Automated Privacy Compliance Auditing as a Service (DRAFT)

Benjamin J. Anderson

University of Wisconsin - Stephens Point

Stephens Point, Wisconsin

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree Of

Master Of Science

in Data Science

December 2020

For my wife and son.

Contents

1 Executive Summary

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum (Anderson, 1983).

2 General Company Description

Cereus is a software-as-a-service company offering privacy compliance auditing tools and solutions for businesses that collect customer information through their websites.

2.1 Unique Value Proposition

Consent involves more than just cookies on your website. It involves all information you share with your vendors and partners including network traffic. Cereus actively scans, monitors, and analyzes your website to identify compliance infringements in accordance with your business rules.

2.2 Mission Statement

To remove privacy barriers between companies and their partners, and enable them to communicate in an efficient, compliant way.

2.3 Vision Statement

Proactively detect all privacy compliance infringements defined by our clients before they reach their customers.

2.4 Values Statement

- Reliability
- Transparency
- Growth

2.5 Company Goals and Objectives

The primary goal of Cereus is to establish itself as a thriving company that leads the privacy industry by providing automated, insightful, audits to ensure that companies and their partners are sharing customer information in accordance to their business rules.

2.6 Business Philosophy

Some companies spend millions to establish a privacy program. Everyone else gets Cereus.

2.7 Industry Overview

The privacy technology industry is a rapidly growing field. In 2020, it is the fastest growing technology sector that includes the fastest growing company in the U.S (Hughes, 2020). As more consumers become impacted by massive data breaches in which sensitive, personally identifying information (PII), is exposed; consumer awareness and the call for data processing regulations are expected to be on the rise.

There are already governmental regulations impacting businesses in the U.S. The Health Insurance Portability and Accountability Act (HIPAA) highly regulates patient information and how it is stored (for Disease Control & Prevention, 2018). The California Privacy Protection Act (CCPA) regulates the selling of user data collected by a business for consumers in the state of California (California, 2018). The Children's Online Privacy Protection Act (COPPA) imposes requirements on website operators on collecting information from children under the age of 13 years old (FTC, 1993). Lastly, the General Data Protection Regulation (GDPR) gives European Union citizens the right to manage their information any business has collected on them and requires explicit consent before information is collected (of the European Union, 2018).

With all these regulations, foreign and domestic, companies that collect information from their customers are relying on the assistance of privacy technologies to operate within the bounds of new regulations and meet consumer privacy expectations (Meehan, 2019). This has attracted massive funding, and, in July of 2019, OneTrust raised $200 million in a Series A investment, TrustArc raised $70 million Series D, Privitar raised $40 million series B, and BigID raised $30 million Series B (Wood, 2019).

In four years since its founding, as of August, 2020, OneTrust is valued at $2.7 billion (Hughes, 2020).

2.8 Market Segment Overview

Any business that operates a website and collects data and analytics on their customers is subject to the regulations in which the customer originates. This also applies to the jurisdiction in which the business operates. Cereus can provide auditing for companies with a single website, to large enterprises with hundreds. It may prove difficult for small businesses with a single website to justify the expense of privacy audits when they are not often the subject of privacy lawsuits (LaNou, 2020). Cereus's primary focus will be medium to large organizations maintaining multiple websites.

2.9 Legal Form of Ownership

Cereus will be organized as a small business corporation (S corporation). As an S corporation, Cereus will be allowed to collect funding and pass off any business profits and expenses to its shareholders without the additional taxes applied to C corporations.

2.10 Guiding Principles

The "Living Principles for Design" framework (Hamlett, 2020) was applied to outline how Cerues can maintain a sustainable design while achieving the company's objectives along the following dimensions:

2.10.1 Environment

The direct environmental impact of Cereus is expected to be minimal. Cereus will provide software-as-a-service (SAAS) and will rely on cloud services to manage its operations. Cloud services, such as Amazon Web Services (AWS), are composed of large computer networks in which infrastructure is shared with other AWS customers (Amazon, 2020). Cereus's physical hardware is limited to the machines required to manage services running on the cloud platforms.

With Cereus's services operating in the cloud, a central office space for employees is not required and will further reduce the company's environmental impact.

2.10.2 People

The societal impact of Cereus and its services are restricted to the transparency of the companies that use it. Cereus offers detailed reports from its audits that can provide insights into how customer information is shared between a website and its partners. If companies choose to share these reports, their customers will better understand how their information is used in exchange for the services the company provides. This has the potential to improve the relationship between a company and their customers – possibly making them more apt to sharing personally identifying information.

2.10.3 Economy

Cereus's operations are expected to reduce the amount of time required to conduct a compliance audit against websites. These actions will minimize the manual auditing cost and likelihood of Cereus's customers being subject to privacy lawsuits. Cereus's customers can

then focus and dedicate more resources towards achieving their goals and growing their business. The overall economic impact of Cereus is limited to the the actions of its customers and is expected to be minimal.

2.10.4 Culture

Cereus has the potential to influence organizations to be more transparent about the sharing of information on their customers with their partners. Traditionally, data processing and sharing are often confidential and kept internal; but with privacy becoming a concern for consumer – transparency will soon be an expectation (Meehan, 2019).

3 Products and Services

3.1 Description of Products and Services

Cereus aims to become the leader in cutting-edge privacy compliance auditing tools. These tools will assist our customers to quickly and efficiently identify privacy and compliance issues on their websites. Most of Cereus's solutions will be offered as software-as-a-service, though professional services will also be available.

3.1.1 Professional Services

Cereus's professional services will serve a less technical clientele or those requiring guidance on privacy regulations applicable to their operations. Cereus will consult configuration the client to identify the best plan for the client and assist with the configuration of Cereus's tools to reflect the client's needs. Training services for Cereus's products will also be provided.

3.1.2 Compliance Auditing

The Cereus compliance auditing system is a series of processes configured by the user to ensure their websites meet the compliance standards that they have defined. It can be configured to scan per the businesses development cycle, automatically through the API services, or manually through the Cereus user interface. There are five main components to the compliance auditing system: the confirmation system, crawler, rules engine, report, and recommendation engine.

3.1.2.1 Confirmation System

The Cereus confirmation system, Figure 3.1, prevents organizations from conducting audits on domains that they do not own or manage. This will ensure Cereus's customers cannot use the auditor to evaluate their competitors websites and privacy practices. When a new property (website) is added through the Cereus user interface, the user will be guided through a series of processes to confirm ownership of the domain before conducting an audit is authorized. This process is dependent on the plan the customer is subscribed to. Any attempt to scan an unconfirmed domain will be rejected.

The free tier is a highly restricted plan that limits the capabilities of the auditor. This tier is intended for less technical website administrators running a small website through a content management system (CMS). For initial confirmation of ownership of the domain, free tier users will be required to own an email address associated with the domain they are requesting to audit. In many instances, an email such as "webadmin@example.com" are

dedicated to the management of the domain. Once the customer creates the property in the Cereus user interface, they can then request a validation email be sent to their inbox with a confirmation link. When confirmed, the customer will be provided an HTML metadata tag to be included on the pages they would like to be scanned.

In the event the property on the free tier expires or is sold, Cereus will no longer be able to audit the website due to the metadata tags not being present on the site.

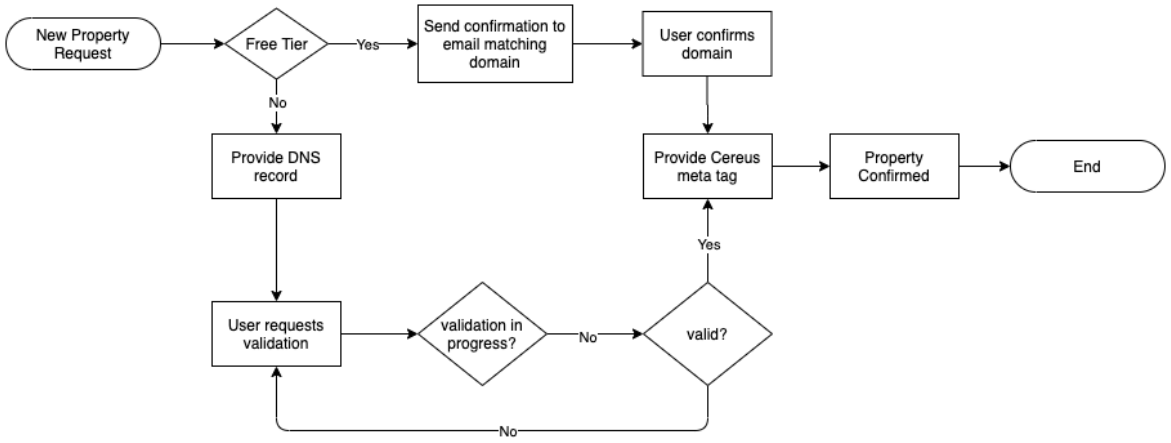

Figure 3.1: Property confirmation system - The Cereus website confirmation system to ensure the organization owns the domain prior to scanning.

All additional tiers offer unrestricted access to Cereus's services, which can provide insights into the privacy operations of the company. To confirm ownership of the property, a TXT DNS record defined by Cereus's systems will be provided to the client. In the event a domain is acquired by a new party, through the sale or expiration of the domain, Cereus will lock access to previous reports and disable auditing services when the TXT DNS record is no longer present. The initial client who set up the domain in Cereus will be notified of the change and can re-validate if the DNS record was removed by mistake.

3.1.2.2 Crawler

The Cereus crawler is a highly configurable network and cookie monitor that extracts information from websites. This includes all network traffic, cookies set on the page, call-trace information, performance data, request type, response status codes, headers, and redirect flags. This data is then processed by the rules validator and aggregated by the report generator. Cereus can then make recommendations on actions to take based on the audit.

Configuration

The crawler can be configured by the client to automatically scan a website based on the company's software release schedule. Once the data has been processed by the rules validator and the report has been generated, the customer will receive an email notification with the audit report. Audits may also be conducted manually through the Cereus user interface or triggered through the API services.

Cereus's customers may also specify a geographic location for which the crawler will originate from. Depending on the geographic location of the visitor, privacy regulations can differ and the company may apply a different set of business rules. Running the crawler in a targeted geographic location allows Cereus's customers to validate their website's behaviors.

Data Extraction and Transformation

The network and cookie data are extracted in a semi-structured format that can be translated to a flat SQL tables for processing by the report generator. Data received by the proxy server can be consolidated with debug information sent by the browser. Figure 3.2 outlines the data collection and transformation process.

The crawler captures the requested URL for users to construct rules on the domains, locations, and query parameter. The requested URL is broken down into the requested protocol, domain, path, query, and location hash. This fragmentation will significantly reduce the amount of processing required by the rules and recommendation engines. It also allows customers to optimize their reports by establishing aggregation or exclusion rules based on part of the URL. In the instance data is sent to the server, for example, in a POST request: the data is converted to a HashMap and stored as JSON in the metadata column.

All request and response header information, depending on the browser meta event, is formatted as JSON and stored in the headers column. Signals such as the Do Not Track header signal (DNT) can be sent from the browser to indicate that the user would prefer privacy rather than personalized content (Mozilla, 2020). Site partners may respond to the DNT header, or some other setting, that the customer can monitor with the Cereus rules engine.

3.1.2.3 Rules Engine

Organizations have the option to establish rules associated with a network request or cookie to determine whether or not they meet compliance standards. These rules can be configured to target specific geographic location to determine if an entire URL, domain, protocol, query path, or parameter, based on the specified condition, meets compliance expectations (Figure 3.3).

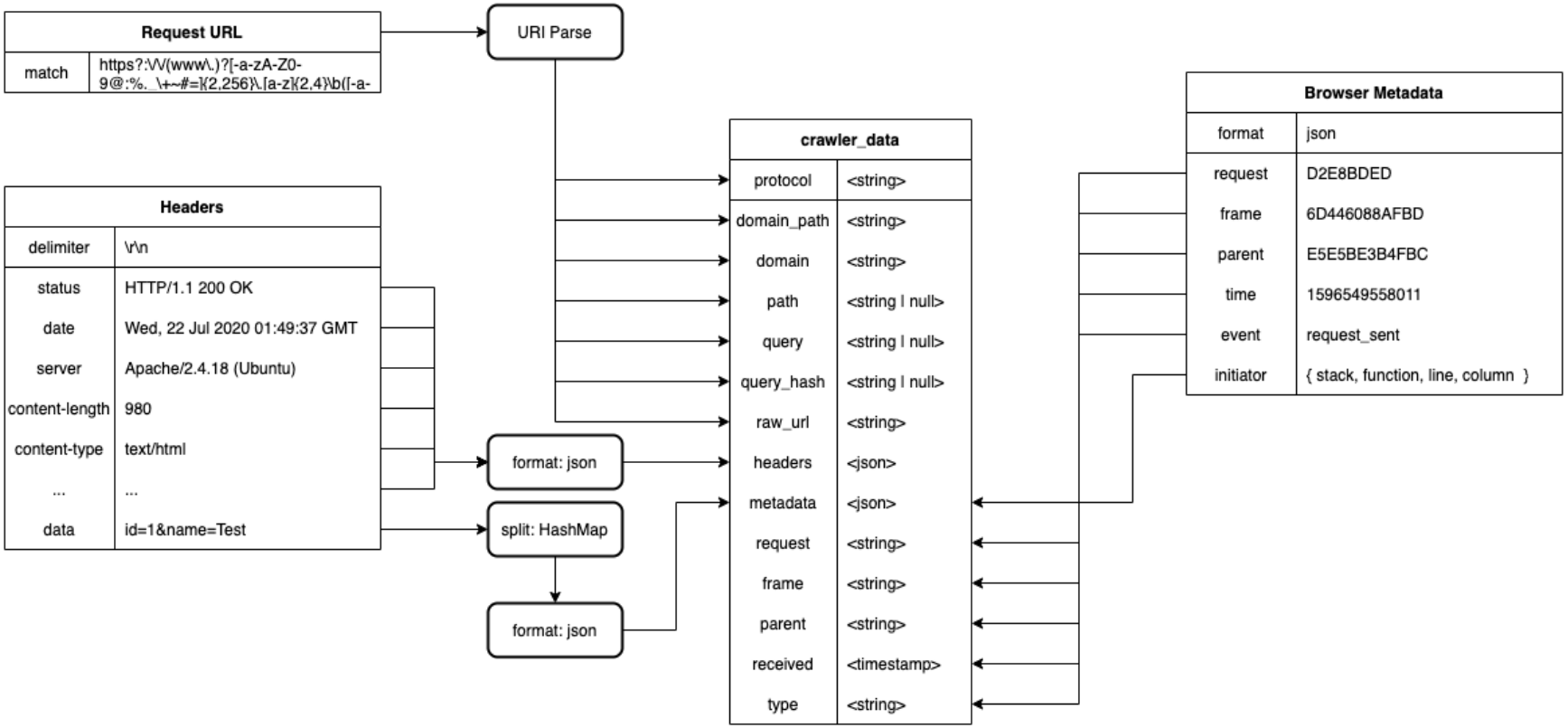

Figure 3.2: Crawler data transformation - The extraction and transformation of data received by the Cereus crawler.

Multiple conditions may be applied to a single URL or cookie as an OR conditional. Rules may also be grouped to establish an AND conditional between two or more rules. Cereus's rules engine supports matches for values that: contain a value, equal a value, does not contain a value, does not equal a value, or whether or not the value matches a regular expression. Based on specified conditions, the user can specify whether or not to flag the request as compliant or not.

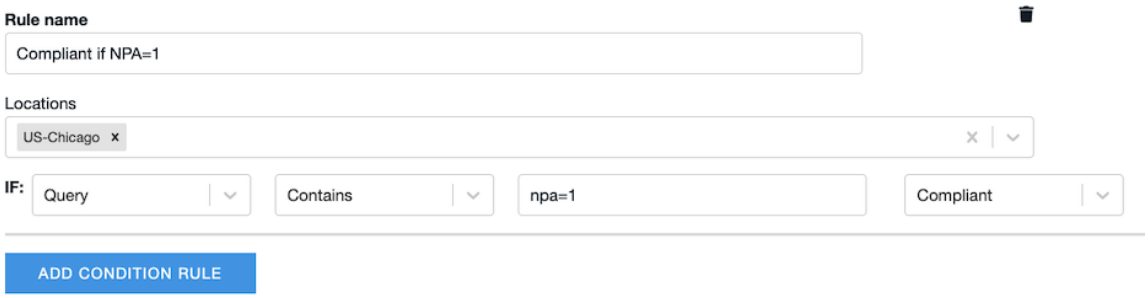

Figure 3.3: Rules definition interface - The Cereus rules engine can be configured to determine whether or not a request meets compliance standards based on the conditions specified by the user.

These rules, by default, are set at an organizational scope. All properties under the organization, when the rules engine processes a crawl, will have the same rules applied. Rules may also be overrode at a property level when exclusions are needed.

3.1.2.4 Reporting

The Cereus audit report formats the network traffic and associated cookies in a clean, tabular, format. This initial overview provides insights into the requests made on the site, response status code, the type, size of the information exchanged, the amount of time for the request to complete, and whether or not the request met compliance expectations (Figure 3.4). Any information that has no data or rules associated with it appears as a question mark (?) icon.

Each row can be broken down to dive into the information associated with the request. The deep dive includes the headers associated with the request, a cache of the original rules associated with the request (and their validation status), query parameters, data sent to the server, cookies, and the initialization chain. Web administrators and compliance managers will be able to quickly reference the audit report and identify where the compliance infringement originated.

The report filters can be used to dynamically query information from it. Users can check for specific URLs, whether or not requests were compliant, the associated category with a URL, or the page in which the information was found. Customers will receive a notification

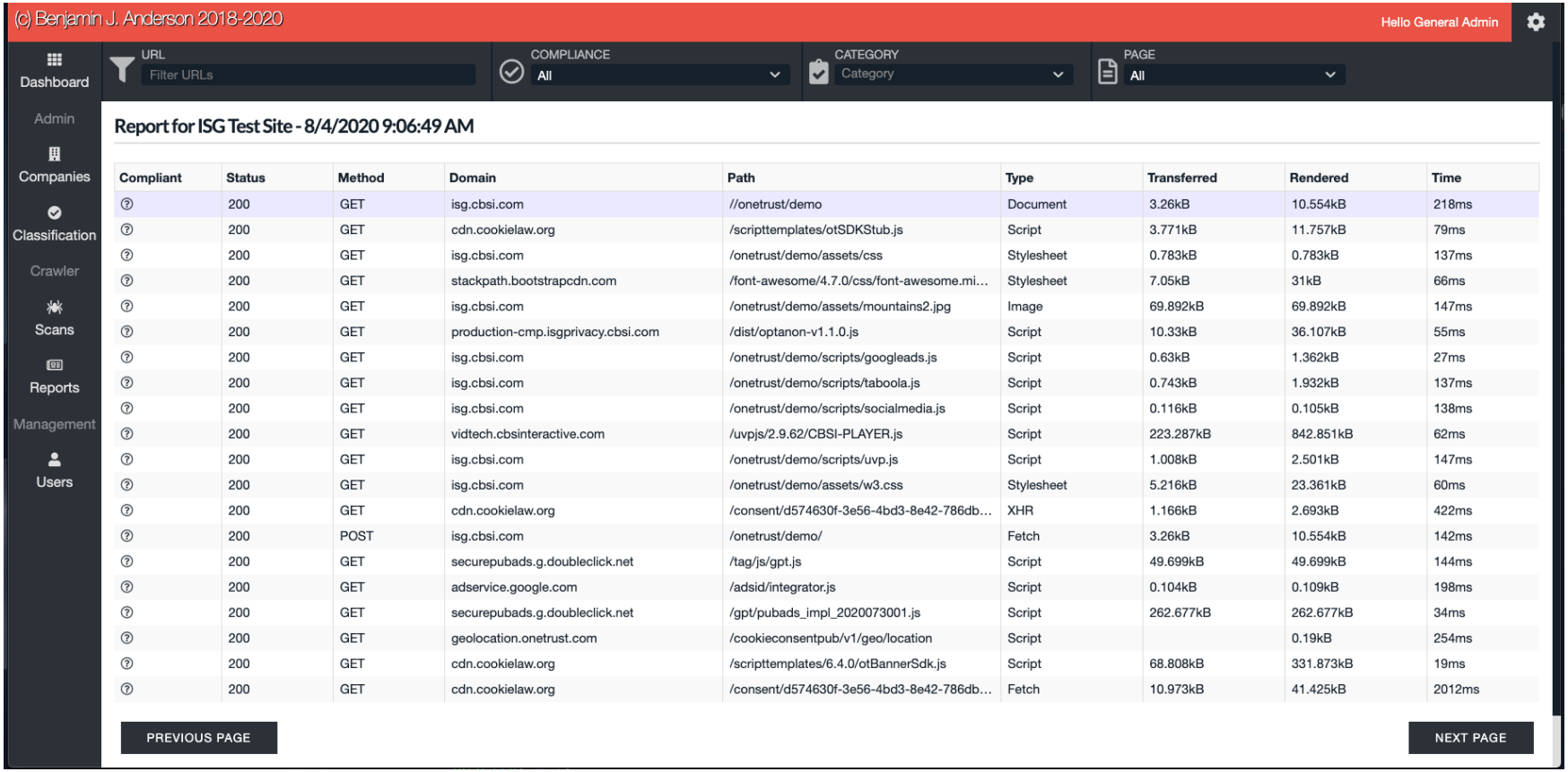

Figure 3.4: Cereus audit report - The Cereus audit report provides clean, familiar, interface for privacy compliance teams.

when the audit identifies requests out of compliance and can use the filter functionality to quickly pull up the request and rules information.

3.1.2.5 Recommendation Engine

Cereus can make suggestions for requests and cookies that have yet to be classified by the organization. This classification system is powered by the categorizations of requests and cookies by other organizations. The system will also crawl the domains, paths, and cookie hosts to grab meta information to improve the accuracy of the recommendations. A risk score will also be assigned to the requests and cookies based on whether or not other organizations have flagged it as necessary to their operations.

Risk Score

When a new request or cookie is identified on a website, through an initial or later scan, an associated risk score with allowing it to load in a list of geographic locations. This score is merely a suggestion based on the operations of other organizations and no action needs to be taken.

$$s_{risk} = \frac{r_n}{R}$$

The risk score is computed as the number of records classified as necessary r_n divided by the total number of records R. This will always result in a ratio between 0 and 1 in which intervals of $\frac{1}{3}$ will determine if the request will be rated as: low, intermediate, or high risk.

To reduce the possibility of organizations incorrectly flagging a request due to the scoring system, the risk score will only be provided when a sample size of at least 20 organizations have classified the request or cookie.

Categorization Recommendation

Cereus will provide categorization recommendations for requests and cookies based on meta information extracted from the request or cookie's origin. The recommendation engine will also incorporate organization classification information pertaining to the request or cookie. Classification within Cereus's internal systems are expected to be single words or small phrases, much like meta tag information present on a web page. This information can best be represented as a bag of words. There's no context to meta tag data or the collection of classifications entered by users, so the representation of language or order has no meaning (Manning, 2008).

The recommendation engine is powered by a Bernoulli document model. This model takes a document and partitions it into a feature vector of binary elements. If a word is found in

the document, it will receive a value of 1, otherwise 0. This document model does not take into account the frequency of a word, but whether or not the word is present. This allows us to calculate the probability of a word occurring in a document with a specific classification, as well as taking into account the probability of it not occurring.

To save reduce the computational power required to provide recommendations, the model calculates estimated probabilities for a request or cookie belonging to a category.

We'll let $\hat{P}(w_t|C_k)$ define the estimated probability that the word w_i occurs in a document, D, with the classification (C) k. The estimated probability that the word w_i not occurring is $1 - \hat{P}(w_i|C_k)$. V will represent our models vocabulary and v consist of the feature vector of our document. The product of the probability of each item (i) in our feature vector occurring or not occurring will determine the overall estimated probability of our document being classified as class k $(\hat{P}(v|C_k))$.

$$\hat{P}(D|C_k) = \prod_{i=1}^{V} [v_i \hat{P}(w_i|C_k) + (1 - v_i)(1 - \hat{P}(w_i|C_k))]$$

There are two parameters for this model: the probabilities of each word in the document class $(\hat{P}(w_i|C_k))$ and its prior probabilities $P(C_k)$. We can estimate probability that a word w_i occurs in a document is the number of documents n classified as k divided by the total number of documents N classified as k.

$$\hat{P}(w_i|C_k) = \frac{n_k(w_i)}{N_k}$$

Where the prior probability of class k can be estimated as the relative frequency of documents containing class k.

$$\hat{P}(C_k) = \frac{N_k}{N}$$

3.1.3 Notifications and Alarms

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

3.1.4 API Services

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco

laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

3.2 Competitive Advantages and Disadvantages, Company Strengths, and Core Competencies

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

3.2.1 Unfair Advantage

Todo.

3.3 Pricing Structure

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

3.3.1 Free

Todo.

3.3.2 Standard

Todo.

3.3.3 Professional

Todo.

3.3.4 Enterprise

Todo.

3.4 Industry Background

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

3.5 Target Market Segment

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

List of Figures

References

Amazon. (2020). *What is aws.* Retrieved 2020-09-23, from https://aws.amazon.com/what-is-aws/

Anderson, J. R. (1983). *The architecture of cognition.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

for Disease Control, C., & Prevention. (2018, 09 14). Retrieved 2020-09-30, from https://www.cdc.gov/phlp/publications/topic/hipaa.html

FTC. (1993, 04 27). Retrieved 2020-09-30, from https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule

Hamlett, P. (2020). *Your roadmap for sustainable design.* Retrieved 2020-09-23, from https://www.aiga.org/roadmap/

Hughes, J. T. (2020, 08 12). *Reflecting on the growth of the privacy industry.* Retrieved 2020-09-23, from https://iapp.org/news/a/reflecting-on-the-growth-of-the-privacy-industry/

LaNou, C. (2020, 09 22). Personal interview.

Legislature, C. (2018, 09 24). Retrieved 2020-09-30, from https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1121

Manning, C. D. (2008). *Introduction to information retrieval.* Cambridge University Press.

Meehan, M. (2019, 11 26). *Data privacy will be the most important issue in the next decade.* Retrieved 2020-09-23, from https://www.forbes.com/sites/marymeehan/2019/11/26/data-privacy-will-be-the-most-important-issue-in-the-next-decade/

Mozilla. (2020, 05 21). Retrieved 2020-09-27, from https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/DNT

of the European Union, C. (2018, 05 23). Retrieved 2020-09-30, from https://gdpr-info.eu/

Wood, N. (2019, 05 11). *New privacy tech industry attracts massive funding.* Retrieved 2020-09-23, from https://fpf.org/2019/07/11/new-privacy-tech-industry-attracts-massive-funding/