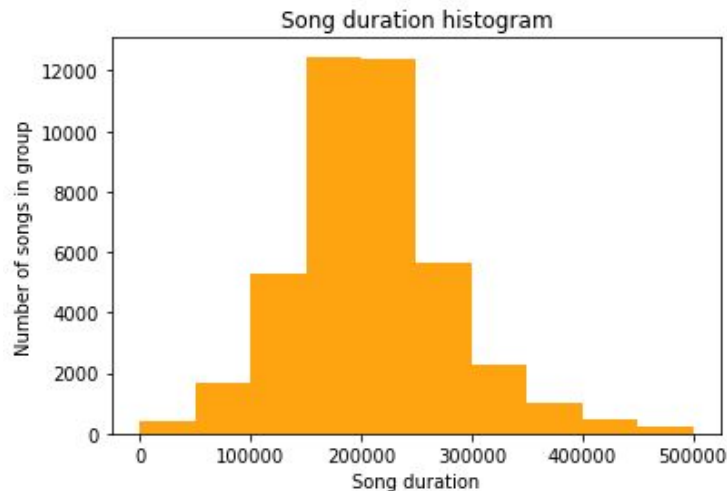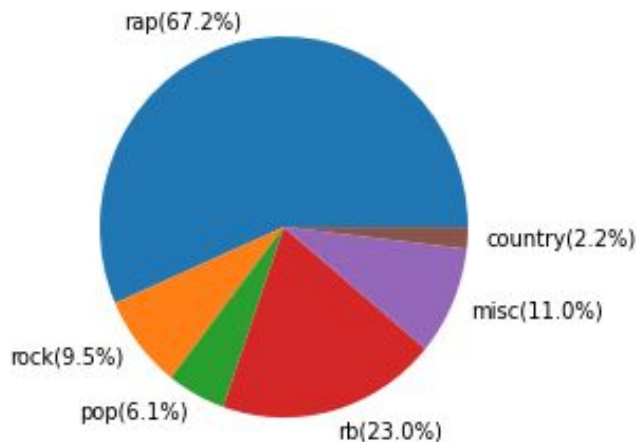# Spotify Track Statistics and Lyrics Analysis

An Information Processing and Retrieval Study
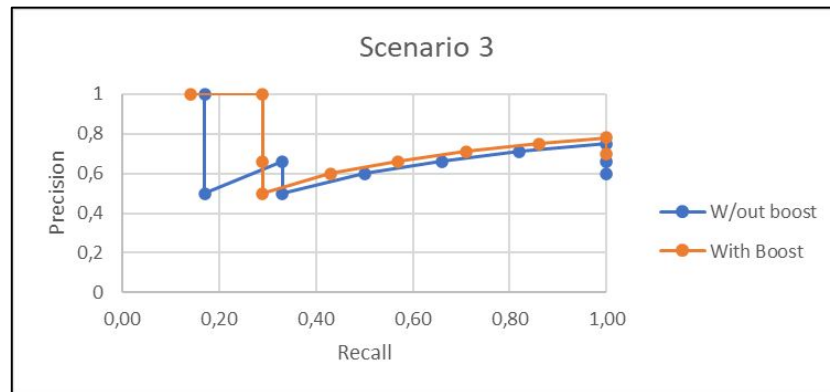
Introduction - Milestone 1

- Spotify Track Statistics and Lyrics Analysis
- Original dataset + Spotify API extra data
- Pipeline: duplicate removal, random trimming, column selection.

Introduction - Milestone 2

- Dataset is multi-language. Many filter options not viable.
- Normal filters applied + Beider-Morse in the artist.
- Many types of boosts applied(Regular Field Boost in Example)
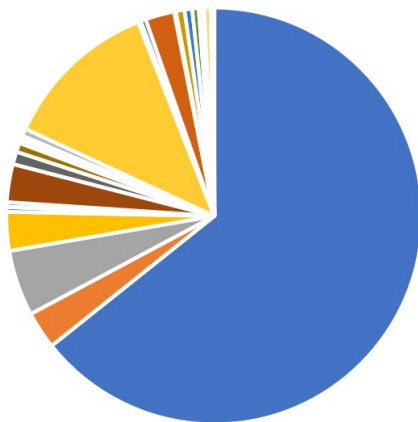
Option	Values
q	lyrics: (summer beach hot sea) title: (summer beach hot sea)
q.op	OR
fl	lyrics, artist, title, album_name
qf	lyrics^3 title
mm	2
ps	3
defType	dismax

Scenario 3

(Precision vs Recall chart: W/out boost, With Boost)

Language Conversion

Solution: Convert the original dataset in an 'english only'.

The advanced natural language processing library 'spaCy' was used to identify the language in each record.

'en'	'fr'	'es'	'de'
'no'	'lt'	'da'	'pt'
'pl'	'nl'	'ca'	'so'
'hu'	'cs'	'id'	'ro'
'sw'	'sq'	'hr'	'it'
'af'	'sv'	'tr'	'fi'
'tl'	'et'	'sl'	'UNKNOWN'
'sk'	'vi'	'lv'	'cy'

New Schema

New filters added to the schema:

- EnglishMinimalStemFilterFactory & EnglishPossessiveFilterFactory
- NGramFilterFactory
- KStemFilterFactory
- SynonymGraphFilter

SGF	text	auto	automobile	machine	motorcar	gondola	railcar	car
	raw_bytes	[61 75 74 6f]	[61 75 74 6f 6d 6f 62 69 6c 65]	[6d 61 63 68 69 6e 65]	[6d 6f 74 6f 72 63 61 72]	[67 6f 6e 64 6f 6c 61]	[72 61 69 6c 63 61 72]	[63 61 72]
	start	0	0	0	0	0	0	0
	end	3	3	3	3	3	3	3
	positionLength	1	1	1	1	1	1	1
	type	SYNONYM	SYNONYM	SYNONYM	SYNONYM	SYNONYM	SYNONYM	<ALPHANUM>
	termFrequency	1	1	1	1	1	1	1
	position	1	1	1	1	1	1	1
	keyword	false	false	false	false	false	false	false

Highlights

Besides the previous sections improvements, we also included highlighting in the search system. This allows us to identify parts of the document matching the query to be referenced in the response.

Re-Evaluation

Scenario 4

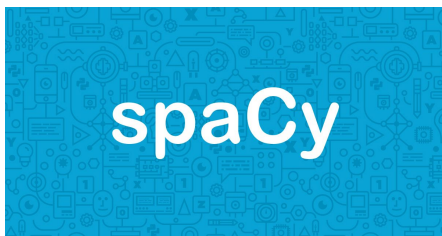- Searching for summer songs

Option	Values
q	lyrics: (summer beach hot sea) title: (summer beach hot sea)
q.op	OR
fl	lyrics, artist, title, album_name
qf	lyrics^3 title
mm	2
ps	3
defType	dismax

Regular Boost

Artist - Title	Relevance
Louis Dunford - Summer in the Manor	R
Flower Boy - Lemonade	R
Hewhocorrupts - Linguistic Violations	N
James Bourne - Alone In Paradise	N
Frank Turner - Little Life	R
Diego Mar - Second Chance	R
SWENDAL - Sunrise In Miami	R
Diego Mar - Second Chance	R
Spain - World Of Blue	R
punii+ - Battle	N

Upgraded Boost

Artist - Title	Relevance
MOUNTAINKING - RED SEA DANCE	R
Ryan Robinette - Hello Summer	R
NerdOut - Summertime	R
ARY - The Sea	R
Hannah Hart -Summer Jam	R
Stephen Schwartz - Simple Joys	R
BZN - Hilee Hilay	R
Death Cab for Cutie - Coney Island	R
Danny B - AwhYeahOohWeE	R
Jaicko - Caribbean Girl	R

Scenario 4 - Evaluation

- Searching for summer songs

The top 10 from the upgraded schema are all relevant! A significant increase in performance comparing to the previous one
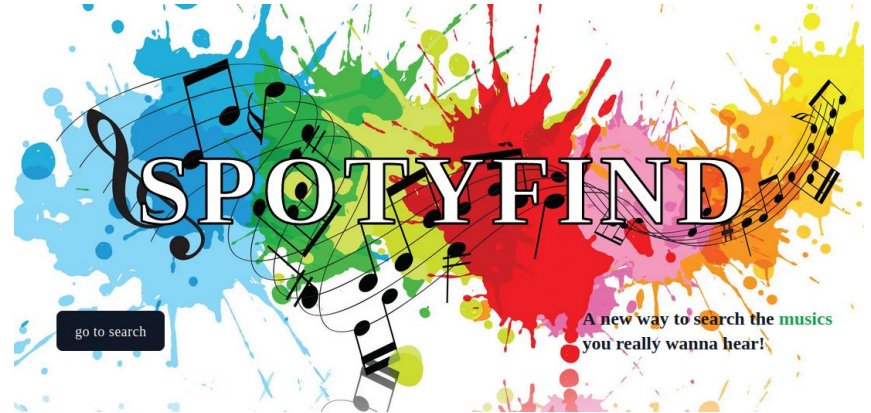

Comparison old boosts vs new boosts
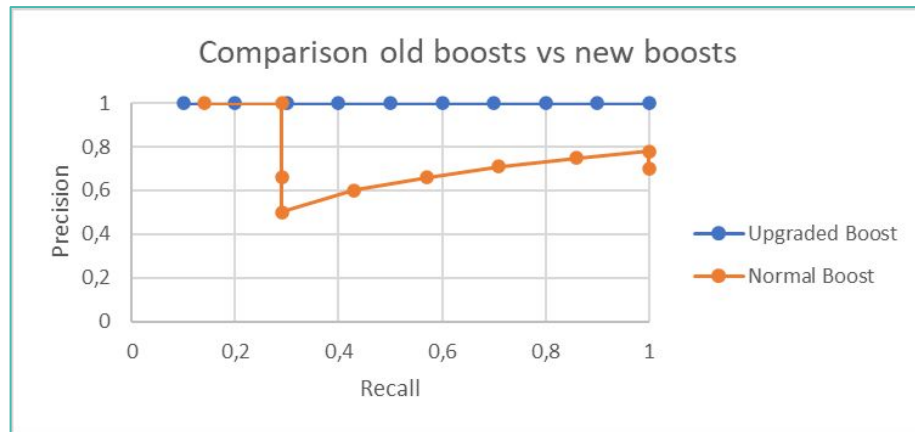
Metric	W/out boost	W/ boost
Average Precision	0.79	1
Precision at 10 (P@5)	0.70	1

User Application

Interface

An application was developed to provide the user with a solution for the complex Solr interface:

- Simple and intuitive

- A search field is present, and the several fields of each track are displayed clearly when used

Conclusions

Conclusions

Since the first steps in the project, the search system has evolved into a solid platform, following a positive growing trend. The main learnings taken from the work developed were:

- The data selection and preparation is fundamental;

- Improvement opportunities to the system are endless, from filter usage, to the interface perks and relevance management algorithms. There is always room to improve;