Pattern-based Vs distant-supervised Relation Extraction from Wikipedia in the Music domain

Artificial Intelligence

Renzo Arturo Alva Principe
746799

Why do i like this artist? (and how to find similar ones?)

Many DBs and KGs about music:

- Discogs: physical & digital releases, artists and labels
- Genius: interpretation of song lyrics
- MusicBrainz: artists and recorded works metadata
- Musixmatch: largest lyrics platform

but...

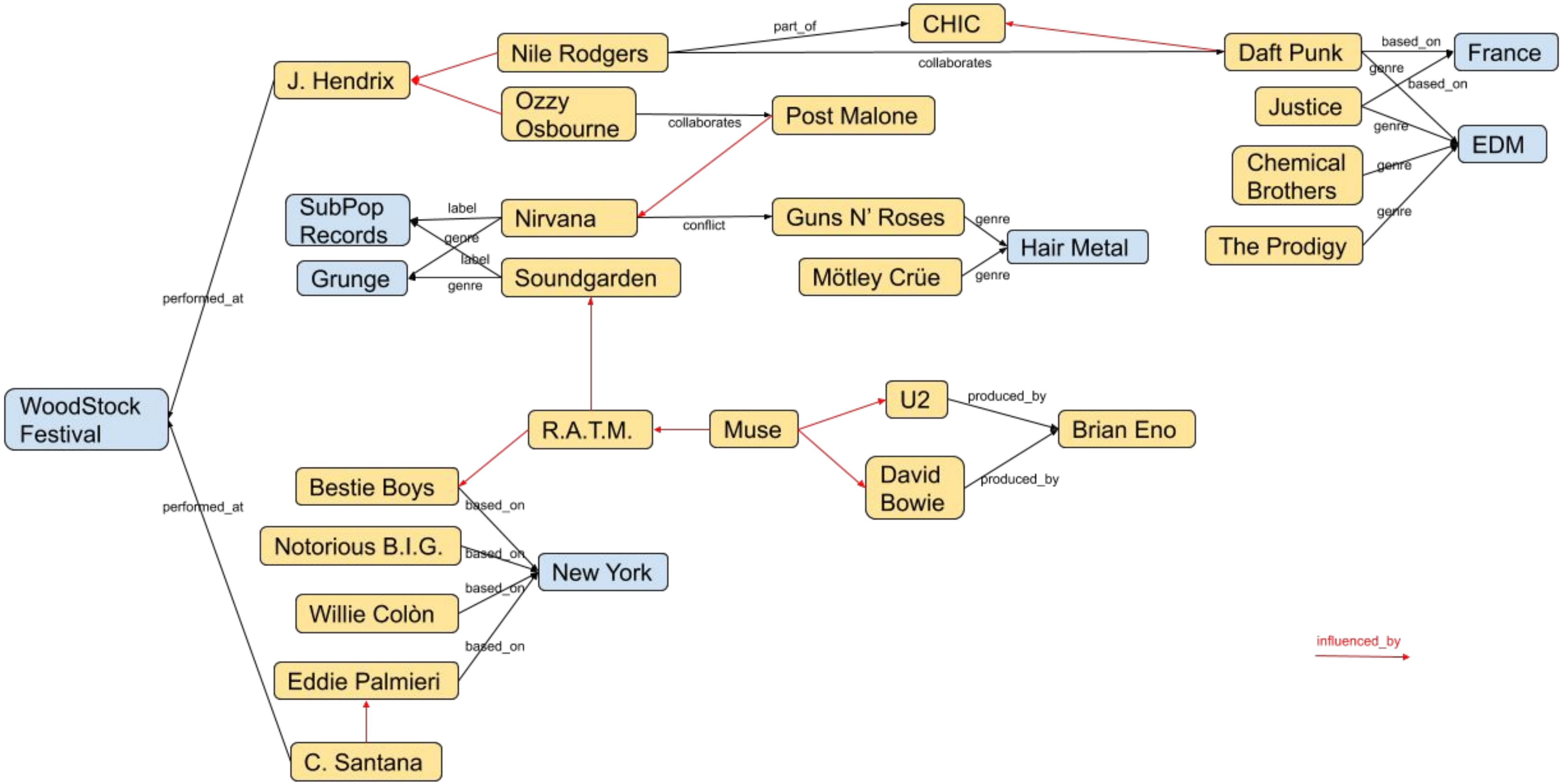
- focused on releases and labels
- no attention in relations between artist (band members, influences, etc)
- useful to answer questions about “what” and not “why” and hard to go beyond basic metadata

Wikipedia is a rich source of information. Artists’ articles contain valuable information useful for the historical/musical characterization of an artist.

Music Understanding Search Engine (MUSE)

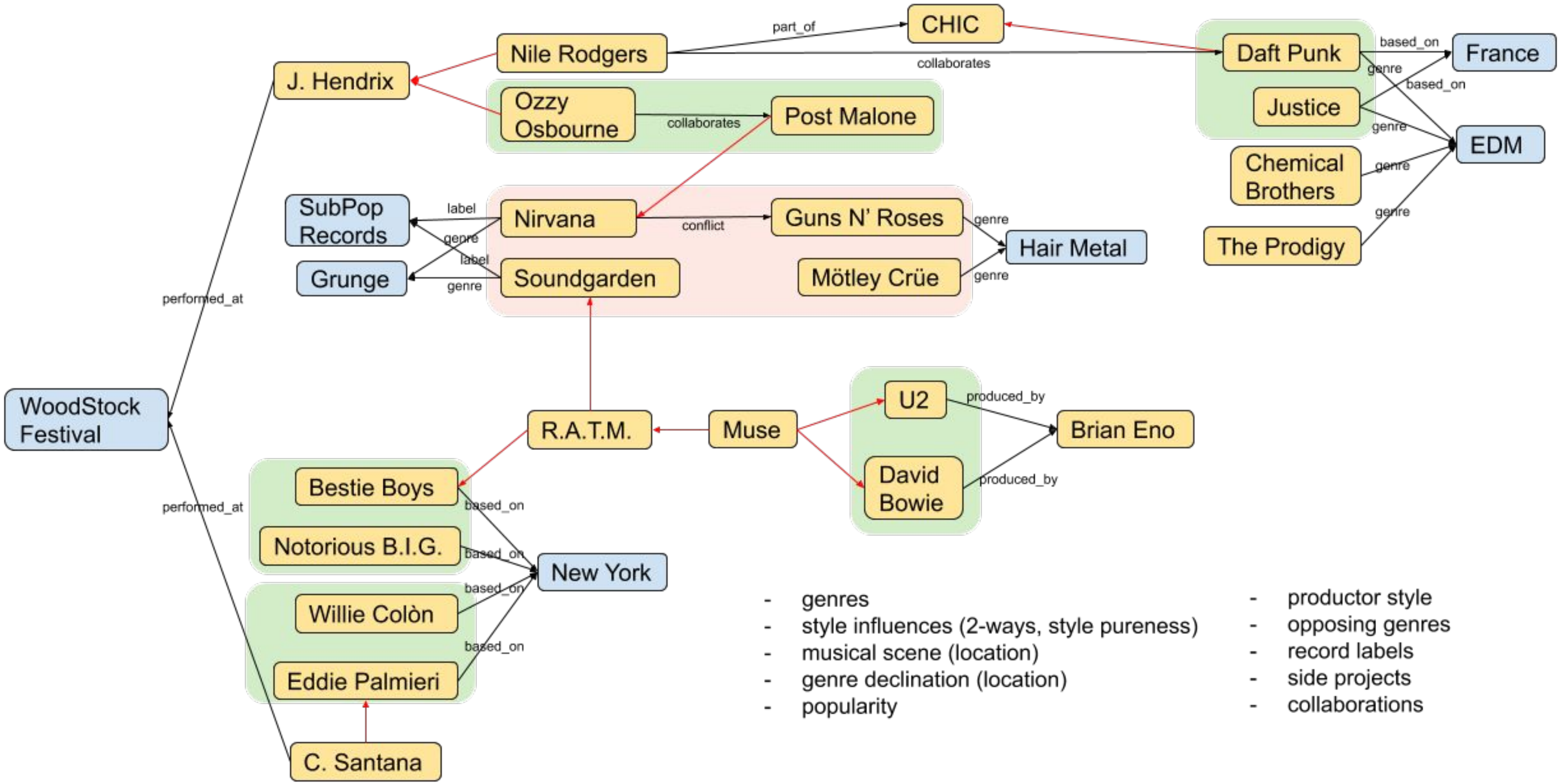
It is a platform that focuses on the artist and the relations between them in order to support the user in a more aware and analytical way of listening and understanding artists, musical genres in a more social/historical perspective

MUSE graph



J. Hendrix
Nile Rodgers
Ozzy Osbourne
part_of
CHIC
collaborates
Daft Punk
based_on
France
genre
based_on
Justice
genre
EDM
Chemical Brothers
genre
genre
The Prodigy
collaborates
Post Malone
SubPop Records
label
Nirvana
genre
label
Grunge
genre
Soundgarden
conflict
Guns N' Roses
genre
Hair Metal
Mötley Crüe
genre
performed_at
WoodStock Festival
performed_at
R.A.T.M.
Muse
U2
produced_by
Brian Eno
David Bowie
produced_by
Bestie Boys
based_on
Notorious B.I.G.
based_on
New York
based_on
Willie Colòn
based_on
Eddie Palmieri
C. Santana
influenced_by

MUSE graph - highlights



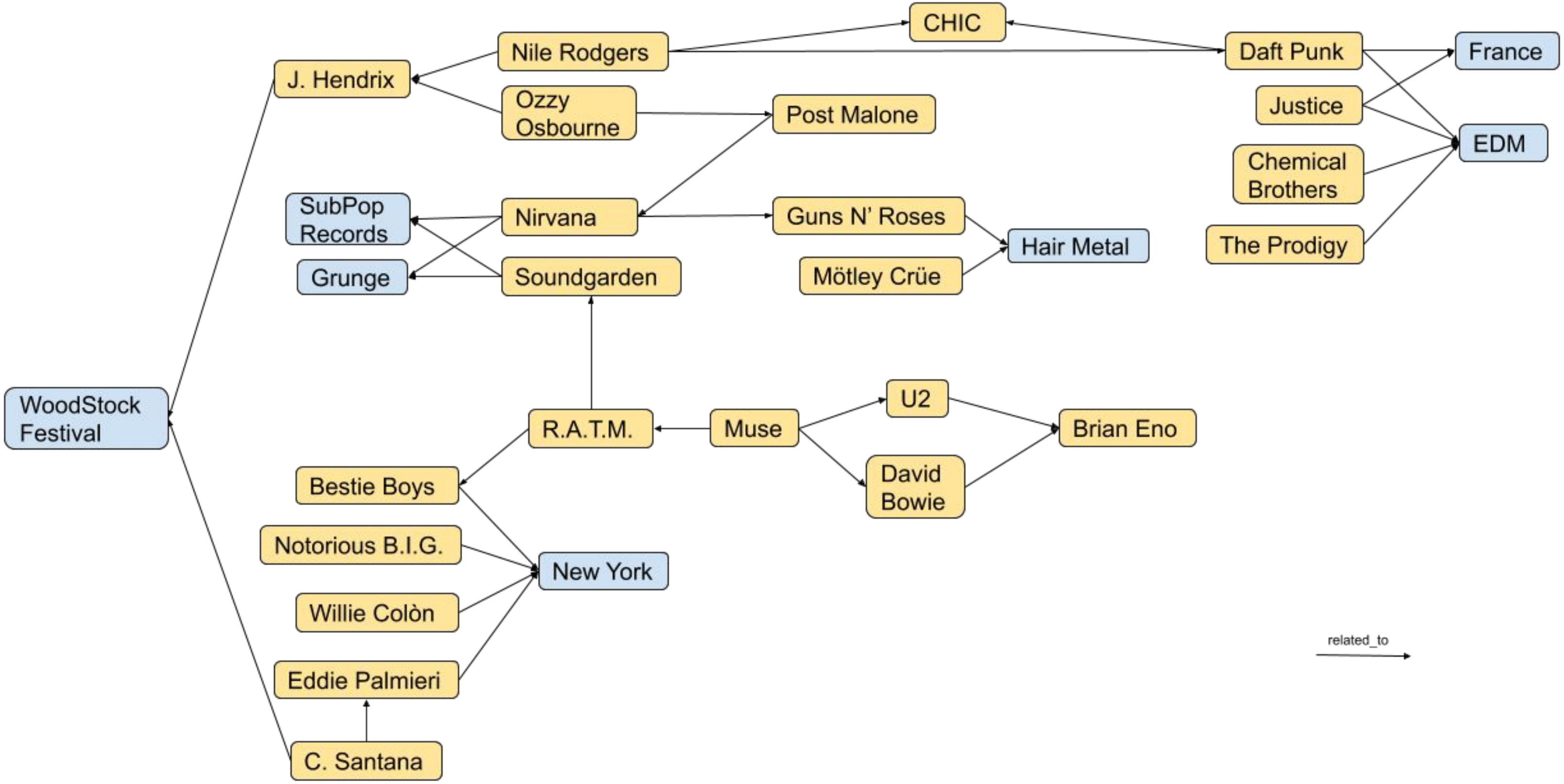
- genres
- style influences (2-ways, style pureness)
- musical scene (location)
- genre declination (location)
- popularity

- productor style
- opposing genres
- record labels
- side projects
- collaborations

Why MUSE?

- music lineage exploration for lazy people: forget about wikipedia long articles reading to understand artist style origins
- music lineage exploration for more curious ones: facts and highlights
- support for reccomendation systems

Start point


CHIC
Nile Rodgers
Daft Punk
France
J. Hendrix
Ozzy Osbourne
Post Malone
Justice
EDM
Chemical Brothers
SubPop Records
Nirvana
Guns N' Roses
Hair Metal
The Prodigy
Grunge
Soundgarden
Mötley Crüe
WoodStock Festival
R.A.T.M.
Muse
U2
Brian Eno
David Bowie
Bestie Boys
Notorious B.I.G.
New York
Willie Colòn
Eddie Palmieri
C. Santana
related_to

Problem addressed

influenced_by relation:

- Wikidata: instances <1000
- Inflooenz.com: many artists' influences and followers

Billie Eilish

- Inflooenz: Skylar Grey, Frank Ocean, Earl Sweatshirt, Marina & the Diamonds
- Wikipedia:

"She has cited Tyler, the Creator, Childish Gambino and Avril Lavigne as major musical and style influences for her and other influences include Earl Sweatshirt, Amy Winehouse, the Spice Girls, and Lana Del Rey"
"She has also named Rihanna as an inspiration (...)"

Aim of this course project:

Evaluate different approaches of RE in order to extract instances of influenced_by relation from Wikipedia and complete the available ones from Inflooenz.

RE techniques (1)

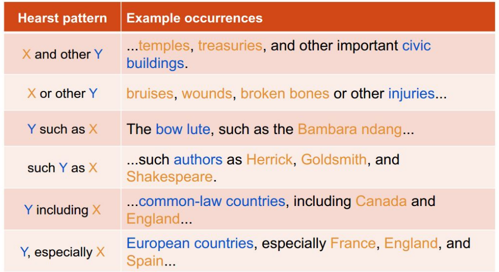
Hand-built patterns (Hearst, 1992): *

- high precision
- low recall (poor generalization)
- lot of work to think of all possible patterns
- don’t want to do his for every relation

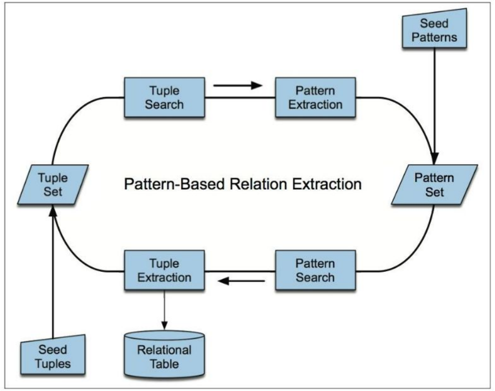
Bootstrapping methods (Brin, 1992):

starting from seed instances and patterns this virtuous circle find new patterns and instances in large amounts of text

- better recall
- semantic drift



Hearst pattern	Example occurrences
X and other Y	...temples, treasuries, and other important civic buildings.
X or other Y	bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y, especially X	European countries, especially France, England, and Spain...



RE techniques (2)

Supervised:

- not limited to one relation per model
- uses features like POS, NER, dependency tree, etc
- can get high accuracies with enough hand-labeled training data,
- build the training set is expensive

Distant Supervised (Mintz, 2009): *

KB instead of labeled data.

positives → For each instance of the relation in the KB, select sentences that match these tuples.

negative → select sentences with couples of entities where the relation doesn’t hold on the KB.

Featurize and train a classifier

- less manual effort but strong assumption
- we cannot learn relations that are not in the KB

Hand-build patterns (1)

The first sentence in a wikipedia pages report in most cases the same information in a standard order.

*"**Philip Glass** (born January 31, 1937) is an American composer and pianist"*

→ patterns work excellently in *type* and *nationality* extraction.

different instead for *influenced_by:*

- NER, POS and dependency tree tags make patterns more general than raw tokens
- 2 patterns for *type* and *nationality* Vs 14 patterns for *influenced_by*

Why not bootstrapping?

X *influenced_by* Y appears 1 time in wikipedia -> no virtuous circle

Hand-build patterns (2)

Manual analysis highlighted that target sentences:

- main verbs: cite, refer, list, mention, credit, claim, states, named
- relevant nouns: inspiration, influences
- verb, nouns and other sentence parts follows patterns

Note that using "influence" as a verb has an opposite meaning than *influenced_by*

```
{'LEMMA': {"IN": ["cite", "refer", "list", "mention", "credit", "claim"]}, "POS": 'VERB'},
{"OP": "*"},
{'LEMMA': 'be'},
{"OP": "*"},
{'LEMMA': 'influence', "POS": 'NOUN'}
```

match:

*“****Michael Jackson*** *claimed that* ***James Brown*** *was his main influence”*

Distant supervision - data set

Support KB (supervise)… ⟶ Inflooenz

Assumption: if there is a dedicated section then positive sentences must be there and outside ones must be negative + Inflooenz supervision

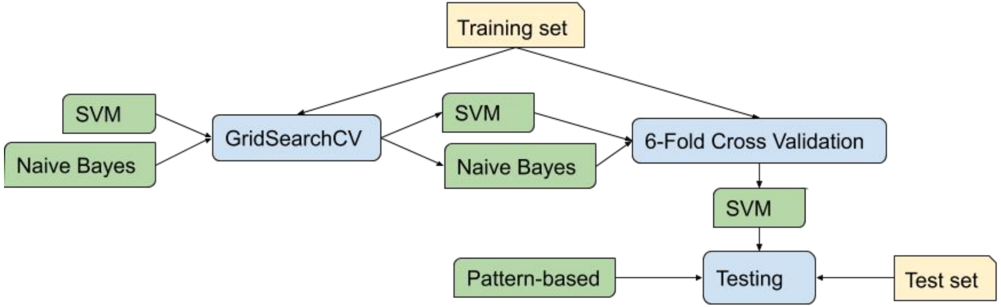
- positive examples (26746):
sentences inside “Style” AND named at least one artist AND that artist is an influencer according to Inflooenz
- negative examples (462):
sentences outside “Style” (but there exists one) AND named at least one artist

(balanced) dataset: 924 sentences

Distant supervision - training

Training:	85% ⟶ 785 sentences (balanced)
Test:	15% ⟶ 139 sentences (balanced)
Text preprocessing:	‘“‘, ‘\n’, ‘\t’ removing
Features:	raw tokens + lemmatized tokens + TFIDF
Models:	Naive Bayes, SVM

POS tagging, lowercase, stop-words and artist names removing ⟶ lower performance



Models comparision - Test

Model	Training set	Tes set	Accuracy	Precision	Recall	F-measure
SVM	auto	auto	.89	.88	.91	.90
patterns	-	auto	.64	.90	.36	.51
SVM	auto	manual	.76	.64	.90	.75
patterns	-	manual	.80	.96	.52	.68
SVM	manual	manual	.80	.75	.74	.75
patterns	-	manual	.80	.96	.52	.68

Auto -> **30%** positives are actually negatives

Distant Supervision is ok but hand-labeled data is still better

Anyway, correct 30% of positives is much better than label data from scratch

Models comparison- Predictions analysis (1)

Distant Supervision:

TP: good generalization (easy patterns & hard to code patterns)

"Turner's sound is reminiscent of that of Warne Marsh, but he also has elements of John Coltrane in his playing."

"Gerard Way said to Rolling Stone, we love bands like Queen, where it's huge and majestic, but also bands like Black Flag and the Misfits, who would go absolutely crazy"

"Frank Iero cites the punk band Lifetime as a big influence."

FN cause: hard expressions

"Adebimpe covered the Pixies song Mr. Grieves under the TV on the Radio moniker at the beginning of his career, layering his voice over forty times."

FP cause: band list and keywords such as “included” are often present in positives:

"The success of Nevermind provided numerous Seattle bands, such as Alice in Chains, Pearl Jam, and Soundgarden, access to wider audiences"

"The ten people involved included musician friends, record company executives, and one of Cobain's closest friends, Dylan Carlson."

Models comparison- Predictions analysis (2)

Pattern-based:

FP: actually is an annotation error!

"Inside Out was touted as a typical Britpop record, and was influenced by the Libertines, Thin Lizzy, the Police and containing elements of the 60s British pop movement"

FN: due to poor generalization of the model, even if almost fit the patterns it's not sufficient

"Other influences that Vedder has cited include Bruce Springsteen, John Mellencamp (...)"

" Their primary musical influences as stated by the band members are Helmet, Tool, and The Cure."

Instances contribution to Inflooenz

MUSE #nodes:	207.286
realted_to instances:	1.880.001
Inflooenz instances:	44.144
pattern-based instances:	38.368
distant-supervised instances:	181.930
pattern-based (new) instances:	36.620 (+83 %)
distant-supervised (new) instances:	>137.786 (+312 %)

Technologies

- NLP: Python & Spacy
- Databases: Mongodb (input data), neo4j (output graph)
- Web-application: Django

Future Works

- extract more relations
- Try bootstrapping approach outside Wikipedia
- extend RE to musical genres and tracks/albums in order to improve highlights and integrate MUSE with other resources like musicbrainz and discogs
- organize/clean artist types and build an ontology to make inferences like:

  Matthew_Bellamy influenced_by Tom_Morello
  Matthew_Bellamy frontman Muse
  then
  Muse influenced_by Tom_Morello

Thank You