Dust NLP

alan@mentalresonance.com

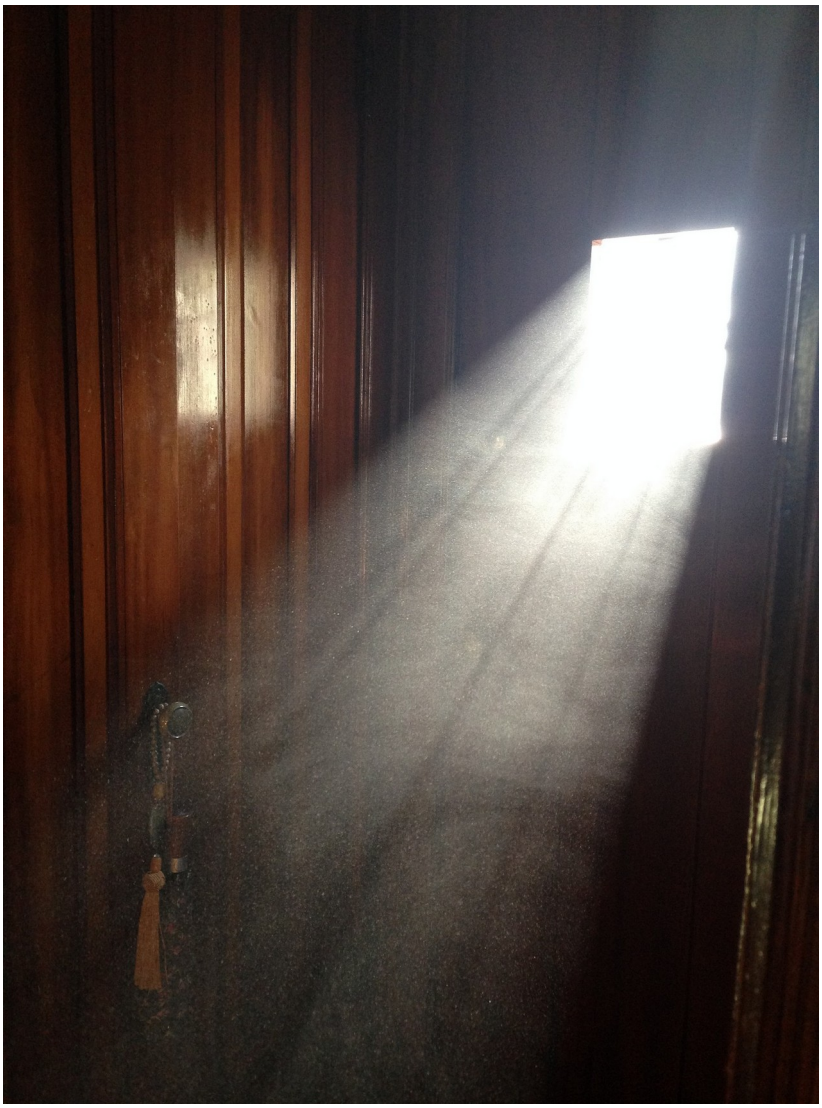
Version 1.0.2

Dust is a JVM based framework for building distributed, scalable applications based on the Actor paradigm.

Dust comes in several modules:

- **dust-core –** the heart of Dust which implements Actors, Actor persistence, several core Actor building blocks and the basics of the Entity framework, built on Dust, for digital twinning and similar applications.
- **dust-http** – library for creating Actors which participate in http/REST style interactions
- **dust-html** – library for adding html processing to Actors
- **dust-nlp** – library for Natural Language Processing using LLMs
- **dust-feeds** – library of Actors for constructing pipelines of feed processors, e.g. RSS feeds.

This document discusses dust-nlp, which uses Dust core and http libraries, so familiarity with the code and idioms in these libraries is assumed.



1 Introduction

Computer-based Natural Language Processing (NLP) has been around since the early 1950's, and early tools were usually grammar-based. However the last few years have seen a totally different approach which began with the development of embeddings.

Essentially an embedding is a map from words → very large dimensional vector space[1]. This mapping is made by analyzing vast[2] amounts of text. An amazing property of embeddings is that they seem to capture the 'meanings' of words. So using simple calculations on the vectors they can complete the following "Man is to king as woman is to …".

While embeddings are powerful it was the recent development of Transformers that rocked the world. In particular Generative Pretrained Transformers (GPTs) have led to the rapid development of large language models (LLMs) which do one simple trick, but do it rather well.

A LLM takes a list of words and, based on its training data[3] predicts what the next word will be. It adds this word to the list and repeats this process on the new list producing a new word. And so on. Simple and yet startlingly effective.

LLMs and Artificial Intelligence have become, rightly or wrongly, synonymous. Dust NLP is a small Dust Actor library which allows embeddings and LLMs to fit easily and idiomatically into the Dust platform.

2 Embeddings

Dust embeddings are a lightweight idiomatic Actor framework around some externally supplied Embedding engine. We have explicit support for the Huggingface embeddings engine via the HFEmbeddingAPIServiceActor.

To get the embeddings for a text you send this Actor an EmbeddingsRequestResponseMsg containing the text to be embedded. The Actor is configured with a 'chunk size' which indicates the maximum length of text to use to get a single embedding. HFEmbeddingAPIServiceActor chunks the provided text into a sequence of complete sentences whose total size is the largest that fits into chunk size. It repeats for the remaining text but it reuses the last sentence of the previous chunk so the chunks overlap each other by one sentence.

Each chunk of sentences is sent to the embedder and the result added to a list of Embedding instances in the EmbeddingsRequestResponseMsg (an Embedding contains the text chunk

1 If you don't know what a vector is don't worry. Think of it as a fixed length series of numbers. In embeddings the length (called the dimension of the space) is often 500 or higher.
2 Internet-sized vast
3 Again internet-sized vast

and its embedding vector). When all chunks have been processes the completed EmbeddingsRequestResponseMsg is sent back to the requester.

The EmbeddingDistance class contains an implementation of Cosine Similarity which measure the ‘distance’ between two vectors. A test shows how to embed a relatively large document and look up the ‘best fit’ chunk given a query.

3 GPTs

Many GPT implementations follow the OpenAI API quite closely so we have a Service Actor, GenericGptAPIServiceActor which receives GenericGptRequestResponseMsgs. These messages contain the prompt, model information, token limit, temperature etc. The Actor sends the request off to the API and fills GenericGptRequestResponseMsgs with the returned utterance (or error message etc).

For local instances of Ollama we have a OllamaRequestResponseMsg which works with GenericGptAPIServiceActor, and ChatGptRequestResponseMsg which works with ChatGptAPIServiceActor for OpenAI support. The latter handles the API key and accounting information.

ChatGPTUtils and ChatUtils contain some useful methods for extracting information from LLM utterances (parsing lists etc).

4 News Reader Test

4.1 Overview

NewsPipelineTest is not so much a test but more an example of how various Dust components can be plugged together to create a non-trivial application. It uses all of the Dust repos released so far: dust-core,html,http,feeds and nlp.

The goal of the application is simple. Monitor RSS news feeds for news about a topic I am interested in and give me a summary of each relevant article.

The test uses the topic: ‘Electric vehicle charging’.

A moments though reveals at least two issues to be resolved:

1. How do we know which RSS feeds to use?
2. Even if we have a useful RSS feed how do we select only those articles I have expressed interest in?

Not surprisingly the answers are:

1. Ask Chat GPT
2. Ask Chat GPT

But we know LLMs really want to answer your questions and are prone to hallucination so we will need to check the RSS feeds really exist …

The core code for the News feed Actor is:

```
String topic = "Electric Vehicle charging"

void preStart() {
	actorOf(RssLocatorServiceActor.props()).tell(
		new RssLocatorServiceActor.RssFeedFinderMsg(topic), self
	)
}

ActorBehavior createBehavior() {
	(message) -> {
		switch(message) {
			case VerifyFeedsMsg:
				ActorRef newsPipe = actorOf(PipelineActor.props([
	PipelineFeedHubActor.props([], []),
	ServiceManagerActor.props(ContentFilterPipeServiceActor.props(topic, []), 1),
	ServiceManagerActor.props(SummarizerPipeServiceActor.props(), 1),
	LogActor.props()
					], [
						'feeds',
						'filter',
						'summarizer',
						'logger'
					]),
				'newspipe')

				((VerifyFeedsMsg)message).verifiedUrls.each {
					newsPipe.tell(new PipelineFeedHubActor.AddFeedMsg(it),
					self)
				}
				break

			default:
				log.error "??? $message"

			}
		}
	}
}
```

The preStart() method fires up a RssLocatorServiceActor and sends it a message with our topic in it. This Actor will reply with a VerifyFeedsMsg which contains a list of known-good feeds relevant to our topic.

When our default Behavior gets this message it creates a pipeline and sends an AddFeed message to the pipeline for every known good feed. It then sits there.

4.2 Pipeline

The pipeline has 4 stages:

1. FeedHubActor – this simply fires up an RSS Feed Actor for each url delivered by the AddFeedMsg. As discussed in dust-feeds these Actors periodically visit their RSS feed, download new documents and send the documents on down the pipe.

2. ContentFilterPipeServiceActor – this takes a document title and asks the LLM if it matches the topic, if it does the document is passed on down the pipe.
3. SummarizerPipeServiceActor – this takes the document, extracts the core content (e.g. if this is a web page it removes superfluous ‘fluff’ and asks the LLM to summarize it. The result summary along with the document title are sent on down the pipe to
4. LogActor – which simply displays the text – i.e. matching article titles and summaries.

Since 2 & 3 have to keep a track of the document across interactions with the LLM we use the Dust ServiceActor idiom. i.e. an Actor is created to process one document then it is destroyed. Since we limit the service pool sizes to 1 this also applies some natural throttling to the process.

4.3 RssLocatorServiceActor

RssLocatorServiceActor extends HttpClientActor and so has a request() method which can make http requests. When RssLocatorServiceActor gets a RssFeedFinderMsg it sends off a request to the LLM and asks for a list of appropriate RSS feed urls.

When it receives the response it iterates through the list of urls (by pumping messages) and for each attempt to retrieve the content and parse it as an RSS feed. If we succeed we add the url to a growing list, which list we eventually send off for further processing.

RssLocatorServiceActor is a Service Actor and so would usually be under a ServiceManager, but for purposes of the demo it is only go to receive a message once.

```
ActorBehavior createBehavior() {
	(message) -> {
		switch(message) {
			case RssFeedFinderMsg:
				RssFeedFinderMsg msg = (RssFeedFinderMsg)message
				originalSender = sender
				context.actorSelection('/user/chatgpt').tell(
					new RssFeedRequestMsg(msg.query),
					self
				)
				break

			case RssFeedRequestMsg:
				RssFeedRequestMsg msg = (RssFeedRequestMsg)message
				List<String> urls = ChatUtils.numericList(msg.utterance)
				self.tell(new VerifyFeedsMsg(urls), self)
				break

			case VerifyFeedsMsg:
				verifyFeedsMsg = (VerifyFeedsMsg)message
				if (verifyFeedsMsg.urls.isEmpty()) {
					originalSender?.tell(verifyFeedsMsg, null)
					context.stop(self)
				} else
					try {
						request(verifyFeedsMsg.urls.removeFirst())
					} catch (Exception e) {
						log.warn e.message
					}
				break

			/*
			 * Verify the feed. Does the page exist and is it a feed ?? ChatGPT
			 * fails in both directions. Simply try to parse result as a feed
			 */
			case HttpRequestResponseMsg:
				HttpRequestResponseMsg msg = (HttpRequestResponseMsg)message

				if (null == msg.exception && msg.response.successful) {
					String url = msg.request.url().toString()
					try {
						new SyndFeedInput().build(
						new XmlReader(
							msg.response.body().byteStream())
						)
						verifyFeedsMsg.verifiedUrls << url
						log.info "$url is a feed !!"
					}
					catch (Exception e) {
						log.warn "$url exists but is not an RSS feed!"
					}
				}
				self.tell(verifyFeedsMsg, self)
				break


				default: super.createBehavior().onMessage(message as Serializable)
			}
		}
	}
```

4.4 Log from a sample run

14:18:52:665 [Test worker] INFO com.mentalresonance.dust.core.actors.ActorSystem - Started ActorSystem: Test on port null

14:18:54:720 news.RssLocatorServiceActor - https://www.greencarreports.com/rss is a feed !!

14:18:55:184 news.RssLocatorServiceActor - https://www.electrive.com/feed/ is a feed !!

14:18:55:502 news.RssLocatorServiceActor - https://www.carscoops.com/category/green-cars/feed/ exists but is not an RSS feed!

14:18:55:661 news.RssLocatorServiceActor - https://www.teslarati.com/feed/ is a feed !!

14:18:56:022 news.RssLocatorServiceActor - https://www.evannex.com/blogs/news.atom is a feed !!

14:18:56:347 news.RssLocatorServiceActor - https://chargedevs.com/feed/ is a feed !!

14:18:56:692 news.RssLocatorServiceActor - https://www.electrek.co/feed/ is a feed !!

Oct 05, 2024 2:18:56 PM okhttp3.internal.platform.Platform log

WARNING: A connection to https://insideevs.com/ was leaked. Did you forget to close a response body? To see where this was allocated, set the OkHttpClient logger level to FINE: Logger.getLogger(OkHttpClient.class.getName()).setLevel(Level.FINE);

Oct 05, 2024 2:18:56 PM okhttp3.internal.platform.Platform log

WARNING: A connection to https://www.autoblog.com/ was leaked. Did you forget to close a response body? To see where this was allocated, set the OkHttpClient logger level to FINE: Logger.getLogger(OkHttpClient.class.getName()).setLevel(Level.FINE);

14:18:56:917 news.RssLocatorServiceActor - https://www.autoevolution.com/rss/green-cars.xml exists but is not an RSS feed!

14:18:56:953 news.PipelineFeedHubActor - /user/news/newspipe/feeds/ adding RSS feed https://www.greencarreports.com/rss

14:18:56:955 news.PipelineFeedHubActor - /user/news/newspipe/feeds/ adding RSS feed https://www.electrive.com/feed/

14:18:56:955 news.PipelineFeedHubActor - /user/news/newspipe/feeds/ adding RSS feed https://www.teslarati.com/feed/

14:18:56:956 news.PipelineFeedHubActor - /user/news/newspipe/feeds/ adding RSS feed https://www.evannex.com/blogs/news.atom

14:18:56:956 news.PipelineFeedHubActor - /user/news/newspipe/feeds/ adding RSS feed https://chargedevs.com/feed/

14:18:56:957 news.PipelineFeedHubActor - /user/news/newspipe/feeds/ adding RSS feed https://www.electrek.co/feed/

14:18:57:038 com.mentalresonance.dust.feeds.rss.RssFeedPipeActor - Processing 30 new entries from RSS feed https://www.electrive.com/feed/

14:18:57:042 com.mentalresonance.dust.feeds.rss.RssFeedPipeActor - Processing 10 new entries from RSS feed https://www.teslarati.com/feed/

14:18:57:058 com.mentalresonance.dust.feeds.rss.RssFeedPipeActor - Processing 100 new entries from RSS feed https://www.electrek.co/feed/

14:18:57:080 com.mentalresonance.dust.feeds.rss.RssFeedPipeActor - Processing 10 new entries from RSS feed https://chargedevs.com/feed/

14:18:57:175 com.mentalresonance.dust.feeds.rss.RssFeedPipeActor - Processing 15 new entries from RSS feed https://www.greencarreports.com/rss

14:18:57:260 com.mentalresonance.dust.feeds.rss.RssFeedPipeActor - Processing 30 new entries from RSS feed https://www.evannex.com/blogs/news.atom

14:19:00:324 com.mentalresonance.dust.core.actors.lib.LogActor - /user/news/newspipe/logger/ got message

Title: Enel X Way North America ceases operations

Summary: Enel X Way North America planned to operate two million charging points in North America by 2030, including 10,000 fast chargers. However, the subsidiary of the Italian energy group Enel has abruptly announced its withdrawal from the US and Canada due to financial challenges and market dynamics. The shutdown includes discontinuing the Enel X Way software landscape, leaving customers without connectivity services and access to charging processes. ChargeLab has stepped in to take over the commercial network JuiceBox, offering migration options for affected commercial site hosts before Enel X Way servers are permanently disabled.

from /user/news/newspipe/

14:19:01:639 com.mentalresonance.dust.core.actors.lib.LogActor - /user/news/newspipe/logger/ got message

Title: EVgo loan for over $1 billion for network expansion in the US

Summary: EVgo plans to install 7,500 new electric vehicle chargers across ten states by 2030, in alignment with the National Electric Vehicle Infrastructure Formula Program. This initiative, supported by a partnership with the Biden-Harris administration, aims to enhance EV charging accessibility in communities in need. The expansion is expected to create over 1,000 jobs and is contingent on meeting specific technical, legal, environmental, and financial requirements. Additionally, EVgo recently partnered with General Motors to establish 400 fast-charging points with high capacity in key US metropolitan areas, with the first location set to open next year.

from /user/news/newspipe/

14:19:03:392 com.mentalresonance.dust.core.actors.lib.LogActor - /user/news/newspipe/logger/ got message

Title: Tesla shares Supercharger Network performance and growth stats in Q3 2024

Summary: In the third quarter of 2024, Tesla reported significant growth in its Supercharger Network, adding 2,800 stalls and delivering 1.4 TWh of energy, resulting in savings of over 150 million gallons of gasoline and offsetting more than 3 billion pounds of CO2. Industry watchers estimate that Tesla has built around 62,400 Superchargers globally. Despite initial concerns following layoffs in the Supercharger team, CEO Elon Musk reassured the community of continued network growth with a focus on achieving 100% uptime and expanding existing locations.

from /user/news/newspipe/

14:19:13:059 com.mentalresonance.dust.core.actors.lib.LogActor - /user/news/newspipe/logger/ got message

Title: Lightning fast charger: Fortescue 6 MW DCFC for electric heavy equipment

Summary: Fortescue and Liebherr have partnered to develop a groundbreaking 6 MW DC fast charger for massive haul trucks, capable of charging a 1,900 kWh battery in under 30 minutes. This charger is part of a $4

billion deal for 475 zero-emission Liebherr machines, including autonomous battery-electric trucks, electric excavators, and battery-powered loaders. The collaboration aims to decarbonize mining activities globally, with Fortescue's 2030 Real Zero target focusing on eliminating carbon emissions from Australian iron ore operations by 2030.

from /user/news/newspipe/

14:19:14:277 com.mentalresonance.dust.core.actors.lib.LogActor - /user/news/newspipe/logger/ got message

Title: Get your EV questions answered at Drive Electric Week, continuing thru Sunday

Summary: Drive Electric Week, organized by various associations, commenced with nearly 200 events both online and in-person, celebrating electric vehicles. The event, now in its 14th year, offers a platform for prospective EV buyers to engage with owners, sharing experiences and tips. Ranging from small meetups to larger festivals, these events provide insights into EV ownership, with some online sessions available. Check local listings for upcoming events and diverse EV showcases, as Drive Electric Week continues through Sunday, October 6.

from /user/news/newspipe/

14:19:15:632 com.mentalresonance.dust.core.actors.lib.LogActor - /user/news/newspipe/logger/ got message

Title: Caterpillar presents dynamic charging system for mining trucks

Summary: Caterpillar introduces Cat Dynamic Energy Transfer (DET) system for intermediate charging of battery-only and hybrid mining vehicles, featuring a power module, rail system for dynamic charging, and transfer arms. The mobile rail system can adapt to various mine layouts, allowing trucks to charge while driving uphill, enhancing uptime and operational efficiency. Caterpillar aims to reduce operating costs and emissions, offering flexibility for future needs. The system integrates with Cat MineStar Command for autonomous hauling, catering to the growing trend of large autonomous vehicles in mining operations.

from /user/news/newspipe/

14:19:17:134 com.mentalresonance.dust.core.actors.lib.LogActor - /user/news/newspipe/logger/ got message

Title: Swissport accelerates electrification of ground services at airports

Summary: Swissport, the world's largest operator of airport ground support equipment, plans to transition to electric ground handling vehicles by 2027. The Swiss company aims to electrify various airport services, including baggage transport vehicles, conveyor belts, forklift trucks, and service vehicles. With an investment of over a billion euros in the next decade, Swissport aims to increase the proportion of electric, zero-emission equipment to support sustainability goals and reduce supply chain emissions for airlines. Despite challenges with charging infrastructure, Swissport is making progress in fleet electrification at major European hubs like Zurich and Amsterdam airports.

from /user/news/newspipe/

14:19:18:475 com.mentalresonance.dust.core.actors.lib.LogActor - /user/news/newspipe/logger/ got message

Title: Siemens unveils power distribution switchboard for EV fast chargers

Summary: Siemens has launched the RapidSBx switchboard to facilitate rapid EV charging station deployment, catering to commercial switchboards. The UL891 switchboard offers a streamlined design for charging stations in various settings, including urban and remote locations. It provides flexibility, durability, and easy installation with customizable options and a rugged outdoor enclosure. Safety features include arc mitigation compliance and tamper resistance, making it compatible with the NEVI program and EUSERC certified. Available through US distributors and channel partners.

from /user/news/newspipe/

14:19:19:729 com.mentalresonance.dust.core.actors.lib.LogActor - /user/news/newspipe/logger/ got message

Title: Tesla shares impressive data point about its Supercharger network

Summary: Tesla reported delivering an impressive 1.4 TWh of electricity through its Supercharger network in Q3, showcasing significant growth despite a decrease in quarterly stall deployment. The company's revenue from the network, estimated at $490 million to $700 million, reflects its expanding reach and increasing utilization, driven in part by the inclusion of non-Tesla EVs. While Tesla's DC fast-charging station deployment costs remain low, operational expenses are notable due to peak charges, emphasizing the importance of renewable energy sources for CO2 offsetting.

from /user/news/newspipe/

14:19:21:164 com.mentalresonance.dust.core.actors.lib.LogActor - /user/news/newspipe/logger/ got message

Title: First Ionna station flips gas station into EV "Rechargery"

Summary: Ionna broke ground on its first EV fast-charging site in Apex, North Carolina, on land formerly occupied by a gas station. The site, named the "Rechargery," will feature 10 covered parking bays with Combined Charging Standard (CCS) and North American Charging Standard (NACS) connectors capable of up to 400 kW of power. Ionna plans to establish a network of 30,000 high-power EV fast chargers across North America by 2030, backed by seven automakers. The site will offer amenities such as a driver's lounge, food and beverages, bathrooms, wifi, and outdoor pet-friendly areas.

from /user/news/newspipe/

14:19:22:546 com.mentalresonance.dust.core.actors.lib.LogActor - /user/news/newspipe/logger/ got message

Title: GM opens access to 17,000 Tesla Supercharger stations

Summary: GM drivers in the US can now purchase GM-approved North American Charging Standard (NACS) DC adapters through GM vehicle brand apps, granting access to over 17,800 Tesla Superchargers. The adapters will be available in Canada later this year, allowing drivers to locate Superchargers, check station status, initiate charges, and pay for sessions. GM plans to use multiple suppliers for the adapters, expanding access to over 231,800 public Level 2 and DC fast chargers in the US and Canada. GM aims to grow this network further through infrastructure deployment in communities and key travel corridors, supporting their commitment to an all-electric future.

from /user/news/newspipe/

14:19:51:067 com.mentalresonance.dust.core.actors.lib.LogActor - /user/news/newspipe/logger/ got message

Title: Rare Tesla Wall Connector discounts from $420, Anker 90,000mAh station return lows, Juiced JetCurrent Pro e-bike $700 off, more

Summary: Today's Green Deals feature discounts on various eco-friendly products, including Tesla's Universal and Standard Wall Connector EV charging stations starting from $420, Anker's C300 90,000mAh Power Stations from $150, Juiced's JetCurrent Pro Foldable e-bike at $2,099, Hiboy's EX6 Step-Thru Fat Tire e-bike for $800, EcoFlow's flash sale on power station bundles, and Goal Zero's Alta 80 Portable Electric Fridge and Freezer at a new low price. These deals offer significant savings on sustainable products for environmentally conscious consumers.

from /user/news/newspipe/

14:19:56:153 com.mentalresonance.dust.core.actors.lib.LogActor - /user/news/newspipe/logger/ got message

Title: Would you pay extra for bidirectional charging?

Summary: To enhance home resilience and reduce power outages, consider adopting solar energy with a battery storage system through EnergySage. This free service connects you with reputable solar installers offering competitive prices, ensuring quality solutions and savings of 20-30%. By accessing personalized quotes and unbiased Energy Advisers, you can make informed decisions without receiving sales calls until you choose an installer and share your contact information. Start your solar journey with EnergySage today.

from /user/news/newspipe/

…… [snip] ……..

14:21:22:718 com.mentalresonance.dust.core.system.GuardianActor - Guardian shut down

14:21:22:719 com.mentalresonance.dust.core.system.SystemActor - /system shut down

14:21:22:719 com.mentalresonance.dust.core.system.UserActor - /user shut down

14:21:22:720 com.mentalresonance.dust.core.system.DeadLetterActor - /system/deadletters shut down