Task 1: Simple text-based similarity and retrieval

JULIA WEISSENBRUNNER, DANIELA EBERHARD, DANIEL LINDHUBER, RICHARD HOANG, and SEBASTIAN WINDSPERGER, Johannes Kepler University, Austria

Additional Key Words and Phrases: text-based retrieval, qualitative analysis, Music4All-Onion dataset

ACM Reference Format:
Julia Weißenbrunner, Daniela Eberhard, Daniel Lindhuber, Richard Hoang, and Sebastian Windsperger. 2023. Task 1: Simple text-based similarity and retrieval. 1, 1 (November 2023), 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The main objective of this year's MMSR class is to create a content-based music retrieval system that supports multiple modalities.

Hence, the first step resolves around creating different text-based music retrieval systems and qualitatively comparing their outputs. The system should use a subset of the Music4All-Onion dataset and receive a song as a query, which then outputs a list of songs including their title and artist that are similar to the input song. Additionally, files that contain textual features of each song should facilitate the creation of the retrieval systems.

For the task at hand four different systems were implemented:

- `random baseline`: Independent of the query, the system returns N random songs.
- `Text-based(cos-sim, tf-idf)`: Using cosine similarity as a similarity measurement on the tf-idf weights of each song, the system returns N songs that are most similar to the input track.
- `Text-based(cos-sim, BERT)`: Using cosine similarity as a similarity measurement on the BERT embeddings of each song, the system returns N songs that are most similar to the input track. Here the BERT dataset was chosen because it captures in combination with the cosine similarity the semantic similarity better.
- `Text-based(cos-euc, word2vec)`: Using the Euclidean similarity as a similarity measurement on the word2vec embeddings of each song, the system returns N songs that are most similar to the input track. The Euclidean similarity uses the inverse of the Euclidean distance to get a value in the same range as the cosine similarity. It differs from the cosine similarity by using the distance instead of the angle between two vectors to compute the similarity. Here it was chosen as an alternative to the cosine similarity due to the nature of word2vec embeddings, words that are similar to each other are also closer to each other in the vector space. Hence, the Euclidean similarity seems like a reasonable choice.

Authors' address: Julia Weißenbrunner; Daniela Eberhard; Daniel Lindhuber; Richard Hoang; Sebastian Windsperger, Johannes Kepler University, Altenberger Str. 69, Linz, Austria.

2 IMPLEMENTATION DETAILS

For the implementation of the retrieval systems, Jupyter Notebook together with Python was chosen as a platform and programming language to efficiently write code and visualize results. Jupyter helps in visualizing the datasets nicely as tables, while there are dozens of neat libraries for Python that facilitate the work with datasets, such as pandas. Before writing the methods for each system, the datasets were first read into DataFrames which represent tabular data structures in pandas. The systems were then implemented as follows:

- `random baseline`: DataFrames provide a built-in function called `sample`, which as the name suggests returns N random samples.
- `Text-based retrieval`: For the text-based retrieval systems, a generic method was implemented which accepts an info set, a feature set, song, artist, number of rows to return and a similarity function as arguments. First, the query song is retrieved from the info set, which holds the metadata of all songs, by searching by the artist and song name. The id of the query song is then used to get the corresponding row in the feature set. The similarities between the query song and other songs are computed with the given similarity function. The next step is very crucial, because the row indexes of the features do not match with the indexes of the info set, the similarities are added as a new column to the feature set instead. The two sets are then merged by their song id with the rows being sorted by the similarity column in descending order. The first $N + 1$ rows with the highest similarity get returned, because the first song corresponds to the query track, it gets skipped.

3 QUALITATIVE ANALYSIS

This section covers the qualitative analysis of three different queries (search tracks) and their respective results for the different retrieval methods.

3.1 Methodology

The three search tracks have been selected based on familiarity by all participants. The results have been computed with the implementation described in Section 2.

3.2 Results

To find the genres of the songs, we used the Tool "Music Genre Finder" of Chosic[1]. This Tool shows how Spotify assigns songs to genres.

[1]Available at: https://www.chosic.com/music-genre-finder/

3.2.1 Random Baseline. We can see from the result images below that it always choose random songs and therefore the songs don't fit together either.

	id	artist	song
278	1TdDpmXA1QWe0F1T	Green Day	St. Jimmy
9560	wf4gB1ShcR6STumY	Slipknot	Wait and Bleed
5457	Xbnc3y7iGCDkJ2Be	David Bisbal	Premonición
2585	FipfdKTiPklugYMy	Europe	Rock the Night
8811	sAA5dJb3sNXFrMcm	Dirty Projectors	Cool Your Heart
748	4Rq3yuzopV6EOMrp	Piotr Rogucki	Wizja dźwięku
4827	TZ2IPg8YwxDCcyua	Evanescence	The Only One
1061	6NNmYh1zeixjYgqx	Stevie B	Spring Love
6450	dvfjEpKVTKFa4ykG	Katatonia	Sold Heart
7729	lYZDqRYVV573BRz4	mewithoutYou	Julia (or, 'Holy to the LORD' on the Bells of ...

Fig. 1. "Jingle Bells" result list of Baseline retrieval from the Artist "Frank Sinatra"

	id	artist	song
9730	xYV6Ac6m4otPj7sD	Manilla Road	The Veils of Negative Existence
6684	fI0rk5LL5l6NyCeh	Mazzy Star	California
3598	M4WUwd7i57cfQvZW	Manowar	Kingdom Come
4469	RQJ8biiC0Ed7MKyY	Shakira	Sale el Sol
4750	T4XEwusk9mHBs2mi	Alter Bridge	In Loving Memory
2826	HAyKV0B6Ty1rA6e4	Lisa Germano	Candy
1937	BeqFueRPlZbqOpHD	Metric	Calculation Theme
2372	ELY2iJ98YMasAAW1	Charon	Colder
3596	M4IrvSFYcTDYOZpu	The Pussycat Dolls	Beep
3396	Kpbzp06WpfKQWu5v	Inna	My Dreams

Fig. 2. "Shape of You" result list of Baseline retrieval from the Artist "Ed Sheeran"

	id	artist	song
9783	xvXrMH2YnlH77Y5i	Depeche Mode	Behind the Wheel
7938	mrmtGytnUDYGOUOH	The Black Crowes	Bad Luck Blue Eyes Goodbye
10090	zyzILCQvVeUFIINi	Crowded House	When You Come
5392	X9UmKzxyEDZ0Oscc	Enya	One by One
7711	lUgSuAl7kvy7MSon	New Kids on the Block	You Got It (The Right Stuff)
3265	K1HscSosu2LAfq59	Mr.Kitty	Neglect
1484	8qsKHsA7akNyFo0N	Living Colour	I Want to Know
5193	Vx7KpqO40dAp7SY5	Kraftwerk	Radioactivity - 2009 Remastered Version
7858	mNQNaz6dTkZUtnA3	Bob Dylan	Three Angels
746	4QtPdgfD37ulMNZa	Sylvan Esso	Hey Mami

Fig. 3. "Natural" result list of Baseline retrieval from the Artist "Imagine Dragons"

3.2.2 Text-based(cos-sim, tf-idf). The following results were computed with a cosine similarity function against a feature frame in the tf-idf format. It has to be noted that the tf-idf feature appears to be non-exhaustive, which can lead to poor accuracy or false positives.

The Results for the Song "Jingle Bells" from Frank Sinatra seem very accurate, because in the data set there is another song, which is called also "Jingle Bells" and it is retrieved as the result with the highest similarity. The other songs like "Hear the Bells" or "Saved by the Bell" were retrieved, because of the lyrics and their title ("Bell" in the lyrics and title). Very interesting is that songs like "Michelle" from the Beatles is also retrieved even though the lyrics are not really similar, but it sounds very similar (Same Genre).

	id	artist	song	sim
9160	u8bj2RyzoYZ99dWB	Gwen Stefani	Jingle Bells	0.963528
6081	blZ9zSQBqOMxcPhN	Vanessa Carlton	Hear the Bells	0.601615
6540	eU30OjpKt9zzV6R6	Lil Xan	Saved by the Bell	0.526900
3249	JtwyzoBa2N48HsHo	The Beatles	Michelle	0.507277
9709	xSbRgzIgvXuoeLpL	The Black Heart Procession	Your Church Is Red	0.457886
7142	hxGDHGbn3KTlfp6d	Metallica	For Whom The Bell Tolls - Remastered	0.407539
3130	J4onmjAmjdnYYbpX	The Faint	Southern Belles In London Sing	0.372241
7878	mTbTSXakQDcIH7MK	Dire Straits	Portobello Belle	0.365819
4116	PK2m4Mc7MnPT6az8	Gregory and the Hawk	Voice Like a Bell	0.331044
8397	pZIjIGqH2GBH2U3X	The Free Design	Kites Are Fun	0.317380

Fig. 4. "Jingle Bells" result list of Text-based(cos-sim, tf-idf) retrieval from the Artist "Frank Sinatra"

The Results for the Song "Shape of You" from Ed Sheeran seem very promising in terms of the genre. Nearly each resulting song can be associated with the pop genre, which "Shape of you" is a part of. Only the song "Cheree" as a rock/punk song doesn't fit.

	id	artist	song	sim
7095	hk8cDn49nRGqyAFI	Christina Aguilera	Ven Conmigo (Solamente Tú)	0.581465
7727	lYWQPI2TFoh8nSTC	Blur	Tender	0.551355
10012	zPnDmTeGC4vikUDP	Natalia Lafourcade	No Viniste	0.499672
5782	ZmdJLhQm8gtsOkjW	Clap Your Hands Say Yeah	Over and Over Again (Lost and Found)	0.466910
292	1YOagKsFqTS8BDK2	Zara Larsson	I Would Like	0.463639
2405	EZJM2B4piUzzlJPo	Vanessa Hudgens	Come Back to Me	0.459581
10023	zT3GPtC2T6AyfG9b	Suicide	Cheree	0.423330
2272	DdVnX8c2v5NJtBu4	Aventura	Cuando volveras	0.418954
2634	G2otV6WAmea6VB1f	Steps	Heartbeat	0.416584
9012	tM6vR5sdaje8tjmU	Jenny Hval	Female Vampire	0.416305

Fig. 5. "Shape of You" result list of Text-based(cos-sim, tf-idf) retrieval from the Artist "Ed Sheeran"

The results for the song "Natural" from the band Imagine Dragons are very accurate in terms of genre. The first seven results are a very good fit, only the last four seem to deviate from the search song. Although they are in the ten result tracks, they do not really have any similarities with "Natural". For example, the seventh track is soul and the ninth song reggae. Although these appear to be complete misses, the lyrics are still somewhat similar at certain points. Combined with the fact that all of them belong to the lower half of the results, this is more a problem of ordering than of result quality.

	id	artist	song	sim
5124	VTiQufVGka4dXkST	Napalm Death	Hierarchies	0.544006
9428	vjkaT5dKknSZY00H	Vanessa Paradis	Natural High	0.512680
6744	fcqqxxck3btykgKT	Woods of Ypres	Keeper of the Ledger	0.503596
3501	LXg9nms2tgD14h3B	Papa Roach	Blood Brothers	0.496838
4050	OvWDkevn5dCLHvrW	Turnover	Super Natural	0.436332
7444	jxocCDUjcOYHAoXx	Gilbert O'Sullivan	Alone Again (Naturally)	0.425614
9847	yMRvpBNGnVzi5N6s	Carole King	(You Make Me Feel Like) A Natural Woman	0.410265
2558	FWgz1hVIq59Ve6SY	PJ Harvey	On Battleship Hill	0.328914
293	1YUiDKKSD7GIcXyx	Forfun	Cósmica	0.325034
9127	tyH1H0Aoy6l2E3bN	Wye Oak	It Was Not Natural	0.317386

Fig. 6. "Natural" result list of Text-based(cos-sim, tf-idf) retrieval from the Artist "Imagine Dragons"

3.2.3 Text-based(cos-sim, <feature>). The following result were computed with a cosine similarity function against a feature frame in the bert-format.

The Results for the Song "Jingle Bells" from Frank Sinatra seem not that accurate as with the tf-idf. The First result is again "Jingle Bells" from Gwen Stefani. The second result, "Hellhound On My Trail" sounds very similar and could be a similar genre. There are also some Christmas-Songs retrieved, which makes sense, because Jingle Bells is also a Christmas-Song. The last three tracks are more Pop-Songs and no Christmas-Songs, probably because the Christmas-Songs which were retrieved at the beginning are more in the pop genre.

	id	artist	song	sim
9160	u8bj2RyzoYZ99dWB	Gwen Stefani	Jingle Bells	0.951122
6427	doTmvQlJVL1JRO4V	Robert Johnson	Hellhound On My Trail	0.662801
4621	SJZTstFdLSYvbRAi	Change	The Glow of Love	0.623144
5661	YzXWwWKeFMKNgkU7	Hot Chip	You Ride, We Ride, In My Ride	0.617038
1585	9ScGeeaW8XcxgePd	Kelly Clarkson	Every Christmas	0.603126
9989	zHozLx4GhJsG7xLJ	Cyndi Lauper	Christmas Conga	0.598624
3442	LArarDy0SyTJDoIZ	Eric Clapton	Circus	0.597165
5927	afmSDk2caOd8CCfx	B*Witched	Rollercoaster	0.592372
6176	cItkoIyGIr9LrIJS	Beirut	Elephant Gun	0.590946
5907	aYfhVF6MlwbLPm0i	Rihanna	We Ride	0.590712

Fig. 7. "Jingle Bells" result list of Text-based(cos-sim, bert) retrieval from the Artist "Frank Sinatra"

Also for "Shape of You" the results don't seem as genre accurate as with the tf-idf. There are some "pop" songs like "Jump", "Bing Bing" and "I'm In Love" but most of the other songs are "r&b" and "urban contemporary" songs like "Lovergirl" and "The Way You Make Me Feel". It is noticeable that another "Ed Sheeran" song appears. Furthermore, two of the similar songs are from the same artist "Teena Marie".

	id	artist	song	sim
8951	sxnkfQ7yHuBOZFjJ	Girls Aloud	Jump	0.837990
3693	MhYLWu5NwwmAzrnq	Mary J. Blige	I'm In Love	0.837520
3315	KJyVTXfwdjUGtj5m	Boyzone	Love Is A Hurricane	0.832975
3190	JSW7OaMUVh17yN8J	Michael Jackson	The Way You Make Me Feel	0.831327
2141	CqHIvWSho0esW9se	Crayon Pop	Bing Bing	0.831041
4368	QlwiTDUXdw6TlKMf	Big Mountain	Baby I Love Your Way	0.830039
10074	zu435lwGtNr9JjMF	Teddy Pendergrass	Turn off the Lights	0.823441
1504	8xFD1UO8nr1qwOg6	Ed Sheeran	Thinking Out Loud	0.823009
1876	BGka5j0KKRDSzj1N	Teena Marie	Lovergirl	0.822292
6799	g1Fkyx0T1Ylcw49Z	Teena Marie	Ooo La La La	0.820115

Fig. 8. "Shape of You" result list of Text-based(cos-sim, bert) retrieval from the Artist "Ed Sheeran"

The results for the song "Natural" from the band Imagine Dragons are very accurate in terms of genre. Every single song fits the search criteria, only slightly deviating in the songs emotional pitch. For this search query, this is the most accurate result out of all presented methods.

	id	artist	song	sim
1778	AiE9Cln05EjE1dOf	Within Temptation	The Heart of Everything	0.822411
5543	YCl58AvJwcQ5F7jY	Dream Theater	Learning to Live	0.822307
5753	ZctViWjWYSA62gVU	Draconian	Rivers Between Us	0.819205
3421	L2gnt8dqlkmRjefp	Fallujah	The Void Alone	0.818282
1142	6tRWuK7iWbu3wkFf	Joe McElderry	Ambitions	0.816821
6481	e6Os8czgZksMBWXw	Foster the People	Pseudologia Fantastica	0.816118
7313	j63Pu2SUnO4OASD9	Mystery Jets	Someone Purer	0.813864
5297	Wb1ilt34MSoFeLrh	Of Monsters and Men	Hunger	0.812177
4862	TnDbCfYqdBuEN1h2	My Chemical Romance	Burn Bright	0.811639
4797	TMKMcGJdBPDL2acD	David Bowie	Teenage Wildlife	0.811501

Fig. 9. "Natural" result list of Text-based(cos-sim, bert) retrieval from the Artist "Imagine Dragons"

3.2.4 Text-based(<similarity>, <feature>). We used as similarity euc-sim and as feature word2vec.

The results for the Song "Jingle Bells" from Frank Sinatra are again as for the BERT-Feature not that accurate than the td-idf. The first result is again "Jingle Bells" from Gwen Stefani. The second and the third result are christmas songs, which is fitting for the query. But "The Bomb", "Walpurgisnacht" and "Celina" are really not fitting to the query. "The Bomb" is in the Pop Rap genre. "Walpurgisnacht" is in the Medieval Rock genre. "Celina" is in the Polish Rock genre. At the end we have more Rock Tracks, which are more similar to the original track than the three tracks mentioned before.

	id	artist	song	sim
9160	u8bj2RyzoYZ99dWB	Gwen Stefani	Jingle Bells	0.951122
6427	doTmvQlJVL1JRO4V	Robert Johnson	Hellhound On My Trail	0.662801
4621	SJZTstFdLSYvbRAi	Change	The Glow of Love	0.623144
5661	YzXWwWKeFMKNgkU7	Hot Chip	You Ride, We Ride, In My Ride	0.617038
1585	9ScGeeaW8XcxgePd	Kelly Clarkson	Every Christmas	0.603126
9989	zHozLx4GhJsG7xLJ	Cyndi Lauper	Christmas Conga	0.598624
3442	LArarDy0SyTJDoIZ	Eric Clapton	Circus	0.597165
5927	afmSDk2caOd8CCfx	B*Witched	Rollercoaster	0.592372
6176	cItkoIyGIr9LrIJS	Beirut	Elephant Gun	0.590946
5907	aYfhVF6MlwbLPm0i	Rihanna	We Ride	0.590712

Fig. 10. "Jingle Bells" result list of Text-based(euc-sim, word2vec) retrieval from the Artist "Frank Sinatra"

The results for "Shape of You" for this feature are again genre similar. Only the song "Moody's Mood for Love" is not a pop song, it is more a "soul" song. If you listen to it you will realise that the style of the song is not that far away from "Shape of you", so it somewhat fits. It is noticeable that two Spanish appear as similar songs. The songs "Mi soledad y yo" and "Si Tu No Vuelves" are two "mexican/latin pop" songs.

	id	artist	song	sim
8951	sxnkfQ7yHuBOZFjJ	Girls Aloud	Jump	0.837990
3693	MhYLWu5NwwmAzrnq	Mary J. Blige	I'm In Love	0.837520
3315	KJyVTXfwdjUGtj5m	Boyzone	Love Is A Hurricane	0.832975
3190	JSW7OaMUVh17yN8J	Michael Jackson	The Way You Make Me Feel	0.831327
2141	CqHIvWSho0esW9se	Crayon Pop	Bing Bing	0.831041
4368	QlwiTDUXdw6TlKMf	Big Mountain	Baby I Love Your Way	0.830039
10074	zu435IwGtNr9JjMF	Teddy Pendergrass	Turn off the Lights	0.823441
1504	8xFD1UO8nr1qwOg6	Ed Sheeran	Thinking Out Loud	0.823009
1876	BGka5j0KKRDSzj1N	Teena Marie	Lovergirl	0.822292
6799	g1Fkyx0T1Ylcw49Z	Teena Marie	Ooo La La La	0.820115

Fig. 11. "Shape of You" result list of Text-based(euc-sim, word2vec) retrieval from the Artist "Ed Sheeran"

The results for the song "Natural" from the band Imagine Dragons are somewhat accurate. The first and third tracks are pop rock tracks which fit perfectly into the style of "Natural" but at second place is a heavy metal song that is objectively wrong. At 0.83 similarity, it is no surprise that the lyrics match at certain points, but nonetheless the track is a false positive in our opinion. The lower the similarity gets, the more the songs turn into rap/hip-hop which can be seen as a "soft" match, as it fits the style of the band but only to a certain extent the current song. In summary, the results are not as accurate as the other methods but cover a larger spectrum of possible similarities.

	id	artist	song	sim
1778	AiE9Cln05EjE1dOf	Within Temptation	The Heart of Everything	0.822411
5543	YCl58AvJwcQ5F7jY	Dream Theater	Learning to Live	0.822307
5753	ZctViWjWYSA62gVU	Draconian	Rivers Between Us	0.819205
3421	L2gnt8dqIkmRjefp	Fallujah	The Void Alone	0.818282
1142	6tRWuK7iWbu3wkFf	Joe McElderry	Ambitions	0.816821
6481	e6Os8czgZksMBWXw	Foster the People	Pseudologia Fantastica	0.816118
7313	j63Pu2SUnO4OASD9	Mystery Jets	Someone Purer	0.813864
5297	Wb1ilt34MSoFeLrh	Of Monsters and Men	Hunger	0.812177
4862	TnDbCfYqdBuEN1h2	My Chemical Romance	Burn Bright	0.811639
4797	TMKMcGJdBPDL2acD	David Bowie	Teenage Wildlife	0.811501

Fig. 12. "Natural" result list of Text-based(euc-sim, word2vec) retrieval from the Artist "Imagine Dragons"

3.2.5 General qualitative analysis. It is interesting to note that in no query except one (Fig. 8), the same artist appears in the queries. In addition to that, if the same Artist does appear, it does not appear as first choice.

It is also noticeable that in the first two queries by word2vec, songs in other languages were also played as a result.