Information Retrieval - 2021-2-F1801Q110
Pirola Lorenzo - matricola 816418

Personalized Search Engine for microblog

Introduction

Aim and Strategy

- **Aim:** design and build a search engine for a micro-blog
- **Platform:** bonsai.io (Elasticsearch)
- **Crawler and Search Engine:** Elasticsearch Java REST client and Twitter4J
- **User interface:** Spring and Thymeleaf

Dataset

Users

The content of the micro-blog comes from **six** different Twitter users:

Innovation & Technology

- Bill Gates — Bill & Melinda Gates foundation, Microsoft
- Tim Cook
- Sundar Pichai — Google, Alphabet

US Politics

- Barack Obama
- Joe Biden
- Hillary Clinton

Dataset

Crawling

Users data and their tweets were collected using a **crawler** and they were stored as **documents** in Elasticsearch.

Given an username, the crawler collected the user's data and his last **1000** tweets, alternating:

- A **download** phase of 10 tweets (**retweets** are excluded).
- An **indexing** phase.
- A **break** of 10 seconds.

Dataset
Database

- Two distinct indices: one to store **tweets** and one to store **users**.
- Both of them were distributed on a **single shard** with **no replicas**.

Tweet

id: _id
text: text
parsed_text: text
hashtags: keywords array
mentions: keywords array
created_at: date
user.id: keyword
user.name: keyword
user.screenName: keyword

User

id: _id
name: keyword
screenName: keyword
hashtags: keywords array
profile: keywords array

Dataset

Pre-processing

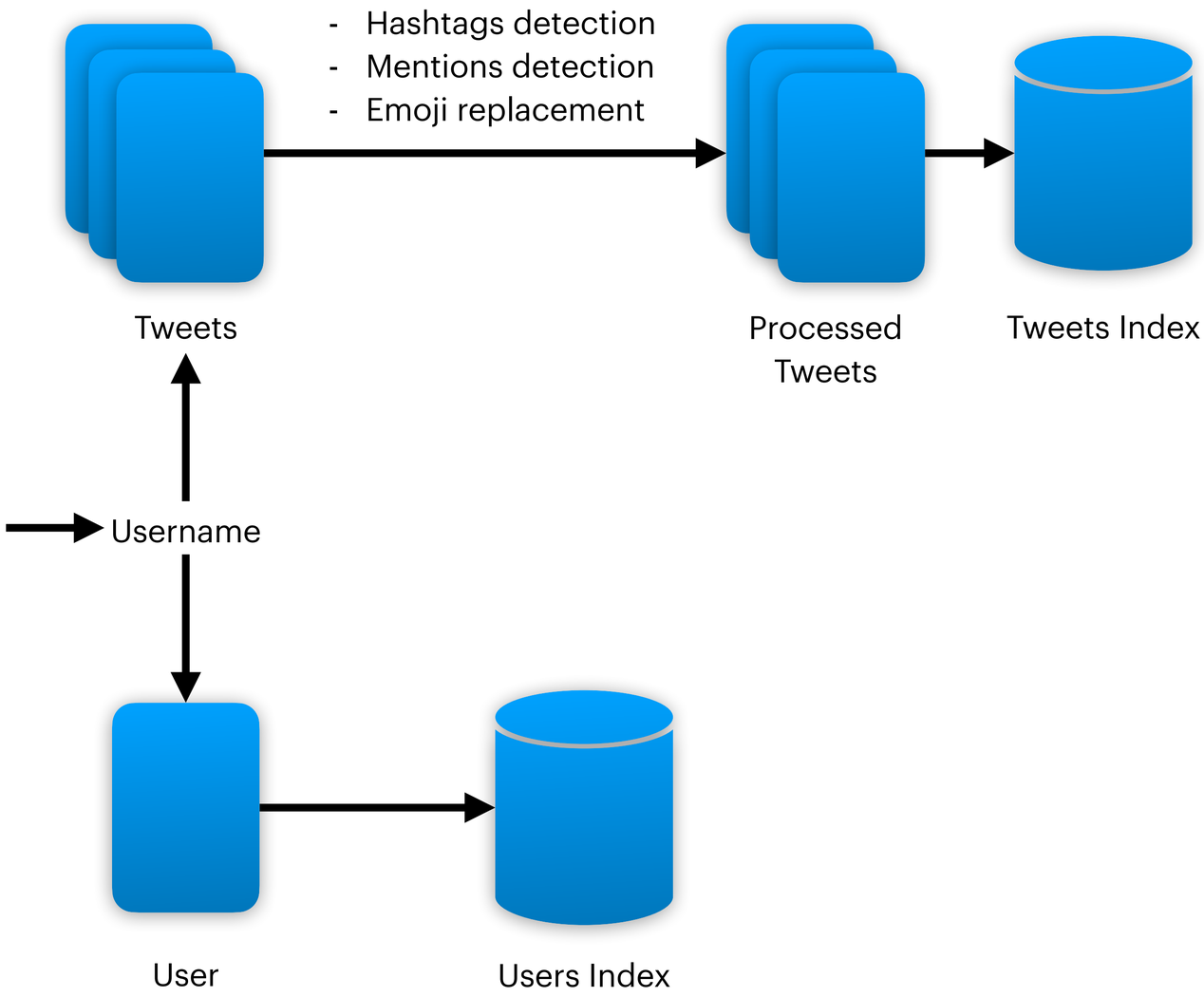
Tweets:

- Hashtags detection (#)
- Mentions detection (@)
- Emoji replacement

  (❤️ to :heart:)

Users:

- Profile and hashtags



Dataset
Statistics

- **Documents:** 4149
- **Average length:** ~ 29 words
- **Publication date:** 9/12/2015 to 23/12/2020

User	Documents	Average num. of words	Oldest Tweet	Newest Tweet
Bill Gates	748	29,295	09-04-2018	22-12-2020
Tim Cook	793	28,996	06-02-2017	17-12-2020
Sundar Pichai	573	24,120	09-12-2015	12-12-2020
Barack Obama	721	33,109	23-09-2016	22-12-2020
Joe Biden	723	27,804	14-10-2020	23-12-2020
Hillary Clinton	591	27,619	08-05-2020	22-12-2020
Summary	4149	28,688	09-12-2015	23-12-2020

Search Engine
Indexing

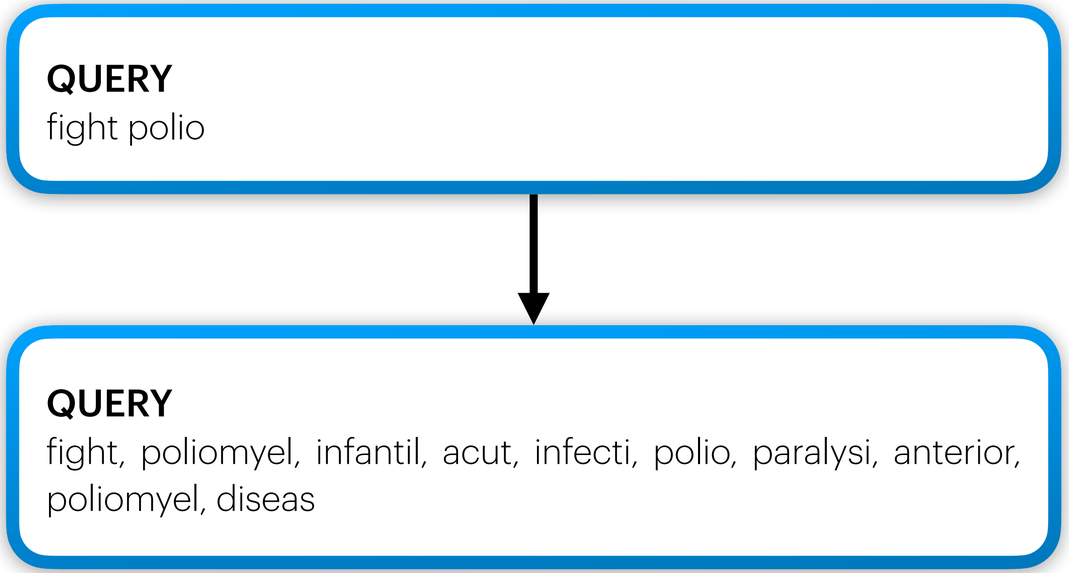
custom_analyzer:

- UAX URL Tokenizer
- Two char filters (: and _)
- Token filters:
  - Lowercase
  - Apostrophe
  - Stop
  - Porter stem

Search Engine
Indexing

custom_search_analyzer:

- UAX URL Tokenizer
- Two char filters (: and _)
- Token filters:
  - Lowercase
  - Apostrophe
  - Stop
  - **Synonym filter (new)**
  - Porter stem

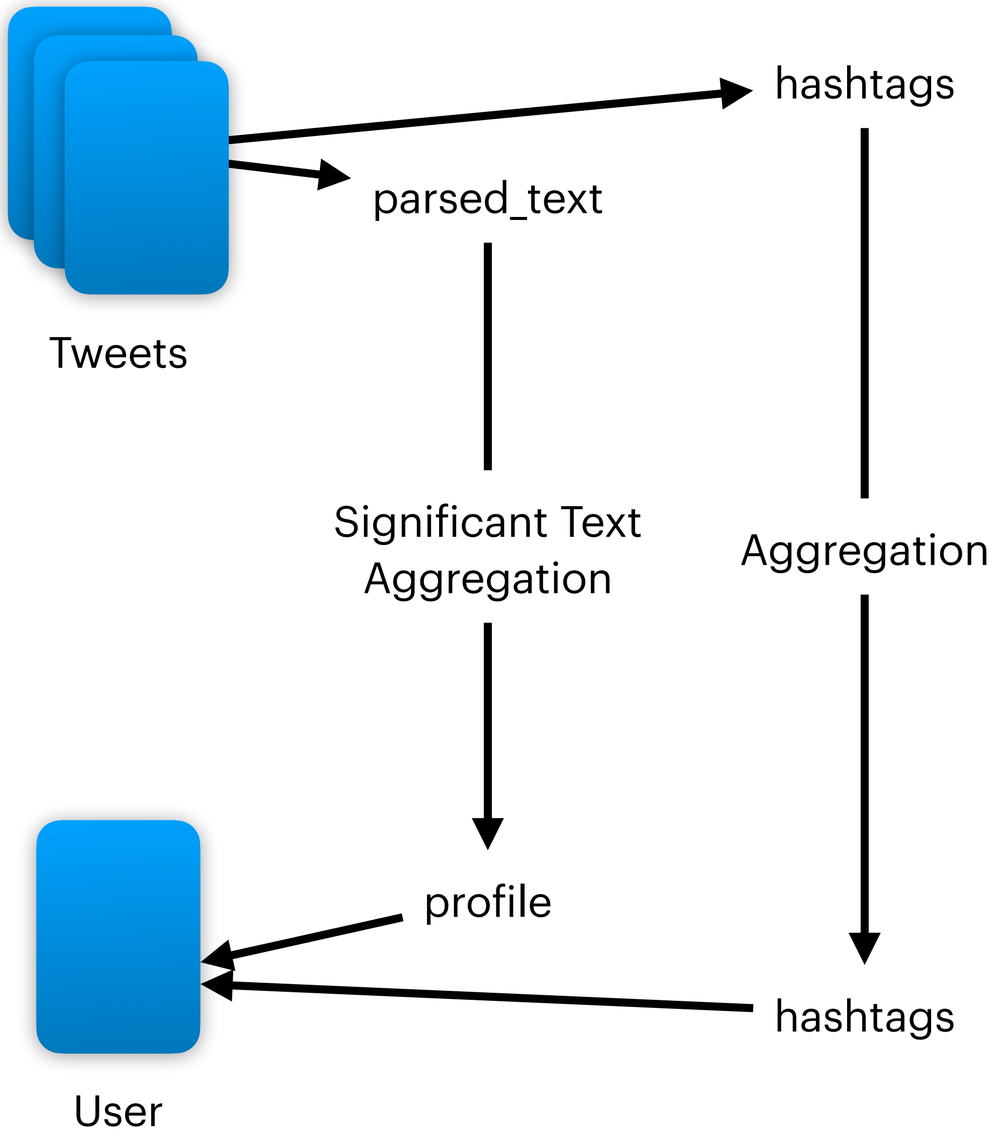


Search Engine

Personalized search

It uses a **pre-processing approach** with a keyword-based query expansion:

- **Profile:** most significant 30 terms
- **Hashtags:** most popular 10 hashtags



Search Engine

Search

Dynamic composition of a **boolean query:**

- Search over the ***parsed_text*** field using a Query String query
- Search over the ***hashtags*** or ***mentions*** fields with a Term query
- Filters over the ***created_at*** field by a Range query
- Synonym-based query expansion through the ***custom_search_analyzer***
- Keyword-based query expansion via the user profile.

Search over text

Search over hashtags and mentions

Filters over publication date

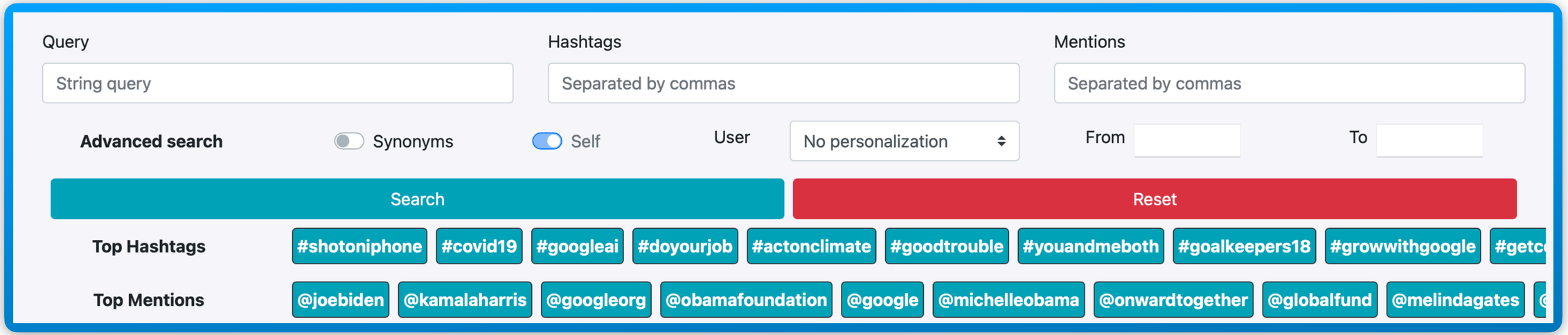
Search using synonyms

Personalized search

User cases

User Interface

- Fields needed to perform a simple or advanced search
- Lists of the most popular hashtags and mentions



User cases

Textual search on a specific field

Query: **earth**

Hashtags:

Mentions:

1. TIM COOK
Wishing everyone a happy #EarthDay2018! We all share this planet and a duty to protect it. 🌎🌍🌏

2. TIM COOK
We're celebrating **Earth** Day here at Apple. Thanks @ziggymarley for an inspired performance! 🎶🌏 https://t.co/cuNvbzdbiA

3. TIM COOK
Companies have a responsibility to use their innovation and agility to lead on the climate crisis. Thank you to Ceres for their work and for this award — and to @LisaPJackson and the team for driving us all forward. 🌎🌍🌏

4. TIM COOK
😀🌎🌍🌏📆 Happy #WorldEmojiDay! 🎉 We've got some 😎 new ones to show you, coming later this year! 👀👇 https://t.co/xBR9ZJ7l4g https://t.co/fhDrr4J5KG

User cases

Textual search on a combination of fields

Query: **portrait**

Hashtags: **#iphone**

Mentions: **@oyuelaphoto**

1. TIM COOK
Portrait Lighting on the new iPhone 8 Plus by **@oyuelaphoto** is amazing. Can’t wait to see what our customers do with the new **#iPhone**! https://t.co/5CdfEUDHCU

User cases

Personalized Search

Query: **fight** Hashtags: Mentions: User: **none**

1. HILLARY CLINTON
The single most important **fight** of our times, the **fight** that makes it possible to wage every other **fight**, is the **fight** to protect the right to vote.

2. JOE BIDEN
Donald Trump will always **fight** for his wealthy and well-connected friends. I'll always **fight** for you.

3. HILLARY CLINTON
One of the best ways to honor those we lost 19 years ago is by **fighting** for those who survived. It's outrageous that first and second responders are still **fighting** for the health care they deserve.

4. JOE BIDEN
We're **fighting** to ensure every last vote is counted across the country — and we need your help to do it. Chip in to the Biden **Fight** Fund to fuel our election protection efforts: https://t.co/VsuxvtqAFa

User cases

Personalized Search

Query: **fight**

Hashtags:

Mentions:

User: **Bill Gates**

1. BILL GATES

Glad to see this story told. I'm always inspired by the local leaders and health workers involved in the **fight** to #endpolio. https://t.co/WJuWGzcOr5

2. BILL GATES

It's hard to overstate how important finding a better diagnostic is for **fighting** #Alzheimers. That's why I'm investing in new ideas for easier and more accurate diagnosis of the disease. https://t.co/WM9phNmyJo

3. BILL GATES

Some of our most vital partners in the **fight** to #EndPolio are philanthropies like @alwaleed_philan, @bloombergdotorg, Tahir Foundation, Dalio Philanthropies and Ningxia Yanbao Charity Foundation. Their support has helped deliver vaccines and protect millions from polio paralysis.

4. BILL GATES

Melinda and I feel pressure to make every dollar and every day count, which is why we say no to a lot more opportunities than we say yes to. Working with partners around the world to #EndPolio has been a **fight** worth every dollar.

User cases

Personalized Search

Query: **fight** Hashtags: Mentions: User: **Barack Obama**

1. BARACK OBAMA
ICYMI: Read about the historic #ParisAgreement and what it means for the **fight** to #ActOnClimate. https://t.co/eMefgFZk53

2. BARACK OBAMA
This historic step in the **fight** to #ActOnClimate came faster than anyone predicted. https://t.co/W2rtcNXkI7

3. BARACK OBAMA
Today is a historic day in the **fight** to protect our planet for future generations. —President Obama #ActOnClimate https://t.co/x3dJSCYUcj

4. HILLARY CLINTON
The single most important **fight** of our times, the **fight** that makes it possible to wage every other **fight**, is the **fight** to protect the right to vote.

User cases

Personalized Search

Query: **fight** Hashtags: Mentions: User: **Joe Biden**

1. JOE BIDEN
This Domestic Violence Awareness Month, I recommit to ensuring domestic violence is treated like the crime it is, and to ensuring survivors receive the support they need. I've spent my entire career **fighting** domestic abuse — and will continue that fight as president. #DVAM

2. HILLARY CLINTON
The single most important **fight** of our times, the **fight** that makes it possible to wage every other **fight**, is the **fight** to protect the right to vote.

3. JOE BIDEN
Donald Trump will always **fight** for his wealthy and well-connected friends. I'll always **fight** for you.

4. JOE BIDEN
We're **fighting** to ensure every last vote is counted across the country — and we need your help to do it. Chip in to the Biden **Fight** Fund to fuel our election protection efforts: https://t.co/VsuxvtqAFa

User cases

Search with synonyms

Query: **poliomyelitis**

Hashtags:

Mentions:

1. BILL GATES
These global health heroes have helped Bangladesh become a model for other countries for how to respond to **infectious disease** outbreaks: https://t.co/5CEHpFCklp https://t.co/Kl6ADN43ux

2. BILL GATES
For the last 25 years, Dr. Firdausi Qadri, an immunologist and **infectious disease** researcher in Bangladesh, has been working to protect entire communities from cholera epidemics. https://t.co/F2pQilYqry

3. BILL GATES
Our best strategy in the age-old fight against the germs is our collaborative, data-based effort to study the world around us and within us. Our best strategy is science. @MaxCRoser explains how vaccines have dramatically improved humanity's ability to fight **infectious disease**. https://t.co/1xWzzOapYw

4. BILL GATES
Almost a decade ago, I predicted that the world would soon be **polio** free. But we had more **polio** cases in 2018 than in 2017. Even so, I remain optimistic that we could eradicate **polio** soon.

Thank you for your attention