seu navegador/browser habitual e o motor de pesquisa Google para **fazer um conjunto de pesquisas e guardar a 1a página de resultados**. Não serão recolhidos dados pessoais de nenhum participante. Durante a sessão, a cada utilizador será atribuído um identificador, que passará a ser a única forma de identificação do participante, para preservar a privacidade. Estima-se que a experiência tenha uma duração de 25 minutos.

Posso participar na experiência?

Só poderá participar se responder afirmativamente às seguintes perguntas:

- Usa o seu navegador regularmente em modo não “privado/incógnito”? ([Como saber?](#))
- O seu navegador está a fazer armazenamento de cookies? ([Como saber?](#))
- Já fez pesquisas no Google?

O que devo fazer?

1. Criar uma pasta chamada “Exp1” no seu ambiente de trabalho.
2. Abrir uma janela do seu navegador (Nota: não deve estar ligado a nenhuma VPN).
3. Se estiver autenticado ([Como saber?](#)), fazer **logout** da sua conta google ([Como?](#)).
4. Mudar a configuração do **user agent** corrente para: “*Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/90.0.4430.85 Safari/537.36*” ([Como?](#)).
5. Aceder à página inicial da Google.
6. Aguardar pelo investigador para validação das condições necessárias:
 - A janela em que está a realizar as pesquisas **não** é privada/incógnita ([Como?](#))
 - **Não** está autenticado na sua conta google ([Como?](#))
 - A configuração “safe search” está **desativada** ([Como?](#))
 - A região está definida como “Região Atual/Current location” ([Como?](#))
 - O Idioma/Language está definido como “Português (Portugal)” ([Como?](#))
 - O **user agent** está configurado corretamente ([Como?](#))

7. Para cada uma das [20 expressões de pesquisa](#) (pela ordem indicada):
 - Submeter a expressão de pesquisa na caixa de pesquisa do Google
 - Guardar a 1a página de resultados devolvida **(CTRL+S)** ([Como?](#)) na pasta “Exp1”. **Nota:** Ao guardar a página, deve escolher a opção “Guardar com o tipo: **Página Web, Apenas HTML**” (“Webpage, HTML Only”). Em MacOS escolher o formato “Página Fonte”.
8. Verificar que a pasta “Exp1” contém 20 ficheiros.
9. Enviar a pasta “Exp1” para uma pasta comprimida ([Como?](#)).
10. Enviar a pasta comprimida resultante do passo anterior através [deste formulário](#).
11. Dar indicação ao investigador que terminou a experiência.

Expressões de pesquisa

1. the home depot
2. enchilada recipe
3. honey
4. kitchen storage
5. sheds
6. dumpster world
7. prickley pear drink recipes
8. neapolitan mastiff dog
9. recliner slip covers
10. garden tractor pulling in ohio

11. automobile reliability
12. lexus discount parts
13. elder rv montrose mo
14. gmc jimmy recall
15. 67 mustang
16. car wash
17. ford alarm wire diagrams
18. brake pads
19. omega rims
20. titanium rv dealer

Instruções de Apoio

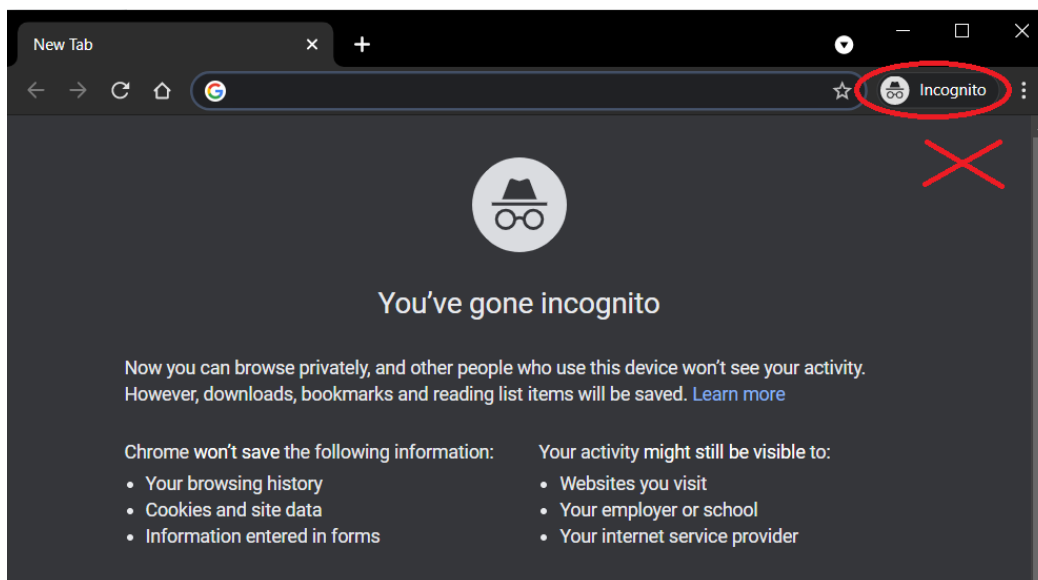
De seguida descrevem-se com mais detalhe alguns dos passos necessários no processo.

Como verificar se a janela é non-private (não incógnita)

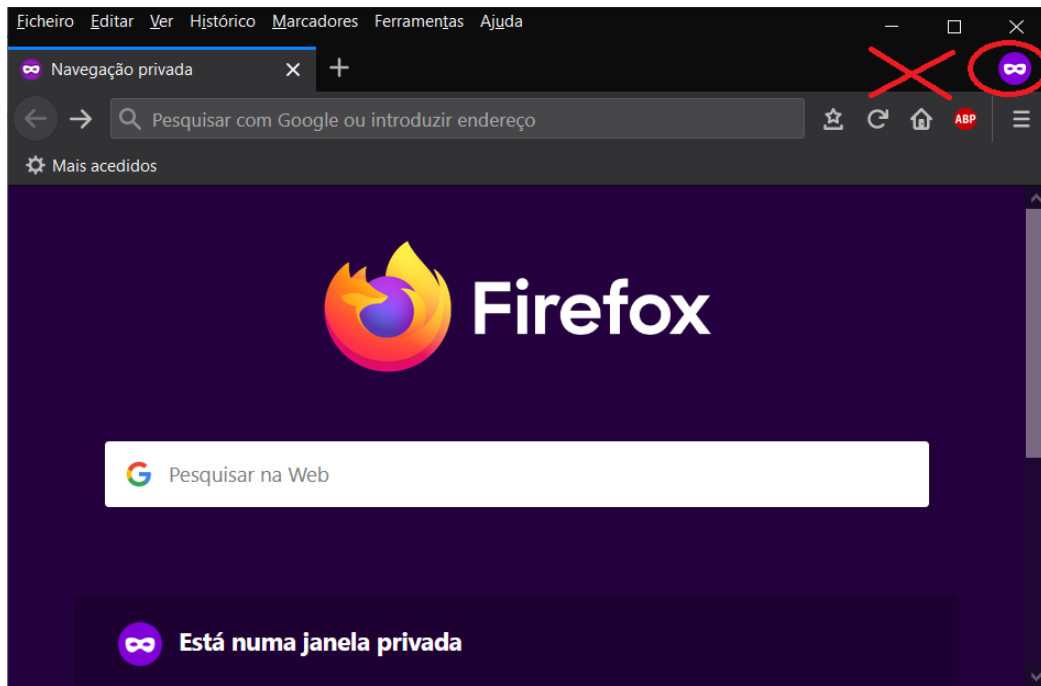
Para que a janela não seja privada, no Chrome, ela **não** pode ter um ícone que diga “Incognito” no canto superior direito, tal como ilustrado na imagem abaixo.

Nos outros browsers, o ícone é um pouco diferente, tal como pode ver nas restantes imagens. [Chrome](#), [Firefox](#), [Edge](#), [Safari](#)

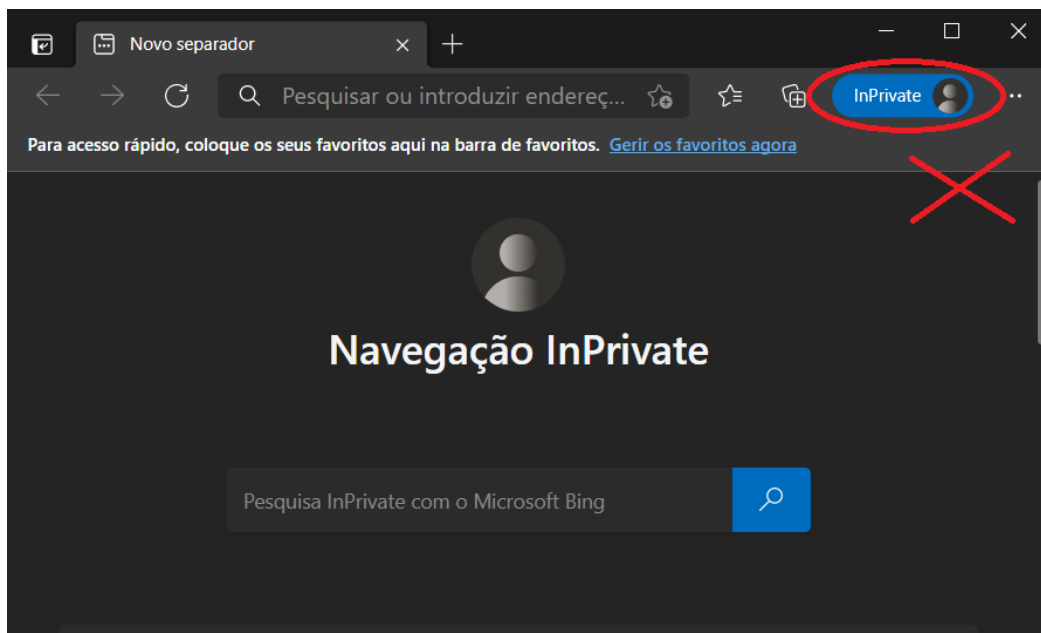
Chrome:



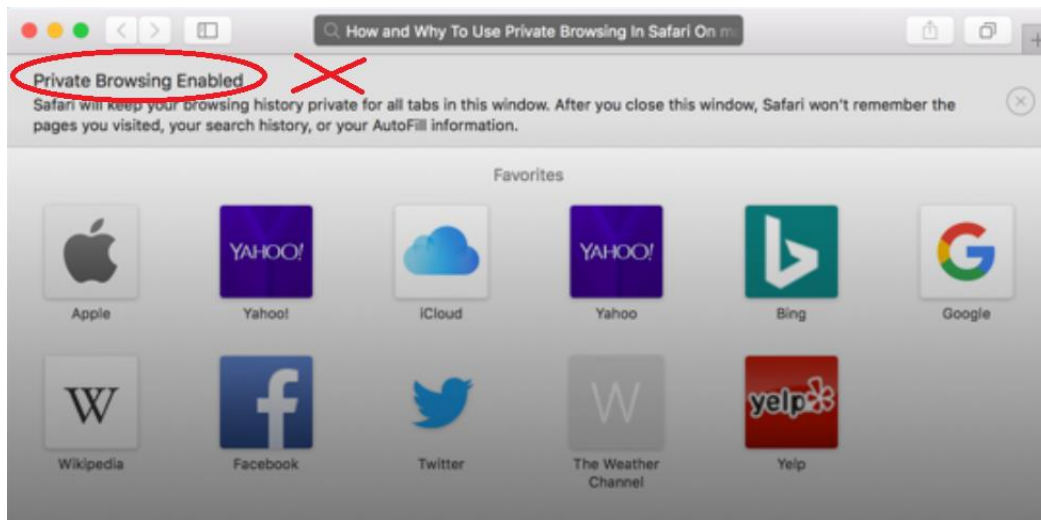
Firefox:



Edge:



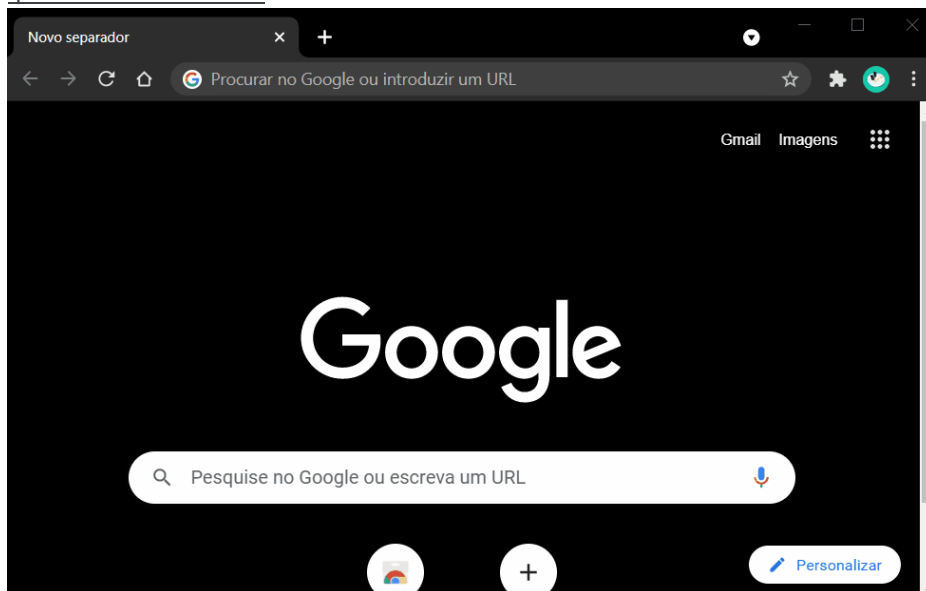
Safari:



Como saber se o navegador armazena cookies?

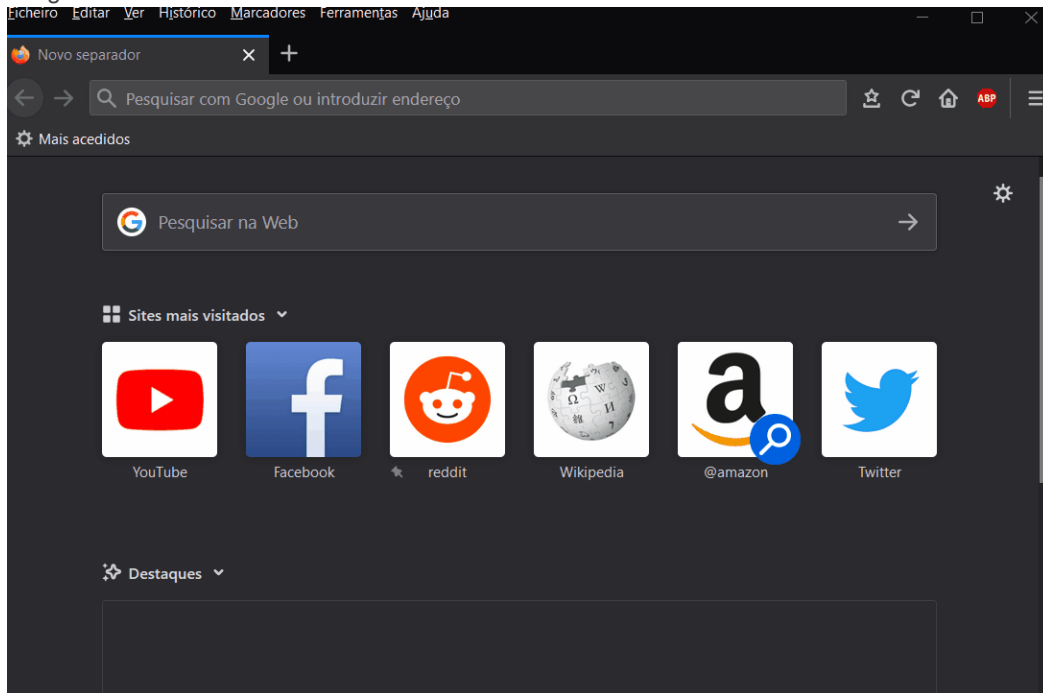
Chrome:

Para saber se o Chrome tem vindo a armazenar as suas cookies, deve seguir a imagem abaixo. As opções "Bloquear todos as Cookies" e "Limpar os cookies e os dados do site quando sair do Chrome" **não** devem estar seleccionadas.



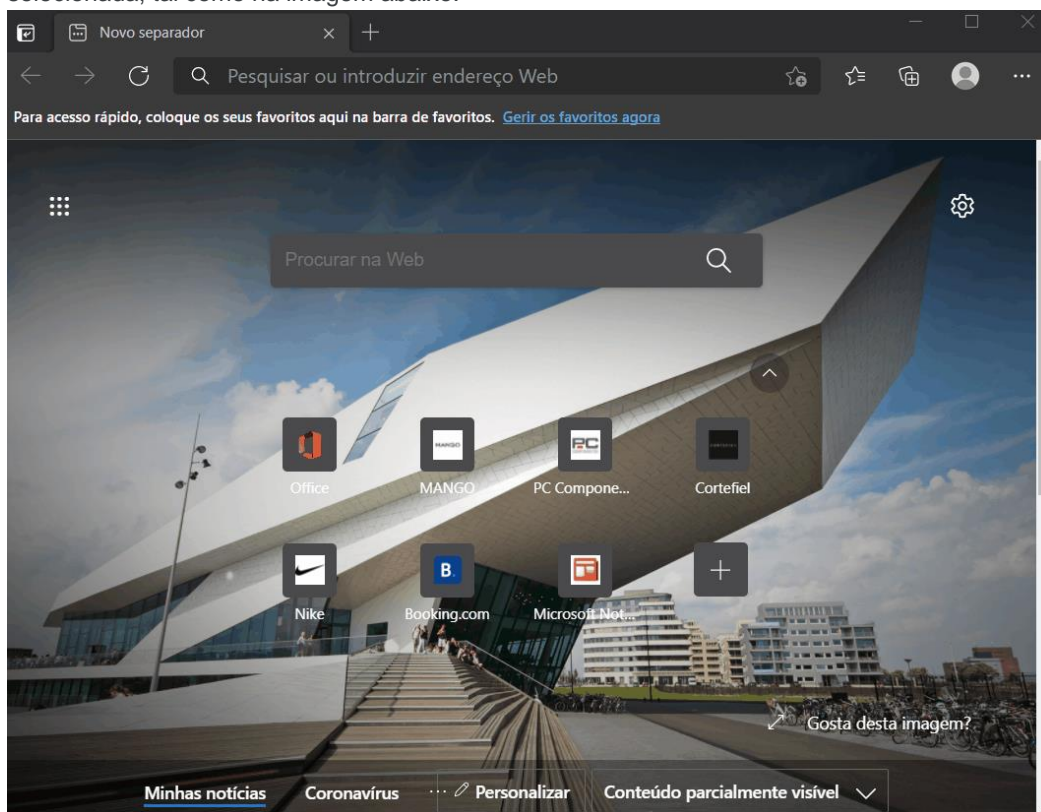
Firefox:

Em Firefox, as opções "Eliminar cookies e os dados de sites quando o Firefox é fechado" e "Utilizar sempre o modo de navegação privada" **não** devem estar seleccionadas, tal como na imagem abaixo.



Edge:

Em Edge, a opção "Permitir que os sites guardem e leiam dados de cookies" **deve** estar selecionada e a opção para limpar cookies e outros dados do site ao fechar **não deve** estar selecionada, tal como na imagem abaixo.

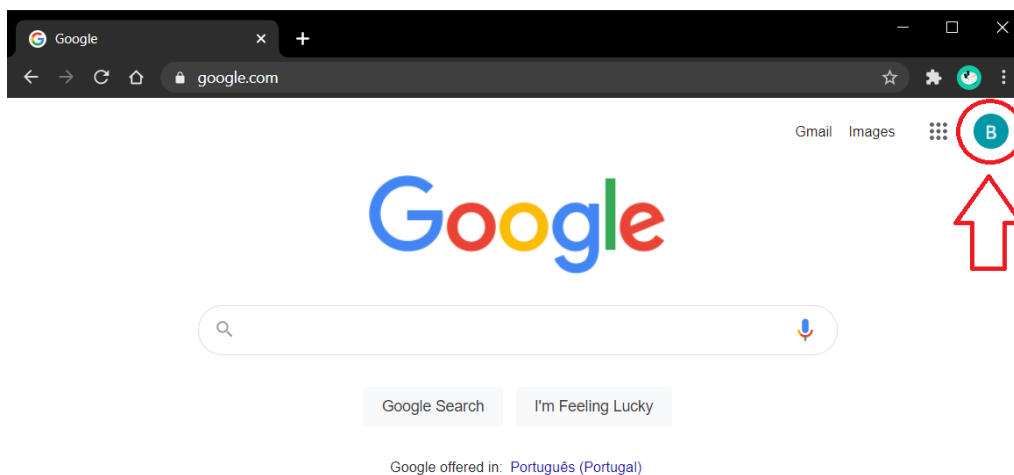


Safari:

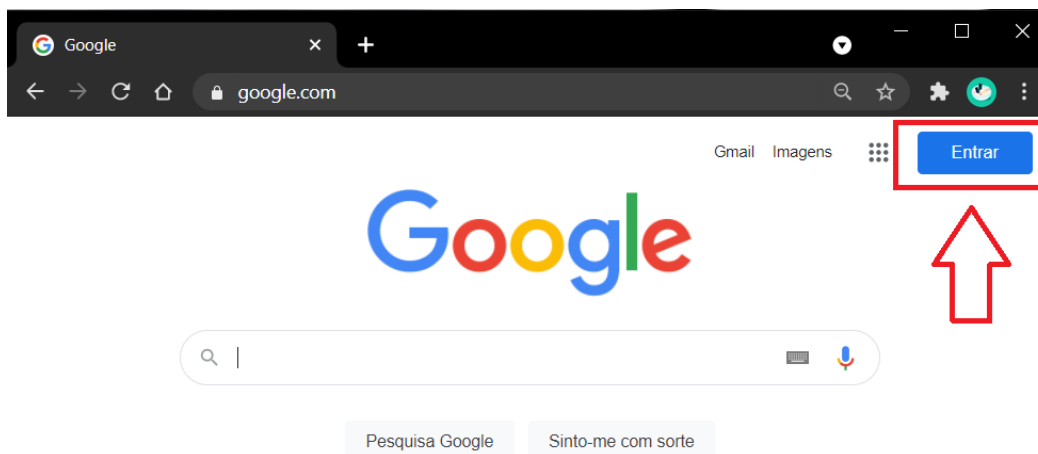
No Safari, aceder a: **Safari** → **Preferências** → **Privacidade**. A definição "**Bloquear todos os cookies**" **não** deve estar selecionada.

Como saber se estou autenticado?

Se estiver autenticado na sua conta Google, um pequeno círculo com uma imagem estará presente no canto superior direito, tal como ilustrado na imagem abaixo.

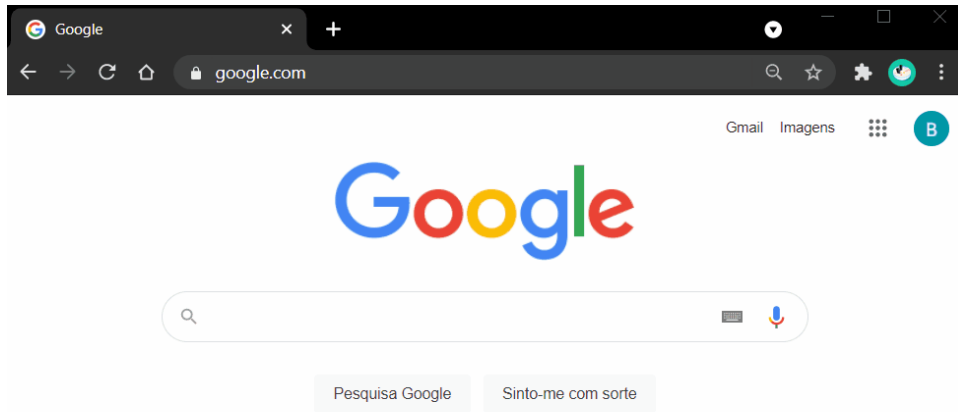


Se não estiver autenticado na sua conta Google, um ícone “Entrar” (“Sign in”) estará presente no canto superior direito, tal como ilustrado na imagem abaixo.



Como fazer logout da sua conta Google?

Seguir as instruções abaixo para fazer logout da sua conta



Como mudar a configuração do User Agent?

Seguir as instruções abaixo para mudar a configuração do User Agent para:
**Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/90.0.4430.85 Safari/537.36**

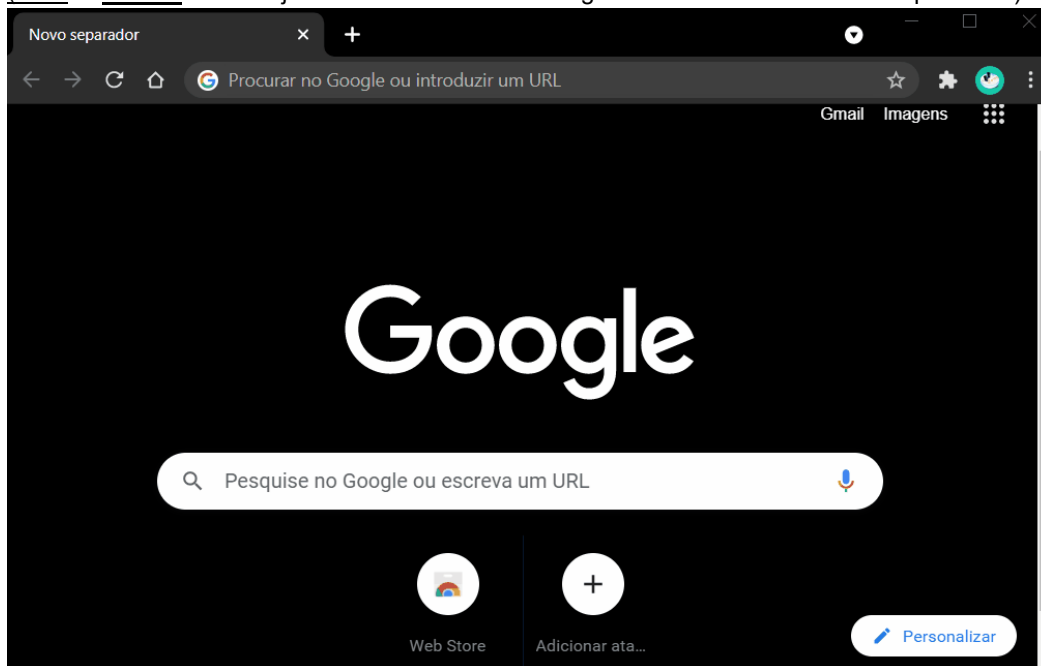
(Nota: pode facilmente reverter este passo depois de terminar a experiência).

[Chrome/Edge](#), [Firefox](#), [Safari](#)

Chrome/Edge:

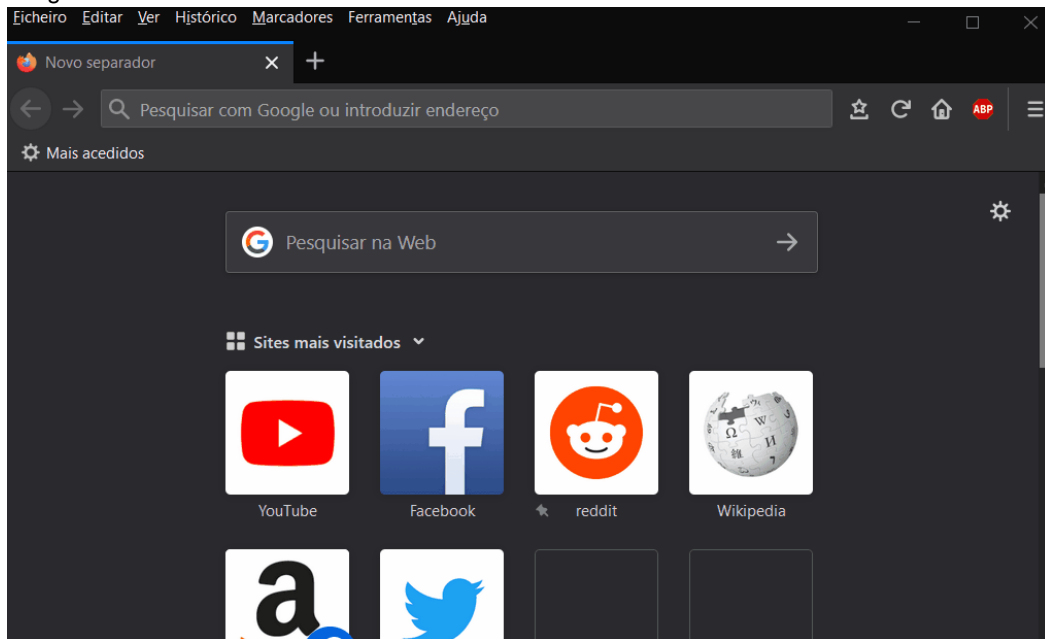
Aceder a: Mais Ferramentas → Ferramentas do Programador → More Tools → Network Conditions (Condições de rede), tal como indicado na imagem abaixo, e seguir os passos.

(Nota: é **crucial** deixar a janela “Ferramentas do Programador” aberta durante a experiência).



Firefox:

Procurar “about:config” e criar um novo item de preferências chamado “general.useragent.override”, onde deve colar o User Agent indicado, tal como mostra a imagem abaixo.

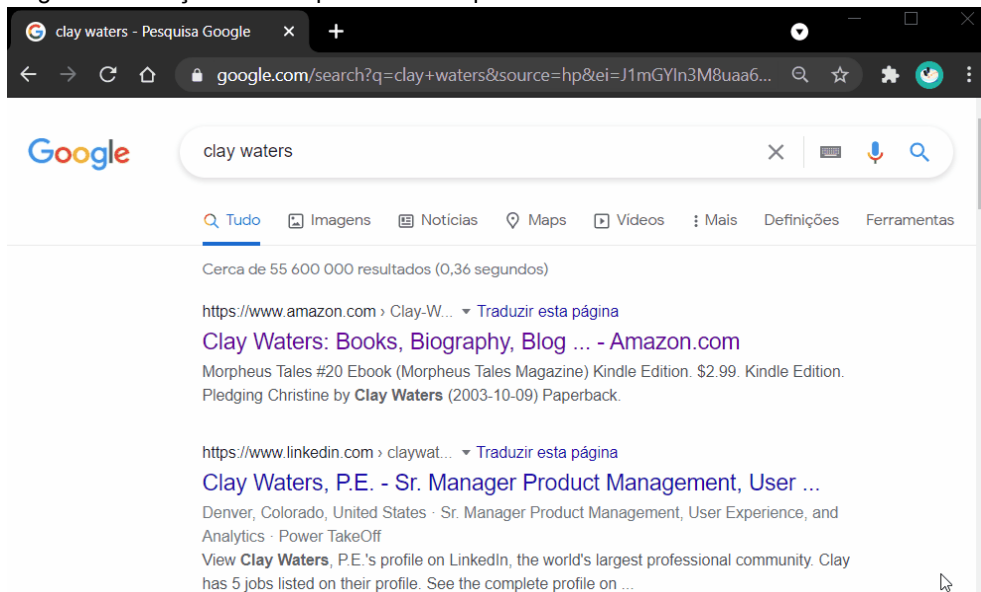


Safari:

Acéder a: **Safari** → **Preferências** → **Avançado**, e seleccionar “**Mostrar menu de Programador na barra de menus**”.

Depois, acéder a: **Programador** → **User Agent** → **Other...**, e colar o User Agent indicado e confirmar.

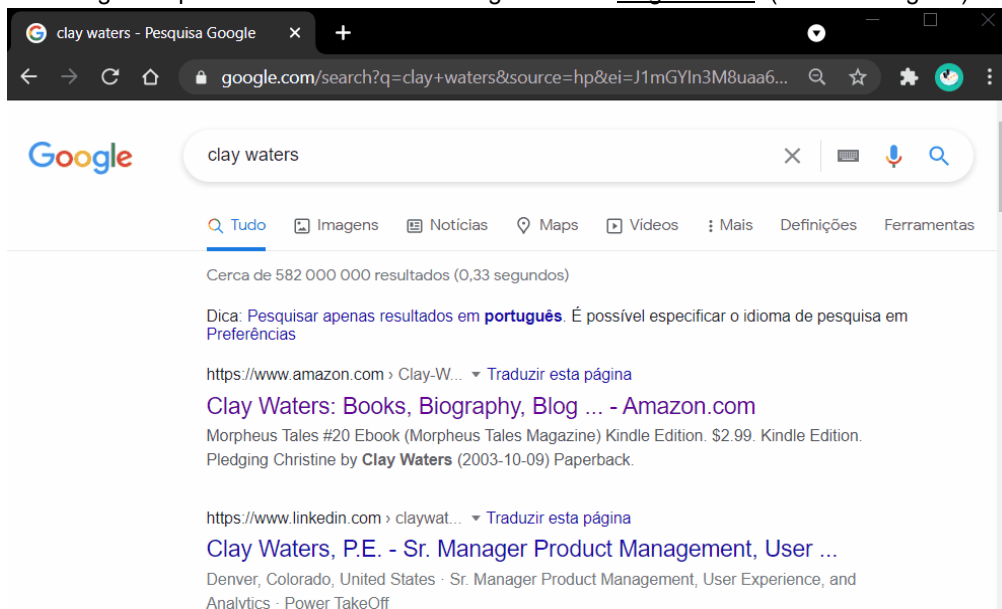
Como verificar que a “safe search” (“pesquisa segura”) está desativada
Seguir as instruções abaixo para verificar que a “safe search” está desativada.



A screenshot of a Google search page for the query "clay waters". The browser's address bar shows the URL "google.com/search?q=clay+waters&source=hp&ei=J1mGYIn3M8uaa6...". The search bar contains "clay waters". Below the search bar, there are navigation options: "Tudo", "Imagens", "Notícias", "Maps", "Vídeos", "Mais", "Definições", and "Ferramentas". The search results indicate "Cerca de 55 600 000 resultados (0,36 segundos)". The first result is from Amazon.com, titled "Clay Waters: Books, Biography, Blog ... - Amazon.com", with a subtitle "Morpheus Tales #20 Ebook (Morpheus Tales Magazine) Kindle Edition. \$2.99. Kindle Edition. Pledging Christine by Clay Waters (2003-10-09) Paperback". The second result is from LinkedIn, titled "Clay Waters, P.E. - Sr. Manager Product Management, User ...", with a subtitle "Denver, Colorado, United States - Sr. Manager Product Management, User Experience, and Analytics - Power TakeOff". A mouse cursor is visible at the bottom right of the page.

Como verificar/mudar a Região Atual

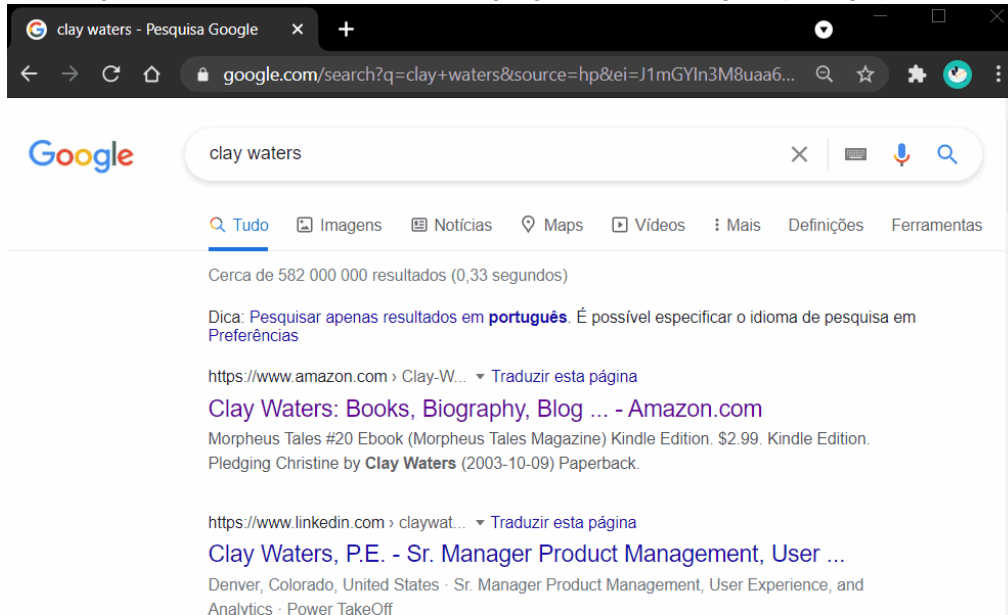
Deve seguir os passos abaixo e definir a Região como “Região atual” (“Current Region”).



A screenshot of a Google search page for the query "clay waters". The browser's address bar shows the URL "google.com/search?q=clay+waters&source=hp&ei=J1mGYIn3M8uaa6...". The search bar contains "clay waters". Below the search bar, there are navigation options: "Tudo", "Imagens", "Notícias", "Maps", "Vídeos", "Mais", "Definições", and "Ferramentas". The search results indicate "Cerca de 582 000 000 resultados (0,33 segundos)". A tip is displayed: "Dica: Pesquisar apenas resultados em português. É possível especificar o idioma de pesquisa em Preferências". The first result is from Amazon.com, titled "Clay Waters: Books, Biography, Blog ... - Amazon.com", with a subtitle "Morpheus Tales #20 Ebook (Morpheus Tales Magazine) Kindle Edition. \$2.99. Kindle Edition. Pledging Christine by Clay Waters (2003-10-09) Paperback". The second result is from LinkedIn, titled "Clay Waters, P.E. - Sr. Manager Product Management, User ...", with a subtitle "Denver, Colorado, United States - Sr. Manager Product Management, User Experience, and Analytics - Power TakeOff".

Como verificar/mudar a Linguagem

Deve seguir os passos abaixo e definir a Linguagem como “Português (Portugal)”



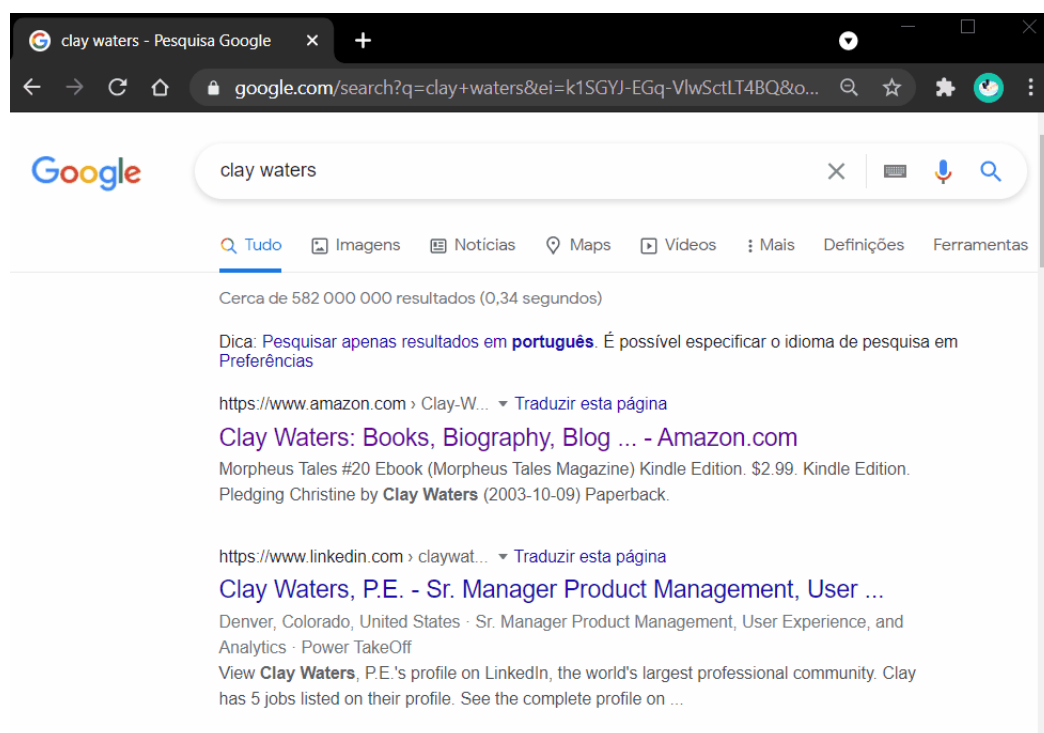
The screenshot shows a Google search results page for the query "clay waters". The browser's address bar shows the URL "google.com/search?q=clay+waters&source=hp&ei=J1mGYIn3M8uaa6...". The search bar contains the text "clay waters". Below the search bar, there are navigation options: "Tudo", "Imagens", "Notícias", "Maps", "Vídeos", "Mais", "Definições", and "Ferramentas". The search results indicate "Cerca de 582 000 000 resultados (0,33 segundos)". A tip in Portuguese suggests searching in Portuguese: "Dica: Pesquisar apenas resultados em português. É possível especificar o idioma de pesquisa em Preferências". Two search results are visible:

- <https://www.amazon.com> › Clay-W... ▾ Traduzir esta página
Clay Waters: Books, Biography, Blog ... - Amazon.com
Morpheus Tales #20 Ebook (Morpheus Tales Magazine) Kindle Edition. \$2.99. Kindle Edition.
Pledging Christine by **Clay Waters** (2003-10-09) Paperback.
- <https://www.linkedin.com> › claywat... ▾ Traduzir esta página
Clay Waters, P.E. - Sr. Manager Product Management, User ...
Denver, Colorado, United States · Sr. Manager Product Management, User Experience, and Analytics · Power TakeOff

Como guardar a página de resultados de pesquisa?

Para guardar uma página de resultados, **usar o atalho de teclado “CTRL+S”** enquanto está na página dos resultados. Alternativamente, pode utilizar o método ilustrado nas imagens abaixo. Ao guardar cada página, deve escolher a opção “Guardar com o tipo: **Página Web, Apenas HTML**” (“Webpage, HTML Only”). **Não altere** o nome dos ficheiros que guardar.

Chrome/Edge/Firefox:

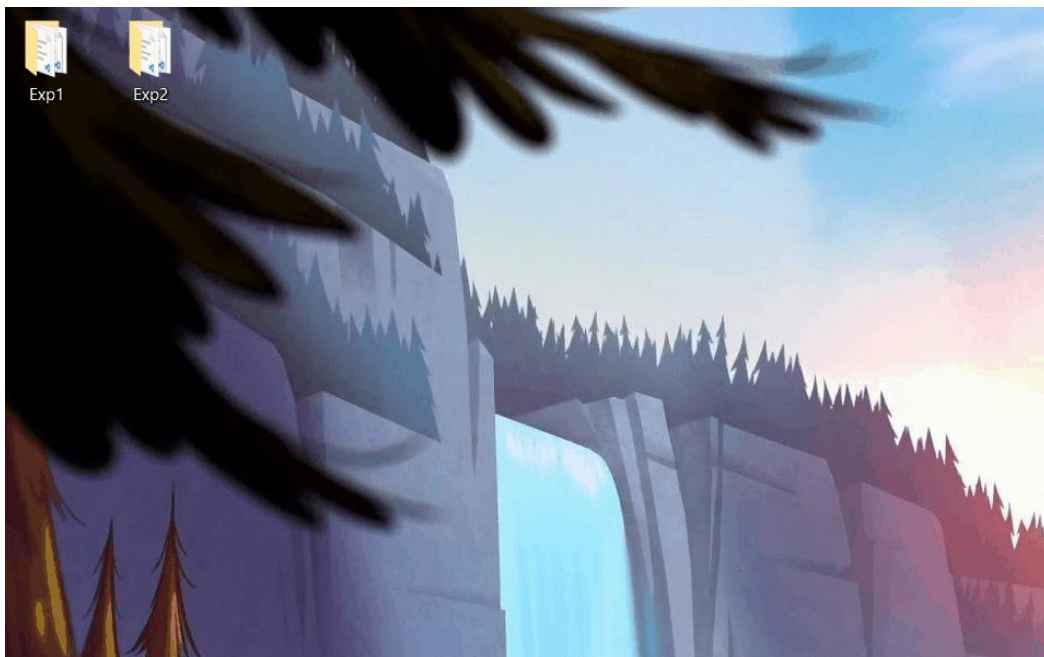


Safari:

Em Safari, deve clicar em “**Ficheiro → Guardar Como**” e deve guardar no formato “**Page Source**” (“**Página Fonte**”)

Como enviar as pastas “Exp1” e “Exp2” para uma pasta comprimida?

Para comprimir as pastas “Exp1” e “Exp2” deve seguir as instruções da imagem abaixo. Se o seu sistema operativo for **macOS** ou **Linux**, em vez de “Enviar para > Pasta Comprimida”, a sua opção será “Compress” (“Comprimir”).



B.2 *Authentication status* Experiment Script:

Influência de estado de autenticação nos resultados de pesquisa

Esta experiência enquadra-se no trabalho associado a uma dissertação do Mestrado Integrado em Engenharia Informática e Computação em que se pretende estudar a mutabilidade de resultados em pesquisas web. Esta experiência foca-se na **influência de cookies e estado de autenticação** nestes resultados.

Durante a experiência, os participantes utilizarão o seu navegador/browser habitual e o motor de pesquisa Google para **fazer um conjunto de pesquisas e guardar a 1a página de resultados**. Não serão recolhidos dados pessoais de nenhum participante. Durante a sessão, a cada utilizador será atribuído um identificador, que passará a ser a única forma de identificação do participante, para preservar a privacidade. Estima-se que a experiência tenha uma duração de 25 minutos.

Posso participar na experiência?

Só poderá participar se responder afirmativamente às seguintes perguntas:

- Já fez pesquisas no Google?
- Tem **uma conta google** criada e que use regularmente? (será necessário que **se lembre da password da sua conta google**, pois será necessário desativar temporariamente as cookies do seu navegador, o que causa a terminação da sessão)

O que devo fazer?

1. Criar uma pasta chamada “Exp1” no seu ambiente de trabalho.
2. **Mover o ficheiro “Cookies” para fora do diretório do seu navegador** ([Como?](#))
(este processo **não** afeta o ficheiro de cookies correntes e **é rapidamente reversível**).
3. Aguardar pelo investigador para validação das condições necessárias:
 - O ficheiro das Cookies **não** está no diretório do navegador
 - (Voltar a abrir o navegador, aceder à página inicial da Google e **autenticar-se na sua conta Google**)
 - Mudar a configuração do **user agent** corrente para: *Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/90.0.4430.85 Safari/537.36* ([Como?](#))
 - A janela em que está a realizar as pesquisas **não** é privada/incógnita ([Como?](#))

- **Está autenticado** na sua conta google ([Como?](#))
 - A configuração “safe search” está **desativada** ([Como?](#))
 - A região está definida como “Região Atual/Current location” ([Como?](#))
 - O Idioma/Language está definido como “Português (Portugal)” ([Como?](#))
4. Para cada uma das [20 expressões de pesquisa](#) (pela ordem indicada):
 - Submeter a expressão de pesquisa na caixa de pesquisa do Google
 - Guardar a 1a página de resultados devolvida (**CTRL+S**) ([Como?](#)) na pasta “Exp1”. Nota: Ao guardar a página, deve escolher a opção “Guardar com o tipo: **Webpage, HTML Only**” (“Apenas HTML”). Em MacOS escolher o formato “Página Fonte”.
 5. Verificar que a pasta “Exp1” contém 20 ficheiros.
 6. Enviar a pasta “Exp1” para uma pasta comprimida ([Como?](#)).
 7. Enviar a pasta comprimida resultante do passo anterior através [deste formulário](#).
 8. Dar indicação ao investigador que terminou a experiência.
 9. [EXTRA] **No final de toda a experiência**, de modo a repor o ficheiro original das suas cookies, deve substituir o novo ficheiro “Cookies” gerado pelo sistema pelo ficheiro “Cookies” original ([Como?](#)).

Expressões de pesquisa

1. golfcarts
2. olympic diving headlines
3. big balls management
4. charles mccrea
5. travis pastrana
6. lakers rumors
7. wrestling spoilers
8. dicks sporting goods
9. espn nfl 2k5
10. grumman boat

11. biman air
12. travel to germany with kids
13. jet blue airlines
14. travels advantage
15. rio hotel
16. hitlon hotels
17. seabourn pride
18. hiking tours in va
19. cruise jubilee 8 16
20. cheyene mountain resort

Instruções de Apoio

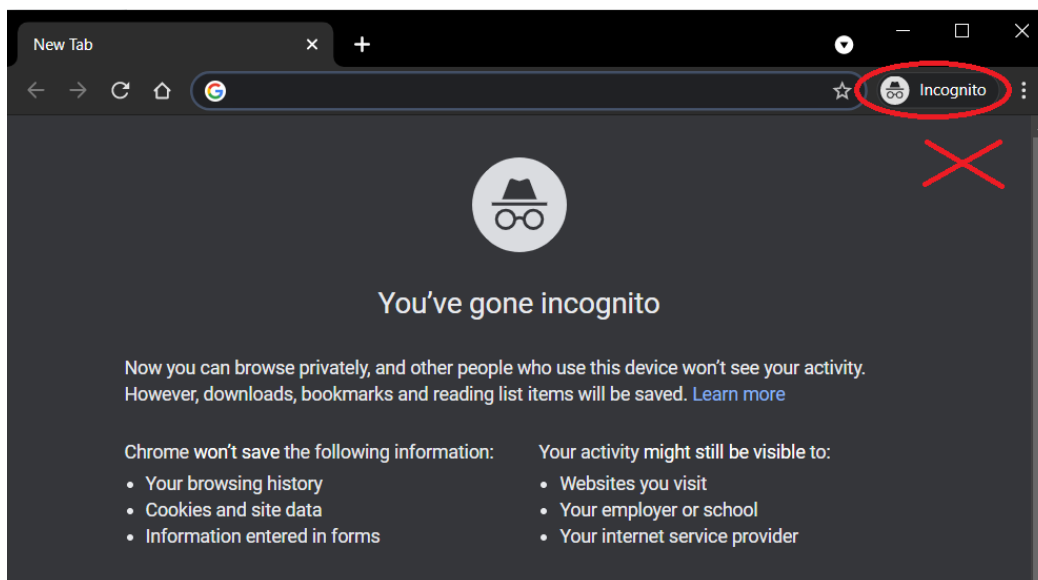
De seguida descrevem-se com mais detalhe alguns dos passos necessários no processo.

Como verificar se a janela é non-private (não incógnita)

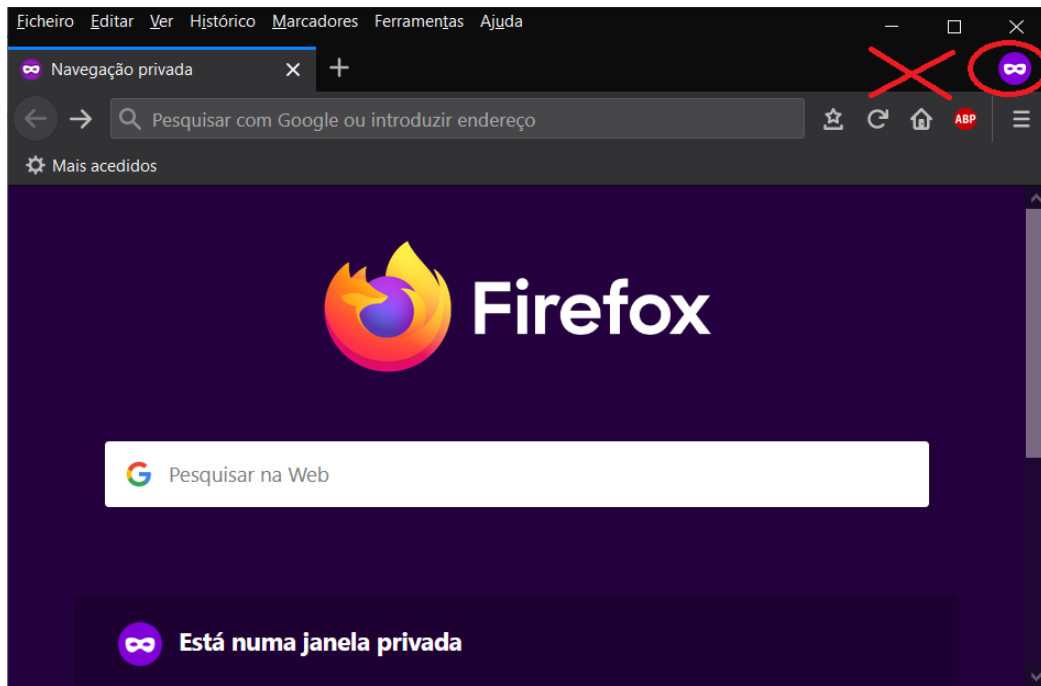
Para que a janela não seja privada, no Chrome, ela **não** pode ter um ícone que diga “Incognito” no canto superior direito, tal como ilustrado na imagem abaixo.

Nos outros browsers, o ícone é um pouco diferente, tal como pode ver nas restantes imagens. [Chrome](#), [Firefox](#), [Edge](#), [Safari](#)

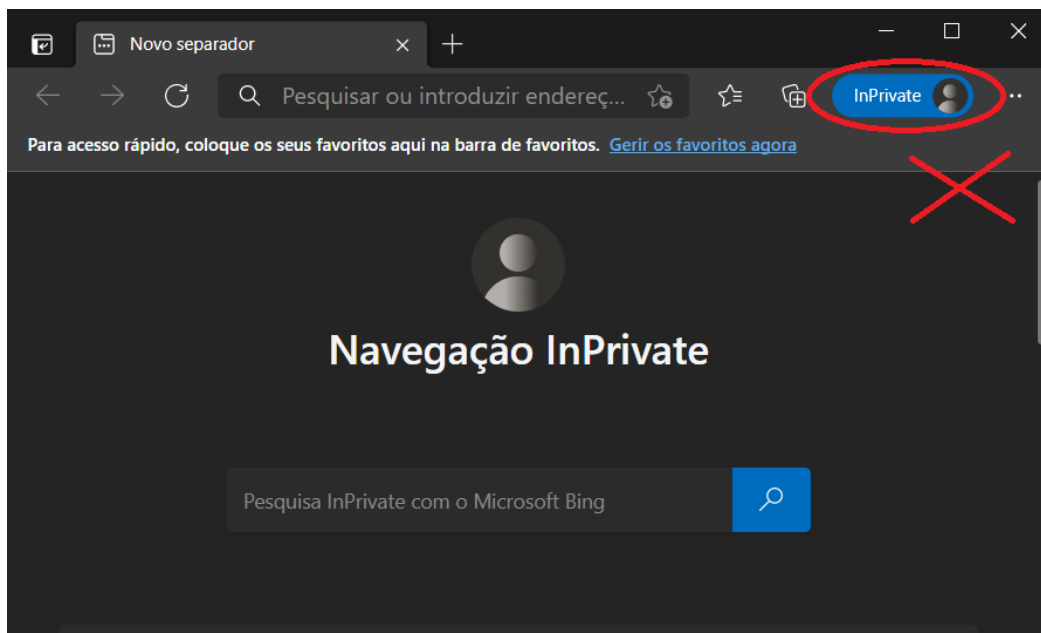
Chrome:



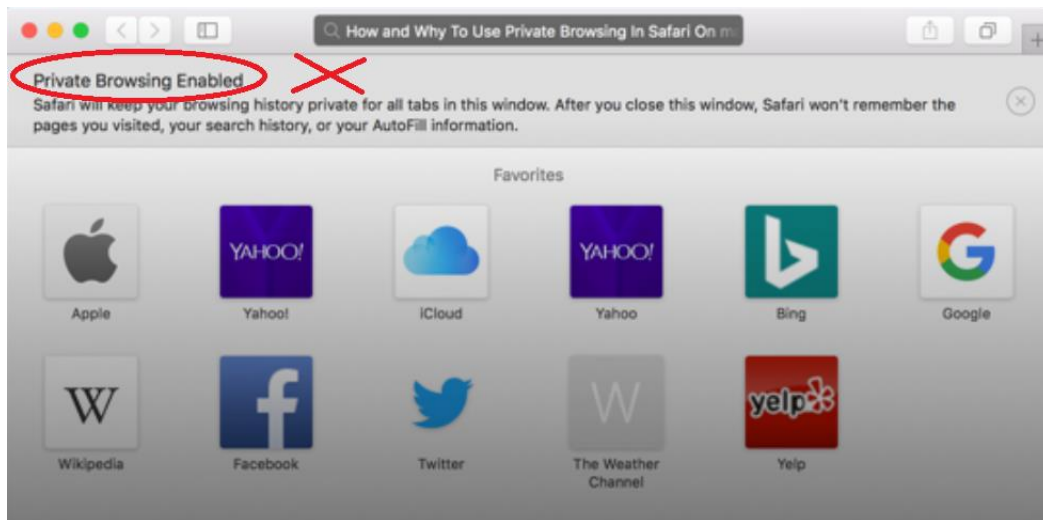
Firefox:



Edge:



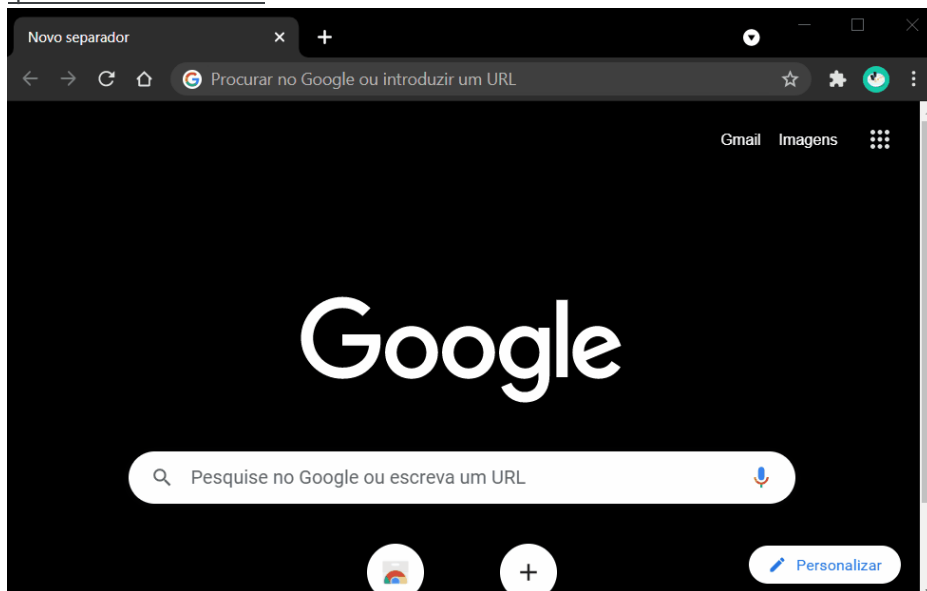
Safari:



Como saber se o navegador armazena cookies?

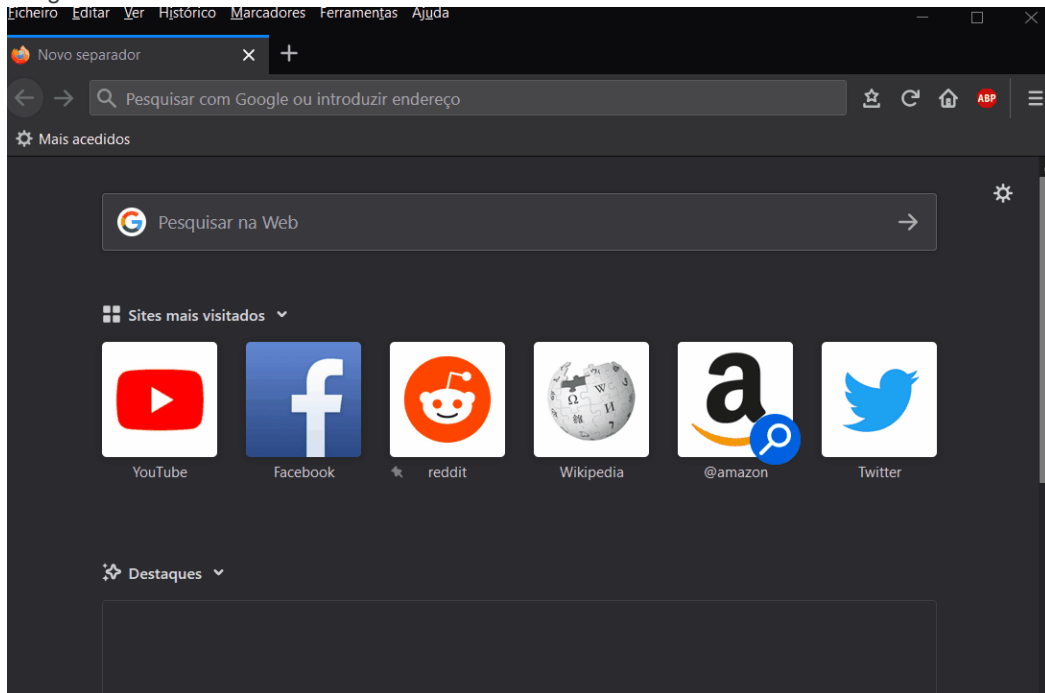
Chrome:

Para saber se o Chrome tem vindo a armazenar as suas cookies, deve seguir a imagem abaixo. As opções "Bloquear todos as Cookies" e "Limpar os cookies e os dados do site quando sair do Chrome" **não** devem estar seleccionadas.



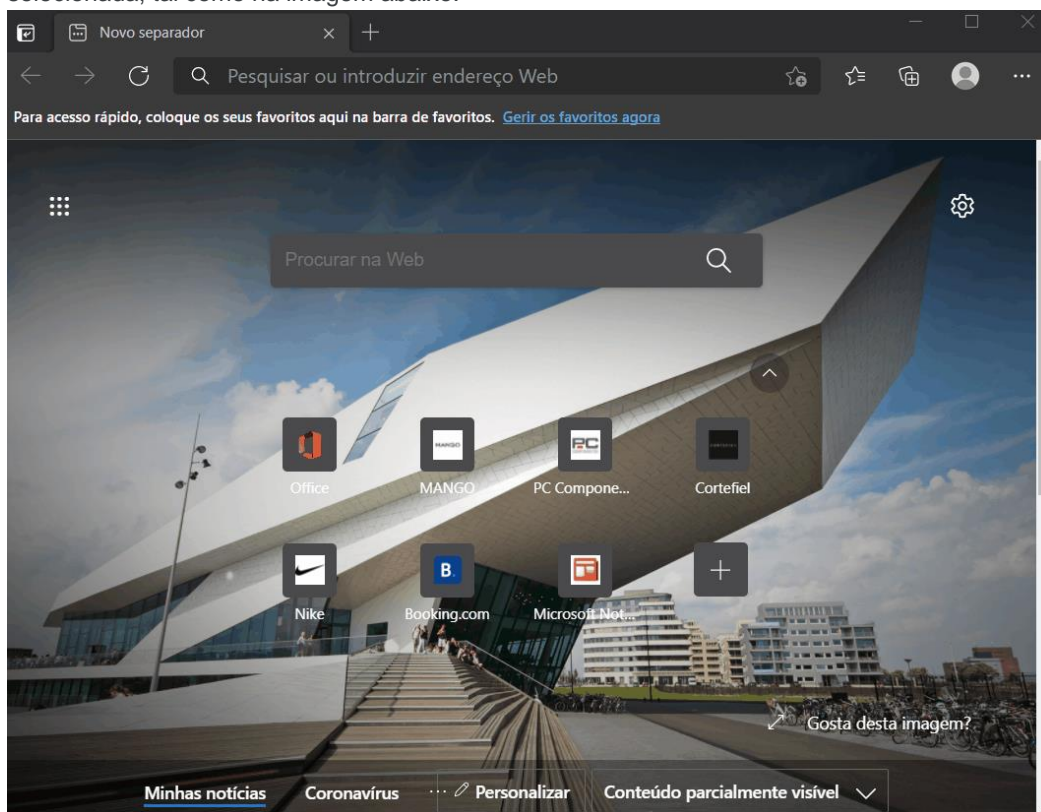
Firefox:

Em Firefox, as opções "Eliminar cookies e os dados de sites quando o Firefox é fechado" e "Utilizar sempre o modo de navegação privada" **não** devem estar seleccionadas, tal como na imagem abaixo.



Edge:

Em Edge, a opção "Permitir que os sites guardem e leiam dados de cookies" **deve** estar selecionada e a opção para limpar cookies e outros dados do site ao fechar **não deve** estar selecionada, tal como na imagem abaixo.

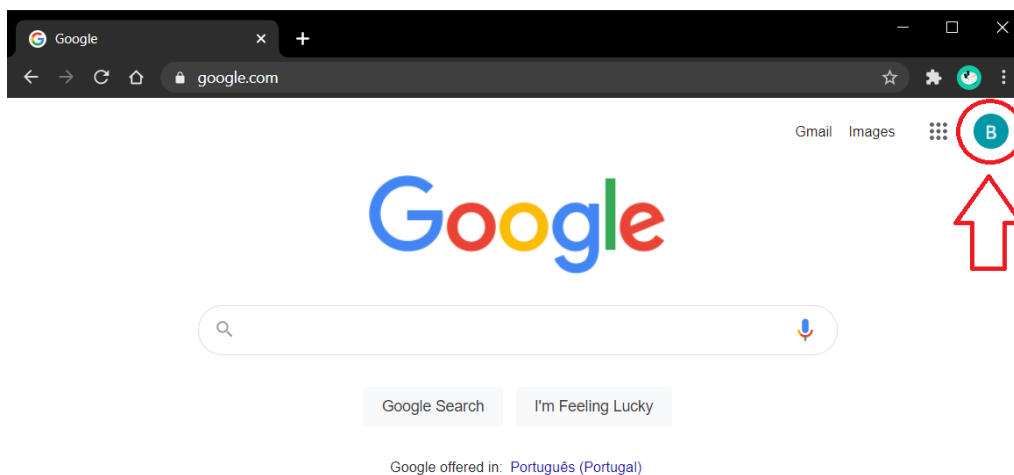


Safari:

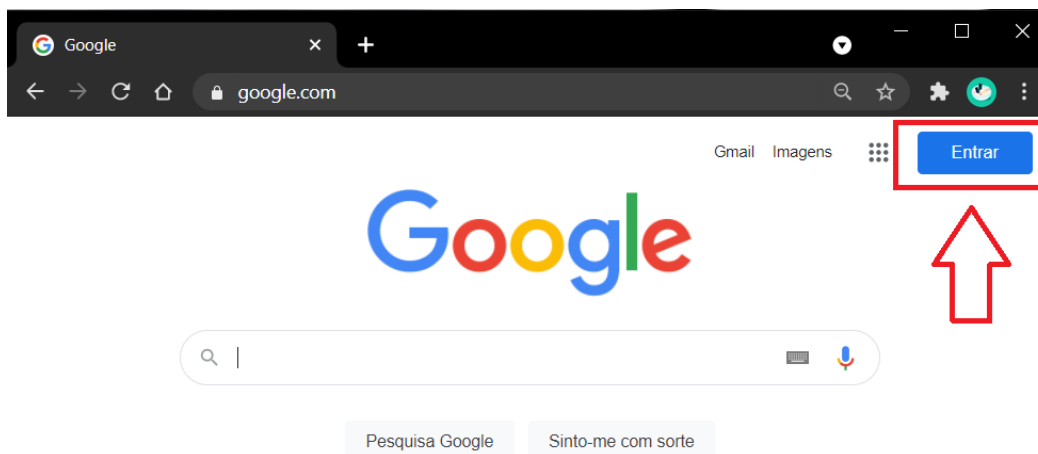
No Safari, aceder a: **Safari** → **Preferências** → **Privacidade**. A definição "**Bloquear todos os cookies**" **não** deve estar selecionada.

Como saber se estou autenticado?

Se estiver autenticado na sua conta Google, um pequeno círculo com uma imagem estará presente no canto superior direito, tal como ilustrado na imagem abaixo.

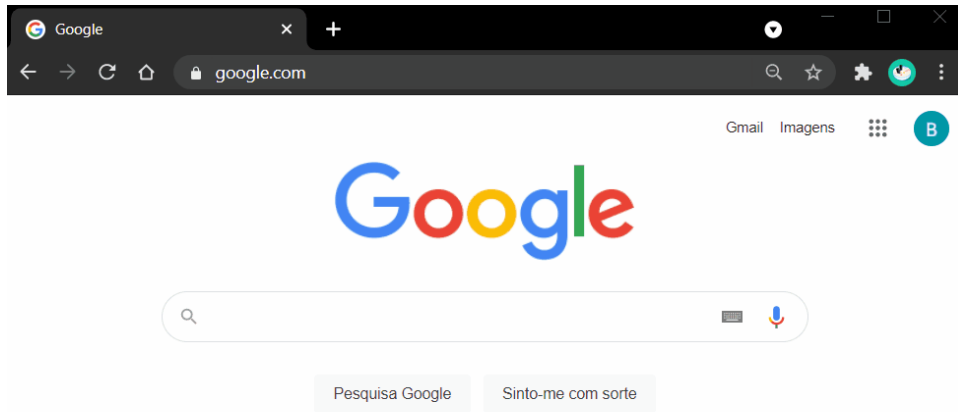


Se não estiver autenticado na sua conta Google, um ícone “Entrar” (“Sign in”) estará presente no canto superior direito, tal como ilustrado na imagem abaixo.



Como fazer logout da sua conta Google?

Seguir as instruções abaixo para fazer logout da sua conta



Como mudar a configuração do User Agent?

Seguir as instruções abaixo para mudar a configuração do User Agent para:
Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/90.0.4430.85 Safari/537.36

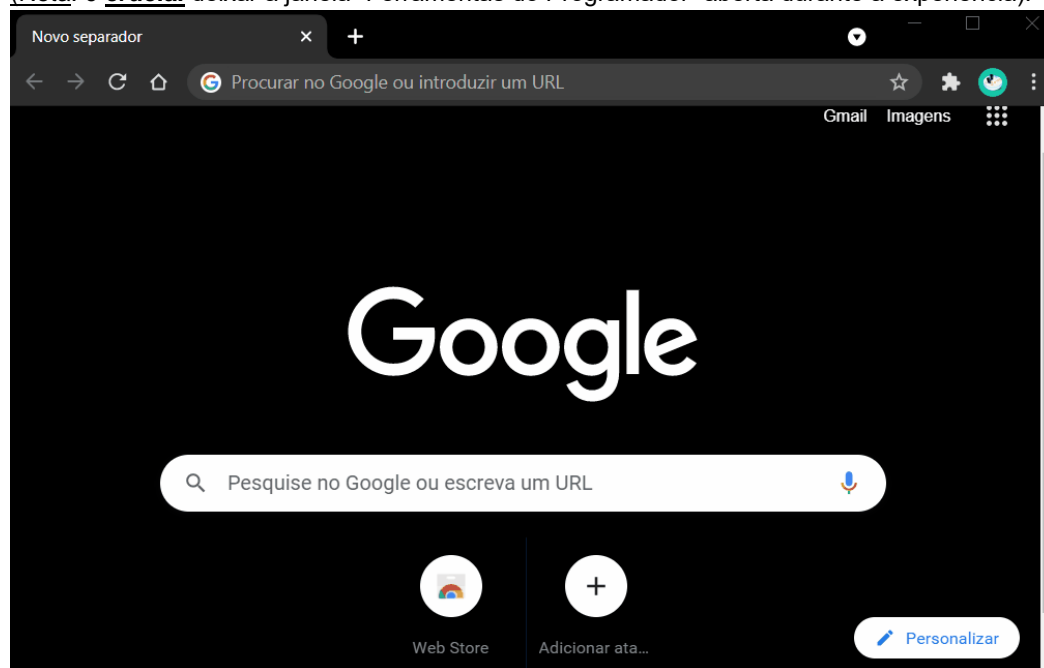
(Nota: pode facilmente reverter este passo depois de terminar a experiência).

[Chrome/Edge](#), [Firefox](#), [Safari](#)

Chrome/Edge:

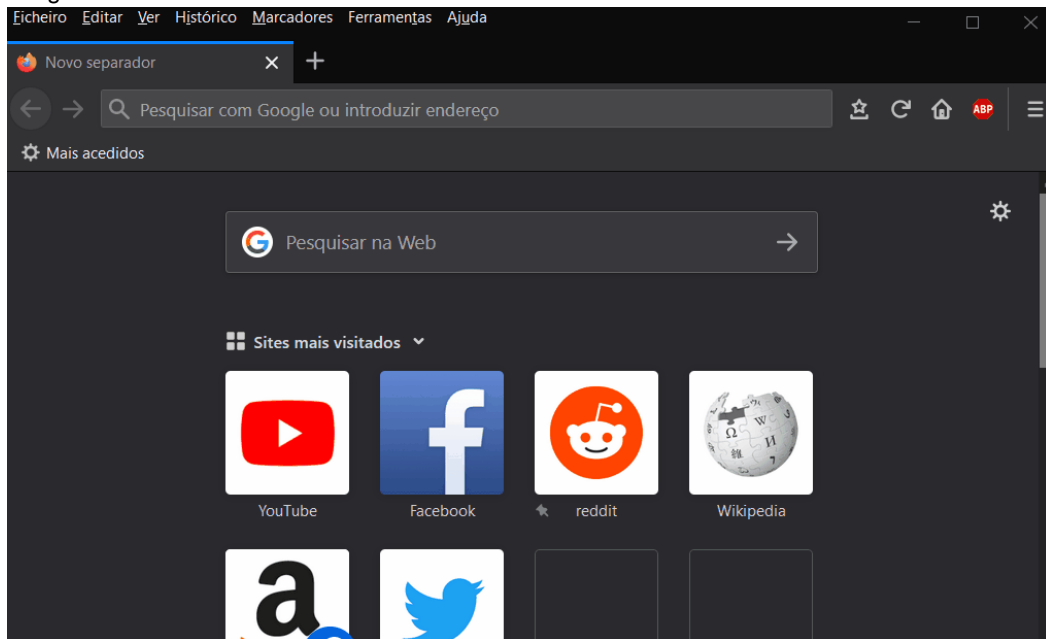
Aceder a: Mais Ferramentas → Ferramentas do Programador → More Tools → Network Conditions (Condições de rede), tal como indicado na imagem abaixo, e seguir os passos.

(Nota: é **crucial** deixar a janela “Ferramentas do Programador” aberta durante a experiência).



Firefox:

Procurar “about:config” e criar um novo item de preferências chamado “general.useragent.override”, onde deve colar o User Agent indicado, tal como mostra a imagem abaixo.

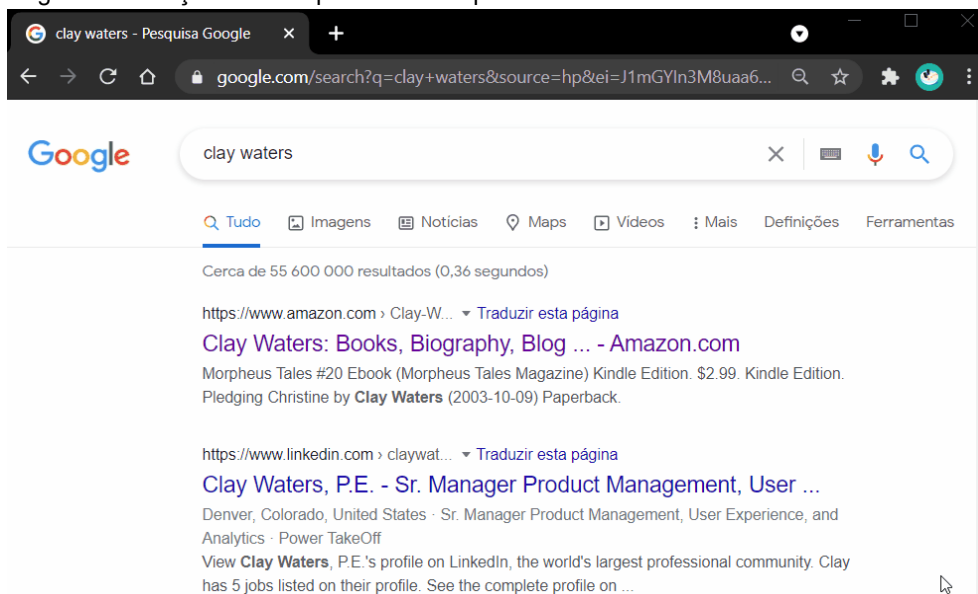


Safari:

Aceder a: **Safari** → **Preferências** → **Avançado**, e selecionar “**Mostrar menu de Programador na barra de menus**”.

Depois, aceder a: **Programador** → **User Agent** → **Other...**, e colar o User Agent indicado e confirmar.

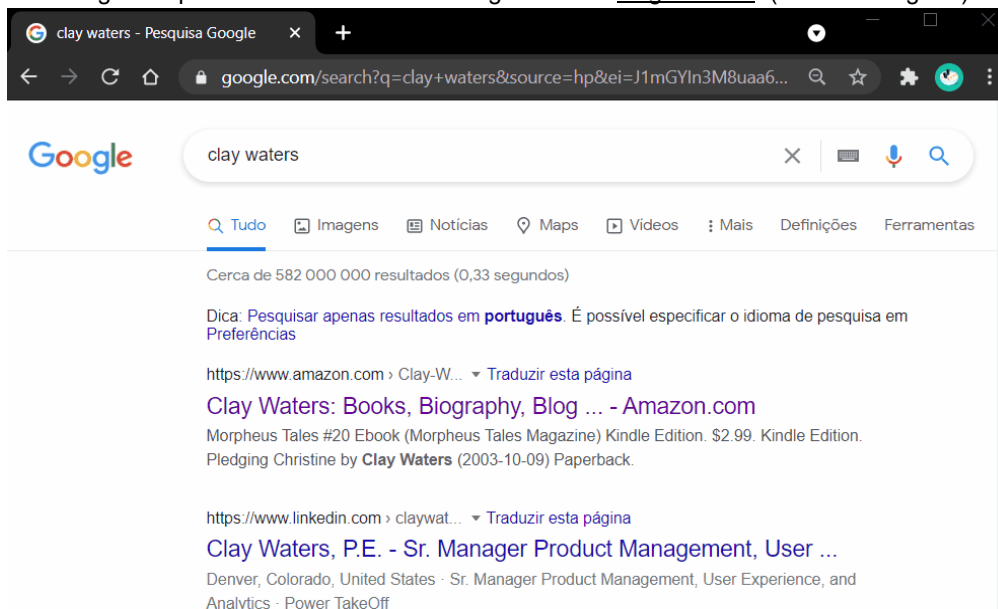
Como verificar que a “safe search” (“pesquisa segura”) está desativada
Seguir as instruções abaixo para verificar que a “safe search” está desativada.



A screenshot of a Google search page for the query "clay waters". The browser's address bar shows the search URL. The search results show approximately 55,600,000 results in 0.36 seconds. The top results include an Amazon link for "Clay Waters: Books, Biography, Blog ... - Amazon.com" and a LinkedIn profile for "Clay Waters, P.E. - Sr. Manager Product Management, User Experience, and Analytics - Power TakeOff".

Como verificar/mudar a Região Atual

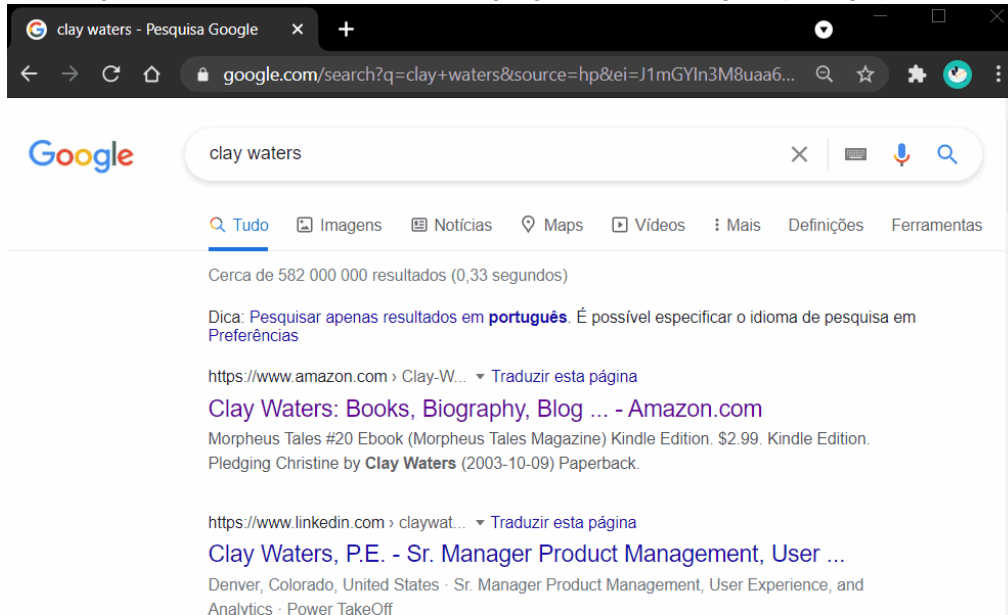
Deve seguir os passos abaixo e definir a Região como “Região atual” (“Current Region”).



A screenshot of a Google search page for the query "clay waters". The browser's address bar shows the search URL. The search results show approximately 582,000,000 results in 0.33 seconds. A tip at the top of the results says "Dica: Pesquisar apenas resultados em português. É possível especificar o idioma de pesquisa em Preferências". The top results are identical to the previous screenshot, including the Amazon link and the LinkedIn profile.

Como verificar/mudar a Linguagem

Deve seguir os passos abaixo e definir a Linguagem como “Português (Portugal)”



The screenshot shows a Google search results page for the query "clay waters". The browser's address bar shows the URL "google.com/search?q=clay+waters&source=hp&ei=J1mGYIn3M8uaa6...". The search bar contains the text "clay waters". Below the search bar, there are navigation options: "Tudo", "Imagens", "Notícias", "Maps", "Vídeos", "Mais", "Definições", and "Ferramentas". The search results indicate "Cerca de 582 000 000 resultados (0,33 segundos)". A tip in Portuguese suggests searching in Portuguese: "Dica: Pesquisar apenas resultados em português. É possível especificar o idioma de pesquisa em Preferências". Two search results are visible: one from Amazon.com for "Clay Waters: Books, Biography, Blog ... - Amazon.com" and another from LinkedIn for "Clay Waters, P.E. - Sr. Manager Product Management, User ...".

clay waters - Pesquisa Google

google.com/search?q=clay+waters&source=hp&ei=J1mGYIn3M8uaa6...

Google

clay waters

Tudo Imagens Notícias Maps Vídeos Mais Definições Ferramentas

Cerca de 582 000 000 resultados (0,33 segundos)

Dica: Pesquisar apenas resultados em português. É possível especificar o idioma de pesquisa em Preferências

https://www.amazon.com › Clay-W... Traduzir esta página

Clay Waters: Books, Biography, Blog ... - Amazon.com

Morpheus Tales #20 Ebook (Morpheus Tales Magazine) Kindle Edition. \$2.99. Kindle Edition.
Pledging Christine by **Clay Waters** (2003-10-09) Paperback.

https://www.linkedin.com › claywat... Traduzir esta página

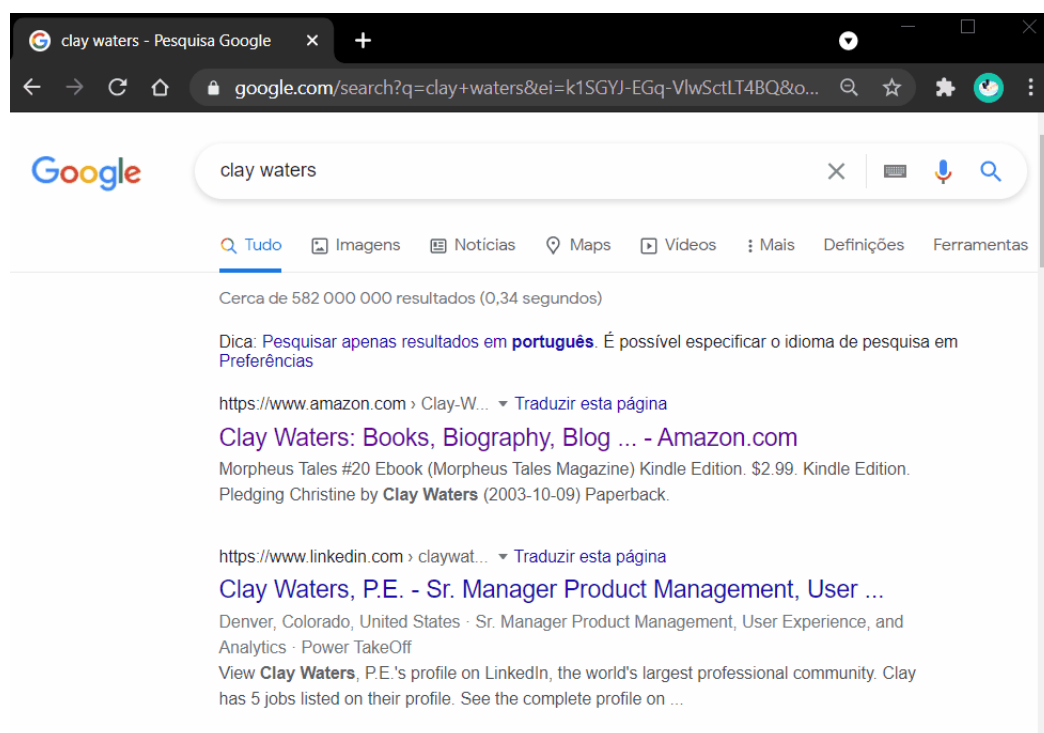
Clay Waters, P.E. - Sr. Manager Product Management, User ...

Denver, Colorado, United States · Sr. Manager Product Management, User Experience, and Analytics · Power TakeOff

Como guardar a página de resultados de pesquisa?

Para guardar uma página de resultados, **usar o atalho de teclado “CTRL+S”** enquanto está na página dos resultados. Alternativamente, pode utilizar o método ilustrado nas imagens abaixo. Ao guardar cada página, deve escolher a opção “Guardar com o tipo: **Página Web, Apenas HTML**” (“Webpage, HTML Only”). **Não altere** o nome dos ficheiros que guardar.

Chrome/Edge/Firefox:

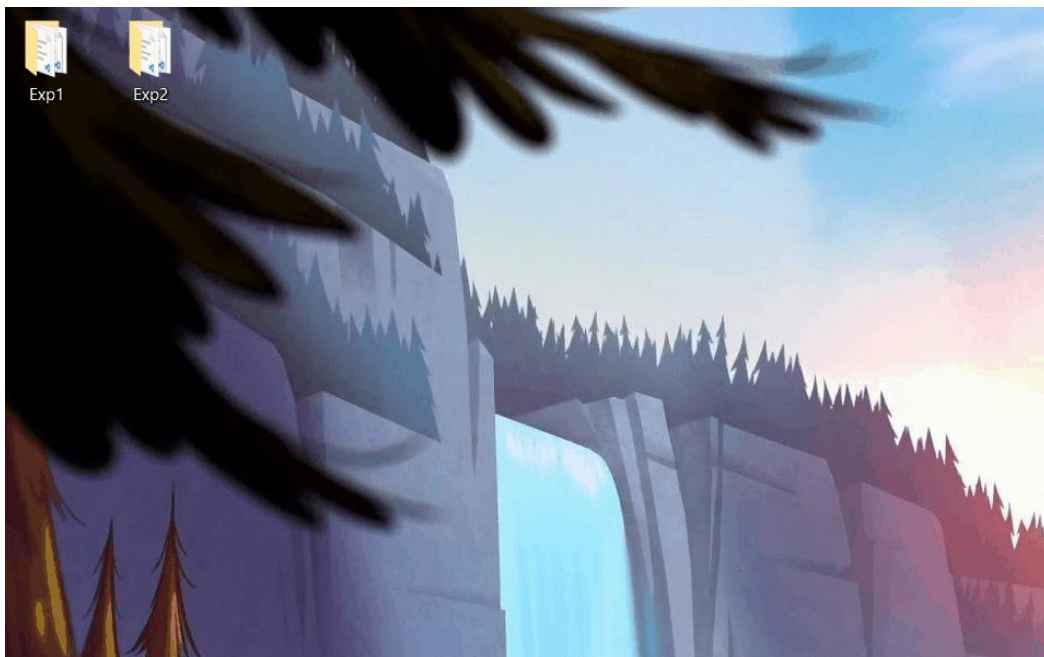


Safari:

Em Safari, deve clicar em “**Ficheiro → Guardar Como**” e deve guardar no formato “**Page Source**” (“**Página Fonte**”)

Como enviar as pastas “Exp1” e “Exp2” para uma pasta comprimida?

Para comprimir as pastas “Exp1” e “Exp2” deve seguir as instruções da imagem abaixo. Se o seu sistema operativo for **macOS** ou **Linux**, em vez de “Enviar para > Pasta Comprimida”, a sua opção será “Compress” (“Comprimir”).

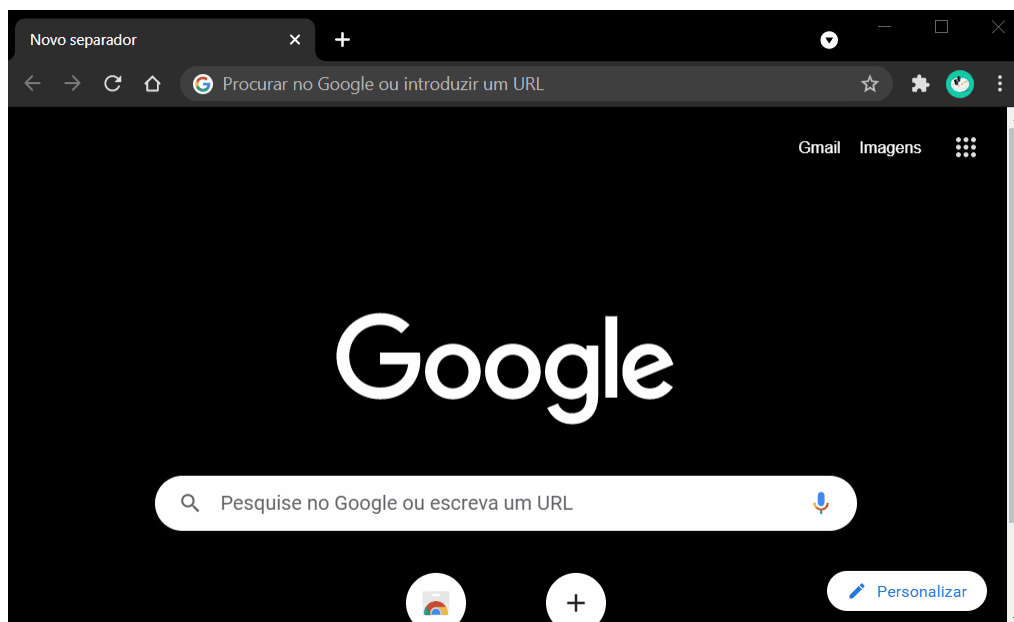


Mover o ficheiro “Cookies” para fora do diretório do seu navegador

1. Em primeiro lugar, abrir uma instância do seu navegador e seguir as instruções abaixo para encontrar e copiar o diretório onde se encontram as cookies.

[Chrome](#), [Edge](#), [Firefox](#), [Safari](#)

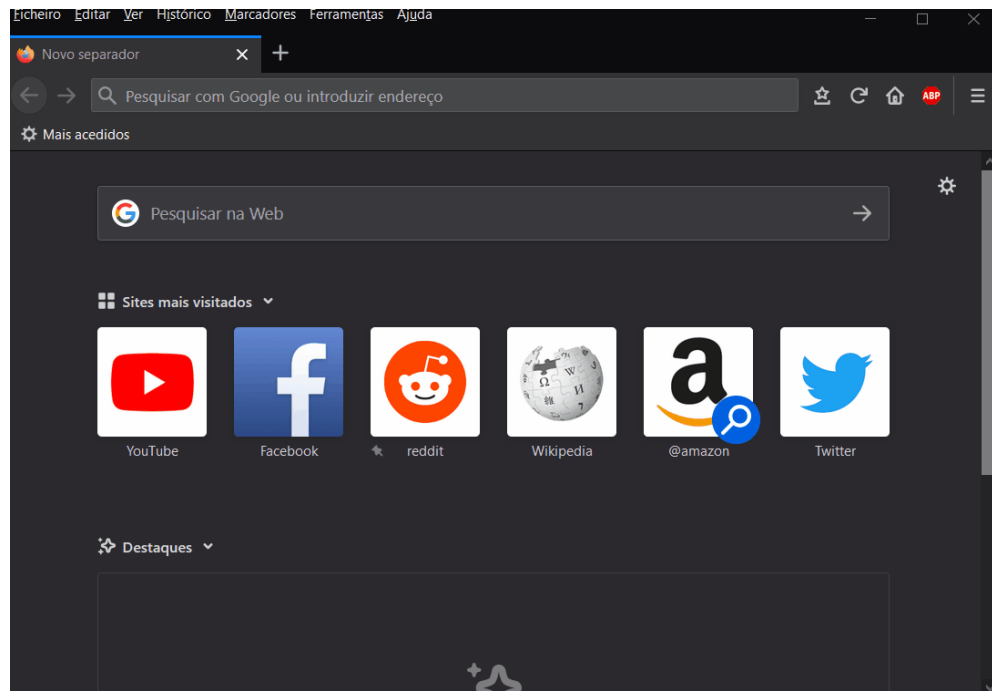
Chrome:



Edge:

(Em [Edge](#), procurar “edge://version/” e proceder do mesmo modo que em Chrome)

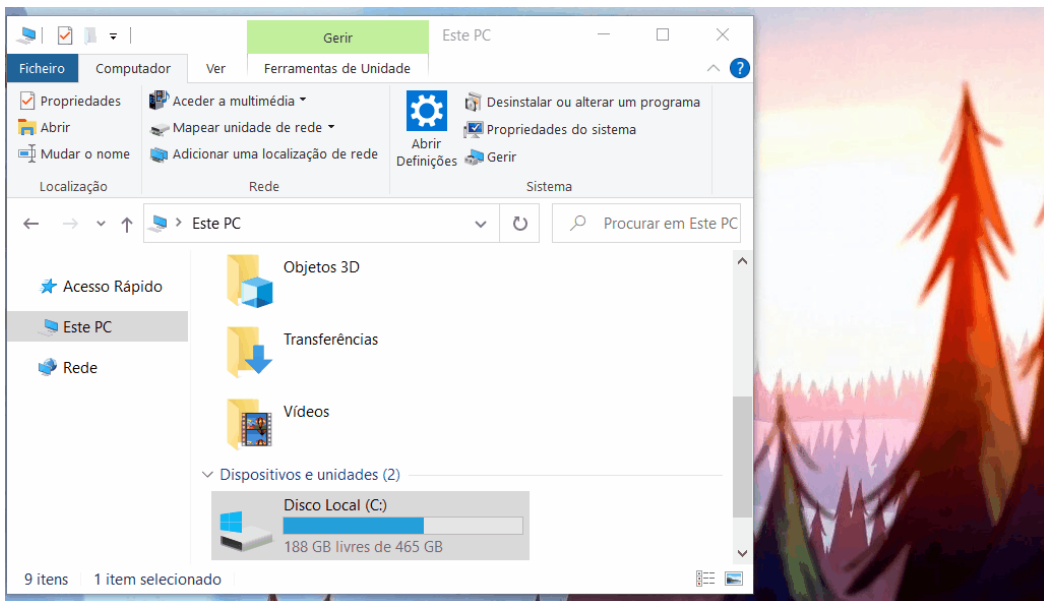
Firefox:



Safari:

Em Safari, o ficheiro de Cookies encontra-se na pasta/diretório "[~/Library/Cookies](#)"

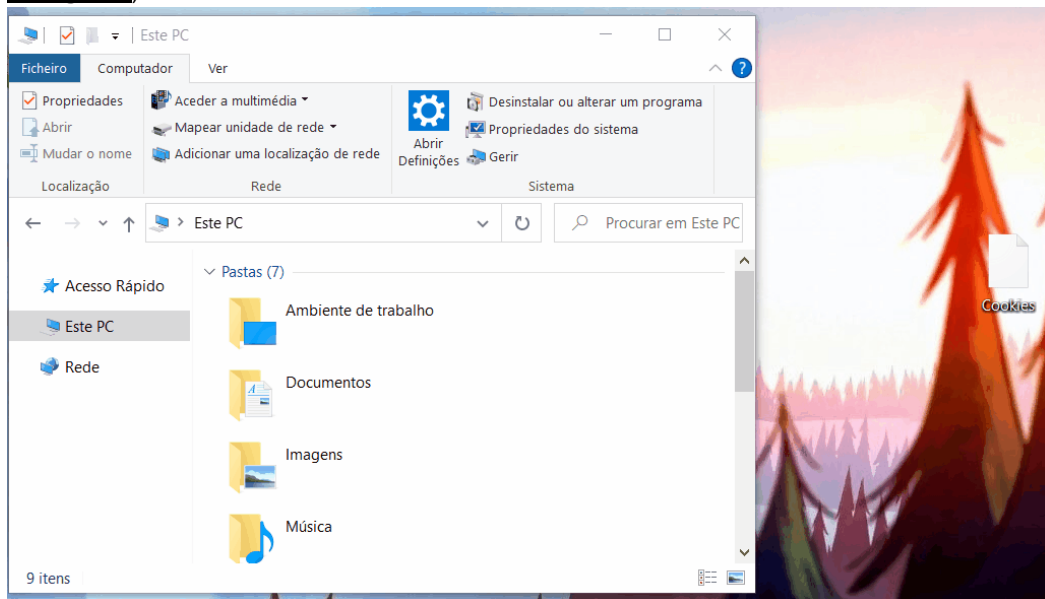
2. De seguida, **fechar todas as instâncias do seu navegador** e mover o ficheiro “Cookies” de acordo com a imagem abaixo (o diretório copiado anteriormente é necessário para aceder à pasta).
(Se precisar de guardar todas as suas tabs abertas antes de fechar o browser, pode usar o comando *Ctrl+Shift+D* (ou *Cmd+Shift+D* em Safari))
- Nota: no caso de **Firefox**, o ficheiro a mover chama-se “**cookies.sqlite**”
- Nota: no caso de **Safari**, o ficheiro a mover chama-se “**Cookies.binarycookies**”



Repor o ficheiro “Cookies” original

(Seguir [estas](#) instruções caso precise de encontrar o diretório do seu navegador novamente, o qual será necessário aqui)

(Antes de seguir as instruções da imagem abaixo, **feche todas as instâncias do seu navegador**)



References

- [1] Rakesh Agrawal, Behzad Golshan, and Evangelos Papalexakis. A study of distinctiveness in web results of two search engines. In *Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion*, page 267–273, New York, NY, USA, 2015. Association for Computing Machinery.
- [2] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. The relationship between ir effectiveness measures and user satisfaction. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07*, page 773–774, New York, NY, USA, 2007. Association for Computing Machinery.
- [3] Ismail Sengor Altıngövdü, Rifat Özcan, and Özgür Ulusoy. Evolution of web search results within years. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11*, page 1237–1238, New York, NY, USA, 2011. Association for Computing Machinery.
- [4] Einat Amitay, David Carmel, Michael Herscovici, Ronny Lempel, and Aya Soffer. Trend detection through temporal link analysis. *J. Am. Soc. Inf. Sci. Technol.*, 55(14):1270–1281, December 2004.
- [5] Ricardo Baeza-Yates and Berthier Ribeiro-neto. Modern information retrieval. 07 1999.
- [6] Ricardo A. Baeza-Yates, Felipe Saint-Jean, and Carlos Castillo. Web structure, dynamics and page quality. In *Proceedings of the 9th International Symposium on String Processing and Information Retrieval, SPIRE 2002*, page 117–130, Berlin, Heidelberg, 2002. Springer-Verlag.
- [7] Judit Bar-Ilan. The lifespan of “informetrics” on the web: An eight year study (1998–2006). *Scientometrics*, 79, 04 2009.
- [8] Steven M. Beitzel, Eric C. Jensen, Ophir Frieder, David D. Lewis, Abdur Chowdhury, and Aleksander Kolcz. Improving automatic query classification via semi-supervised learning. In *Proceedings of the Fifth IEEE International Conference on Data Mining, ICDM '05*, page 42–49, USA, 2005. IEEE Computer Society.
- [9] Tim Berners-Lee and Robert Cailliau. Worldwideweb: Proposal for a hypertext project. *Technical Report, European Laboratory for Particle Physics (CERN)*, 11 1990.
- [10] Krishna Bharat and Andrei Broder. A technique for measuring the relative size and overlap of public web search engines. *Computer Networks and ISDN Systems*, 30(1):379 – 388, 1998. Proceedings of the Seventh International World Wide Web Conference.
- [11] Krishna Bharat and Andrei Broder. Mirror, mirror on the web: A study of host pairs with replicated content. *Comput. Netw.*, 31(11–16):1579–1590, May 1999.

- [12] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web 7, WWW7*, page 107–117, NLD, 1998. Elsevier Science Publishers B. V.
- [13] Carlos Castillo. Effective web crawling. *SIGIR Forum*, 39(1):55–56, June 2005.
- [14] F. Chung. A brief survey of pagerank algorithms. *IEEE Transactions on Network Science and Engineering*, 1(1):38–42, 2014.
- [15] E.G. Coffman, Zhen Liu, and Richard R. Weber. Optimal robot scheduling for web search engines, 1997.
- [16] Kevyn Collins-Thompson, Paul Bennett, Fernando Diaz, Charles L. A. Clarke, and Ellen M. Voorhees. Trec 2013 web track overview. In *Proceedings of the 22nd Text REtrieval Conference (TREC 2013)*, February 2014.
- [17] Miguel Costa, Daniel Gomes, and Mário J. Silva. The evolution of web archiving. *Int. J. Digit. Libr.*, 18(3):191–205, September 2017.
- [18] Silverstein Craig, Monika Henzinger, Hannes Marais, and Michael Moricz. Analysis of a very large altavista query log. *Technical Report 1998-014, Systems Research Center, Compaq Computer Corporation*, 1998.
- [19] Michal Cutler, Yungming Shih, and Weiyi Meng. Using the structure of html documents to improve retrieval. In *USENIX Symposium on Internet Technologies and Systems, USITS'97*, page 22, USA, 1997. USENIX Association.
- [20] Robert Dellavalle, Eric Hester, Lauren Heilig, Amanda Drake, Jeff Kuntzman, Marla Graber, and Lisa Schilling. Going, going, gone: Lost internet references. *Science (New York, N.Y.)*, 302:787–8, 11 2003.
- [21] Konstantina Dritsa, Thodoris Sotiropoulos, Haris Skarpetis, and Panos Louridas. Search engine similarity analysis: A combined content and rankings approach. In Zhisheng Huang, Wouter Beek, Hua Wang, Rui Zhou, and Yanchun Zhang, editors, *Web Information Systems Engineering – WISE 2020*, pages 21–37, Cham, 2020. Springer International Publishing.
- [22] Nadav Eiron, Kevin S. McCurley, and John A. Tomlin. Ranking the web frontier. In *Proceedings of the 13th International Conference on World Wide Web, WWW '04*, page 309–318, New York, NY, USA, 2004. Association for Computing Machinery.
- [23] Dennis Fetterly, Mark Manasse, Marc Najork, and Janet Wiener. A large-scale study of the evolution of web pages. In *Proceedings of the 12th International Conference on World Wide Web, WWW '03*, page 669–678, New York, NY, USA, 2003. Association for Computing Machinery.
- [24] Google. Explore - google trends. Available at <https://trends.google.com/trends/explore>, Accessed last time in Jan 2021, 2021.
- [25] Google. How search organizes information. Available at https://www.google.com/intl/en_us/search/howsearchworks/crawling-indexing/, Accessed last time in Jan 2021, 2021.
- [26] Venkat Gudivada, Vijay Raghavan, William Grosky, and R. Kananagottu. Information retrieval on the world wide web. *Internet Computing, IEEE*, 1:58 – 68, 10 1997.

- [27] A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, WWW '05*, page 902–903, New York, NY, USA, 2005. Association for Computing Machinery.
- [28] Siegfried Handschuh, Steffen Staab, and Raphael Volz. On deep annotation. pages 431–438, 01 2003.
- [29] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. Measuring personalization of web search. In *Proceedings of the 22nd International Conference on World Wide Web, WWW '13*, page 527–538, New York, NY, USA, 2013. Association for Computing Machinery.
- [30] Monika Henzinger. Link analysis in web information retrieval. *IEEE Data Eng. Bull.*, 23:3–8, 01 2000.
- [31] Monika Henzinger. Hyperlink analysis on the world wide web. In *Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia, HYPERTEXT '05*, page 1–3, New York, NY, USA, 2005. Association for Computing Machinery.
- [32] Jimmy, Guido Zuccon, and Gianluca Demartini. On the volatility of commercial search engines and its impact on information retrieval research. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18*, page 1105–1108, New York, NY, USA, 2018. Association for Computing Machinery.
- [33] Kaspersky. What are cookies? Available at <https://www.kaspersky.com/resource-center/definitions/cookies>, Accessed last time in July 2021, 2021.
- [34] Jinyoung Kim and Vitor R. Carvalho. An analysis of time-instability in web search results. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval, ECIR'11*, page 466–478, Berlin, Heidelberg, 2011. Springer-Verlag.
- [35] Sung Jin Kim and Sang Ho Lee. An empirical study on the change of web pages. In *Proceedings of the 7th Asia-Pacific Web Conference on Web Technologies Research and Development, APWeb'05*, page 632–642, Berlin, Heidelberg, 2005. Springer-Verlag.
- [36] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. Location, location, location: The impact of geolocation on web search personalization. In *Proceedings of the 2015 Internet Measurement Conference, IMC '15*, page 121–127, New York, NY, USA, 2015. Association for Computing Machinery.
- [37] Mei Kobayashi and Koichi Takeda. Information retrieval on the web. *ACM Comput. Surv.*, 32(2):144–173, June 2000.
- [38] S. Lawrence and C.L. Giles. Accessibility of information on the web. *Intelligence*, 11(1):32–9, Spring 2000.
- [39] Shao Fen Liang, Siobhan Devlin, and John Tait. Using query term order for result summarisation. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05*, page 629–630, New York, NY, USA, 2005. Association for Computing Machinery.
- [40] B. Miguel and C.T. Lopes. Mutability of web search engines results - data collection and brief analysis [dataset]. Available at <https://doi.org/10.25747/S9YY-W366>, 09 2021. INESC TEC research data repository.

- [41] D.M. Moura. Exploring search engine counts in the identification and characterization of search queries. Master's thesis, Faculty of Engineering of the University of Porto, R. Dr. Roberto Frias, 4200-465 Porto, 2018.
- [42] Mozilla. User-agent. Available at https://developer.mozilla.org/en-US/docs/Glossary/User_agent, Accessed last time in July 2021, 2021.
- [43] Alexandros Ntoulas, Junghoo Cho, and Christopher Olston. What's new on the web? the evolution of the web from a search engine perspective. In *Proceedings of the 13th International Conference on World Wide Web, WWW '04*, page 1–12, New York, NY, USA, 2004. Association for Computing Machinery.
- [44] S. Nunes. State of the art in web information retrieval. 2006.
- [45] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems, InfoScale '06*, page 1–es, New York, NY, USA, 2006. Association for Computing Machinery.
- [46] Yonathan Perez, Michael Schueppert, Matthew Lawlor, and Shaunak Kishore. Category-driven approach for local related business recommendations. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15*, page 73–82, New York, NY, USA, 2015. Association for Computing Machinery.
- [47] Christo Petrov. The stupendous world of google search statistics. Available at <https://techjury.net/blog/google-search-statistics/>, Accessed last time in Jan 2021, 2020.
- [48] Shahzad Qaiser and Ramsha Ali. Text mining: Use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181, 07 2018.
- [49] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975.
- [50] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, August 1988.
- [51] Maya Sappelli, Suzan Verberne, and Wessel Kraaij. Recommending personalized touristic sights using google places. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13*, page 781–784, New York, NY, USA, 2013. Association for Computing Machinery.
- [52] Erik Selberg and Oren Etzioni. On the instability of web search engines. In *Content-Based Multimedia Information Access - Volume 1, RIAO '00*, page 223–236, Paris, FRA, 2000. Le Centre de Hautes Etudes Internationales D'Informatique Documentaire.
- [53] Narayanan Shivakumar and Hector Garcia-Molina. Finding near-replicas of documents on the web. In Paolo Atzeni, Alberto Mendelzon, and Giansalvatore Mecca, editors, *The World Wide Web and Databases*, pages 204–212, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.
- [54] StatCounter. Search engine market share worldwide. Available at <https://gs.statcounter.com/search-engine-market-share>, Accessed last time in January 2021, 2021.

- [55] Thanh Tang, David Hawking, Ramesh Sankaranarayana, Kathleen M. Griffiths, and Nick Craswell. Quality-oriented search for depression portals. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy, editors, *Advances in Information Retrieval*, pages 637–644, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.
- [56] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael A. S. Potts. Information re-retrieval: Repeat queries in yahoo’s logs. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’07, page 151–158, New York, NY, USA, 2007. Association for Computing Machinery.
- [57] Vikas Thada and Vivek Jaglan. Web information retrieval. *International Journal of Computer Applications*, 76:29–32, 2013.
- [58] R. Umagandhi and A.V. Senthil Kumar. Query recommendations and its evaluation in web information retrieval. *ICTACT Journal on Soft Computing*, 5(3):991–8, 04 2015.
- [59] Wikipedia. Usage share of web browsers. Available at https://en.wikipedia.org/wiki/Usage_share_of_web_browsers, Accessed last time in April 2021, 2021.
- [60] Worldometer. Regions in the world by population (2021). Available at <https://www.worldometers.info/world-population/population-by-region/>, Accessed last time in January 2021, 2021.