Processamento e Recuperação de Informação

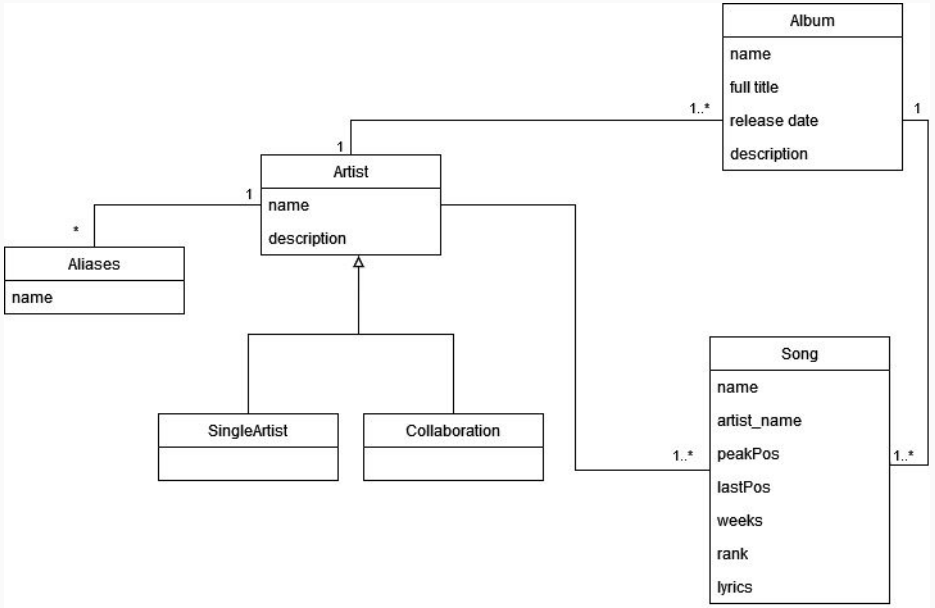
Second delivery

Isla Cassamo - up201808549@g.uporto.pt
Maria Baía - up201704951@g.uporto.pt
Ricardo Nunes - up201706860@g.uporto.pt

Collection and Documents

- **Collection :** All musics, albums and artists obtained in the first milestone and all contained in one core: music.

- **Document Types:** songs, albums, artists.



Collection and Documents

```
song_name, lyrics
    type: descriptiveText
    indexed: true

artist_name
    type: personalName
    indexed: true

music_date
    type: pdate

peakPos, lastPos, weeks, music_rank
    type: pint
```

Listing 1: Simplified version of song schema

```
artist_description
    type: descriptiveText
    indexed: true

artist_name
    type: personalName
    indexed: true
```

Listing 3: Simplified version of artist schema

```
album_name, full_title, album_description
    type: descriptiveText
    indexed: true

artist_name:
    type: personalName
    indexed: true

release_date
    type: pdate
    indexed: true
```

Listing 2: Simplified version of album schema

Information Retrieval

- To perform this task Apache Solr and the EDisMax notation were picked.
- Two systems were developed, System A and System B.
- System A : the base system, designed to match on per defined fields
- System B : similar to system A with the particularity that in this system boosting was implemented to the most relevant fields.
- Two new types “descriptiveText” and “PersonalName” were created to better illustrate the variables at hand.

Indexation Process

Here indexation of the information was done following the pattern bellow:

1. Identify least complex and important fields such as *lastPos and peakPeek* and store these with type *pint* and set indexing to *false*.
2. Date representing fields were stored with type *pdate* but with indexing property set to *true*.
3. The remaining fields, made use of our default field types “descriptiveText” and “PersonalName” as fit best.

Custom Field Types

- “descriptiveText”
  - Standard Tokanizer
  - ASCII folding filter
  - preserveOriginal set to True
  - Lower case filter
  - Filter irrelevant words based on stopwords.txt
  - Stemming filter with applies the Porter Stemming Algorithm
- “artistName”
  - Standard Tokanizer
  - ASCII folding filter
  - preserveOriginal set to True
  - Lower case filter
  - Filter that implements Beider-Morse Phonetic Matching algorithm.

Queries

- The queries are made in order to illustrate the following search topics:
  - Search songs with the main topic “friendship”
  - Search for debut albums of artists in 2000
  - Search for New York born artists with albums in the billboard 200 record chart
  - Search for Albums that won the Grammy Award for Best R&B Album since the '90

Retrieval Process

Information Need #1: Musics where the primary theme is friendship

```
friendship
```

```
System A:
    qf: song_name lyrics

System B:
    qf: song_name^10 lyrics^2
    pf: song_name
```

	System A		System B	
Rank	Song	R	Song	R
1	Reap The Wild Wind	Y	Headlines (...)	Y
2	Headlines (...)	Y	Reap The Wild Wind	Y
3	Freedom Comes, Freedom Goes	N	Freedom Comes, s Freedom Goes	N
4	Could This Be Love	N	Could This Be Love	N
5	Wannabe	Y	Wannabe	Y
6	Buddy	Y	Buddy	Y
7	I Don't Need You	N	I Don't Need You	N
8	I Don't Need You	N	I Don't Need You	N
9	Mirror Man	Y	Mirror Man	Y
10	Simple Kind Of Life	N	Simple Kind Of Life	N
P@10	0.5		0.5	
AP	0.656		0.656	

Table 3. P@10 and AP results

Retrieval Process

Information Need #2: Debut albums of artists in 2000

```
("first album" OR "debut album") AND 2000
```

```
System A:
    qf: release_date album_description

System B:
    qf: release_date^10 album_description^2
    pf: album_description
    ps: 3
```

	System A		System B	
Rank	Album	R	Album	R
1	Madonna by Madonna	N	Madonna by Madonna	N
2	Parachutes by Coldplay	Y	Parachutes by Coldplay	Y
3	Who Is Jill Scott? by Jill Scott	Y	Who Is Jill Scott? by Jill Scott	Y
4	Return of Saturn by No Doubt	Y	Return of Saturn by No Doubt	Y
5	2Ge+her: Again by 2gether	N	Best Of by 50 Cent	N
6	Everyday by Dave Matthews Band	N	Operation Stackola by Luniz	N
7	Dance with Me by Debelah Morgan	N	Come Away With Me by Norah Jones	N
8	Best Of by 50 Cent	N	Mad Season by Matchbox Twenty	N
9	Human Clay by Creed	N	Beware of Dog by Bow Wow	Y
10	Operation Stackola by Luniz	N	Tical 2000: Judgement Day by Method Man	N
P@10	0.3		0.4	
AP	0.445		0.467	

Table 4. P@10 and AP results

Retrieval Process

Information Need #3: New York born artists with albums in the billboard 200 record chart

```
q: billboard 200" AND "New York"
fq: !(artist:"featuring")
```

```
System A:
    qf: artist_description album_description

System B:
    qf: artist_description^10 album_description^5
```

	System A		System B	
Rank	Artist	R	Artist	R
1	G Unit	N	G Unit	N
2	Frank Mills	N	Frank Mills	N
3	Double or Nothing	N	Sharissa	Y
4	Dogg Food	N	N.O.R.E.	Y
5	Sharissa	Y	Post Malone	Y
6	Get Rich or Die Tryin'd	N	Gladys Knight And The Pips	N
7	N.O.R.E.	Y	Automatic Man	N
8	52nd Street	N	A Tribe Called Quest	Y
9	Chapter V	N	AJR	Y
10	Post Malone	Y	Bow Down	N
P@10	0.3		0.5	
AP	0.142		0.341	

Table 5. P@10 and AP results

Retrieval Process

Information Need #4: Albums that won the Grammy Award for Best R&B Album since the '90s

```
q: "grammy" AND "r&b" AND ("win" OR "won")
fq: release_date:[1990-01-01T00:00:00Z TO *]
```

```
System A:
    qf: album_description

System B:
    qf: album_description^2
```

	System A		System B	
Rank	Albums	R	Albums	R
1	Get Lifted	Y	Get Lifted	Y
2	As I Am	Y	As I Am	Y
3	Songs in A Minor"	Y	Songs in A Minor"	Y
4	Toni Braxton	N	Toni Braxton	N
5	Words	N	Words	N
6	Power of Love	N	Power of Love	N
7	The Diary of Alicia Keys	N	The Diary of Alicia Keys	N
8	The Sound	N	The Sound	N
9	Secrets	N	Secrets	N
10	G I R L	N	G I R L	N
P@10	0.3		0.3	
AP	0.623		0.623	

Table 6. P@10 and AP results

Conclusion

- As one took a deeper analysis at the informal collected in milestone 1, it proved imperial to cross information from other sources in order to enriched the quality of the dataset and improve the precision.

- One can then conclude that the final result table included a large dataset of music with complete and coherent data throughout, which can easily be queried to obtain sets of information needed to produce a useful and effective engine for the future platform.

- In the future, in order to further improve these result we intend the enrich the dataset further with more foreign sources, divide the work into multiple cores and improve the precision of the resulting table even further.