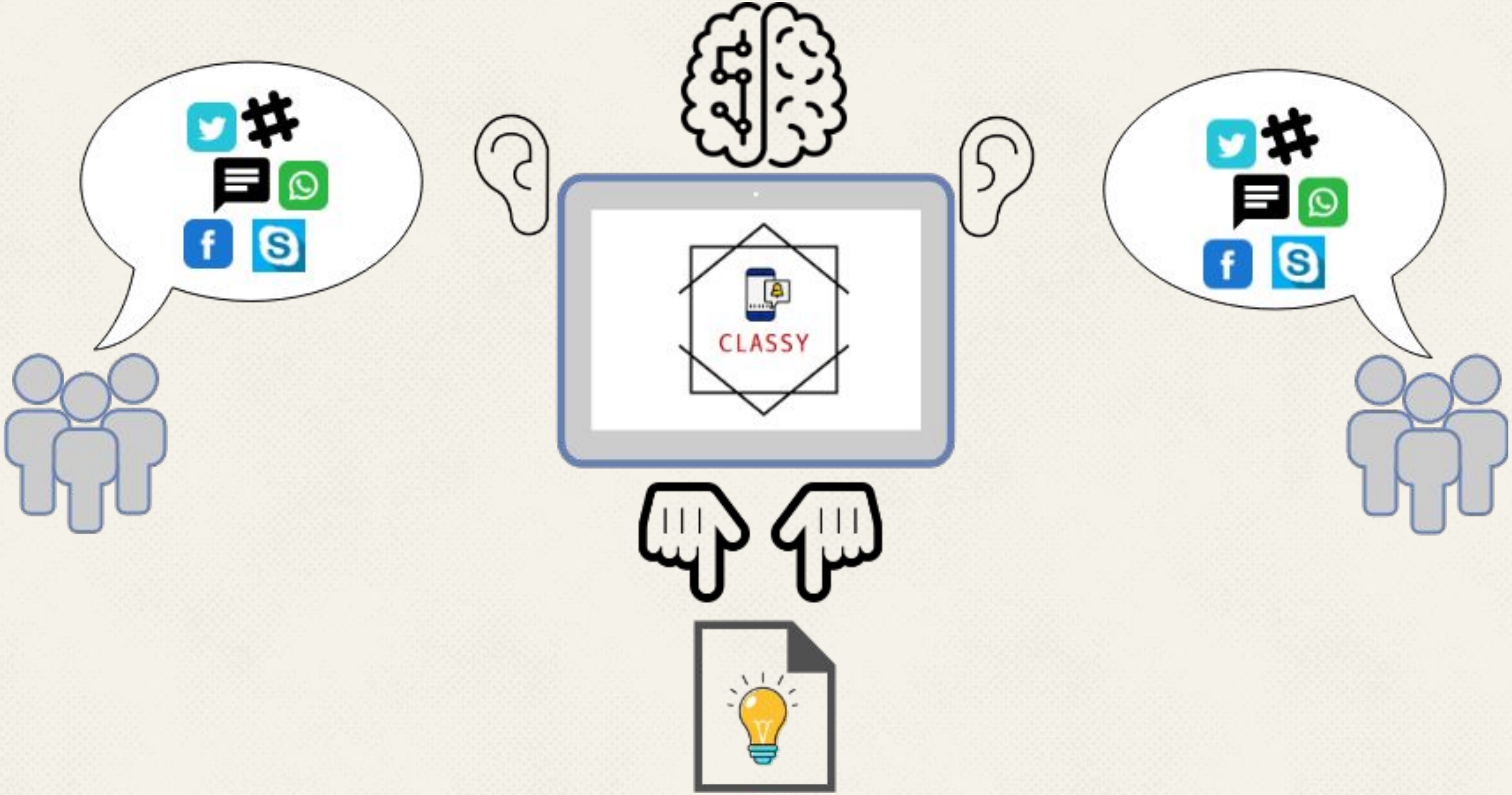
# CLASSY

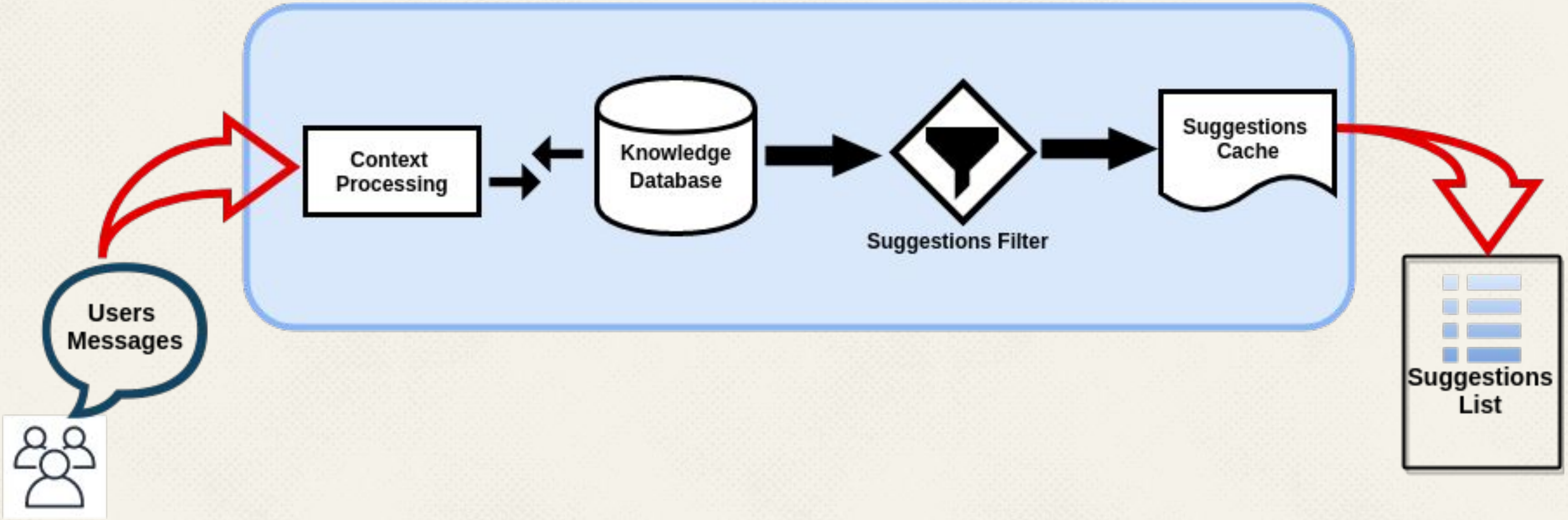
## A Conversational Aware Suggestion System

# Disclaimer

# What for?



# Inside CLASSY



# Context Processing Module

- Find the **keywords**
- Management of the **current context**
- **Implicit query formation**

```
Q1: Have you seen the last report?
Q2: The technical report?
Q3: Yes, the third one.

RTQ1: [last, report]
RTQ2: [technical, report]
RTQ3: [third]

IQ: (last report)^1 OR (technical report)^2 OR (third)^3
```

# Knowledge Database

- **Pre-process and store** information efficiently
- **Searchable information** (through queries)

# Suggestions Filter and Cache

**Post-process** the result list:

- **Filters** on document attributes
- **“Cache” to avoid same suggestions** in consecutive (or almost) time-stamped messages

# A Search Engine Based System

Knowledge database

Apache Solr

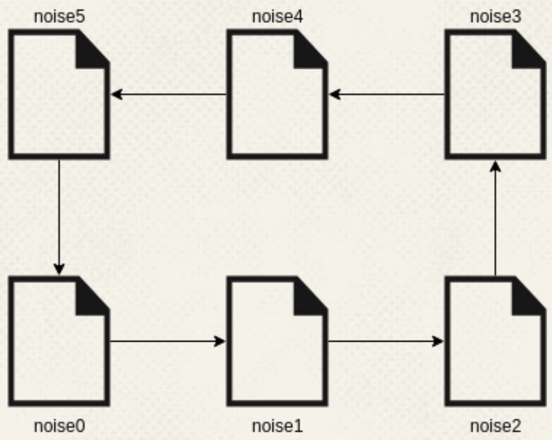
- **Full-text search**
- Real-time indexing
- **Extendable with custom plugins**
- Fault-tolerant
- Scalable

# The Noise File System

More conversational topics → More contexts → More **noise**

Noise* files

**historical registry** of exchanged messages




queries containing noise* files in the result list are **considered noisy**

# Base Approach (w/ noise file system) Results

| Number of noise files | Minimum document score | Total suggestions | Useful suggestions | Accuracy |
|---|---|---|---|---|
| 2 | 20 | 102 | 47 | 46.08% |
| **3** | **20** | **119** | **67** | **56.30%** |
| 4 | 20 | 131 | 69 | 52.67% |

# The Neighborhood Approach

**Goal:** use **semantic similarity** as a mean to get similar contexts

**How?** Using a **distribution profile**

$$P(u) = [\{w_1, o(u, w_1)\}; ...; \{w_i, o(u, w_i)\}]$$

to build **neighborhood vectors**

```
D1: This is the third report.
D2: This report describes the built automatic system.
D3: This cost report gives respect to the costs related with the
   last semester.

RTD1: [third, report]
RTD2: [report, automatic, system]
RTD3: [cost, report, cost, semester]


Report neighborhood(window = 2):
[third: 1, automatic: 1, system: 1, cost: 2, semester: 1]

Report neighborhood(window = 1):
[third: 1, automatic: 1, cost: 2]
```

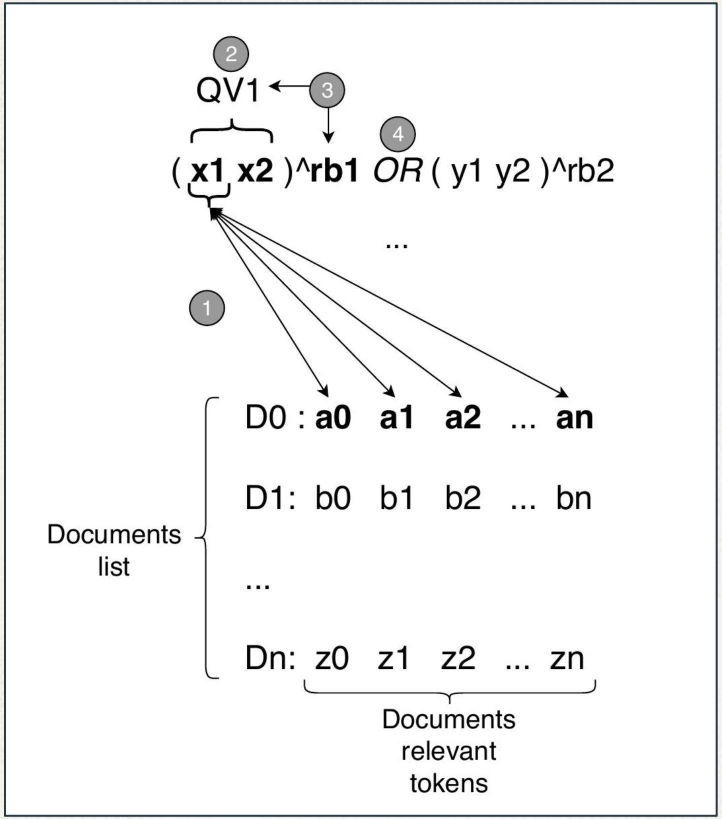
# The Neighborhood Score (NS)

$$NS(u,v) = |u| \times |v| \times cosine(u,v)$$

$$cosine(u,v) = \frac{\sum_{i=1}^{n} o(u,w_i) \times o(v,w_i)}{\sqrt{\sum_{i=1}^{n} o(u,w_i)^2} \times \sqrt{\sum_{i=1}^{n} o(v,w_i)^2}}$$

This score is a **good indicator of how related two tokens are** and inherently how can they **refer to the same context**.

# NS Computation Steps




1. **Neighborhood pair computing**
2. **Intermediate sub-query value computation**
3. **Recency boost**
4. **Final score**

# Neighborhood Approach Results

| Minimum document score | Number of noise files | Window size | Total suggestions | Useful suggestions | Accuracy |
|---|---|---|---|---|---|
| 5 | 3 | 1 | 162 | 34 | 20.99% |
| | | …. | | | |
| 20 | 3 | 1 | 35 | 26 | 74.29% |
| **23** | **3** | **1** | **31** | **25** | **80.65%** |

Although **these results seem optimistic**, the reason for this sudden accuracy increase is the **decrease of contextual suggestions concerning the direct word matches** performed by the system.

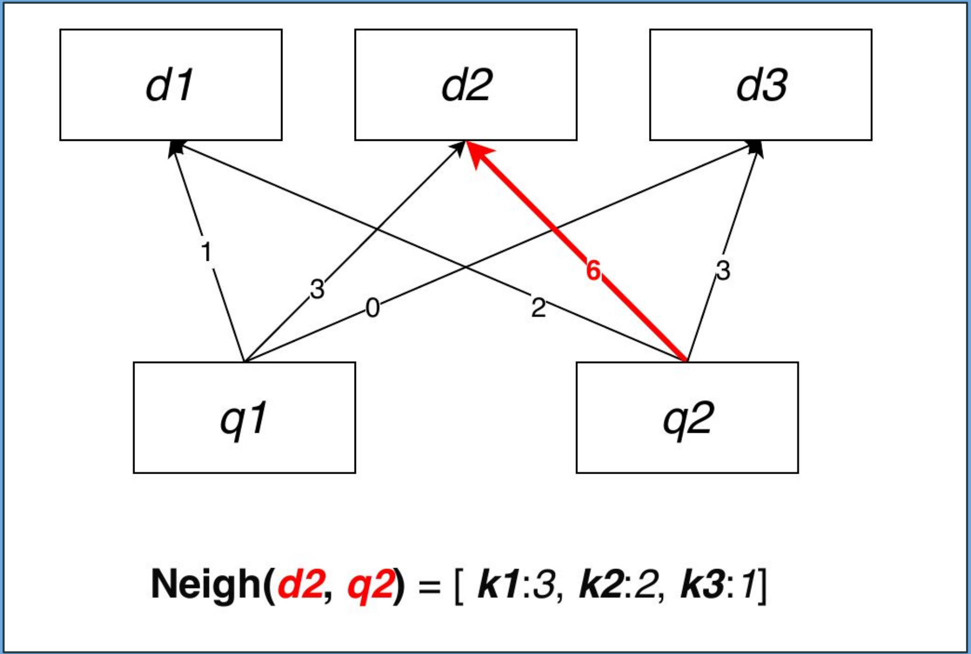
# Using Hybrid Score

*Hybrid score* = Solr score * neighborhood score

| Minimum document score | Number of noise files | Window size | Total suggestions | Useful suggestions | Accuracy |
|---|---|---|---|---|---|
| 40 | 3 | 1 | 142 | 40 | 28.17% |
| | | … | | | |
| 140 | 3 | 1 | 46 | 29 | 63.04% |
| | | … | | | |
| **220** | **3** | **1** | **35** | **27** | **77.14%** |

# The Reinforcement Learning Approach

# The (context) States




The singular context state is composed by **the top *n* tokens** that characterize the **strongest neighborhood connection**

# The (context) States

As each suggestion will have a **defined context regarding the implicit query**, it means that **each implicit query can be described as a set of probabilities** of belonging to each context state:

$$Q = \sum_{i=1}^{n} P(Q|cntx_i)$$

For each implicit query, the **contexts from the top *n* results** will be gathered and **used as the state** in order to find the action that maximizes the reward.

# The Actions

## Possible actions

**Suggest one** of the documents that composed the states vector

## Optimal action

Suggest the document with **higher probability**

# The Probabilities

## Initially :

All documents are **equiprobable**, with probability $\frac{1}{|contexts|}$

## Then:

$$P(state_i, action_j) = P(state_i, action_j) + \begin{cases} reward * alpha & \text{if } suggested \\ \frac{1}{|cntxs|-1} * -reward * alpha & \text{if } not\ suggested \end{cases}$$

# The Rewards

**1**

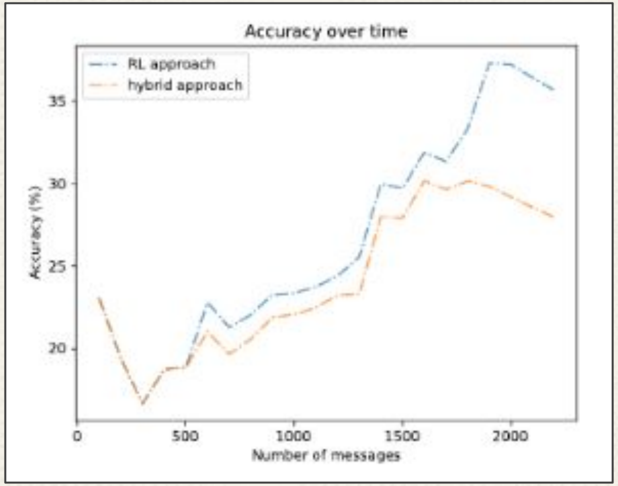
when clicked

**-1**

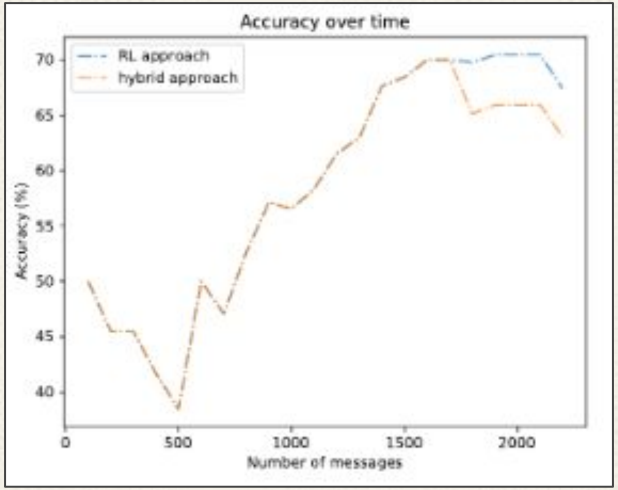
when not clicked during the defined time window

# RL Approach Results

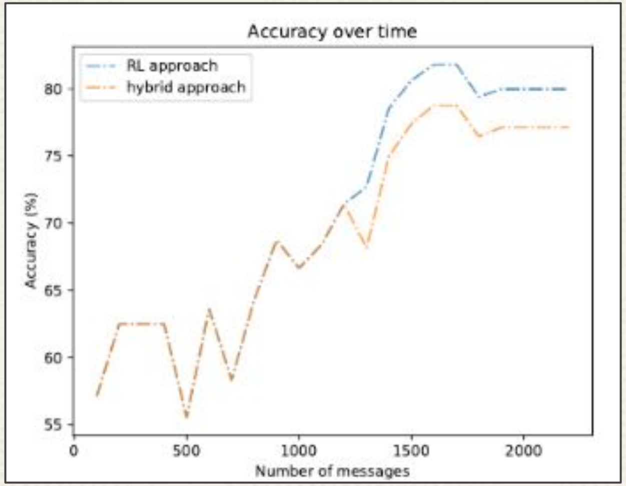
## Minimum document score = 40




## Minimum document score = 140




## Minimum document score = 220



# Trust me, it usually works

# *Discord & Wikipedia A.I* Related Articles

**Hazhor** 07/15/2019

Hi guys, welcome to our channel 😃

Let's talk about some *AI* topics like deep learning?

**CLASSY Agent** BOT 07/15/2019

💼 @here Let me suggest the following topic 💼

💻 TITLE:

Deep learning

🗒️ SUMMARY:

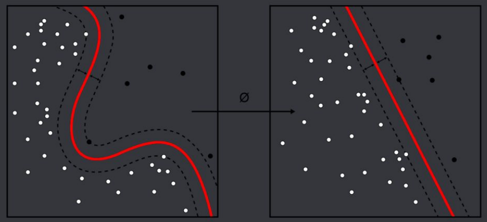
Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on artificial neural networks. Learning can be supervised, semi-supervised or unsupervised.Deep learning architectures such as deep neural networks, deep belief networks, recurrent neural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced results comparable to and in some cases superior to human experts.Artificial Neural Networks (ANNs) were inspired by information processing and distributed communication nodes in biological systems. ANNs have various differences from biological brains. Specifically, neural networks tend to be static and symbolic, while the biological brain of most living organisms is dynamic (plastic) and analog.

🔗 READ MORE:

https://en.wikipedia.org/wiki/Deep_learning

**Deep learning**

Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on artificial neural networks. Learning can be supervised, semi-supervised or unsupervised.Deep learning architectures such as deep ...



# *Hangouts & JIRA* Issues (simulated in *Discord*)

**Ryan F.** BOT 07/15/2019
o JSSIP tem esse problema dos 30s
é o timeout a determinar os ICE Candidates

**Issues Agent** BOT 07/15/2019
💼 @here Let me suggest the following issues 💼

💻 ISSUE_ID:
4195c2f2-0555-49b7-8d71-8c5822b5c85c
🗒️ SUMMARY:
Melhorias na cache de ice candidates
DESCRIPTION:
Mecanismo de evolução progressiva de timeout de negociação de ice candidates

# Future Work

- Test the whole system in a **larger dataset** (more documents and more meaningful messages)
- **Auto-tune system** - automatically find the best initial thresholds (minimum document score, etc)
- **Richer pre-processing** step (with more complex and powerful techniques)
- (add yours here)

CLASSY

# THANKS!

# *Any questions?*

You can find me at:

https://www.linkedin.com/in/diogofferreira/

https://github.com/diogofferreira