Your education platform, powered by

wizenoze

Wizenoze Search White paper

Written by Thijs Westerveld

About the author

Thijs Westerveld PhD

Thijs Westerveld, Director of Science at Wizenoze has has a Ph.D. in information retrieval and a background in search technology and information retrieval. He spent 20 years between academia and Industry and studied at the University of Twente in the east of the Netherlands. Thijs also spent five years doing research at the Center for Mathematics and Computer Science.

Contents

Why do learners find search engines difficult to use? 3

How do search engines work 4

Why is Wizenoze different 7

How does Wizenoze work 8

How does the Wizenoze machine learning model work? 10

How Wizenoze makes search easy for learners 11

Why do learners find search engines difficult to use?

The internet can offer an incredible wealth of information for learners around the world to strengthen them in their understanding of science, democracy, economy, ecology, and freedom. However, it is very hard for learners from primary all the way up to graduate students to find relevant, reliable and readable information in the 'information jungle' that the internet has become. Especially when they use a general-purpose search engine, such as Google.

Most information on the internet is not relevant nor suitable for educational purposes, or too difficult to read. Evaluating the reliability of sources on the internet is also very difficult for most learners.

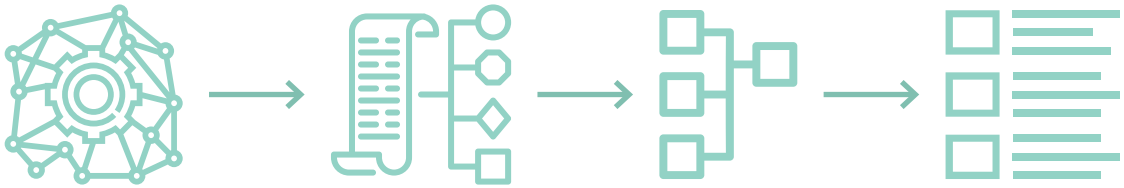
Additionally, learners find it hard to formulate and reformulate queries to get good search results. In a Wizenoze study on primary students' search strategies and difficulties, children described four distinct difficulties in searching for information on the internet: creating search queries, spelling search terms correctly, selecting an appropriate website from a search list, and understanding the language used in search results.

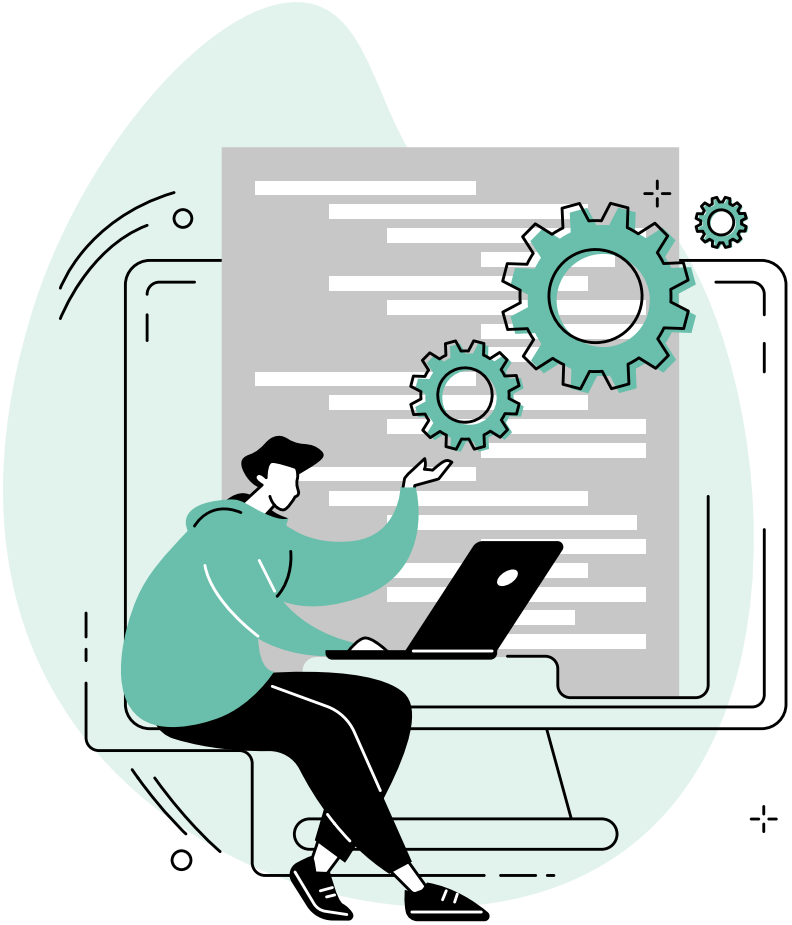
Before we delve into the solutions to overcome these difficulties, let's take a look at how search engines actually work.

How do search engines work?

If you're using the internet, it's more than likely that you're familiar with how to use search engines. You enter the search query, hit enter and a set of results will appear in less than a second. But how does this actually happen? How do search engines such as Google, Bing and Duckduckgo, look at billions of websites and identify the top ten web pages?

Put simply, every search engine has its own copy of the web and most of the work has been done before you even enter your search query. The information can be stored very compactly and efficiently making it possible for the complexity of the work to be completed in advance. The process is broken down into the following stages: collection, extraction, categorisation, and indexing.





How do search engines work?

a. The collection stage

Robots collect pages from the web and store them locally. This process starts from a single or group of URLs. The robot asks for the HTML, downloads it locally, and analyses the HTML file. Within the file, robots look for hyperlinks to other pages on the web. The process is repeated for these pages and this is how it grows and collects more information from the web. All the pages found, from the very first to the last, are downloaded and stored locally.

b. The extraction stage

Once the data has been collected, the HTML file is analysed in order to identify and extract the main article. Text analysis algorithms identify which parts of the page to keep or ignore. For example, a list of keywords may indicate a site menu which should be ingored. On the other hand, a longer block of paragraphs with running text indicates an area of importance, possibly holding the main article. Characteristics like these help our algorithms to identify the main content on the page and ignore menu items, advertisements and references to other articles.

c. The categorisation stage (or metadata extraction)

So what happens next to the information that is found through the page analysis? Robots categorise the information gathered from the analysis. Information can be categorised as:

- Title
- Author
- Text
- Images and videos (multimedia)

d. The indexation stage

Each category of information is stored as a separate component. Next the analysis draws information from the title and main body text and keeps track of the pages where specific terms and phrases appear. They list terms used within the text which are then stored to provide information about the web page - this provides a similar function to the index that you'll find in a book.

What happens to all of the lists?

This process is repeated for all the pages that are collected as a result of that initial HTTP request. The term lists are merged into an index so a robot can quickly find the URLs where a term is mentioned. This means when you enter a query it is not necessary to examine all pages at the moment you type your query. Instead your query is efficiently cross-referenced against the index, providing an immediate list of matching hits. The only intelligence required at this stage is to order the URLs from best to worst match.

How do big search engines like Google and Duckduckgo decide on the best hits?

This algorithm really is the secret sauce of any search engine!
The common factors are based on the key components mentioned on the previous page under C. The Categorisation stage. above. A good result will include:

- All the query terms
- Many of the query terms occurring together in a phrase
- The query terms appear in the important parts of the document (e.g. the title or body text)

What is an example of a poor quality result or hit:

- Query terms are not in the title
- There are fewer occurrences of the query term
- Query terms are separated
- Only one query term matches

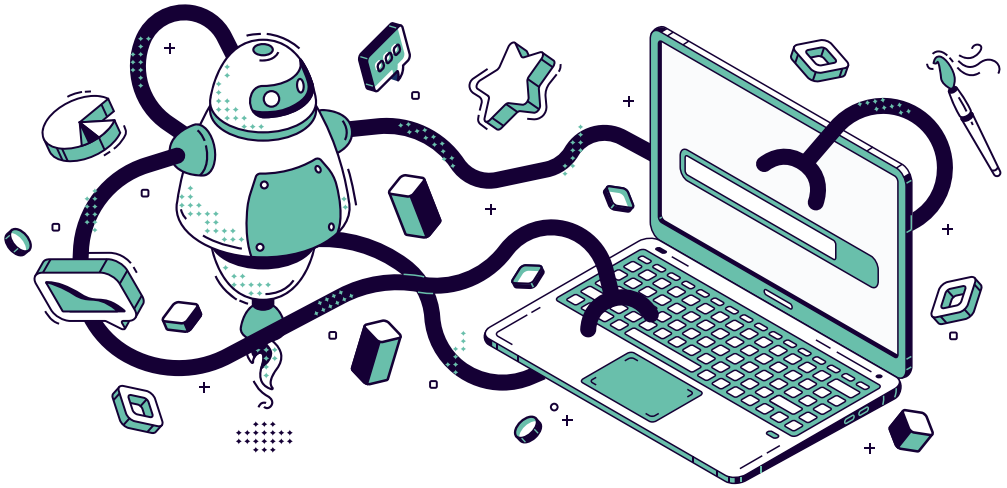
Now that you have a better understanding of how information retrieval and search works, let's take a look at how a Wizenoze search is different from other search engines and why it is a better fit for learners than commercial search engines.

How does Wizenoze technology decide on the best hits?

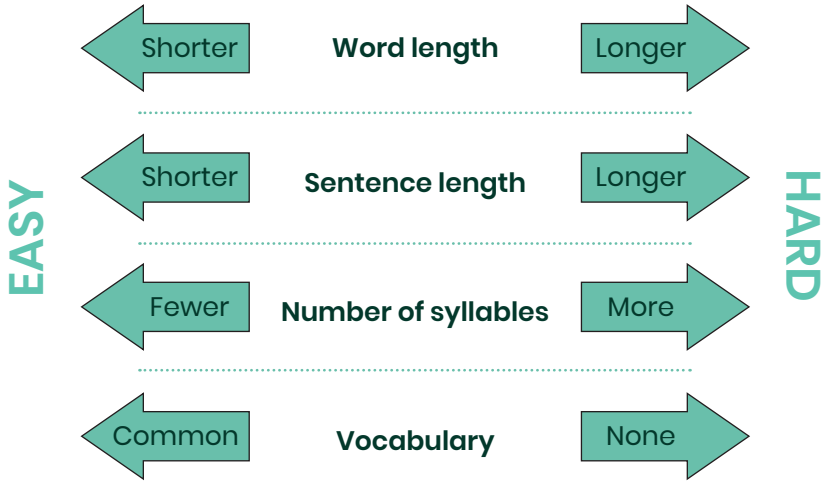
Wizenoze specialises in information for learners in educational environments so it is important that learners have relevant and reliable information. It also needs to be understandable from a readability point of view. There are many different levels of students, and Wizenoze is able to differentiate between results to match these differences. From a Wizenoze perspective, a best hit is not just based on the common factors and components above. It is fine-tuned to meet the individual needs of the learner and it all starts with readability.

Traditional readability metrics are based on just a few simple characteristics.

Since the 1940s linguists have been developing ways to label the readability of textual documents. Traditional readability metricsare based on just a few simple characteristics.



Traditional readability metrics are based on just a few simple text characteristics



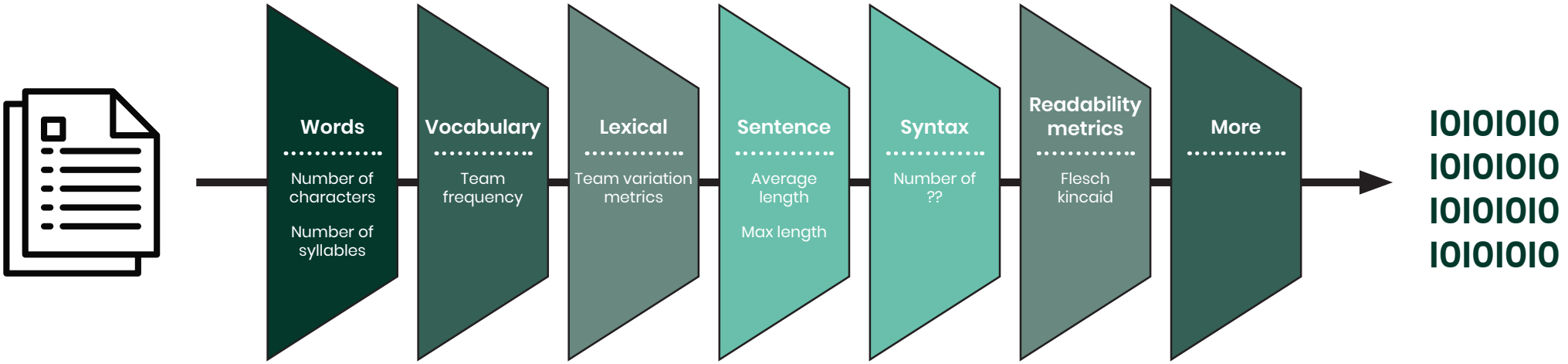
How does Wizenoze technology decide on the best hits?

At Wizenoze, we think that traditional readability metrics represent a limited and biased view of readability classification. For example, if we simply applied the metrics above, if the name of a character in a story changed from four syllables to one, (for example, Donatello to Don!) this would increase the readability level of the story. However, the context of the story still may be too complex for someone to read.

As well as the traditional readability metrics, Wizenoze analyses many other features that influence readability. In fact, Wizenoze algorithms look at 100+ characteristics that are computed from each text.

"In fact, Wizenoze algorithms look at 100+ characteristics that are computed from each text."

We compute a long list of readability features from each document



The above features are stored in a computer-readable format. Wizenoze then uses machine learning to establish a model to determine what is an easier or harder level of text. In this next section, we dive deeper into the AI and machine learning model that supports Wizenoze technology and takes a student's learning to the next level.

How does the Wizenoze machine learning model work?

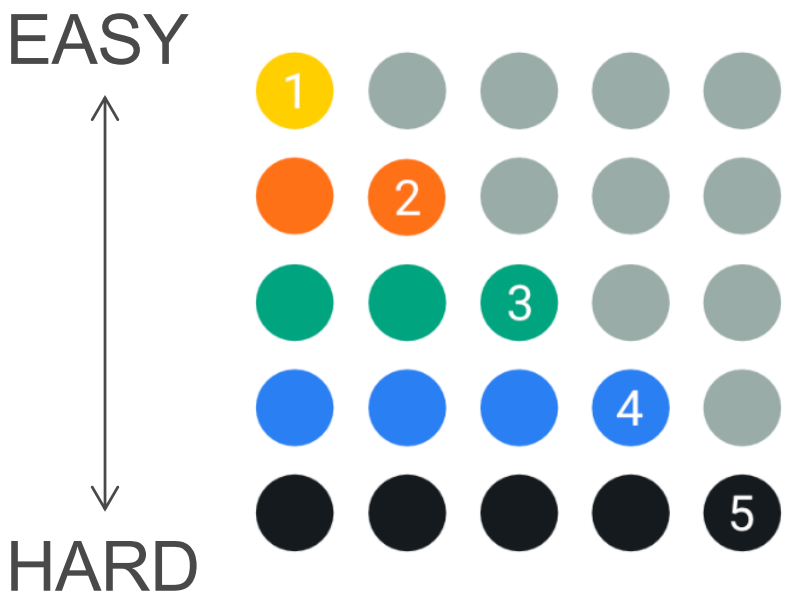
We have trained a machine learning model to determine readability. We started this process by looking at the objective and creating a readability index which provides a scale of one to five, one being the easiest reading level.

To train the machine learning model a set of texts with known readability levels is introduced These readability levels correspond to the readability index and these texts become the examples for the machine learning algorithm. From this point, all texts are compared to the reading levels of these set texts which have been assessed by humans. The readability features (See the readbility index on the right.) in these set texts are computed. The text is then transformed into a set of computer-readable features. This set of features is used to train the machine learning model so that we know which of those features are most important in determining readability. Machine learning can systematically analyse and label text.

This algorithm is then used to classify new documents with an unknown readability classification.

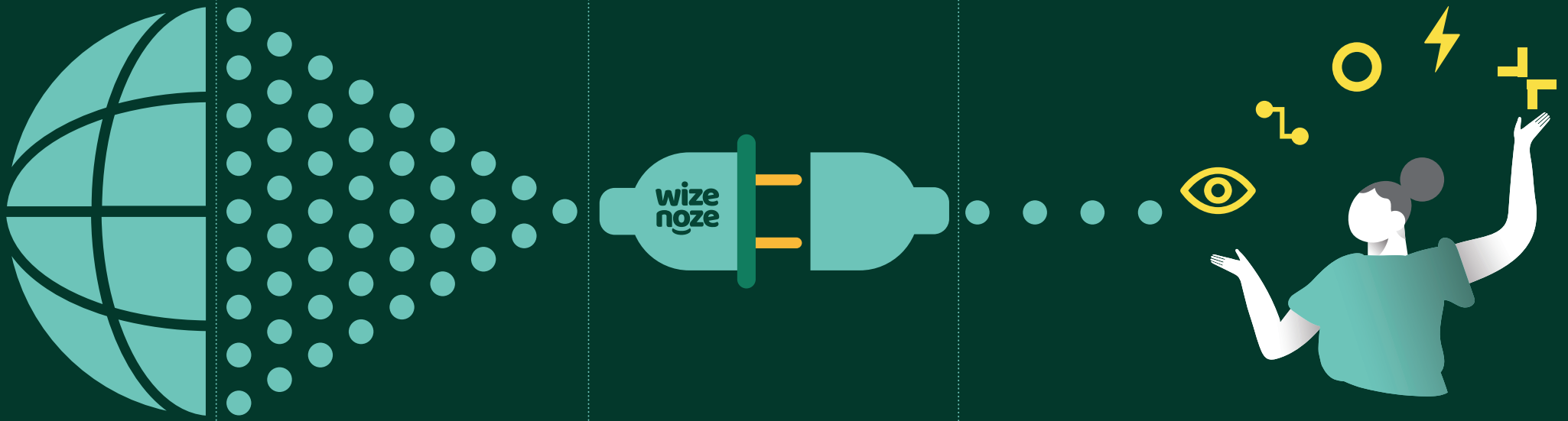
The outcome of this is that Wizenoze is able to indicate to users the reading level of specific sources that appear in a search query term as well as the best hits that other search engines will rank in a search.

The Wizenoze readability Index:



How does Wizenoze make using search easy for learners?

Just as Wizenoze was born from research, we are equally committed to demonstrating evidence-based learning outcomes. To find out more about the impact of Wizenoze solutions on learning, read our latest research: Supporting self-directed learning and exploratory research on the web.



The Internet

Our technology searches the internet for relevant resources.

Filtered content

A.I. technology and our experts determine relevance, reliability,

Bespoke integration

Our technology is then integrated into your platform, specifically matched to your curriculum.

Engaging experiences

Your platform is enriched by giving users access to the world's largest collection of curated educational materials, carefully matched to their needs.

Your education platform, powered by

wizenoze

About Wizenoze

Wizenoze gives learners access to the world's largest curated educational collection of trustworthy internet resources and proprietary content, matched to reading age, curriculum; integrated and fully tailored to specific and regional needs.

Our collection is available via an integration in a LMS or education platform, perfectly matched to the customer's needs. We are always looking for new and trusted (proprietary) content to add to our collection. If your content is digitally available it is easily integrated in the Wizenoze education collection and made available to our customers. Wizenoze have already partnered with Britannica to include over 100,000 multimedia resources and articles across its K12 offering.

If you would like to increase your range of customers it is time to get your collection integrated in the Wizenoze collection.

Founder - Diane Janknegt: diane@wizenoze.com
Country Director - India: Suchindra Kumar: suchindra@wizenoze.com
MENA : Fadi Khalek: fadi@wizenoze.com

www.wizenoze.com