Information Retrieval

A personalized Search Engine for microblog content

Matteo Angelo Costantini - 795125
Dario Gerosa - 793636
Michele Perrotta - 795152

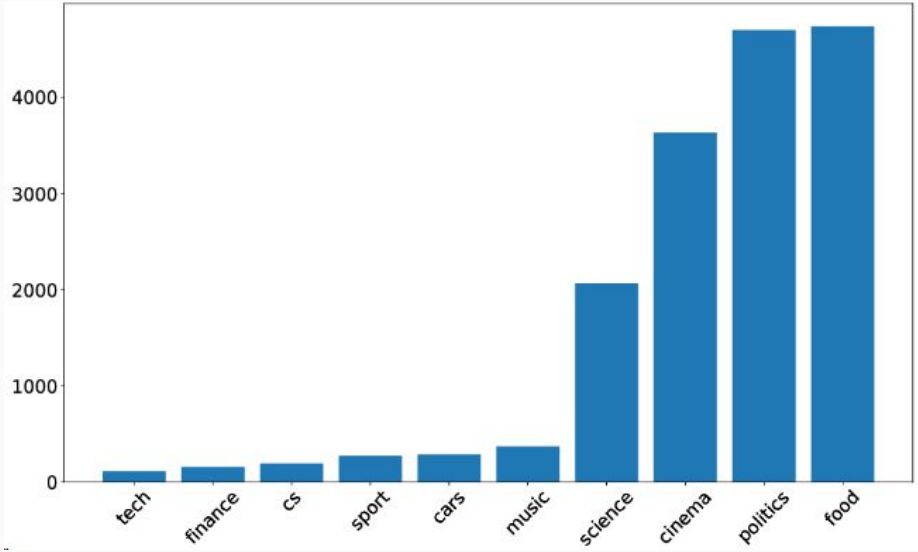
https://github.com/CostantiniMatteo/progetto-ir

Goal

- Crawl tweets dealing with different topics using Twitter APIs
- Create a Search Engine
  - Must support both personalized and not personalized search
  - At least five user profiles with three topic of interest
  - Interest extracted from documents given by the user

Crawler

- Python script using tweepy APIs wrapper
- Multiprocess approach to speed up the crawling process
- Online PostgreSQL database for persistence
- 10 topics



Dataset

- ~ 36 000 000 tweet
- ~ 16 000 users
- Up to 3 200 tweet for each user
- 60 GB of data
- Tweets only in English
- JSON Format

Percentile	Follower
5	226
10	391
20	746
30	1,210
40	1,879
50	2,954
60	4,625
70	7,851
80	14,563
90	39,785
95	105,922
97	231,844
99	1,262,254
99.5	3,120,360
100	104,683,236

Search Engine

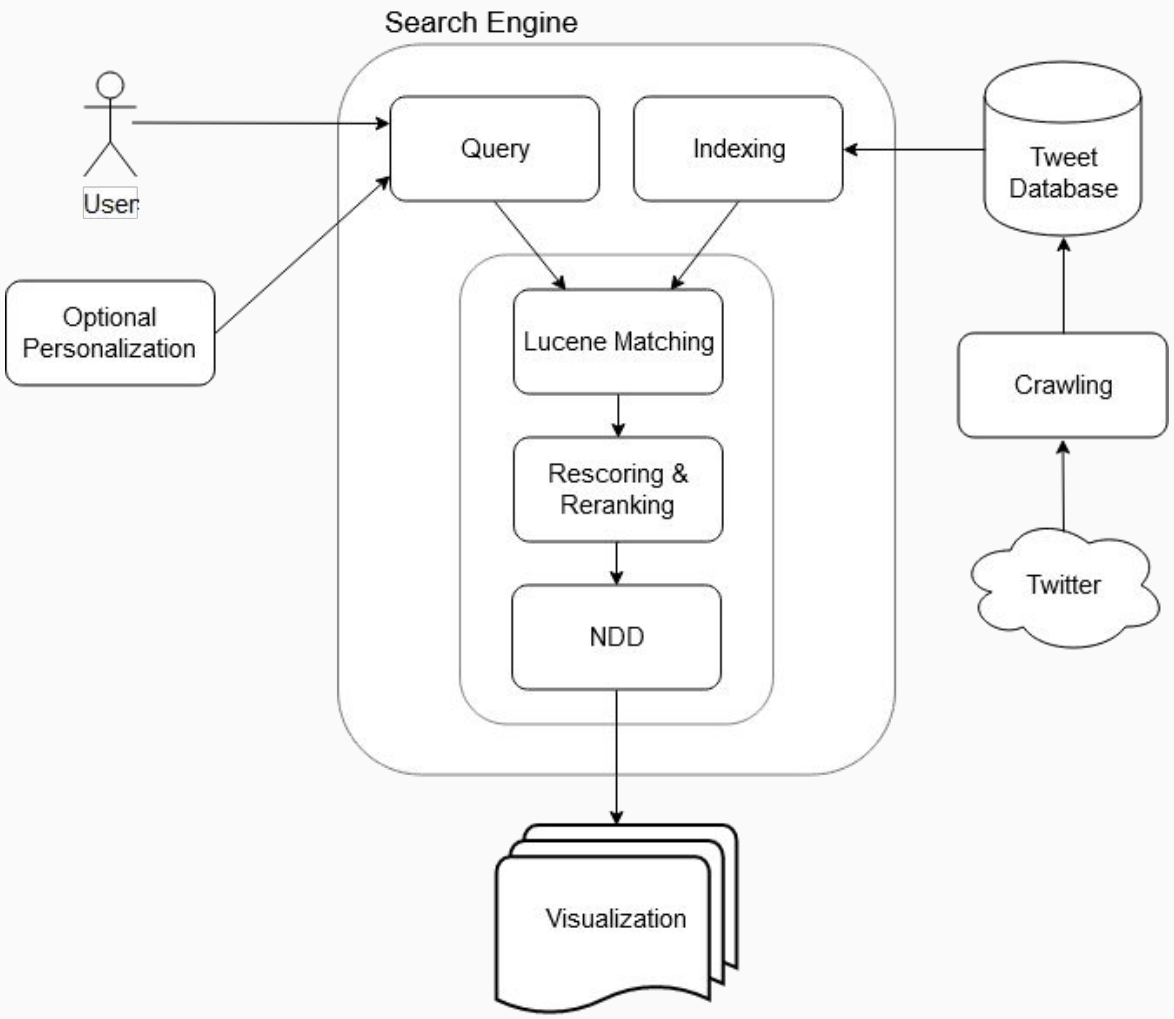
Overview

The Search Engine uses Lucene ***TF- IDF similarity*** to index the documents crawled from Twitter.

Query **personalization** is achieved using *Query Expansion* and ***bag-of-words*** model.

The final output is the product of a ***re-ranking*** phase using a combination of the Lucene matching algorithms with other custom scores followed by a ***Near Duplicate Detection*** step.

A Web App is used to interact with the Search Engine and visualize the results



Indexing

Analyzer:

- Based on Lucene's Classic Analyzer
  - Preserves URLs, emails and numbers separated with hyphens
- Porter's Stemmer
- Removal of URLs and emails
- Stopword removal using Lucene default list
- Tokens normalized to lowercase

Indexer:

- Lucene's Classic Similarity
  - Refinement of the cosine-similarity based on TF-IDF
- Indexed Fields:
  - Text, Hashtags for text search
  - Date for range queries
  - Retweet count, author's data and other for scoring

Queries

Two main types of queries:

1. Based on recent information
   Focused on new tweets
2. Based on the whole information available in the dataset

Characteristics:

- Both queries support personalization based on user interests
- Re-Scoring based on Twitter's nature
- Near Duplicates Detection
- Range queries on Date

Queries - 1

- Boolean Model retrieves relevant documents and chronological ordering of the results
- Selection of *n* most recent documents
- Scoring of the subset of documents using the Vector Space Model with Lucene's Classic Similarity and other Twitter-related factors such as:
    - Retweet rate
    - User's influence
    - Presence of URL

Nagmoti, Rinkesh, Ankur Teredesai, and Martine De Cock. "Ranking approaches for microblog search." *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 2010.

Queries - 2

- Default Lucene's approach: Boolean Model followed by Vector Space Model to score the documents
- Select the *n* most relevant documents
- Re-score & Re-rank the subset using a linear combination of the Similarity Score and other Twitter-related factors such as:
  - Retweet rate
  - User's influence
  - Presence of URL

Nagmoti, Rinkesh, Ankur Teredesai, and Martine De Cock. "Ranking approaches for microblog search." *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 2010.

Scoring

To Re-Score the documents we took into account the score given using the revised cosine similarity computed by Lucene but also other factors such as:

- The number of followers of the author of a tweet
- The number of retweets
- The length of the text
- The presence/absence of an URL
- The tweet being a quote/retweet or not

Scoring - Equations

- Base Lucene Score

$$bS = lw \cdot \frac{s_d}{max_{i \in R}\ s_i}$$

- Follower Score

$$fS = fw \cdot \frac{ufi_d}{ufi_d + ufo_d}$$

- Retweet Score

$$rS = rw \cdot \frac{r_d + fav_d}{max_{i \in R}\ r_i + fav_i}$$

- Quote Score

$$qrS = q_d \cdot qw + r_d \cdot rw$$

- Length Score

$$lS = lw \cdot \frac{l_d}{max_{i \in R}\ l_i}$$

- URL Score

$$uS = uw \cdot u_d$$

Final Score: $S = bS + fS + rS + qrS + lS + uS$

Personalization

Outline

- Document based personalization
- Bag-of-Words model for each topic
  - Topic dependent
- The bag is computed indexing the given documents and either using the whole set or using a subset of most informative (based on TF-IDF) terms
- Query expansion to include the terms defining the interests of the user

Profiles:

- Five users with different topics of interest
- At least three topics for each user
- At least 10 documents for each topic
- A customizable user using the web app

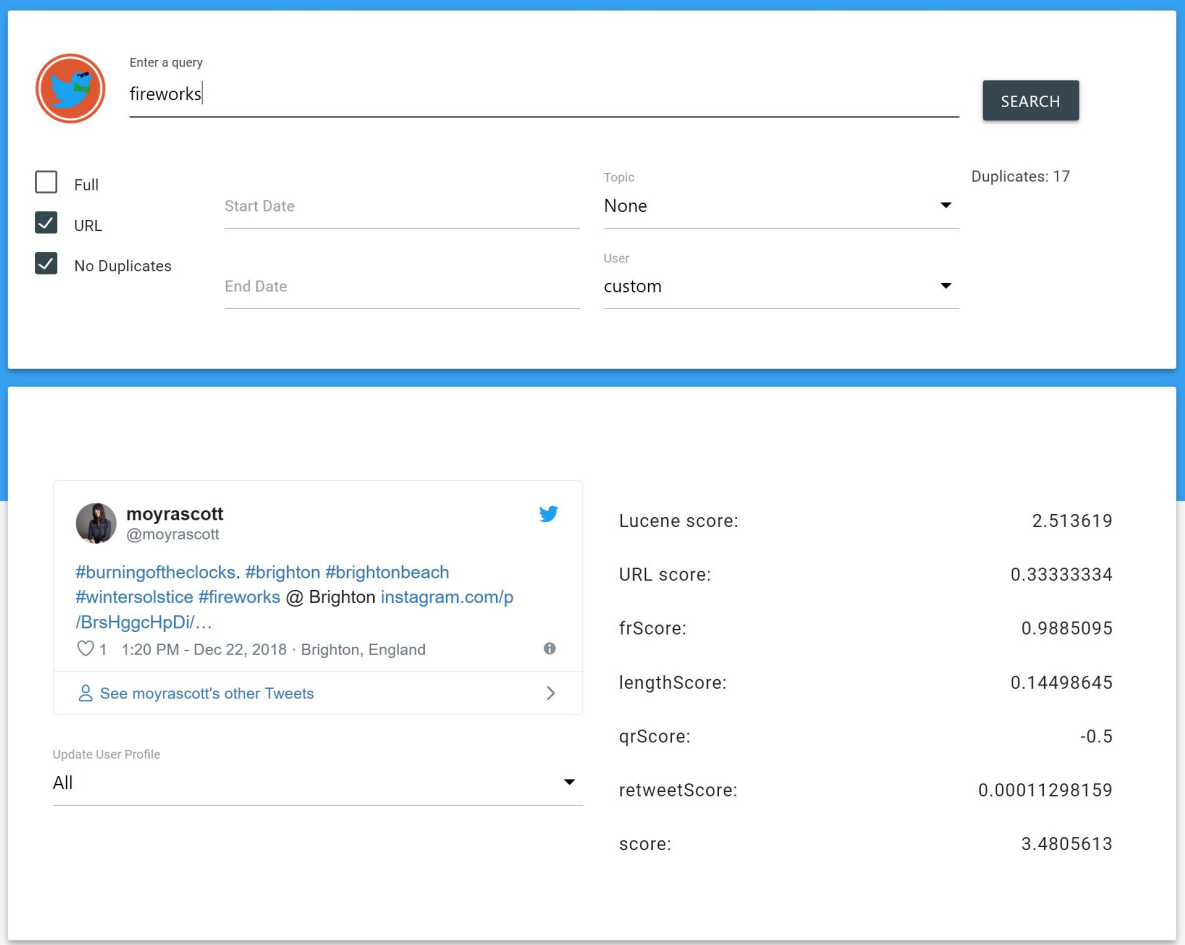
Near Duplicate Detection

- We chose to use the Overlap coefficient since the Jaccard coefficient led to a lot more false negatives when one tweet overlaps with another
- The set used to compare the documents consists of the bi-grams extracted from the text of the tweet.
- Not much slower that Jaccard coefficient approach on a limited number of documents (tested with about 150 documents)
- Threshold at 80% of overlap

Interface

Supports:

- Interactive tweets
- Personalization based on user profile and topic
- Range queries on the Date Fields
- Duplicates detection and filter
- URL scoring
- User profile creation



Query Personalization - Examples

Enter a query
car
SEARCH

- [x] Full
- [x] URL
- [x] No Duplicates

Start Date
End Date

Topic: None
User: custom

Duplicates: 14

Alison Sudol @AlisonSudol
cool kid #seattle #car #vintage instagram.com/p/0EpOW_mh_2/
6 4:36 AM - Mar 11, 2015
See Alison Sudol's other Tweets

DUB Magazine @dubmagazine
Winter fun @_vehiclesdaily_ ghostriderto _tunersdaily_ ferrari_458_ #cars #car #sportscar… instagram.com/p/BORHBP0A3fh/
4 6:24 AM - Dec 21, 2016
See DUB Magazine's other Tweets

DUB Magazine @dubmagazine
cheshire_cars
ONLY the BEST #onyx #onyxconcept #GTX700S #bentley #car #cars #carspotter… instagram.com/p/BN2v2UqgzSo/
2 12:41 AM - Dec 11, 2016
See DUB Magazine's other Tweets

Top Daddies @TopDaddies
What does winter do to your car? @KalTire helps answer this question! Read more: topdaddies.com/spring-tune-ve… #car #cars
2 9:49 PM - Apr 10, 2016
See Top Daddies's other Tweets

Kenneth Warner @KennyWarner
Made my mark on the #Toyota #specialolympics2015 car. Thanks for sponsoring!
---#car #cars… instagram.com/p/-ZMAa7mcNm/
2 5:19 PM - Nov 22, 2015
See Kenneth Warner's other Tweets

Enter a query
car
SEARCH

- [x] Full
- [x] URL
- [x] No Duplicates

Start Date
End Date

Topic: Cars
User: user2

Duplicates: 25

Electric Road Trip @EVehicular
Inside the Tesla electric car factory - Telegraph.co.uk bit.ly/1nyIsvi
8:54 PM - Jan 27, 2016

@Telegraph
Latest news, business, sport, comment, lifestyle and culture from the Daily Telegraph and Sunday Telegraph newspapers and
telegraph.co.uk

See Electric Road Trip's other Tweets

Lalit Kumar Modi @LalitKModi
Check this video out. If True - then it will really #change our #lives 😃😃😃😃🤗#tesla #electric #car #gasoline #h2opower
1,751 10:33 PM - Feb 24, 2017
4,073 people are talking about this

BMW | MotorShowBlog @MotorShowBlog
Why the BMW X Coupé Concept was a groundbreaking car at the #NAIAS back in 2001 - bit.ly/1SFKuFv #BMWNAIAS
2 7:15 PM - Jan 9, 2016
See BMW | MotorShowBlog's other Tweets

GermanCarForum @GermanCarForum
World Premiere - BMW X7 (G07) - More pics and info -> germancarforum.com/threads/bmw-x7… #bmw #x7 #bmwx7 #g07 #xdrive #bimmer #bmwx #bmwm #sav #suv #suvlife #bmwsav #luxury #luxurysuv #luxurylifestyle #coolcars #dreamcar #dreamgarage #car #instacar #carinstagram… dlvr.it/QnZGYz
2 3:17 AM - Oct 17, 2018
See GermanCarForum's other Tweets

Jim Harris | #WEF19 @JimHarris
MILESTONE SURPASSED: Tesla has hit a significant milestone: 100,000 Model 3 electric cars have left the factory, according to Bloomberg. bloom.bg/2Aa4Eop #TeslaModel3 #Tesla #EV #Automotive #auto #ElectricVehicle #Model3 #cars #disruption #disruptiveinnovation #innovation
57 7:07 PM - Oct 14, 2018
33 people are talking about this

Query Personalization - Examples

Enter a query
stock
SEARCH

- [x] Full
- [x] URL
- [x] No Duplicates

Start Date
End Date

Topic: None
User: custom

Duplicates: 15

Scott Minerd @ScottMinerd
The #stock rally seems to confirm that the bull market is intact, and that the correction is probably over.
42 8:39 PM - May 10, 2018
19 people are talking about this

Anmol Singh @DeltaNinety
The Biggest Reasons Why New Stock Traders Fail Miserablyow.ly/7zY830ltmCE#daytrading #trading #stock
2 6:10 PM - Aug 20, 2018
See Anmol Singh's other Tweets

Anmol Singh @DeltaNinety
Becoming a Successful Stock Trader – Do You Have What it Takes?ow.ly/Ujvu30lsxp9#daytrading #trading #stock
2 11:00 PM - Aug 19, 2018
See Anmol Singh's other Tweets

Wes Moss @WesMoss365
Q: Where on the continuum is a happy medium #stock? bit.ly/2p9aRtQ #stocks #investing #money
2 8:05 PM - Sep 30, 2017
See Wes Moss's other Tweets

I Know First @i_Know_First
Cleveland-Cliffs Stock Forecast: Cleveland-Cliffs (NYSE: #CLF) is Back to Basics ow.ly/jnTO30lciqJ #fintech #AI #stock #stockstowatch #stocks #TradingStrategy #trading #investing #markets #stockmarkets #WallStreet #nasdaq
2 2:00 AM - Aug 1, 2018

Cleveland-Cliffs Stock Forecast: Clevela...
This article was written by Isabelle Tao, a Financial Analyst at I Know First. CLF Stock Forecast "Cleveland-Cliffs was founded and
iknowfirst.com

See I Know First's other Tweets

Enter a query
stock
SEARCH

- [x] Full
- [x] URL
- [x] No Duplicates

Start Date
End Date

Topic: Finance
User: user1

Duplicates: 10

StockTwits @StockTwits
What if you invested $10,000 into a popular stock two years ago? How much would you have today? Your answer:

- $NVDA $62,111
- $MU $54,205
- $NFLX $37,148
- $AMD $32,561
- $BA $25,920
- $ADBE $25,204
- $AMZN $22,579
- $AAPL $17,961stocktwits.com/saaketham/mess…

227 4:03 PM - May 16, 2018

Russian Market @russian_market
Roses are red
Violets are blue
Buy Apple stock on long overdue

TP: $200 $AAPL #applestock #aapl #apple #buythefuckingdip @ Paradeplatz instagram.com/p/BrN5X-qHWCk/…
8 7:38 PM - Dec 10, 2018 · Zurich, Switzerland
See Russian Market's other Tweets

Ross Gerber @GerberKawasaki
FAANG is getting old. The future is TAND. Check out my new blog post on tech stocks. $tsla $atvi $nvda $dis #tesla #activision #nvidia #disney forbes.com/sites/greatspe…
98 5:22 PM - Nov 7, 2018

StockTwits @StockTwits
AMD. It was once the stock everyone made fun of.

Now? It's popping off. It just traded at $22/share. These are its highest levels since 2006. 🚀

Over 132,000 people watch $AMD on StockTwits. Congrats. stocktwits.com/symbol/AMD
93 7:31 PM - Aug 23, 2018
33 people are talking about this

Query Personalization - Examples

Enter a query
turtle
SEARCH

- [x] Full
- [x] URL
- [x] No Duplicates

Start Date
End Date
Topic: Cars
User: user2
Duplicates: 1

Anmol Singh
@DeltaNinety
Are You a Turtle or a Rabbit? #turtle #Rabbit #Preppingforsuccess livetraders.com/are-you-a-turt…
2 6:05 PM - Sep 21, 2018
See Anmol Singh's other Tweets

Hilton Head Island
@hiltonheadsc
It's #turtle time! 🐢 Sea turtles, that is. Here's a guide to hatching and nesting season on #HiltonHeadIsland: hiltonheadisland.org/island-time/ou… #LowcountryLife

📷 : jordanski32
52 1:50 AM - Jul 30, 2018
See Hilton Head Island's other Tweets

Shontelle Layne
@Shontelle_Layne
Closeup. Birdseye on a Turtle. #Barbados #BarbadosSand #art #sandsculpture #turtle #tbt 🐢 By the… instagram.com/p/1RdU_CqWux/
1 12:34 AM - Apr 10, 2015
See Shontelle Layne's other Tweets

Grok Learning
@groklearning
Here's a fun and floral coding project to try with your class, using our free Turtle Playground: blog.groklearning.com/turtle-flowers…#programming #turtle #art
5 5:40 AM - Jan 11, 2018
See Grok Learning's other Tweets

Johnny Weir
@JohnnyGWeir
My Happy Place | princessjax210 | #throwbackthursday #turtle #delaware instagram.com/p/BYwXsgOAXdD/
74 11:00 PM - Sep 7, 2017
See Johnny Weir's other Tweets

Enter a query
lollipop
SEARCH

- [x] Full
- [x] URL
- [x] No Duplicates

Start Date
End Date
Topic: Tech
User: user5
Duplicates: 1

Ganesh Mannar
@gansmb
@oneplus receives Android 5.1 #Lollipop update via unofficial #CyanogenMod 12.1: How to install ibt.uk/A006Gb0 #oneplusone
6:08 PM - Mar 28, 2015
See Ganesh Mannar's other Tweets

Android Developers
@AndroidDev
Developers, get ready for Android 5.0 Lollipop! Check back 10/17 for Android 5.0 SDK & updated Nexus preview images. developer.android.com/preview
310 5:20 PM - Oct 15, 2014
Android 9 Pie | Android Developers
developer.android.com
731 people are talking about this

Lollipop time 🍦 #childhoodunplugged #letthembelittle #motherhood #lollipop #septembersun #… https://t.co/3doNxesDAI https://t.co/clgIABpQ4o

Date: 2016-09-13T12:56:48.000+0000

Author: wrensteaparty

Times of India
@timesofindia
Google announces Android Lollipop, Nexus 6 smartphone timesofindia.indiatimes.com/tech/tech-news…
39 6:20 PM - Oct 15, 2014
54 people are talking about this

Thanks for your attention

User 1:
cs: java programmer
cinema: avengers superheroes morricone sergio leone
tech: intel amd ryzen nvidia
politics: trump intel top secret russia
finance: nvidia amd apple microsoft stock

User 2:
tech: apple
music: motorhead ace of spades metallica soad
cars: tesla bmw

User 3:
sport: federer serena williams
science: physics higgs
cars: audi volkswagen

User 4:
sport: golf francesco molinari
music: madonna 50 cents
food: fruit pasta

User 5:
tech: lollipop android
food: lollipop candy marshmallows
sport: manchester united premier league chelsea arsenal