toot.ai | an explainable personal AI assistant

OTTO MÄTTAS, AUTHOR2, AUTHOR3, and AUTHOR4, Utrecht University, The Netherlands

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: data sets, neural networks, knowledge representation, text tagging

ACM Reference Format:
Otto Mättas, Author2, Author3, and Author4. 2021. toot.ai | an explainable personal AI assistant. 1, 1 (March 2021), 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 WHY IS THIS NEEDED?

Many people are not willing to adopt and leverage specific technologies unless they can exercise more control over the outcomes. Added to the lack of timely and globally consistent regulation, this creates a trust gap between companies building intelligent solutions and the intended target audience - the patrons to those companies.

Firstly, businesses who address the growing demand for task automation are often met with personal prejudice and distrust from the patrons. The road to continuous privacy concern has been paved by early Internet giants and their incorporated pricing models which abuse the user base through exploiting the individuals' data shared with them. What once seemed free of charge, has always come with a price. We have always paid and only now are starting to see the cost. We as patrons are used to not paying for great experiences online while do wish to keep our privacy. This calls for a new approach in how we deal with online services, especially services which promise task automation as a result of data mining.

Secondly, businesses building such services have to tackle complex technical regulatory environment. Or rather, the lack thereof. We have many technical standards but no consistency in how we leverage and enforce them. While a side-product of free market, it is starting to seriously inhibit our ability to innovate. A lot of the times, patrons are choosing tools which are familiar to them. Either via colleagues or corporate policy. The chosen tools are not always different to other similar tools in functionality while often come with vendor lock-in. The problem is that vendors are not geared towards sharing advances with their competitors. Today, rightfully so - knowledge and information is business and affects directly the bottom-line of those companies.

As we continue on this path that simplify task handling in a certain way, we continue creating fragmented approaches to the same underlying principle - knowledge sharing. It is time to bring control back to the patrons.

Building blocks for such solutions are already out there, but we do not have the confidence to use them yet. In the simplest sense - we still lack an understanding of how things work. And using machine learning is taking away from the potential of AI solutions rather than adding to it.

As of 2021, there are no commercially available AI Platforms that would allow its patrons to record, validate and act on a problem-solution pair directly. While there are solutions that offer enterprise-users functionality to record knowledge in an easy way, we lack the tools focused on openness. Companies are still pushing the envelope of closed source development. It is time to change this.

2 SOLUTION

We are building an AI Platform-as-a-Service, which can be interacted with directly. There are no limits to using the knowledge incorporated into the platform by its individual patron.

Firstly, an explainable AI assistant allows the patron to easily record a solution to a certain problem with their preferences in mind. As a result, a subjective knowledge base is created over time, task by task, no matter the domain.

Secondly, the assistant gives patron easy and direct access to the recorded knowledge. This is mainly to validate information stored, and to be updated if something does not add up.

Finally, after getting a green light from its patron, the assistant can detect an arising problem, explore the knowledge base for a solution and propose handle those tasks subjectively, after getting a confirmation from its patron/teacher. Essentially, it is a DIY AI kit or a human-readable framework for a happier living, if you will.

2.1 UNIQUE INNOVATION

Research on knowledge representation and reasoning has gone through decades of motions. We argue that it is now ready to be actively used in a hybrid system together with machine learning methods. We bring together human-understandable knowledge with machine-powered efficiency.

3 WHO NEEDS IT?

There are different phases with different focus points. We start small and specific with the potential to expand and have a positive effect on the whole of humanity.

Firstly, we focus on existing technology companies that are struggling with automating small but specific tasks. For example, you can think of Managed Service Providers with highly automated workflows and extensive infrastructure. Right now, there is no way to automatically handle a server disk space issue as there is context to be considered before taking action, every single time. Human reasoning is still needed. Which has created the notion of "legacy equals money for service providers." MSPs in Western Europe will

Authors' address: Otto Mättas, otto@toot.ai; Author2, email2@students.uu.nl; Author3, email3@students.uu.nl; Author4, email4@students.uu.nl, Utrecht University, P.O. Box 80125, Utrecht, Utrecht, The Netherlands, 3508 TC.

XXXX-XXXX/2021/3-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

generate almost 70bn in revenue in 2023, which is around 11 percent of the total ICT market in Western Europe.

Secondly, we focus on people who wish to have a(n effective) personal assistant. After gathering knowledge about specific domains, we can connect the dots. This overarching makes reasoning on the knowledge more comprehensive overall. As we are not bound to a single task anymore, we are also able to look additionally for personal approaches to any challenge. For example, think of navigation apps that would be able to remember your specific route to a recurring destination, rather than follow a general and rigid model. Or compare restaurants. Or drive your self-driving car for you. There are 4.6bn active internet users as of July 2020.

3.1 GO-TO-MARKET

We will collaborate with an international MSP to work on finalising the prototype and features relevant for additional business patrons. The use of learned knowledge is free of charge to them (alone), revenue comes from supporting the implementation. Hourly rate will be at least €200 per support staff. We already have a collaborator willing to work with us in principle.

Next, we can generalise on the learned knowledge by creating domain/task-specific models. These models will be a part of a free trial to invite new businesses to try our system. During establishing trust through conferences (e.g DevOpsDays and RE•WORK), we will add a revenue stream. Namely, we will let our new business patrons (MSPs) subscribe for the ability to teach their examples to improve on the public/trial models, with our support. The subscription will become our primary source for revenue, price will be decided case by case as each business patron brings a different value back to toot.ai in the form of knowledge learned.

Then, we expand the knowledge space from MSPs to all businesses that are wishing and willing to automate any task they have. Be it in agriculture, media, space exploration - we will help build an explainable AI system for their needs by providing the framework (an API to use for any task, no matter the domain). By this point, all knowledge learning and describing automatic, so we step back into the role of moderators and regulators of best practices. In other words, we make sure the trial models are spick and span. Support as a service is still relevant, while the focus will move gradually towards informing people rather than implementing systems.

Finally, with accurate (enough) models and a robust framework for creating and using those models, we open the system to the world and individuals as a personal assistant. Individual patrons are to join a free trial with the global models available. Subscribing allows the user to teach their assistant individually with their real-life examples. Revenue will come mainly from subscriptions at this point as support to business patrons is provided by toot.ai itself, so we stop charging for it. toot.ai will eventually be self-sufficient as supporting the implementation and informing people are tasks like any other.

3.2 STORIES

In this section, we will take a look at our product with a multi-actor perspective, by individuating Jobs that have to be done with relative Epic Stories and User Stories.

Job:

(1) Help me to optimize my time.
Epic Story:
- When I have loads of tasks to do, I want to carry them out efficiently and effectively, so that I will not work overtime.
User Stories:
  - As a CFO of a company, I want to save money on overtime salary, so that I can have a better budget.
  - As a part of a family, I want to finish my work on time, so that I can enjoying more time with my family.
  - As a HR person of a company, I want my colleagues to be healthy, so that I have lower employee turnover.

Job:

(2) Help me to automate mundane or repetitive tasks.
Epic Story:
- When I have an important task to do, I want to share part of my mundane tasks, so that I can better focus on the important one.
User Stories:
  - As a freelancer, I want to have a personal assistant manage my time for me, so that I can focus more on customer requests.
  - As an employee of a consultant agency, I want to give more focus to the important customers without overlooking the regular ones, so that I can satisfy all the customers accordingly to their expectations.
Epic Story:
- When there is a repetitive task assigned to me, I want an assistant to recognize the repetition, so that it can handle it for me.
User Stories:
  - As an engineer, I do not want to deal with repetitive tasks, so that I can develop new features.

Job:

(3) Help me to improve the quality of my work.
Epic Story:
- When I am not capable of doing a task, I want some help to cover the lacking abilities, so that I can improve my work performances and learn from it.
User Stories:
  - As a freelancer, I want to install to my customers the best possible solution even though I don't know which is, so that I can offer the best possible service.
Epic Story:
- When I am solving a problem, I want to know the best solution, so that the final result will be better.
User Stories:
  - As an employee in a non-profit company, I want to solve my problem with the best solution, so that I can make a positive impact to the world.

4 TECHNOLOGY

4.1 FUNCTIONAL FEATURES

4.2 TECHNICAL FEATURES

Nowadays, cloud computing is influencing the IT landscape and becoming a significant economic factor for software producing organisations.

There are many aspects to consider in leveraging cloud technologies. As presented by [Au 2016] the list of benefits includes but is not limited to cost savings, easy implementation, storage capacity, accessibility, reliability and manageability. On the flip side, one has to consider vendor lock-in as the extensive platforms bring vendor-specific tool sets. Additionally, extra care has to be taken with security. Whenever using the internet, security is not guaranteed – even if the cloud service providers can implement the best security standards and industry certifications.

Cloud computing is a fast-growing technology in a non-transparent market with diverse vendors, each of them having their specific services and deployment models as explained by [Farshidi et al. 2018]. Typically, the service portfolios are heterogeneous and combined with complicated service features and pricing models.

Even though the benefits are clear for building a scalable platform, using cloud technologies brings a challenge. Namely, evaluating and selecting the most suitable Infrastructure-as-a-Service Cloud Provider for software producing organisations according to their preferences and requirements.

As concluded by [Farshidi et al. 2018], finding a feasible solution for the Infrastructure-as-a-Service Cloud Provider selection problem based on decision-makers' priorities and requirements requires deep investigation into the documentation of cloud vendors and extensive expert analysis. As our the current researchers have extensive experience with Amazon Web Services, further referred to as AWS, we have chosen to leverage the tools provided by this particular Cloud Provider.

4.2.1 User management. Our platform is available through many channels to different actors. Amazon Cognito lets you add user sign-up, sign-in, and access control to your web and mobile apps quickly and easily. Amazon Cognito scales to millions of users and supports sign-in with social identity providers, such as Apple, Facebook, Google, and Amazon, and enterprise identity providers via SAML 2.0 and OpenID Connect.

Amazon Lex is a service for building conversational interfaces into any application using voice and text. Amazon Lex provides the advanced deep learning functionalities of automatic speech recognition (ASR) for converting speech to text, and natural language understanding (NLU) to recognize the intent of the text, to enable you to build applications with highly engaging user experiences and lifelike conversational interactions. With Amazon Lex, the same deep learning technologies that power Amazon Alexa are now available to any developer, enabling you to quickly and easily build sophisticated, natural language, conversational bots ("chatbots").

AWS IoT Core lets you connect IoT devices to the AWS cloud without the need to provision or manage servers. AWS IoT Core can support billions of devices and trillions of messages, and can process and route those messages to AWS endpoints and to other devices reliably and securely. With AWS IoT Core, your applications can keep track of and communicate with all your devices, all the time, even when they aren't connected.

AWS IoT Core also makes it easy to use AWS and Amazon services like AWS Lambda, Amazon Kinesis, Amazon S3, Amazon SageMaker, Amazon DynamoDB, Amazon CloudWatch, AWS CloudTrail, Amazon QuickSight, and Alexa Voice Service to build IoT applications that gather, process, analyze and act on data generated by connected devices, without having to manage any infrastructure.

Amazon API Gateway is a fully managed service that makes it easy for developers to create, publish, maintain, monitor, and secure APIs at any scale. APIs act as the "front door" for applications to access data, business logic, or functionality from your backend services. Using API Gateway, you can create RESTful APIs and WebSocket APIs that enable real-time two-way communication applications. API Gateway supports containerized and serverless workloads, as well as web applications.

API Gateway handles all the tasks involved in accepting and processing up to hundreds of thousands of concurrent API calls, including traffic management, CORS support, authorization and access control, throttling, monitoring, and API version management. API Gateway has no minimum fees or startup costs. You pay for the API calls you receive and the amount of data transferred out and, with the API Gateway tiered pricing model, you can reduce your cost as your API usage scales.

Amazon Kinesis Data Streams (KDS) is a massively scalable and durable real-time data streaming service. KDS can continuously capture gigabytes of data per second from hundreds of thousands of sources such as website clickstreams, database event streams, financial transactions, social media feeds, IT logs, and location-tracking events. The data collected is available in milliseconds to enable real-time analytics use cases such as real-time dashboards, real-time anomaly detection, dynamic pricing, and more.

4.2.2 Data ingestion. After the user has been authenticated and the necessary information has been made available for the pipeline, we move towards working directly with the data.

We leverage Lambda functions to write our custom functions for data ingestion.

AWS Lambda is a serverless compute service that lets you run code without provisioning or managing servers, creating workload-aware cluster scaling logic, maintaining event integrations, or managing runtimes. With Lambda, you can run code for virtually any type of application or backend service - all with zero administration. Just upload your code as a ZIP file or container image, and Lambda automatically and precisely allocates compute execution power and runs your code based on the incoming request or event, for any scale of traffic. You can set up your code to automatically trigger from 140 AWS services or call it directly from any web or mobile app. You can write Lambda functions in your favorite language (Node.js, Python, Go, Java, and more) and use both serverless and container tools, such as AWS SAM or Docker CLI, to build, test, and deploy your functions.

4.2.3 Knowledge extraction. After we have ingested patron data, we move on to extraction the knowledge from it.

Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to find insights and relationships in text. No machine learning experience required.

There is a treasure trove of potential sitting in your unstructured data. Customer emails, support tickets, product reviews, social media, even advertising copy represents insights into customer sentiment that can be put to work for your business. The question is how to get at it? As it turns out, Machine learning is particularly good at accurately identifying specific items of interest inside vast swathes of text (such as finding company names in analyst reports), and can learn the sentiment hidden inside language (identifying negative reviews, or positive customer interactions with customer service agents), at almost limitless scale.

Amazon Comprehend uses machine learning to help you uncover the insights and relationships in your unstructured data. The service identifies the language of the text; extracts key phrases, places, people, brands, or events; understands how positive or negative the text is; analyzes text using tokenization and parts of speech; and automatically organizes a collection of text files by topic. You can also use AutoML capabilities in Amazon Comprehend to build a custom set of entities or text classification models that are tailored uniquely to your organization's needs.

For extracting complex medical information from unstructured text, you can use Amazon Comprehend Medical. The service can identify medical information, such as medical conditions, medications, dosages, strengths, and frequencies from a variety of sources like doctor's notes, clinical trial reports, and patient health records. Amazon Comprehend Medical also identifies the relationship among the extracted medication and test, treatment and procedure information for easier analysis. For example, the service identifies a particular dosage, strength, and frequency related to a specific medication from unstructured clinical notes.

Amazon Rekognition makes it easy to add image and video analysis to your applications using proven, highly scalable, deep learning technology that requires no machine learning expertise to use. With Amazon Rekognition, you can identify objects, people, text, scenes, and activities in images and videos, as well as detect any inappropriate content. Amazon Rekognition also provides highly accurate facial analysis and facial search capabilities that you can use to detect, analyze, and compare faces for a wide variety of user verification, people counting, and public safety use cases.

With Amazon Rekognition Custom Labels, you can identify the objects and scenes in images that are specific to your business needs. For example, you can build a model to classify specific machine parts on your assembly line or to detect unhealthy plants. Amazon Rekognition Custom Labels takes care of the heavy lifting of model development for you, so no machine learning experience is required. You simply need to supply images of objects or scenes you want to identify, and the service handles the rest.

4.2.4 Knowledge representation. After having extracted concepts from the data, we move towards creating a knowledge base.

We store the data in a graph database. Amazon Neptune is a fast, reliable, fully managed graph database service that makes it easy to build and run applications that work with highly connected datasets. The core of Amazon Neptune is a purpose-built, high-performance graph database engine optimized for storing billions of relationships and querying the graph with milliseconds latency. Amazon Neptune supports popular graph models Property Graph and W3C's RDF, and their respective query languages Apache TinkerPop Gremlin and SPARQL, allowing you to easily build queries that efficiently navigate highly connected datasets. Neptune powers graph use cases such as recommendation engines, fraud detection, knowledge graphs, drug discovery, and network security.

Amazon Neptune is highly available, with read replicas, point-in-time recovery, continuous backup to Amazon S3, and replication across Availability Zones. Neptune is secure with support for HTTPS encrypted client connections and encryption at rest. Neptune is fully managed, so you no longer need to worry about database management tasks such as hardware provisioning, software patching, setup, configuration, or backups.

Additionally, we will be creating a representation in the W3C Web Ontology Language (OWL). It is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things. OWL is a computational logic-based language such that knowledge expressed in OWL can be exploited by computer programs, e.g., to verify the consistency of that knowledge or to make implicit knowledge explicit. OWL documents, known as ontologies, can be published in the World Wide Web and may refer to or be referred from other OWL ontologies. OWL is part of the W3C's Semantic Web technology stack, which includes RDF, RDFS, SPARQL, etc.

4.2.5 Knowledge validation. After having created a knowledge base, we move towards validating the concepts.

Amazon Kendra is an intelligent search service powered by machine learning. Kendra reimagines enterprise search for your websites and applications so your employees and customers can easily find the content they are looking for, even when it's scattered across multiple locations and content repositories within your organization.

Using Amazon Kendra, you can stop searching through troves of unstructured data and discover the right answers to your questions, when you need them. Amazon Kendra is a fully managed service, so there are no servers to provision, and no machine learning models to build, train, or deploy.

For more capabilities, we will use the ELK stack. Amazon Elasticsearch Service is a fully managed service that makes it easy for you to deploy, secure, and run Elasticsearch cost effectively at scale. You can build, monitor, and troubleshoot your applications using the tools you love, at the scale you need. The service provides support for open source Elasticsearch APIs, managed Kibana, integration with Logstash and other AWS services, and built-in alerting and SQL querying. Amazon Elasticsearch Service lets you pay only for what you use – there are no upfront costs or usage requirements. With Amazon Elasticsearch Service, you get the ELK stack you need, without the operational overhead.

We are also sharing updates to the knowledge base directly with the user via a stream. Neptune Streams logs every change to your graph as it happens, in the order that it is made, in a fully managed way. Once you enable Streams, Neptune takes care of availability, backup, security and expiry.

We allow the patrons to validate the knowledge via multiple channels. Amazon Pinpoint is a flexible and scalable outbound and inbound marketing communications service. You can connect with customers over channels like email, SMS, push, or voice. Amazon Pinpoint is easy to set up, easy to use, and is flexible for all marketing communication scenarios. Segment your campaign audience for the right customer and personalize your messages with the right content. Delivery and campaign metrics in Amazon Pinpoint measure the success of your communications. Amazon Pinpoint can grow with you and scales globally to billions of messages per day across channels.

4.2.6 Action. After we have validated the knowledge, we move the information to a data lake.

AWS Lake Formation is a service that makes it easy to set up a secure data lake in days. A data lake is a centralized, curated, and secured repository that stores all your data, both in its original form and prepared for analysis. A data lake enables you to break down data silos and combine different types of analytics to gain insights and guide better business decisions.

However, setting up and managing data lakes today involves a lot of manual, complicated, and time-consuming tasks. This work includes loading data from diverse sources, monitoring those data flows, setting up partitions, turning on encryption and managing keys, defining transformation jobs and monitoring their operation, re-organizing data into a columnar format, configuring access control settings, deduplicating redundant data, matching linked records, granting access to data sets, and auditing access over time.

Creating a data lake with Lake Formation is as simple as defining data sources and what data access and security policies you want to apply. Lake Formation then helps you collect and catalog data from databases and object storage, move the data into your new Amazon S3 data lake, clean and classify your data using machine learning algorithms, and secure access to your sensitive data. Your users can access a centralized data catalog which describes available data sets and their appropriate usage. Your users then leverage these data sets with their choice of analytics and machine learning services, like Amazon Redshift, Amazon Athena, and (in beta) Amazon EMR for Apache Spark. Lake Formation builds on the capabilities available in AWS Glue.

From there, we can create off-the-shelf machine learning models for addressing a certain problem domain.

Amazon SageMaker is a machine learning service that you can use to build, train, and deploy ML models for virtually any use case. The ML tool set brings together all the relevant steps for creating ML models - preparing data, building notebooks, training and tuning models, deploying and managing them in end applications. It comes with a vast range of different products embedded in the kit. We are still looking at which tools we need for our current step.

After incorporating data and machine learning models in the data lake, we give back a solution to the user.

Amazon Personalize enables developers to build applications with the same machine learning (ML) technology used by Amazon.com for real-time personalized recommendations – no ML expertise required.

Amazon Personalize makes it easy for developers to build applications capable of delivering a wide array of personalization experiences, including specific product recommendations, personalized product re-ranking, and customized direct marketing. Amazon Personalize is a fully managed machine learning service that goes beyond rigid static rule based recommendation systems and trains, tunes, and deploys custom ML models to deliver highly customized recommendations to customers across industries such as retail and media and entertainment.

Amazon Personalize provisions the necessary infrastructure and manages the entire ML pipeline, including processing the data, identifying features, using the best algorithms, and training, optimizing, and hosting the models. You will receive results via an Application Programming Interface (API) and only pay for what you use, with no minimum fees or upfront commitments. All data is encrypted to be private and secure, and is only used to create recommendations for your users.

Again, we leverage Amazon Pinpoint to deliver the results.

And finally, we can present the results to the patron via speech.

Amazon Polly is a service that turns text into lifelike speech, allowing you to create applications that talk, and build entirely new categories of speech-enabled products. Polly's Text-to-Speech (TTS) service uses advanced deep learning technologies to synthesize natural sounding human speech. With dozens of lifelike voices across a broad set of languages, you can build speech-enabled applications that work in many different countries.

In addition to Standard TTS voices, Amazon Polly offers Neural Text-to-Speech (NTTS) voices that deliver advanced improvements in speech quality through a new machine learning approach. Polly's Neural TTS technology also supports two speaking styles that allow you to better match the delivery style of the speaker to the application: a Newscaster reading style that is tailored to news narration use cases, and a Conversational speaking style that is ideal for two-way communication like telephony applications.

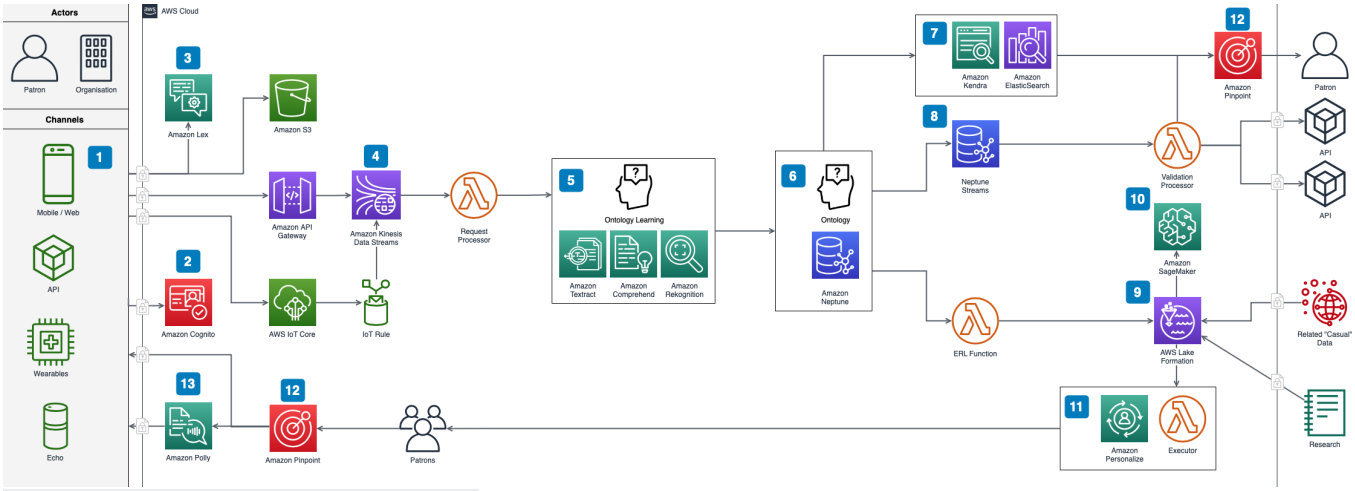
Finally, Amazon Polly Brand Voice can create a custom voice for your organization. This is a custom engagement where you will work with the Amazon Polly team to build an NTTS voice for the exclusive use of your organization. Learn more here.

REFERENCES

Regina Au. 2016. To Cloud Compute, or Not to Cloud Compute? *Innovations in Pharmaceutical Technology* (07 2016), 32–35.

Siamak Farshidi, Slinger Jansen, Rolf Jong, and Sjaak Brinkkemper. 2018. A Decision Support System for Cloud Service Provider Selection Problem in Software Producing Organizations. https://doi.org/10.1109/CBI.2018.00024

A ARCHITECTURE GRAPH



1. toot.ai is available through many channels to different actors. Individual patrons approach individual assistants while organisations approach an environment with the possibility for on-prem knowledge.
2. Add patron sign-up, sign-in, and access control to our web and mobile apps quickly and easily. Supports sign-in with social identity providers and enterprise identity providers via SAML 2.0 and OpenID Connect.
3. Automatic speech recognition (ASR) for converting speech to text, and natural language understanding (NLU) to recognise intent of the text.
4. Real-time data streaming service. All patron preferences together with task requests are captured by continuously ingesting data which is delivered through Amazon API Gateway or AWS IoT Core.
5. Ontology Learning. NLP, Physical documents, Video and Image entity extraction and load into a graph.
6. Graph database service. It supports W3C's RDF and SPARQL. OWL is still needed for proper inference functionality.
7. Knowledge direct validation by patrons. Also used for debugging and continuous improvement.
8. Knowledge update validation. Neptune Streams capture changes to our graph (change-log data) as they happen.
9. AWS Lake Formation stores all our data, both in its original form and prepared for analysis.
10. We leverage Amazon SageMaker for learning ML models to have an off -the-shelf solution.
11. We send back the proper solution to the patron with recommendations for the specific user.
12. We leverage an outbound and inbound marketing communications service. We can connect with patrons over channels like email, SMS, push, or voice.
13. Text-to-Speech (TTS) service synthesises natural sounding speech for connecting to the patrons.