Designing a One-Click Access Information Retrieval System for 1CLICK-2@NTCIR-10

Niek Tax & Dan Ionita
supervised by Djoerd Hiemstra

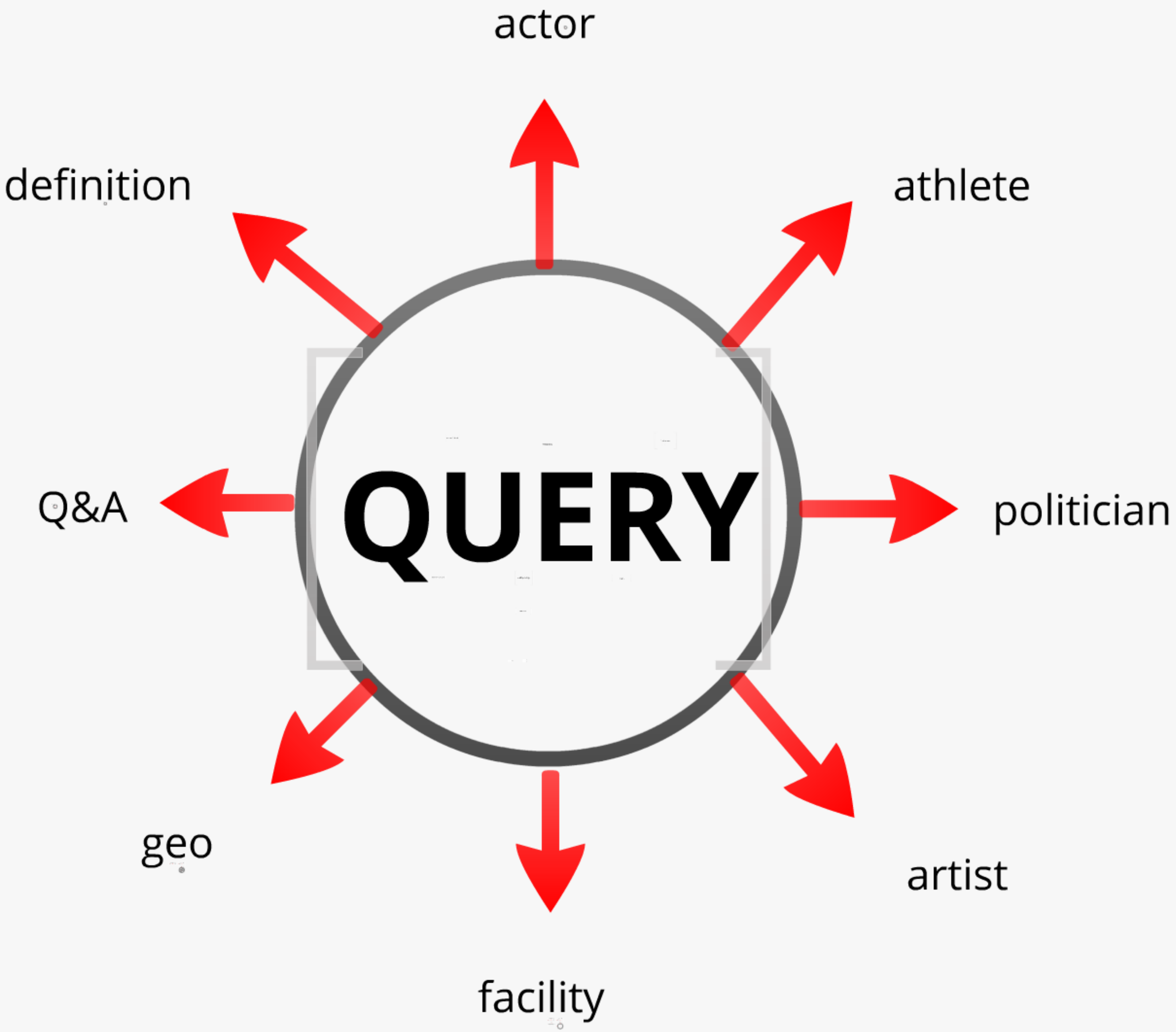
“Classic“ search

1. enter query
2. click search button
3. scan a list of URL's
4. click some URL
5. repeat step 3 and 4

“One-Click“ search

1. enter query
2. click search button

Finding information using a Search Engine



actor

definition

athlete

QUERY

Q&A

politician

geo

artist

facility

[illegible]

Probabilities

Place type code

QUERY

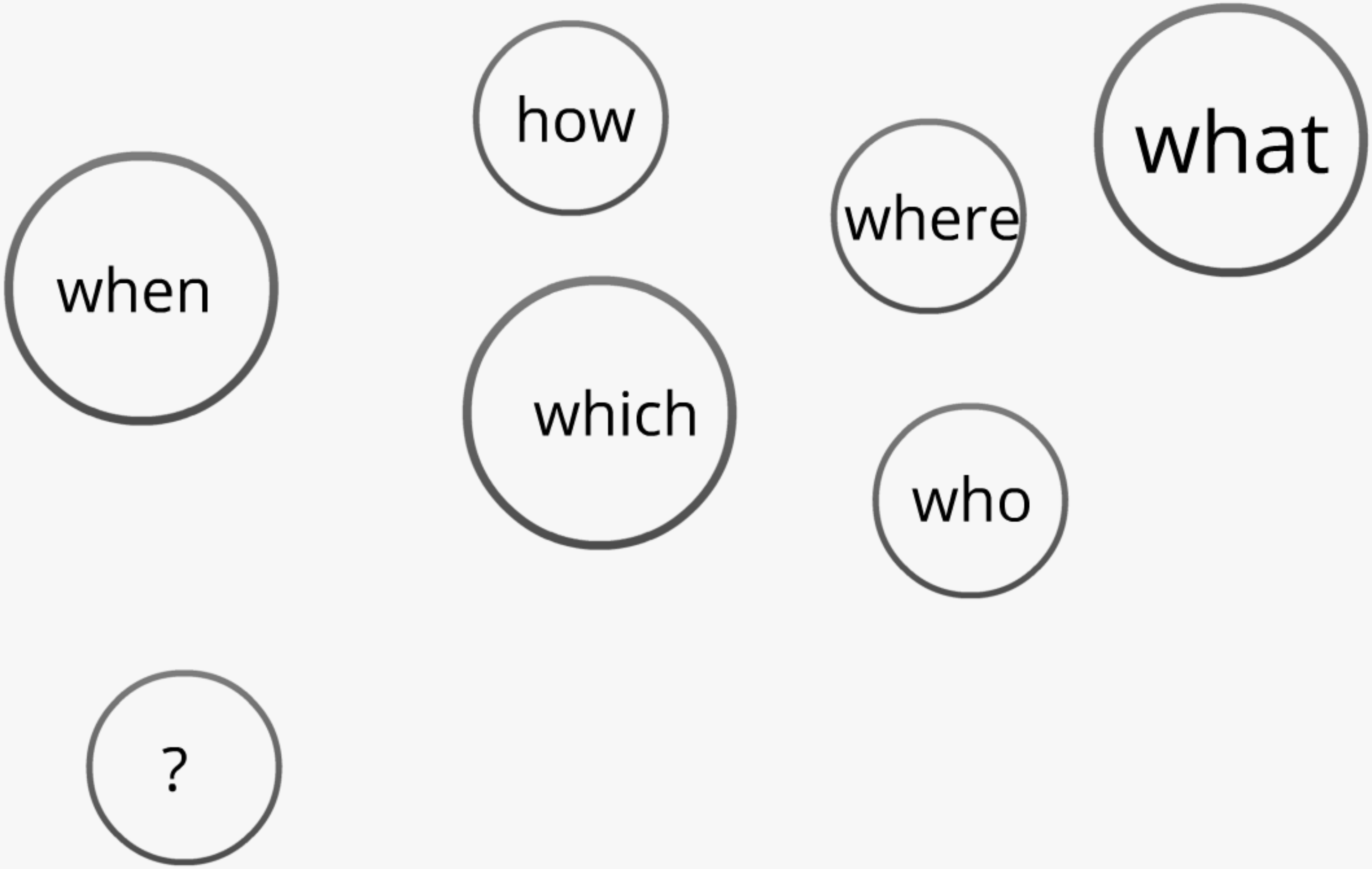
sentence pattern

hasWikipediaPage

length

clue words

sentence pattern



what

how

where

when

which

who

?

hasWikipediaPage

length

clue words

south

west

drive

north

road

east

south

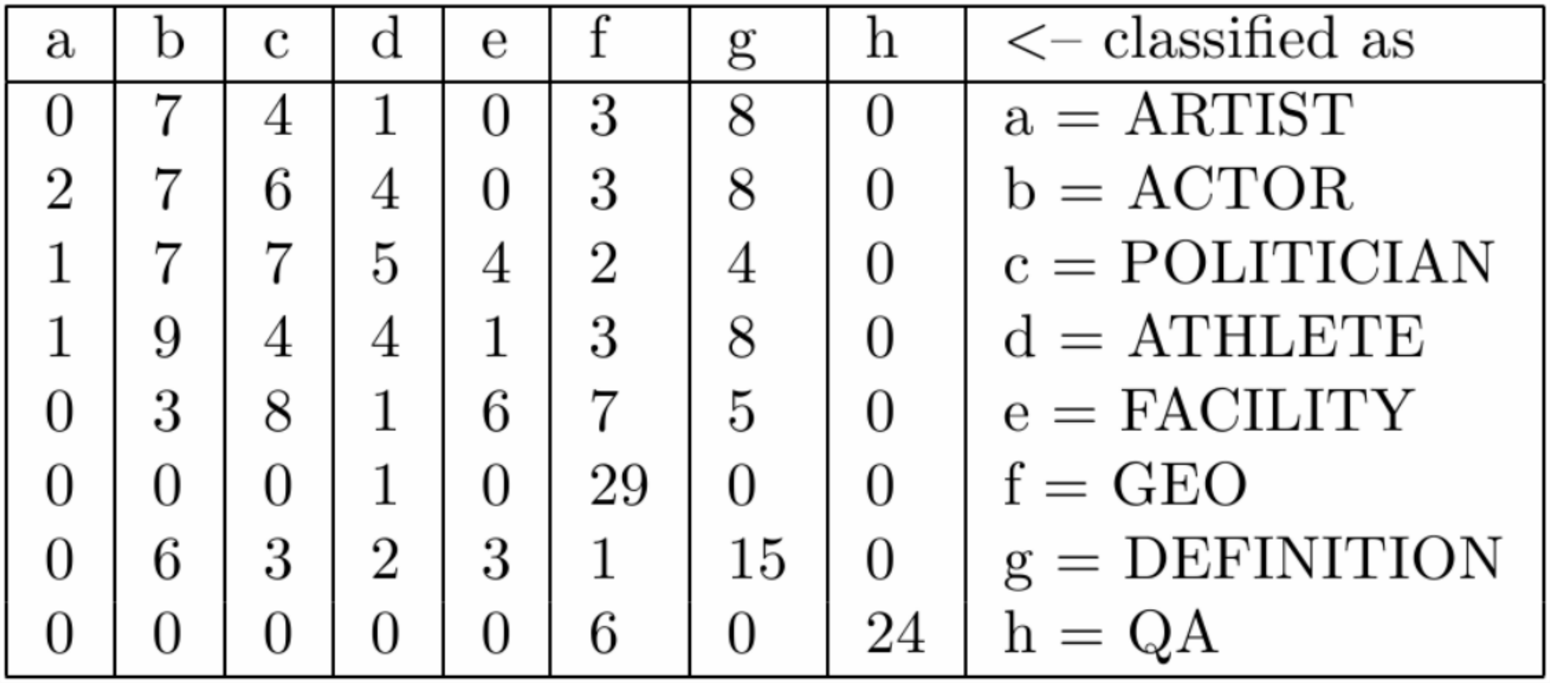
west

street

north

road

east

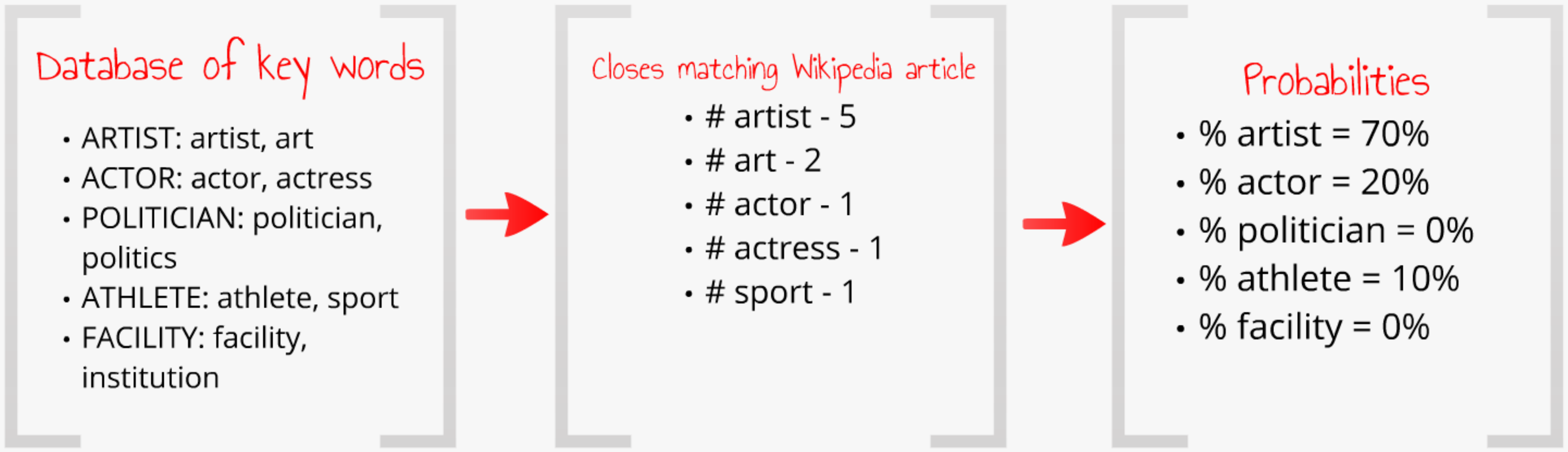


a	b	c	d	e	f	g	h	<– classified as
0	7	4	1	0	3	8	0	a = ARTIST
2	7	6	4	0	3	8	0	b = ACTOR
1	7	7	5	4	2	4	0	c = POLITICIAN
1	9	4	4	1	3	8	0	d = ATHLETE
0	3	8	1	6	7	5	0	e = FACILITY
0	0	0	1	0	29	0	0	f = GEO
0	6	3	2	3	1	15	0	g = DEFINITION
0	0	0	0	0	6	0	24	h = QA

38.3 % correctly classified instances

Probabilities

Learn frequency of keywords in Wikipedia page and then each of those to one of the categories



Database of key words

- ARTIST: artist, art
- ACTOR: actor, actress
- POLITICIAN: politician, politics
- ATHLETE: athlete, sport
- FACILITY: facility, institution

Closes matching Wikipedia article

- # artist - 5
- # art - 2
- # actor - 1
- # actress - 1
- # sport - 1

Probabilities

- % artist = 70%
- % actor = 20%
- % politician = 0%
- % athlete = 10%
- % facility = 0%

Count frequency of keywords in Wikipedia page and link each of them to one of the categories

[redacted]

percentage of real words

Place type code

Yahoo's GeoPlanet API

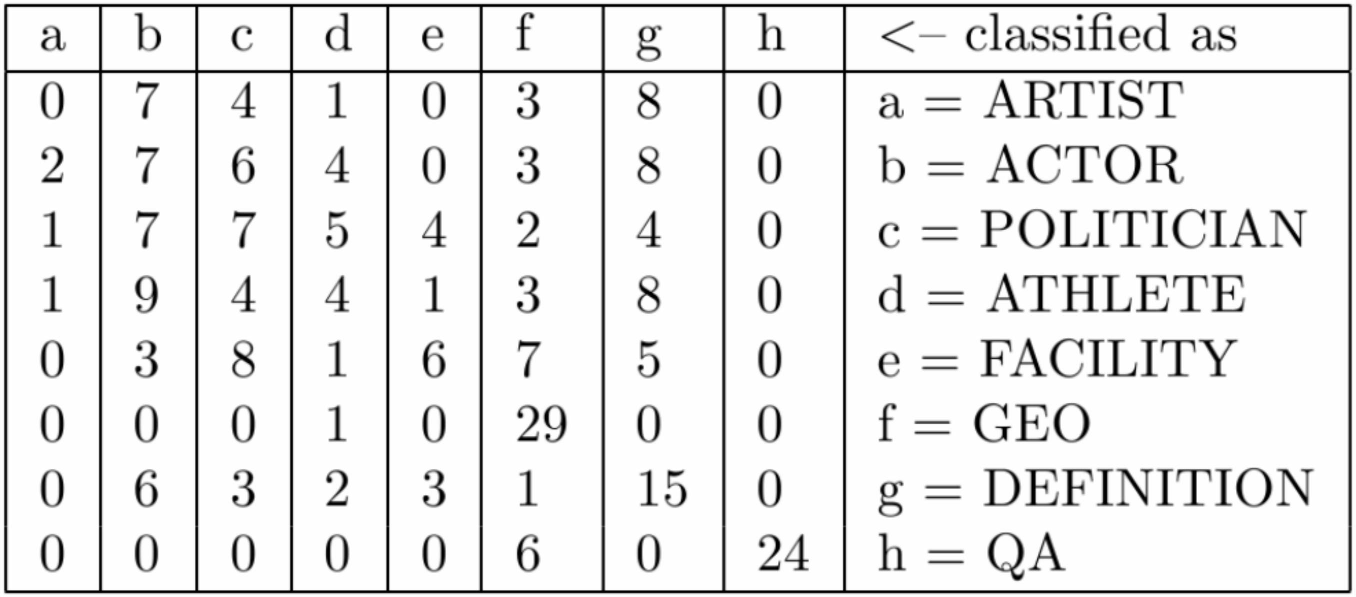
placeTypeName:

- Not a place
- Point of Interest
- Town
- County
- ...

Yahoo's GeoPlanet API →

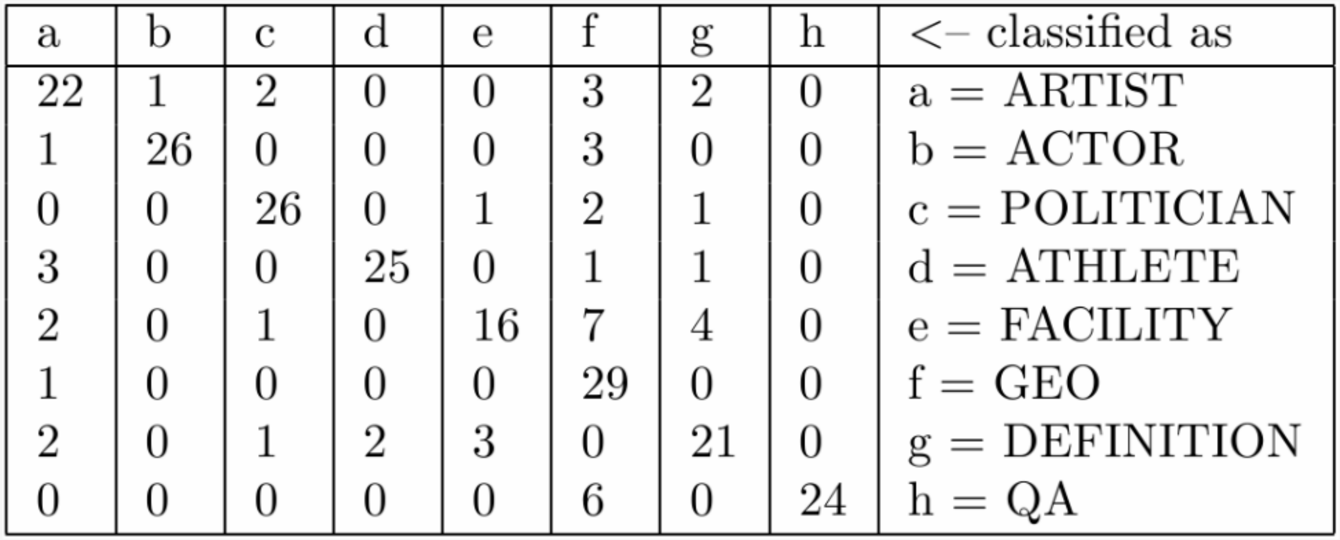
placeTypeName:

- Not a place
- Point of Interest
- Town
- County
- ...



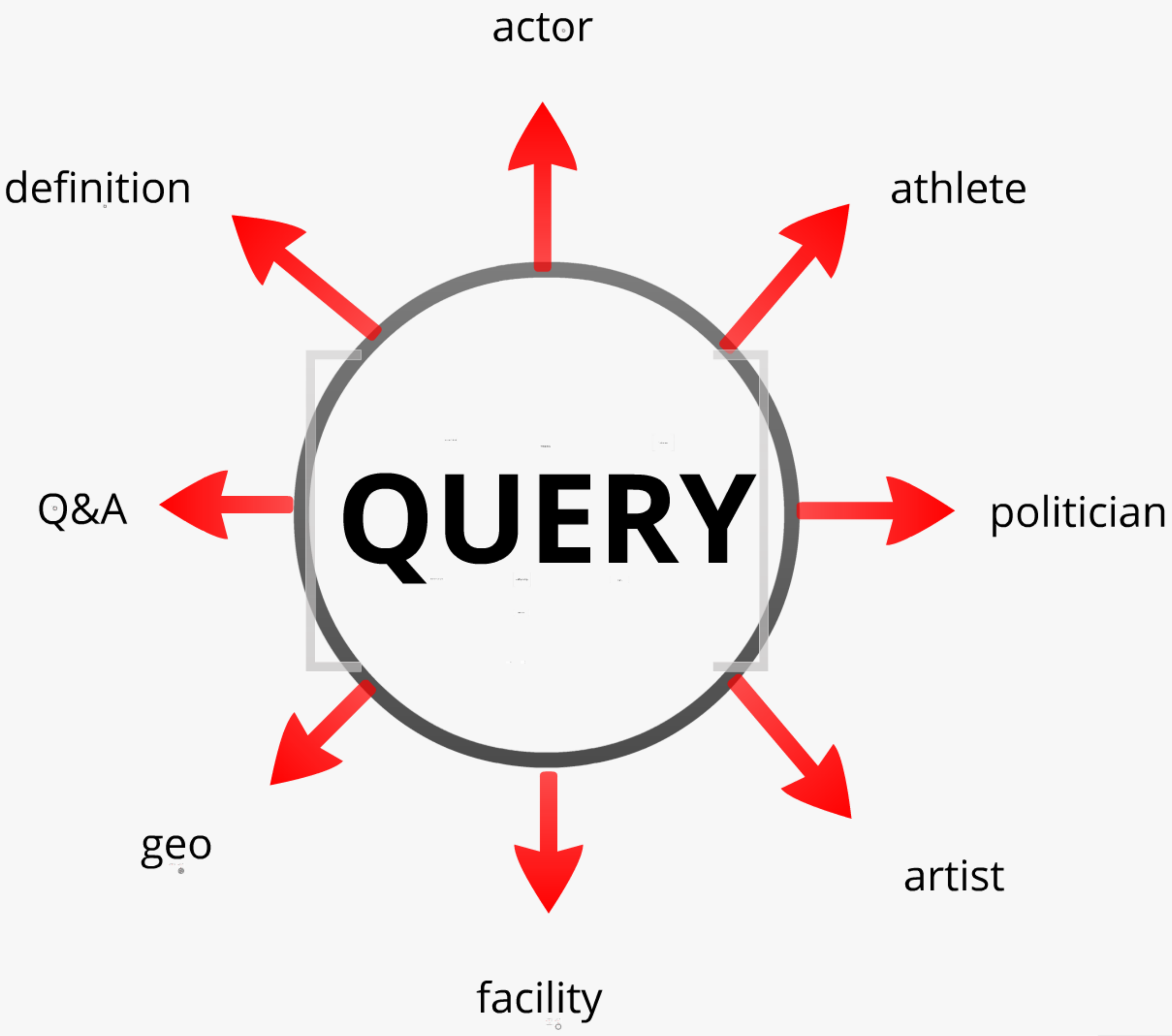
a	b	c	d	e	f	g	h	<– classified as
0	7	4	1	0	3	8	0	a = ARTIST
2	7	6	4	0	3	8	0	b = ACTOR
1	7	7	5	4	2	4	0	c = POLITICIAN
1	9	4	4	1	3	8	0	d = ATHLETE
0	3	8	1	6	7	5	0	e = FACILITY
0	0	0	1	0	29	0	0	f = GEO
0	6	3	2	3	1	15	0	g = DEFINITION
0	0	0	0	0	6	0	24	h = QA

38.3 % correctly classified instances



a	b	c	d	e	f	g	h	<– classified as
22	1	2	0	0	3	2	0	a = ARTIST
1	26	0	0	0	3	0	0	b = ACTOR
0	0	26	0	1	2	1	0	c = POLITICIAN
3	0	0	25	0	1	1	0	d = ATHLETE
2	0	1	0	16	7	4	0	e = FACILITY
1	0	0	0	0	29	0	0	f = GEO
2	0	1	2	3	0	21	0	g = DEFINITION
0	0	0	0	0	6	0	24	h = QA

78.8 % correctly classified instances



QUERY

actor

athlete

politician

artist

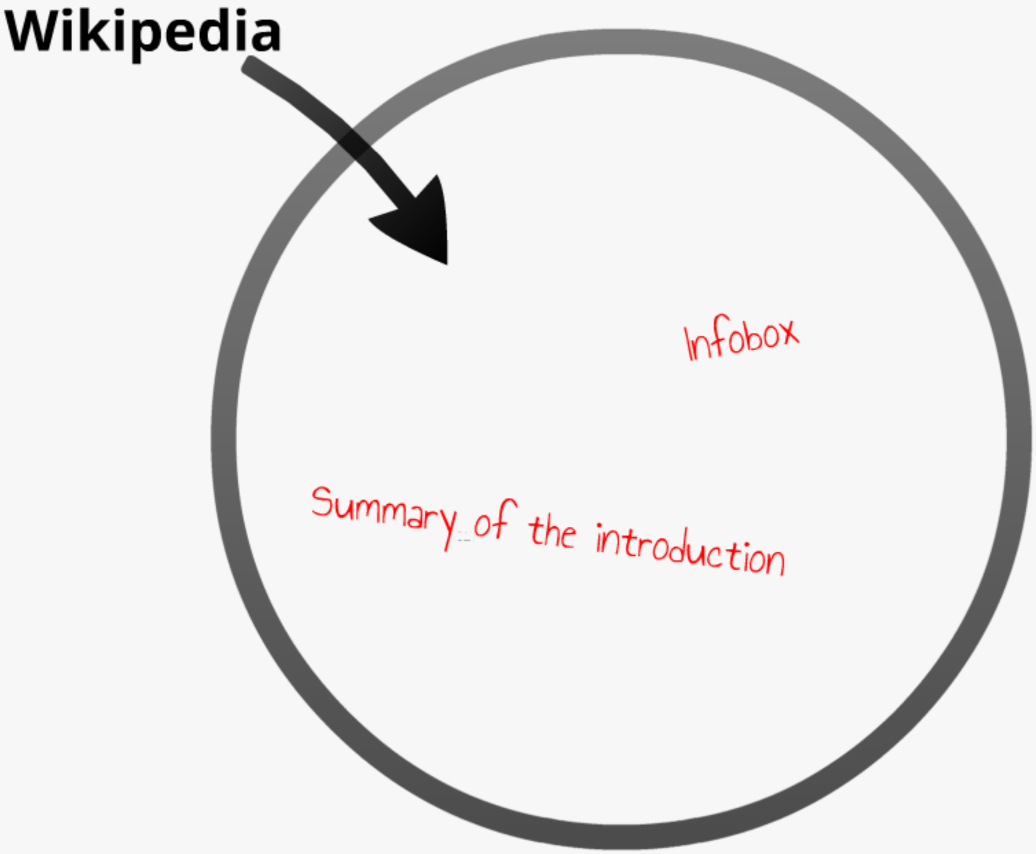
facility

geo

Q&A

definition

actor



Wikipedia

Infobox

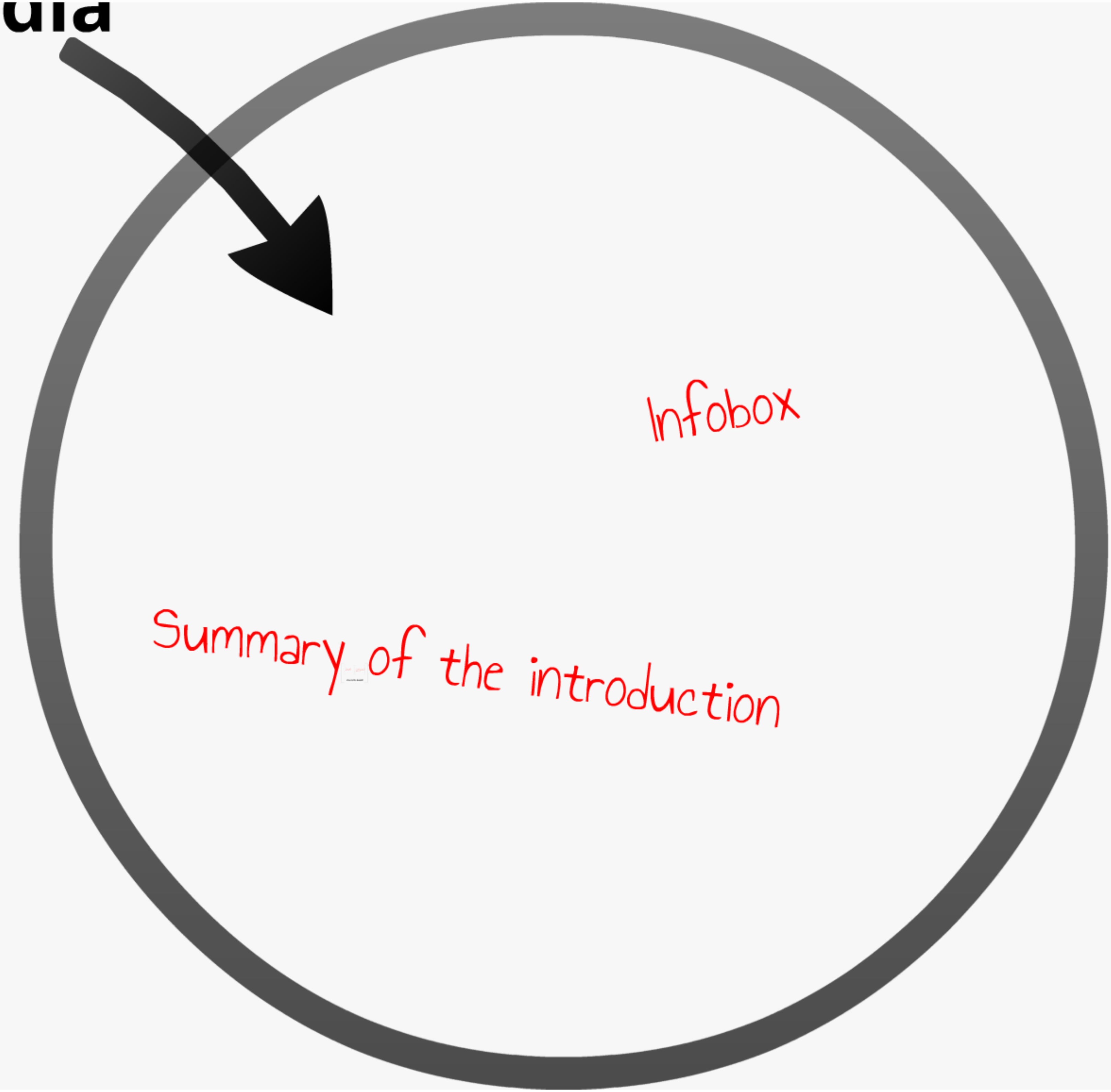
Summary of the introduction

Scrape introduction from Wikipedia artice

Use Freebase.com (huge social database) to find a summary of the Wikipedia page

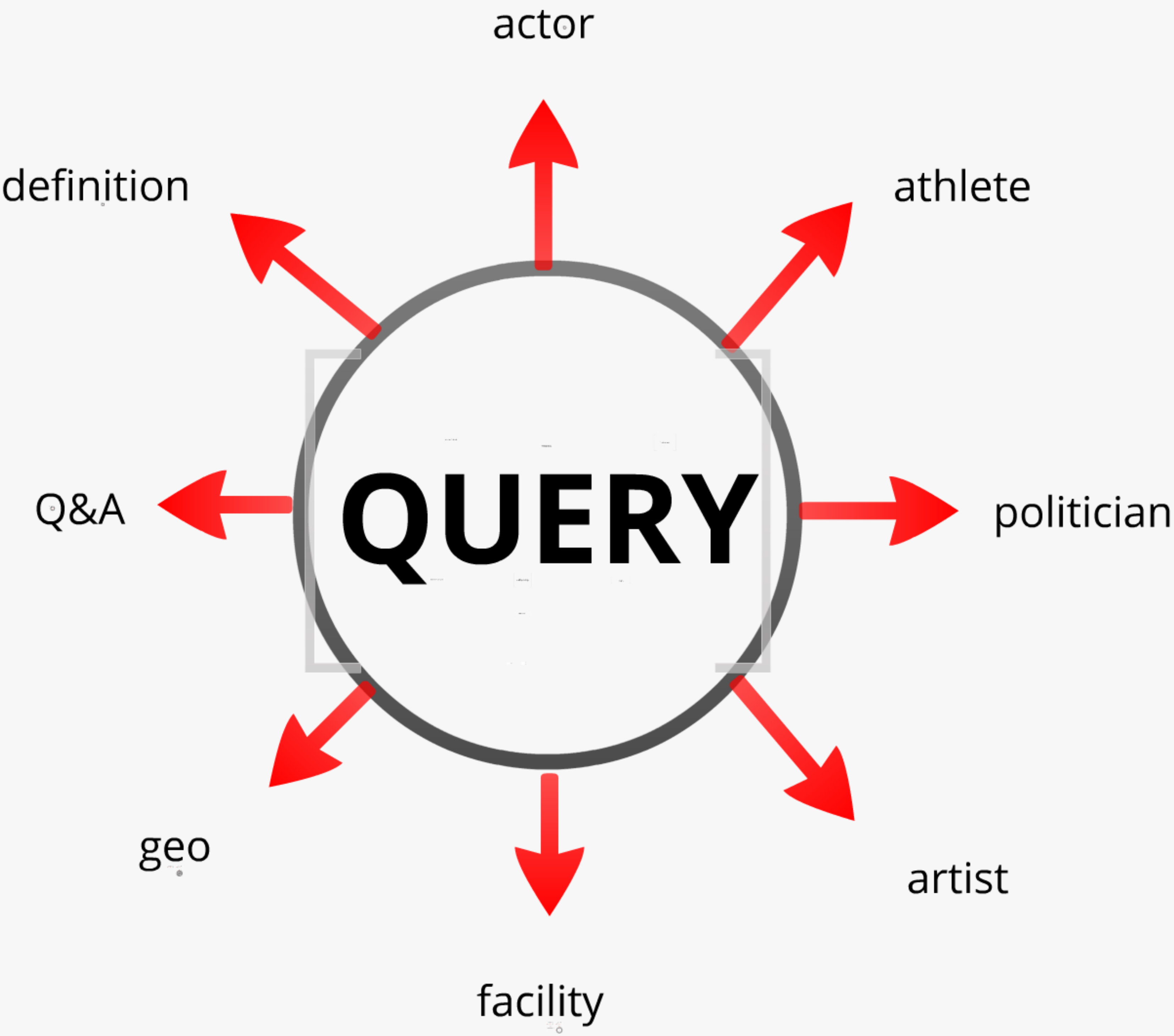
Choose the longest!

kipedia



Infobox

Summary of the introduction



actor

definition

athlete

Q&A

QUERY

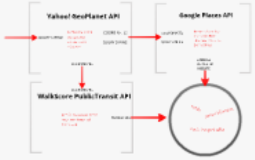
politician

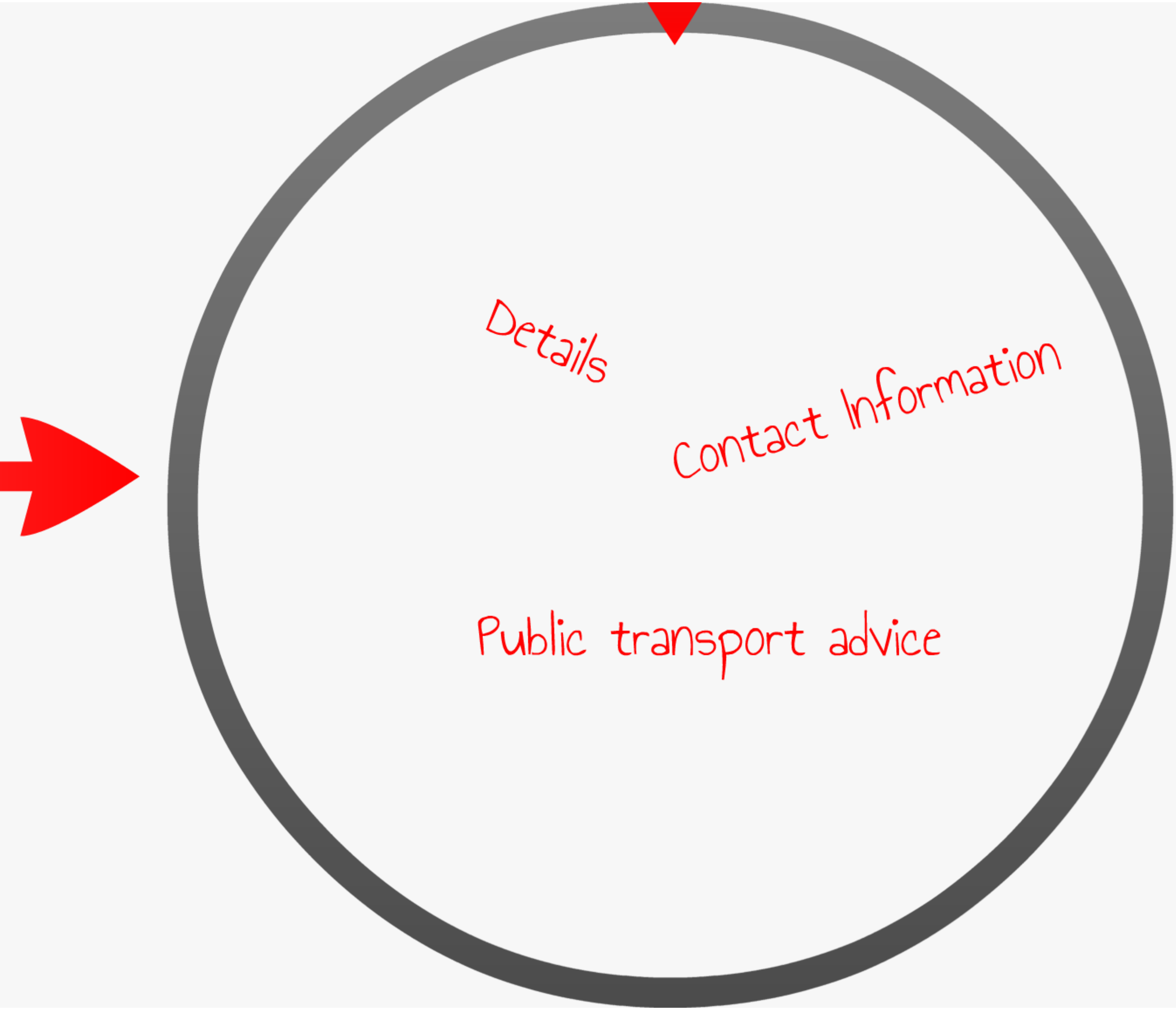
geo

artist

facility

facility


Yahoo! GeoPlanet API
Google Places API
WalkScore PublicTransit API



Details

Contact Information

Public transport advice

Yahoo! GeoPlanet API

QUERY STRING

Identifies the part of the query that contains location information

COORDINATES

QUERY STRING

COORDINATES

Google Places API

COORDINATES

QUERY STRING

Identifies a PLACE close
to the COORDINATES
whose name matches the
QUERY STRING

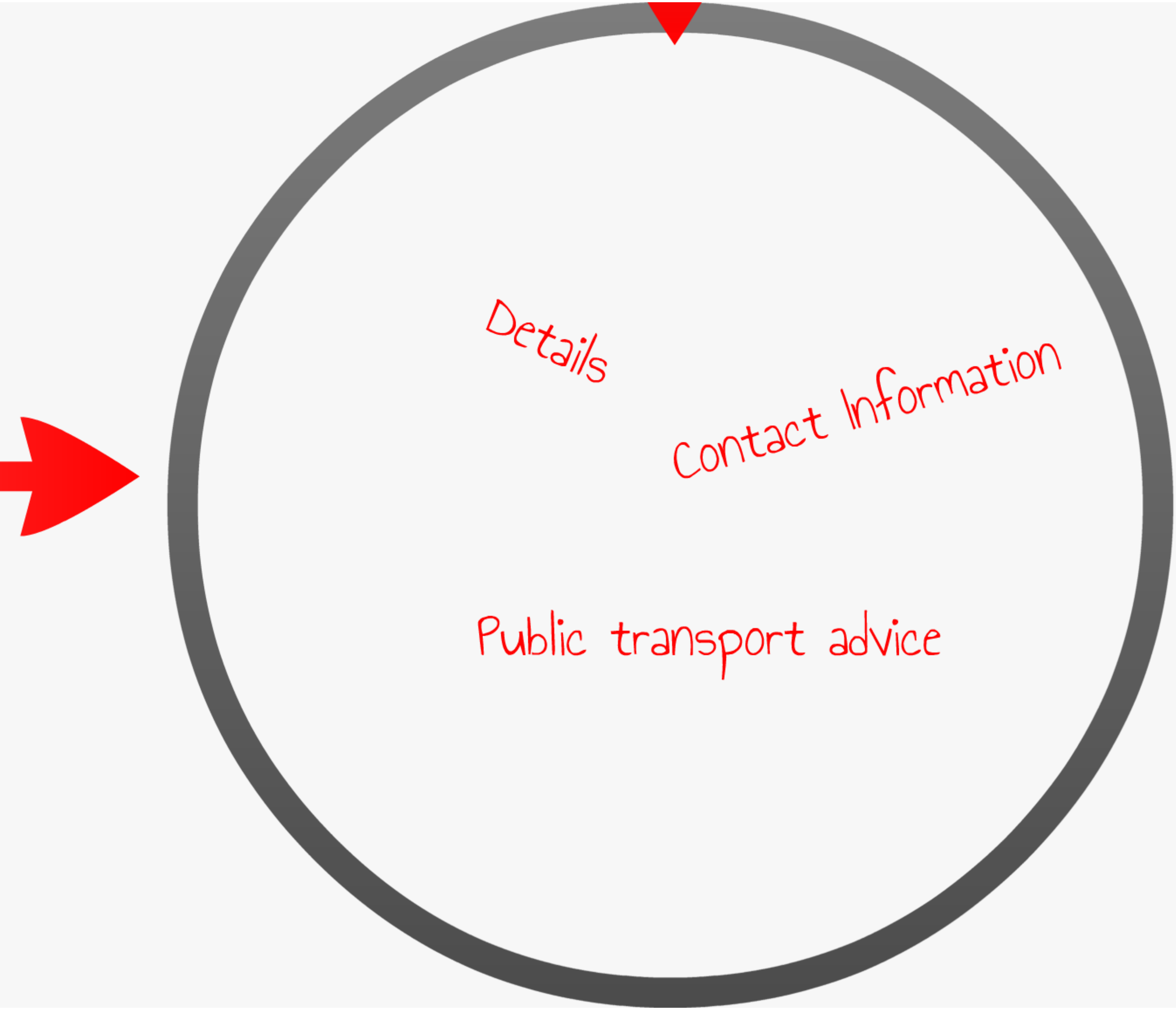
ADDRESS
PHONE NO.
WEBSITE

COORDINATES

WalkScore PublicTransit API

Identifies nearby public transit stops, their distance and lines serviced.

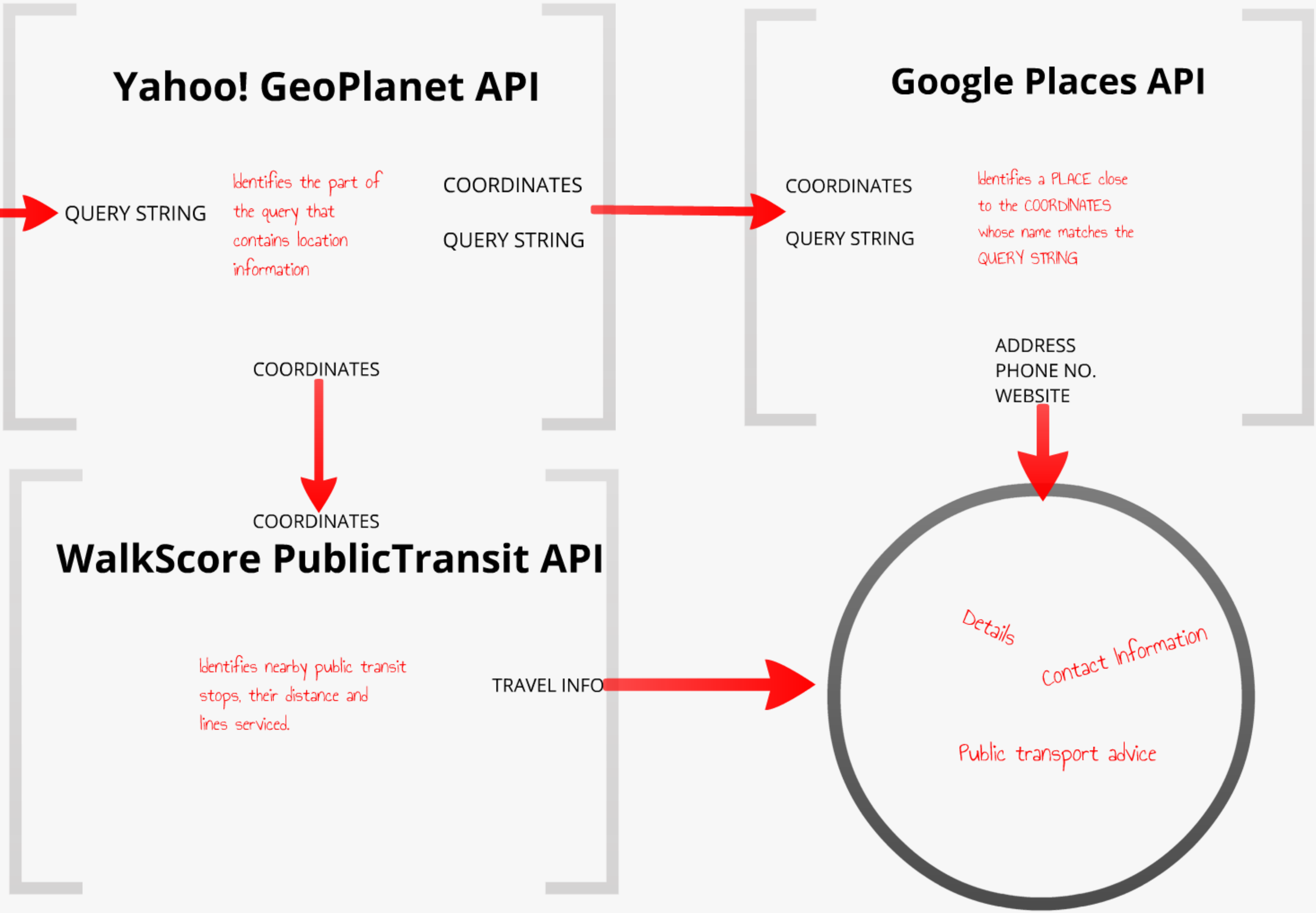
TRAVEL INFO



Details

Contact Information

Public transport advice



Yahoo! GeoPlanet API

QUERY STRING

Identifies the part of the query that contains location information

COORDINATES

QUERY STRING

COORDINATES

Google Places API

COORDINATES

QUERY STRING

Identifies a PLACE close to the COORDINATES whose name matches the QUERY STRING

ADDRESS
PHONE NO.
WEBSITE

COORDINATES

WalkScore PublicTransit API

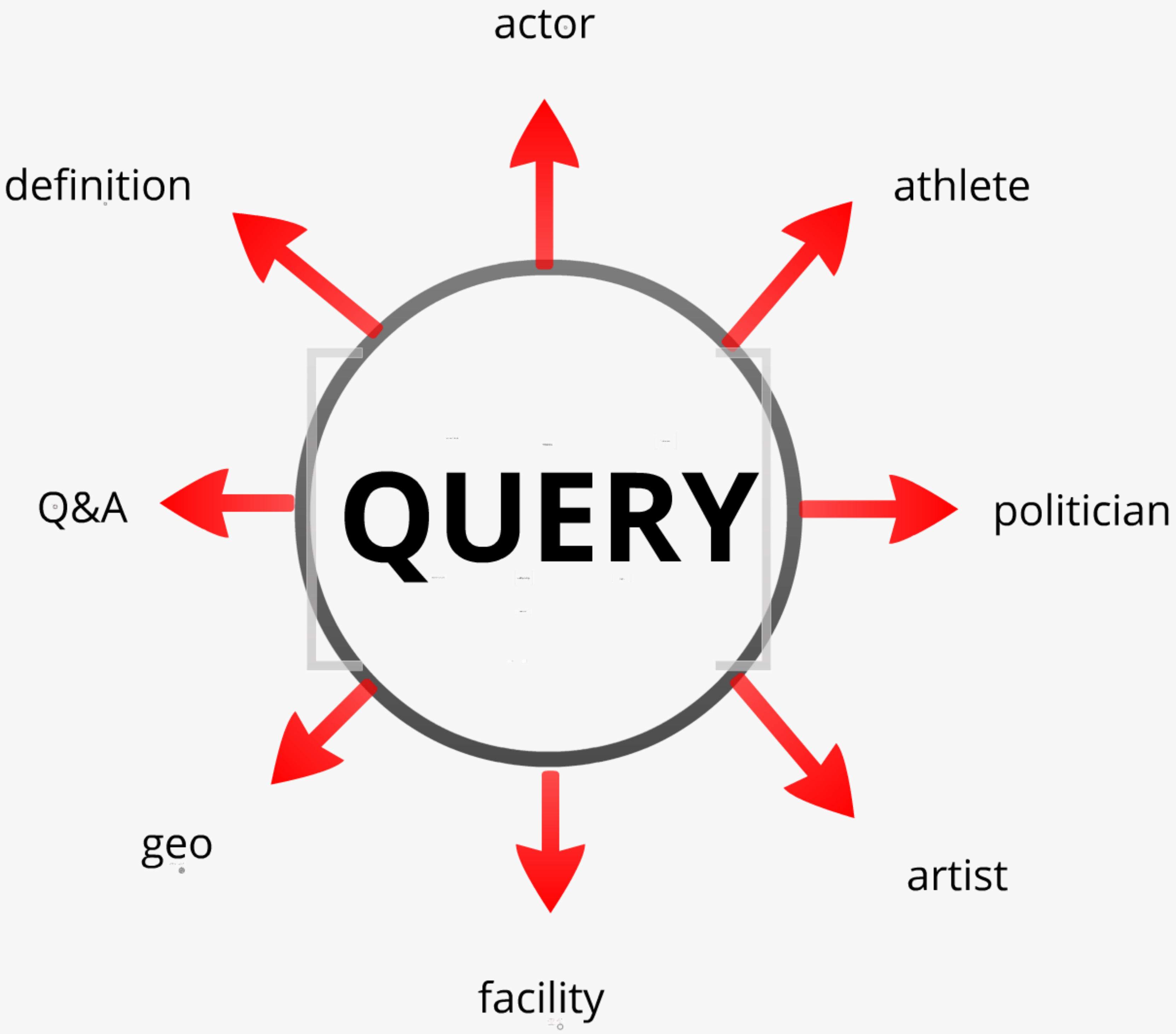
Identifies nearby public transit stops, their distance and lines serviced.

TRAVEL INFO

Details

Contact Information

Public transport advice



QUERY

actor

athlete

politician

artist

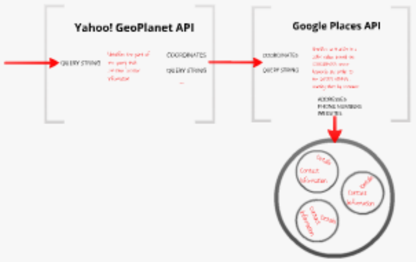
facility

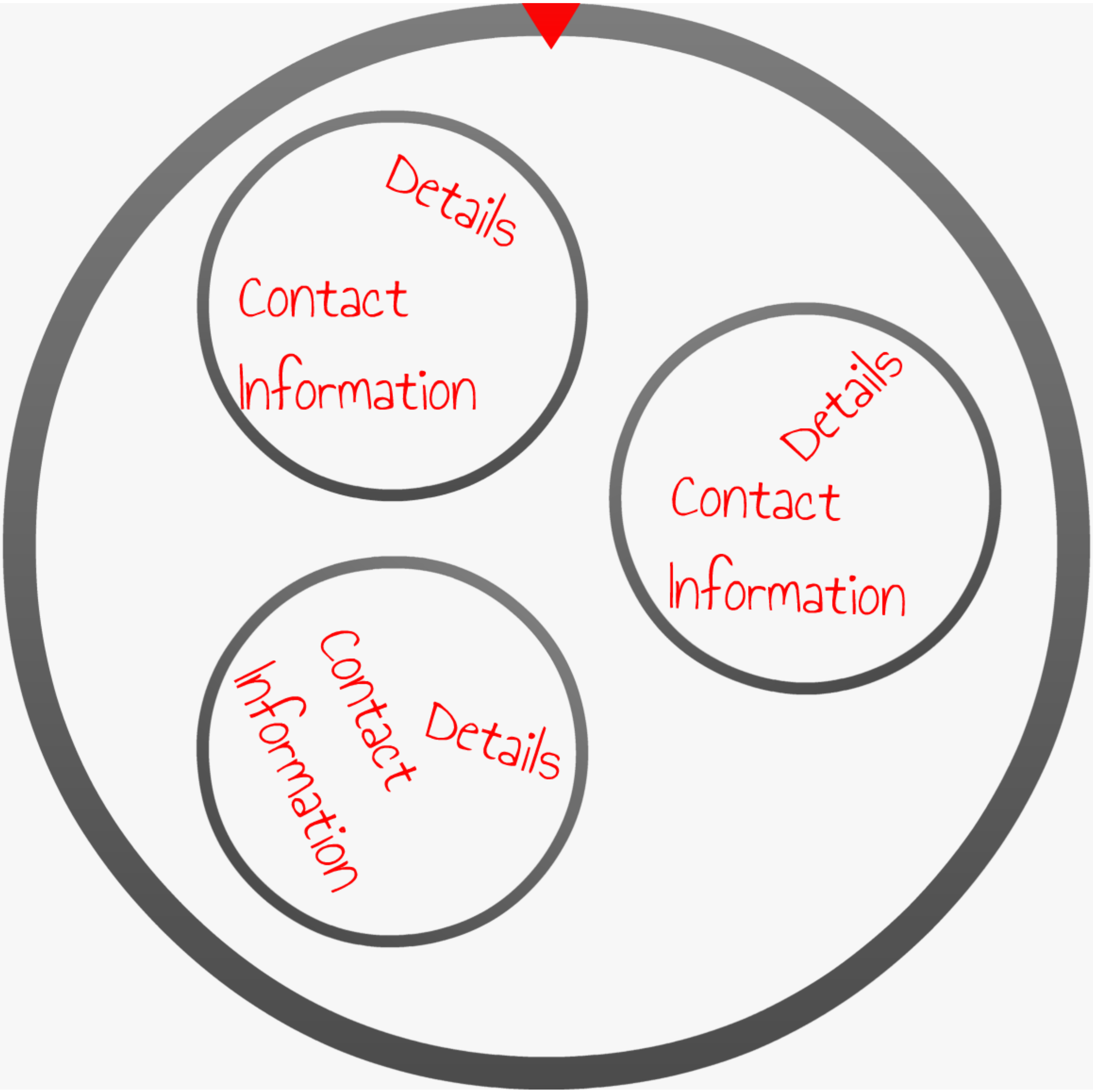
geo

Q&A

definition

geo


Yahoo! GeoPlanet API
QUERY STRING
COORDINATES
QUERY STRING
...
Google Places API
COORDINATES
QUERY STRING
ADDRESSES
PHONE NUMBERS
WEBSITES
Details
Contact information
Details
Contact information
Details
Contact information



Details

Contact Information

Details

Contact Information

Contact Information

Details

Yahoo! GeoPlanet API

QUERY STRING

Identifies the part of the query that contains location information

COORDINATES

QUERY STRING

Infobox

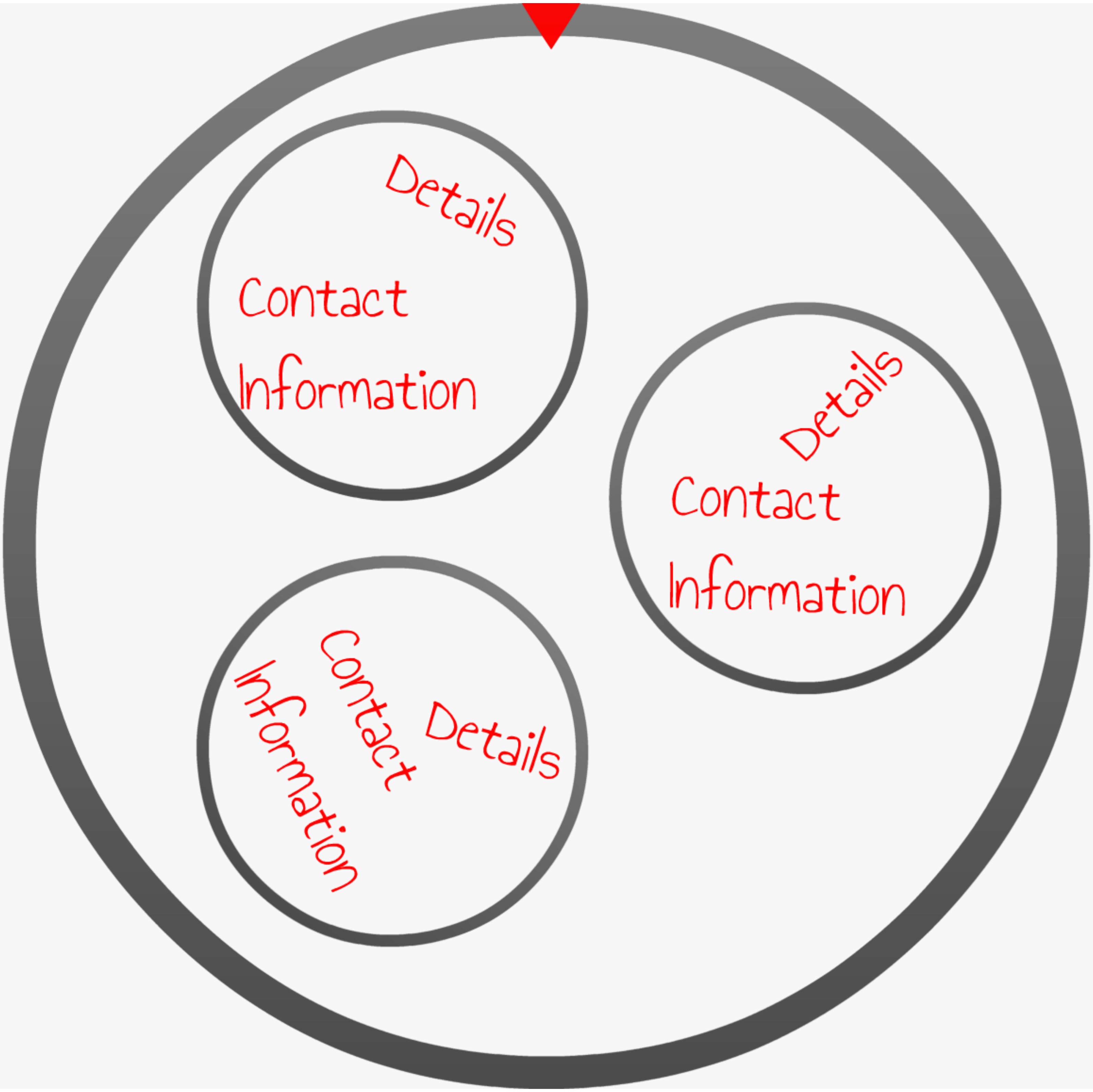
Google Places API

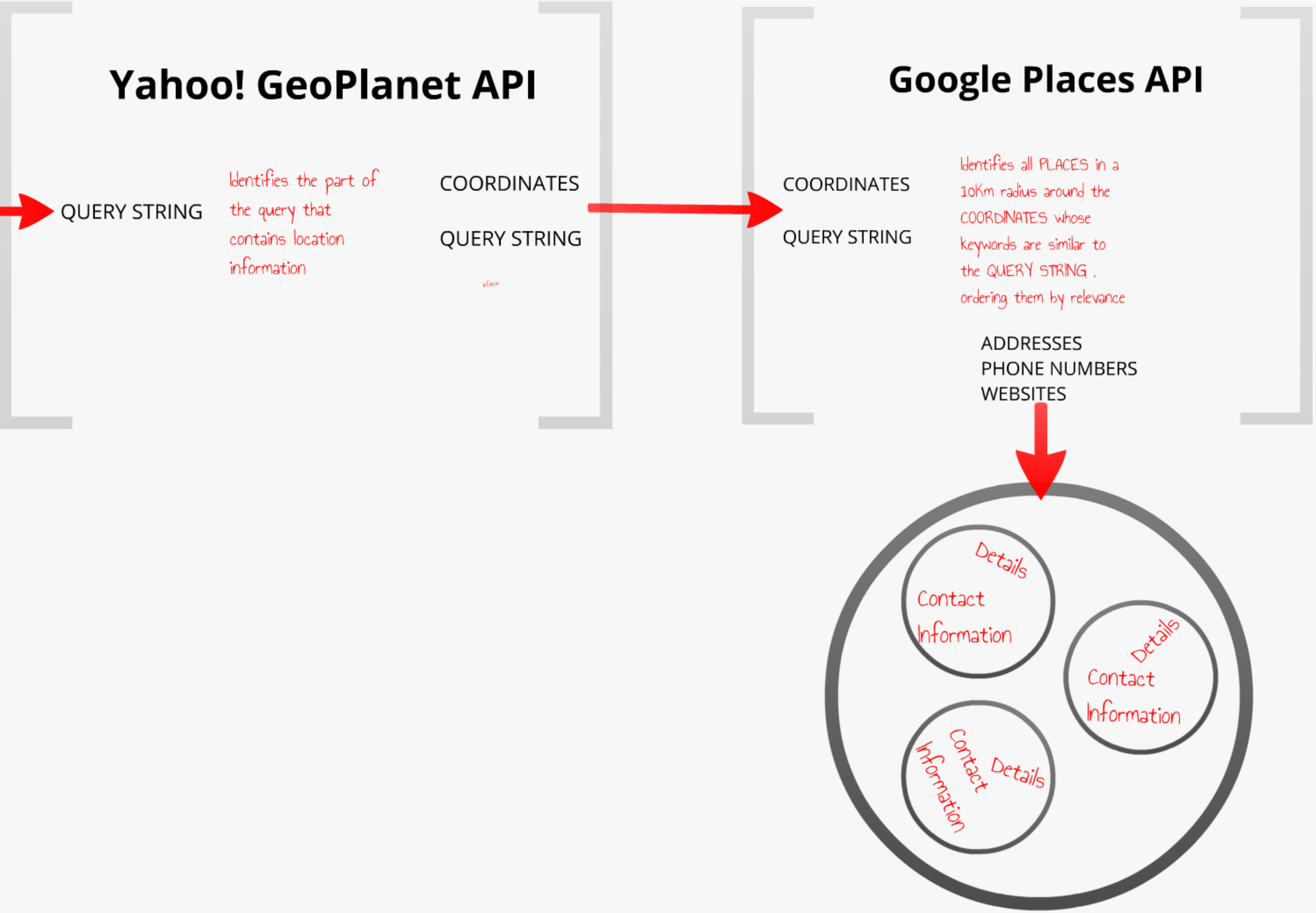
COORDINATES

QUERY STRING

Identifies all PLACES in a 10Km radius around the COORDINATES whose keywords are similar to the QUERY STRING , ordering them by relevance

ADDRESSES
PHONE NUMBERS
WEBSITES


Details
Contact
Information
Details
Contact
Information
Contact
Information
Details



Yahoo! GeoPlanet API

QUERY STRING

Identifies the part of the query that contains location information

COORDINATES

QUERY STRING

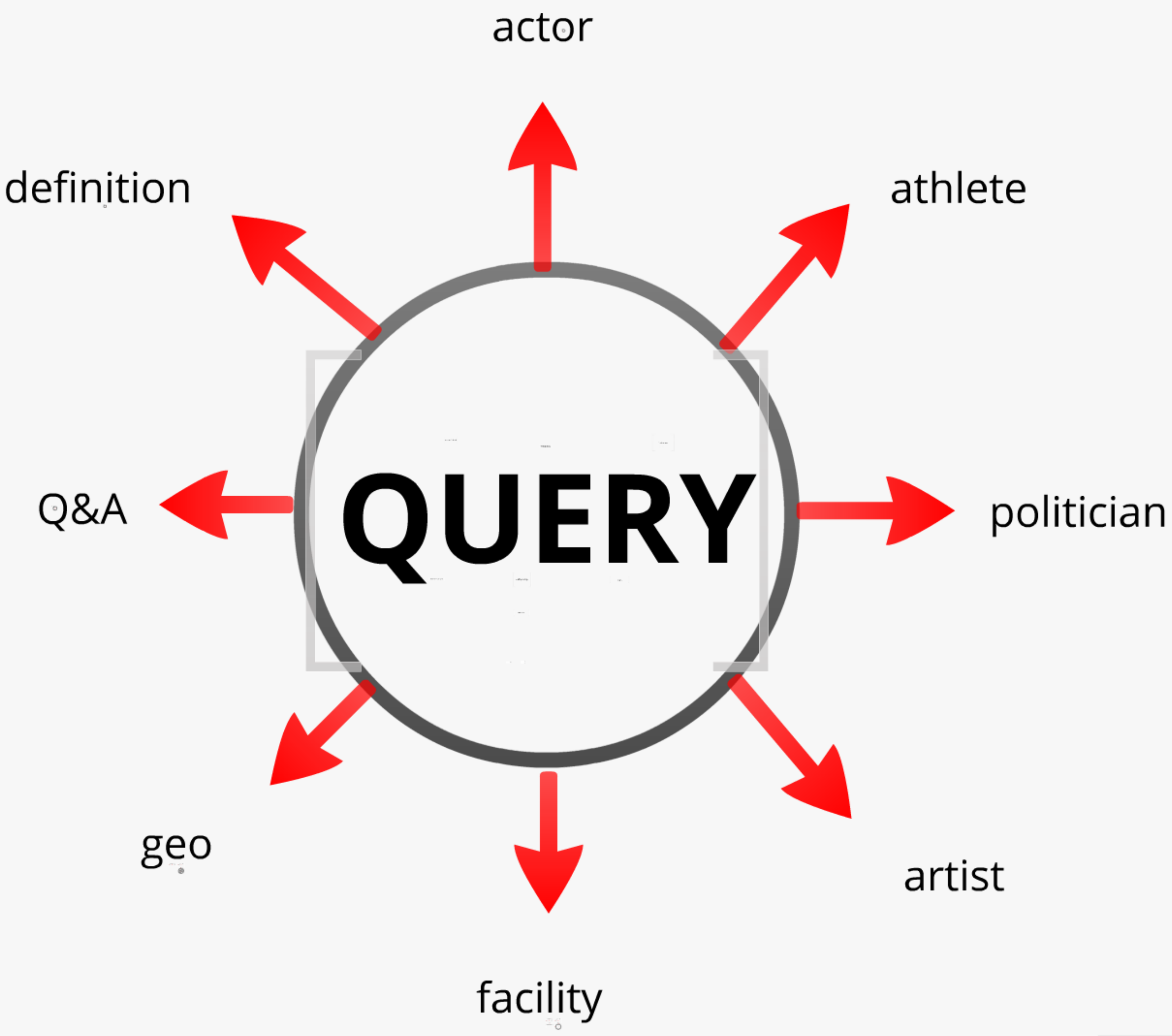
Google Places API

COORDINATES

QUERY STRING

Identifies all PLACES in a 10Km radius around the COORDINATES whose keywords are similar to the QUERY STRING , ordering them by relevance

ADDRESSES
PHONE NUMBERS
WEBSITES



QUERY

actor

athlete

politician

artist

facility

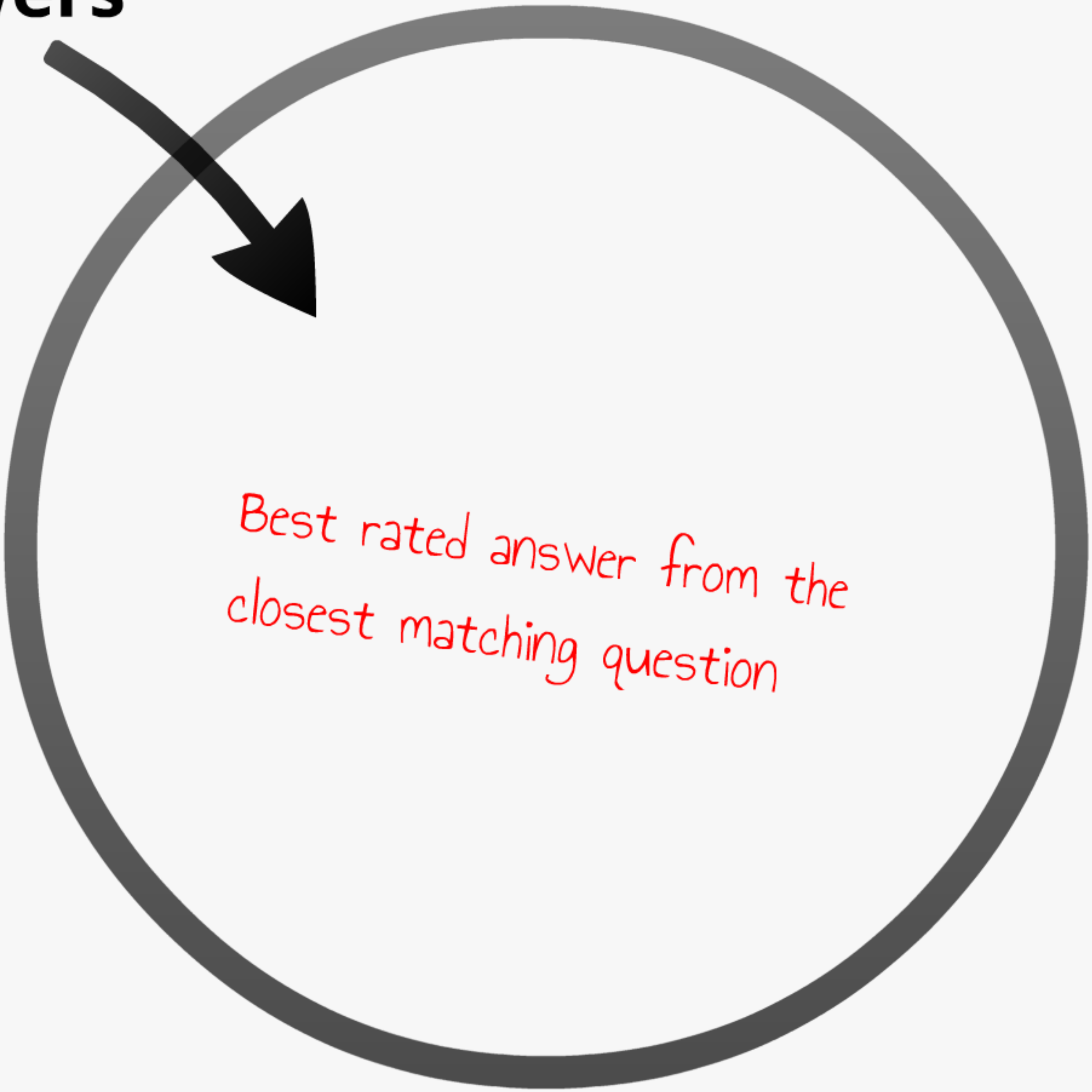
geo

Q&A

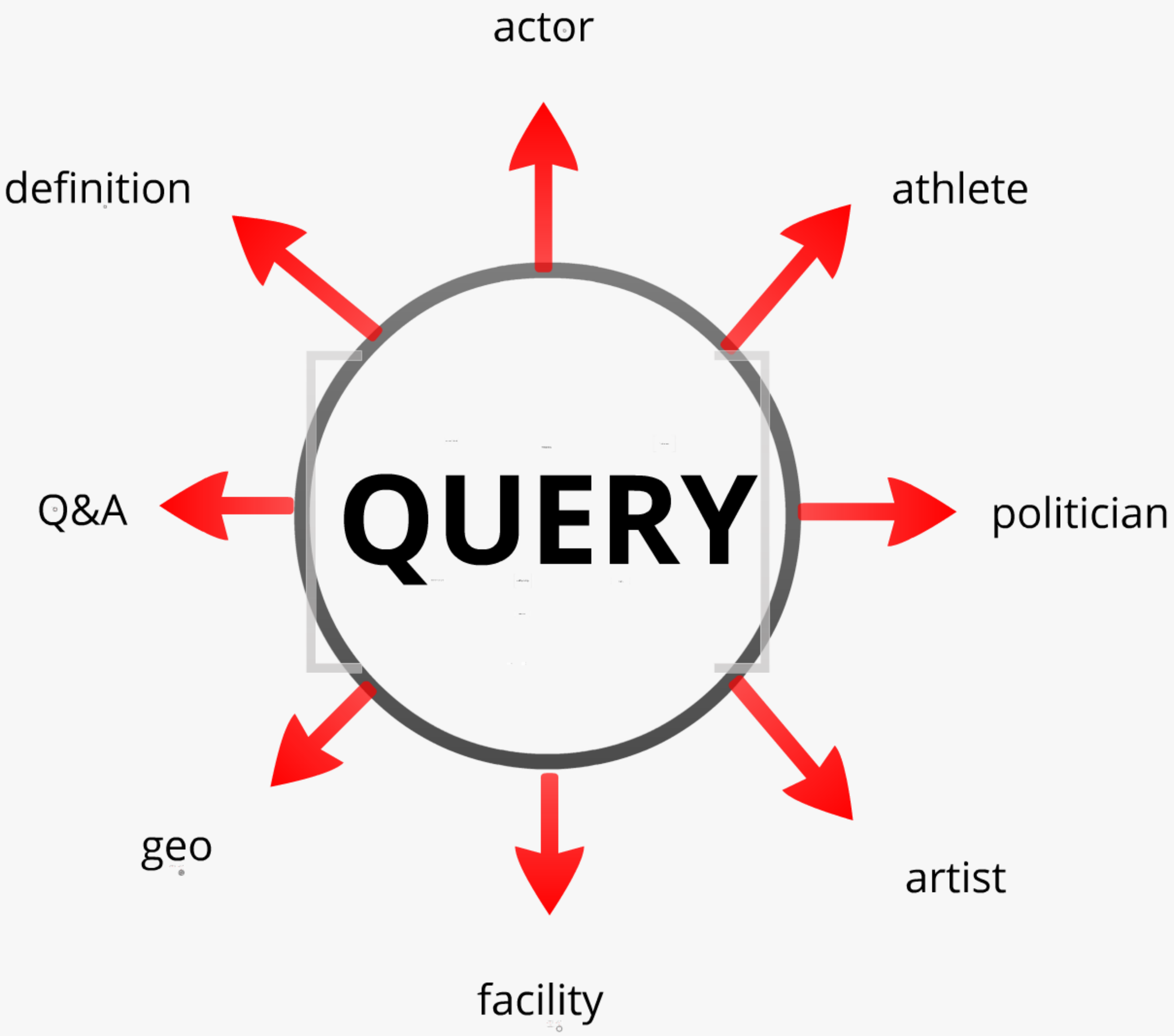
definition

Q&A

Yahoo! Answers



Best rated answer from the closest matching question



QUERY

actor

athlete

politician

artist

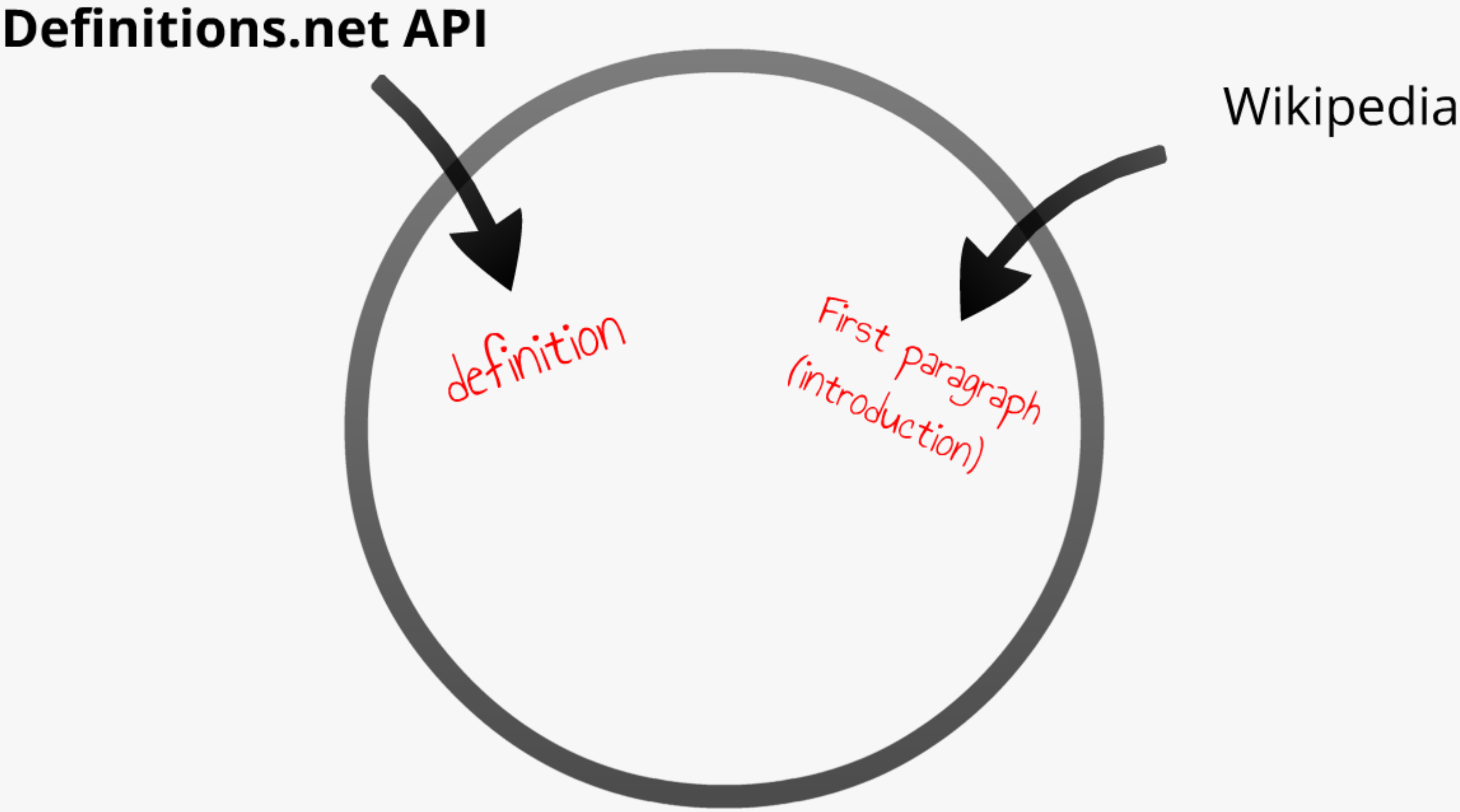
facility

geo

Q&A

definition

definition

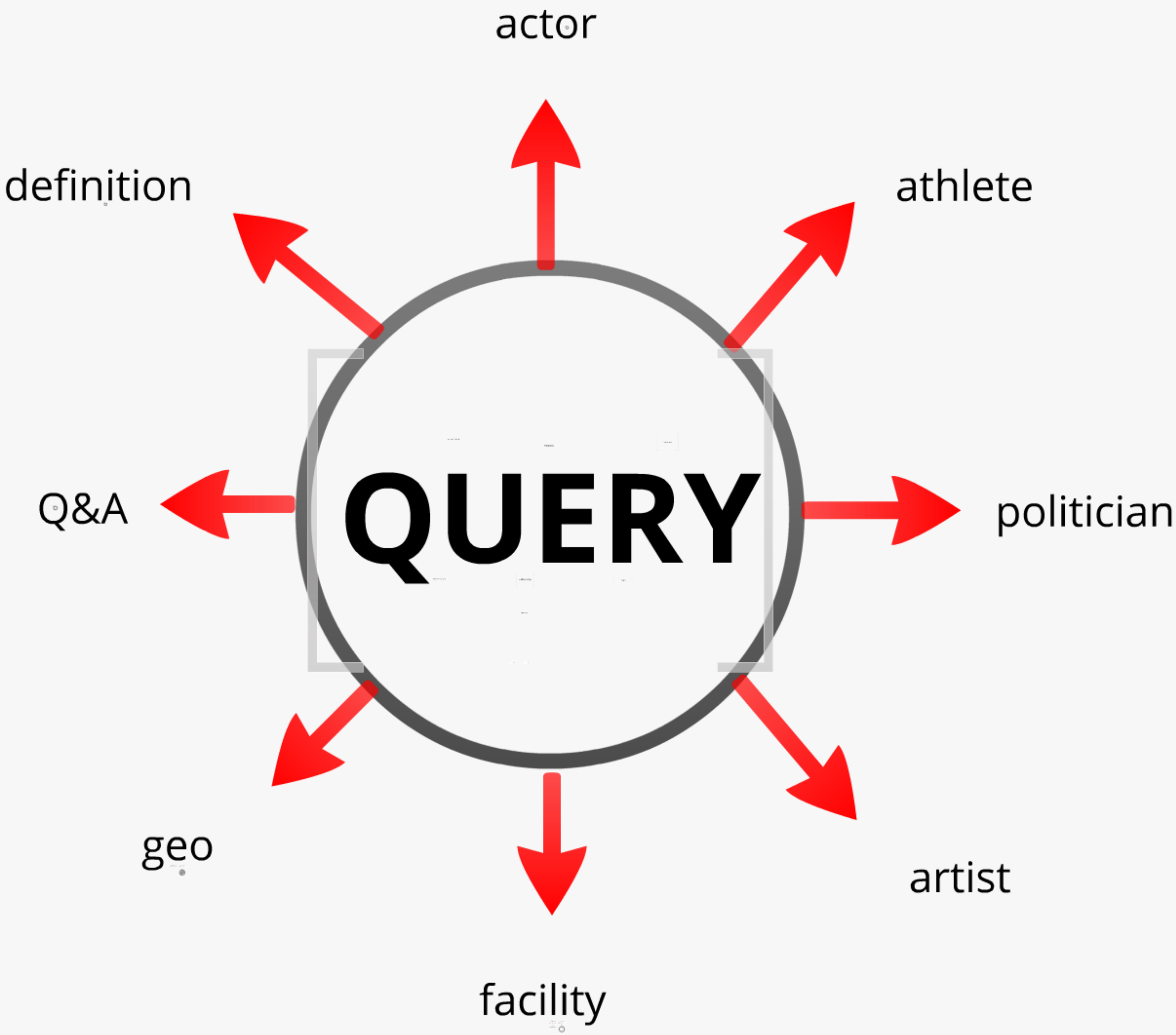


Definitions.net API

Wikipedia

definition

First paragraph (introduction)



QUERY

actor

athlete

politician

artist

facility

geo

Q&A

definition

Evaluation

1. Design a baseline

2. Ask user to enter query

3. Let user pick the most relevant result

Baseline system using Google

Google used as "one-click" ?

Implementation:
Concatenate Google result snippets from the first result page

Evaluation

1. Design a baseline

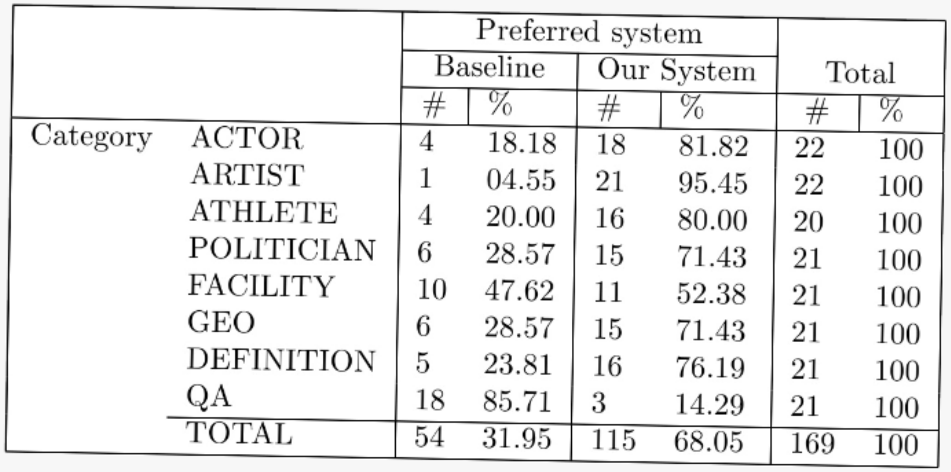
2. Ask user to enter query

3. Let user pick the most relevant result

Results

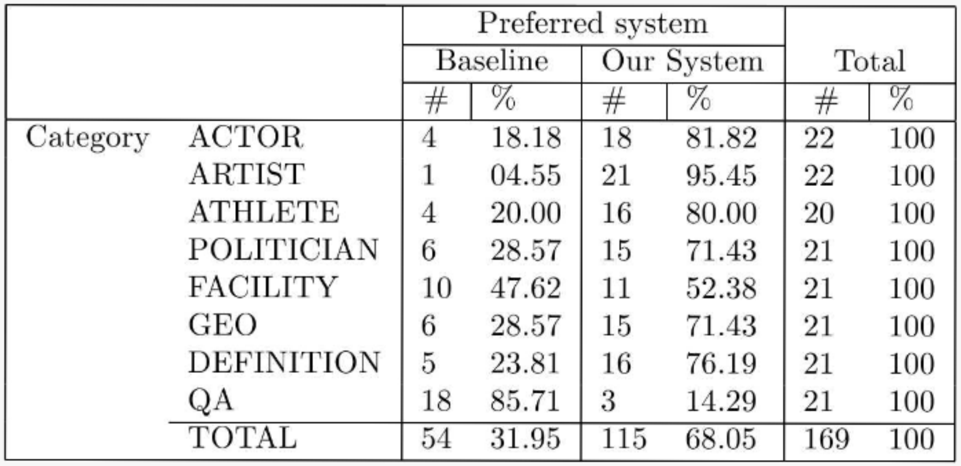
1CLICK system was better for 68% of queries!

(from a total of 169)



		Preferred system				Total	
		Baseline		Our System			
		#	%	#	%	#	%
Category	ACTOR	4	18.18	18	81.82	22	100
	ARTIST	1	04.55	21	95.45	22	100
	ATHLETE	4	20.00	16	80.00	20	100
	POLITICIAN	6	28.57	15	71.43	21	100
	FACILITY	10	47.62	11	52.38	21	100
	GEO	6	28.57	15	71.43	21	100
	DEFINITION	5	23.81	16	76.19	21	100
	QA	18	85.71	3	14.29	21	100
	TOTAL	54	31.95	115	68.05	169	100

Discussion



		Preferred system					
		Baseline		Our System		Total	
		#	%	#	%	#	%
Category	ACTOR	4	18.18	18	81.82	22	100
	ARTIST	1	04.55	21	95.45	22	100
	ATHLETE	4	20.00	16	80.00	20	100
	POLITICIAN	6	28.57	15	71.43	21	100
	FACILITY	10	47.62	11	52.38	21	100
	GEO	6	28.57	15	71.43	21	100
	DEFINITION	5	23.81	16	76.19	21	100
	QA	18	85.71	3	14.29	21	100
	TOTAL	54	31.95	115	68.05	169	100

1CLICK system performs well for most categories:

- Very well for PERSON-type queries

Especially for ARTIST

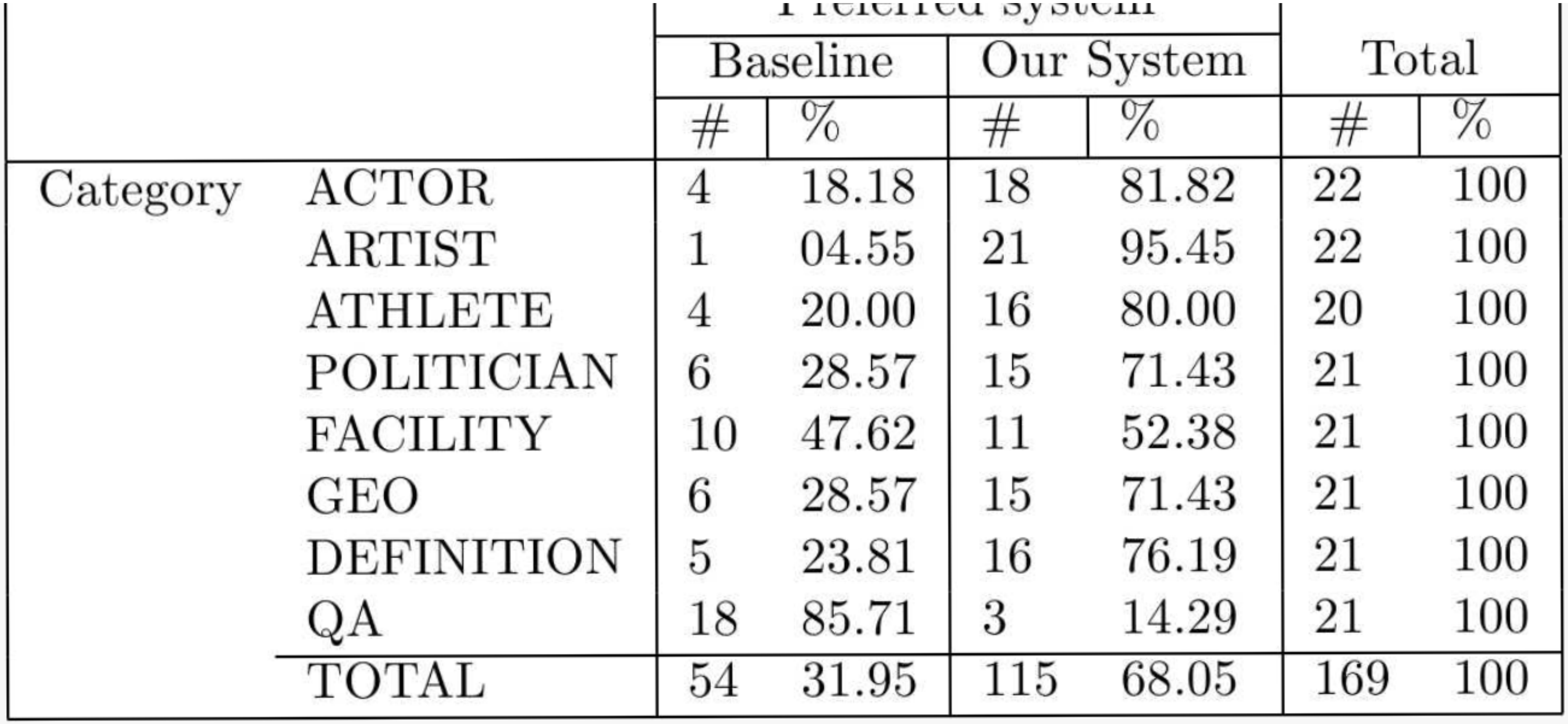
- Average for for FACILITY
- BAD for QA.

eries!

al of 169)

Thank Wikipedia!

Human selected, summarized and structured information about EVERYTHING



		Preferred system				Total	
		Baseline		Our System			
		#	%	#	%	#	%
Category	ACTOR	4	18.18	18	81.82	22	100
	ARTIST	1	04.55	21	95.45	22	100
	ATHLETE	4	20.00	16	80.00	20	100
	POLITICIAN	6	28.57	15	71.43	21	100
	FACILITY	10	47.62	11	52.38	21	100
	GEO	6	28.57	15	71.43	21	100
	DEFINITION	5	23.81	16	76.19	21	100
	QA	18	85.71	3	14.29	21	100
	TOTAL	54	31.95	115	68.05	169	100

1CLICK system performs well for most categories:

- Very well for PERSON-type queries

Especially for ARTIST

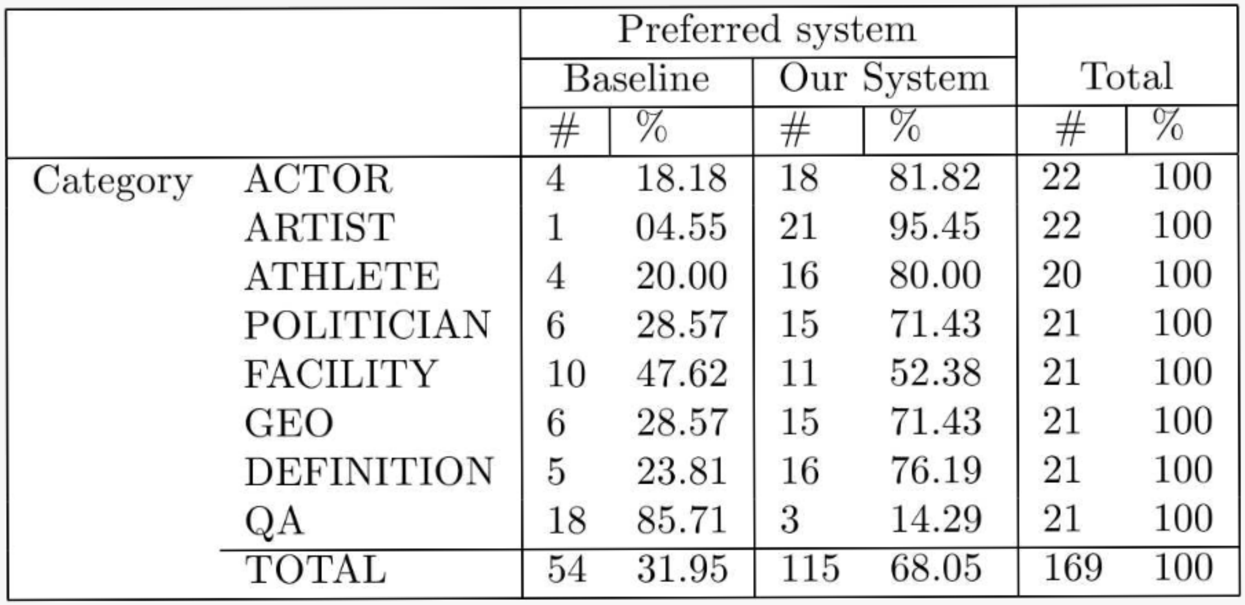
- Average for for FACILITY

- BAD for QA.

Public transit info only available in the US...

sometimes classified incorrectly → wrong information

Google geolocation services don't always play nice with Yahoo...



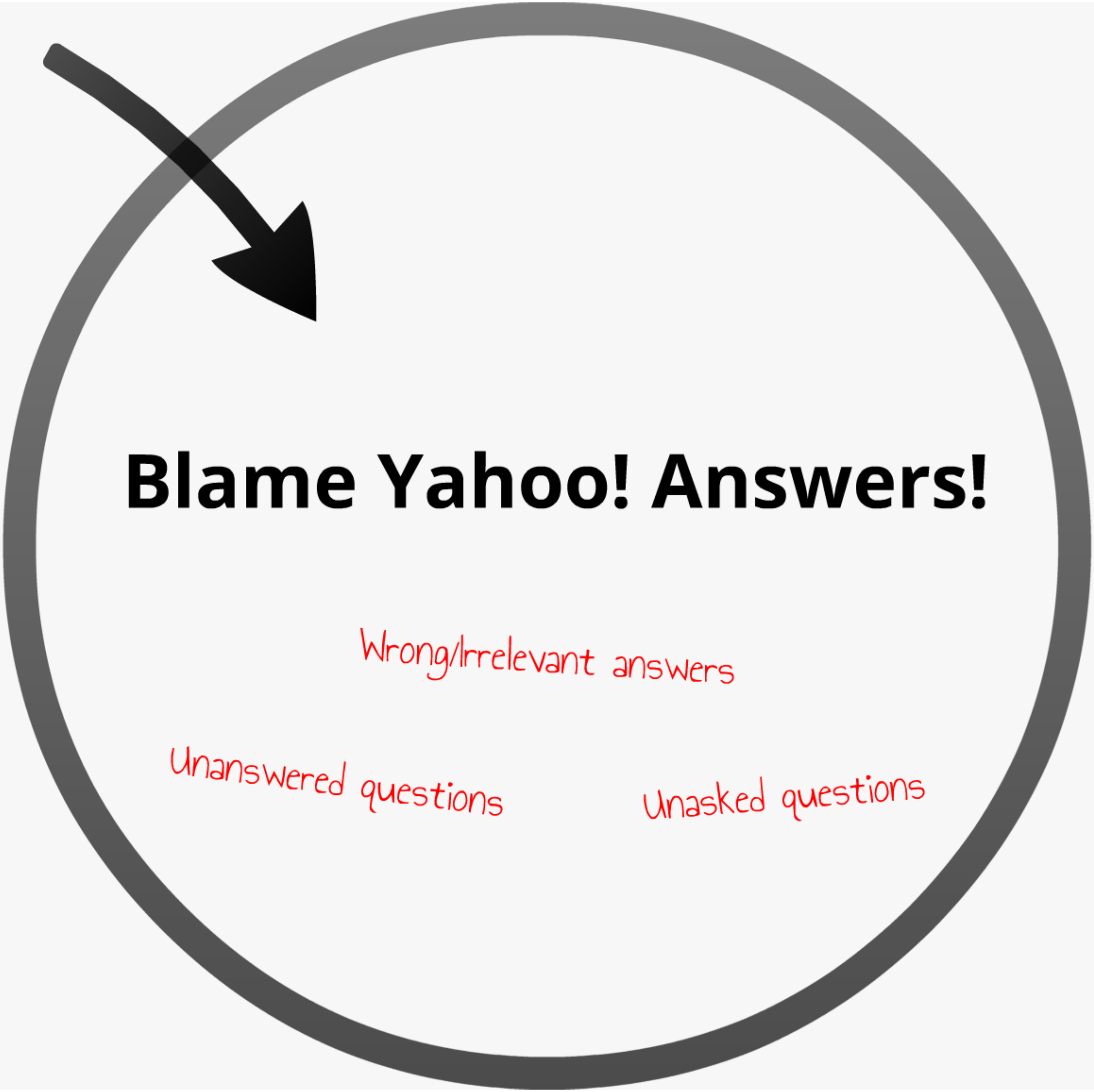
		Preferred system				Total	
		Baseline		Our System			
		#	%	#	%	#	%
Category	ACTOR	4	18.18	18	81.82	22	100
	ARTIST	1	04.55	21	95.45	22	100
	ATHLETE	4	20.00	16	80.00	20	100
	POLITICIAN	6	28.57	15	71.43	21	100
	FACILITY	10	47.62	11	52.38	21	100
	GEO	6	28.57	15	71.43	21	100
	DEFINITION	5	23.81	16	76.19	21	100
	QA	18	85.71	3	14.29	21	100
	TOTAL	54	31.95	115	68.05	169	100

1CLICK system performs well for most categories:

- Very well for PERSON-type queries

Especially for ARTIST

- Average for for FACILITY
- BAD for QA.



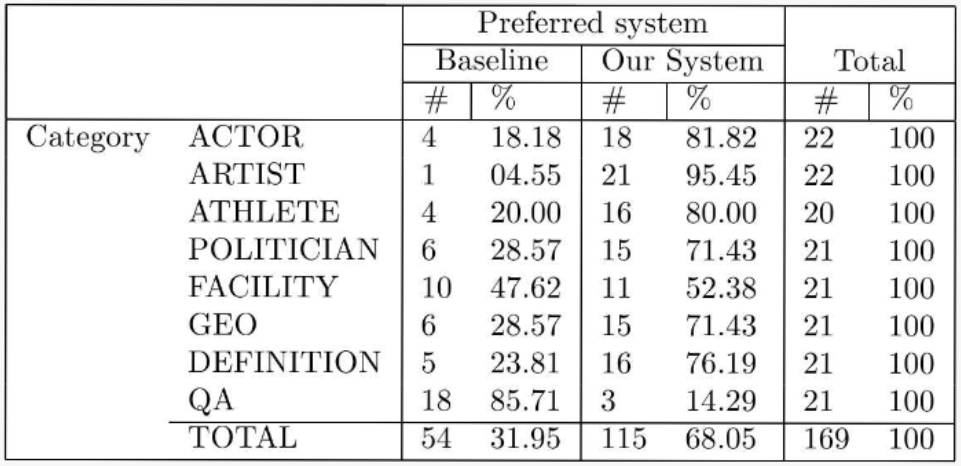
Blame Yahoo! Answers!

Wrong/Irrelevant answers

Unanswered questions

Unasked questions

Discussion



		Preferred system				Total	
		Baseline		Our System			
		#	%	#	%	#	%
Category	ACTOR	4	18.18	18	81.82	22	100
	ARTIST	1	04.55	21	95.45	22	100
	ATHLETE	4	20.00	16	80.00	20	100
	POLITICIAN	6	28.57	15	71.43	21	100
	FACILITY	10	47.62	11	52.38	21	100
	GEO	6	28.57	15	71.43	21	100
	DEFINITION	5	23.81	16	76.19	21	100
	QA	18	85.71	3	14.29	21	100
	TOTAL	54	31.95	115	68.05	169	100

1CLICK system performs well for most categories:

- Very well for PERSON-type queries

Especially for ARTIST

- Average for for FACILITY

- BAD for QA.

eries!

al of 169)

Use the evaluation to improve our system for the official 1CLICK@NTCIR challenge

Improving QA:

- Find a better question answering service: Evi.com
  - Searches multiple QA sources
  - Includes API
- Keep Yahoo Answers as a backup / alternative answer source

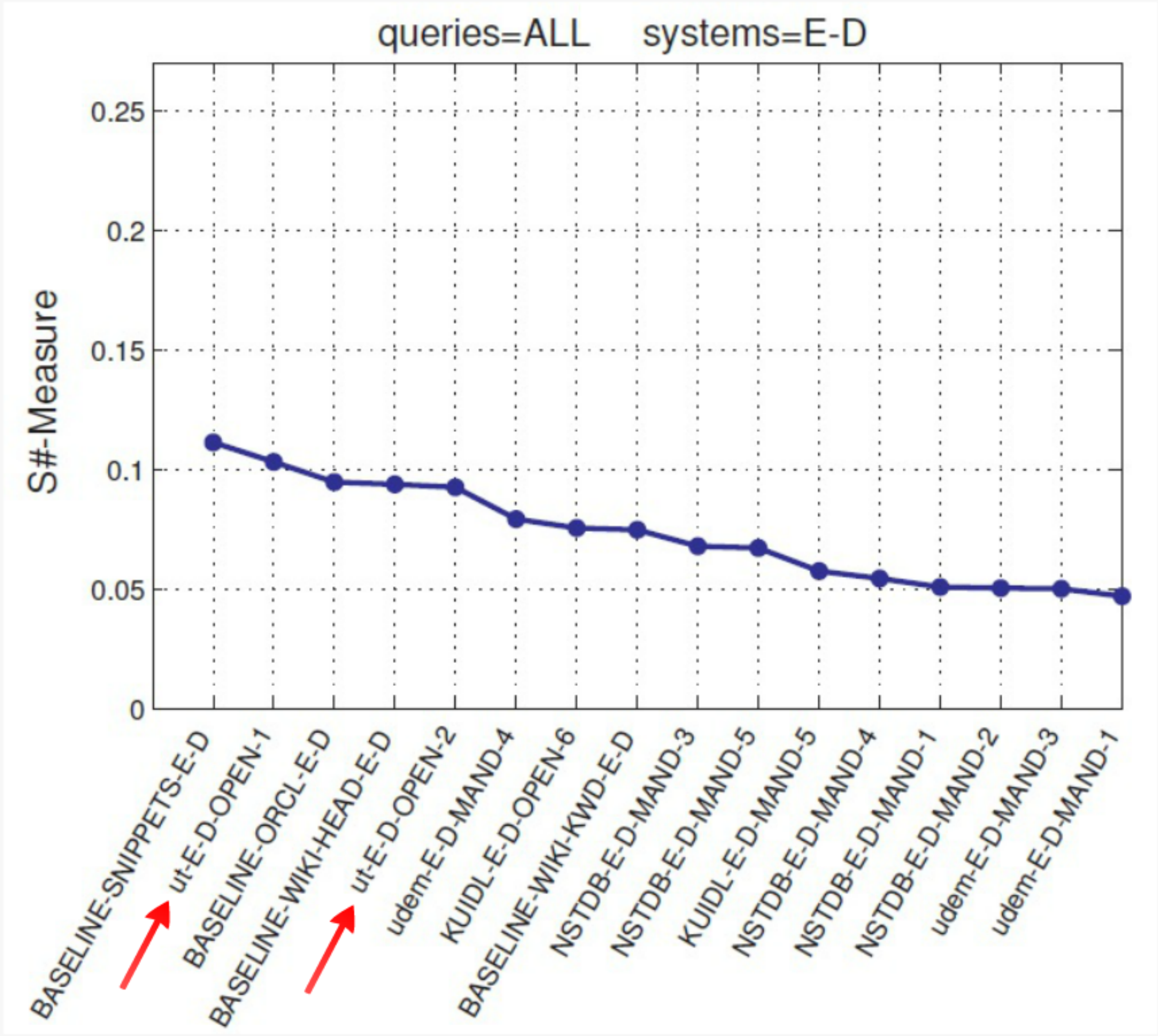
Improving classification:

- Use evaluation queries for training
- Try out different classification models and configurations
  - Increased accuracy to 89%!

- Evaluation query structure != new official query structure
  - Accuracy dropped drastically…

Query structure changed!

- Improving PERSON-like queries:
    - Attempt to split query into person + specifics
    - Use person name to find a Wiki page (same as before)
    - Use full-text search (Lucene) to identify sentences refering to the specifics



Official NTCIR results

S# measure takes into account length, amount position and order of relevant strings

- ut-E-D-OPEN1 only uses partial Wikipedia matches as features
- ut-E-D-OPEN2 uses both partial and full matches as features

Top performing runs in English Desktop 1CLICK-2

Thank you for your attention!

QUESTION-TIME!