Information Retrieval & Text Mining

Marius Popescu

popescunmarius@gmail.com

2021 - 2022

Project 2

Textract 2018 Machine Learning Hackathon: Task 1

Textract 2018 — Sparktech

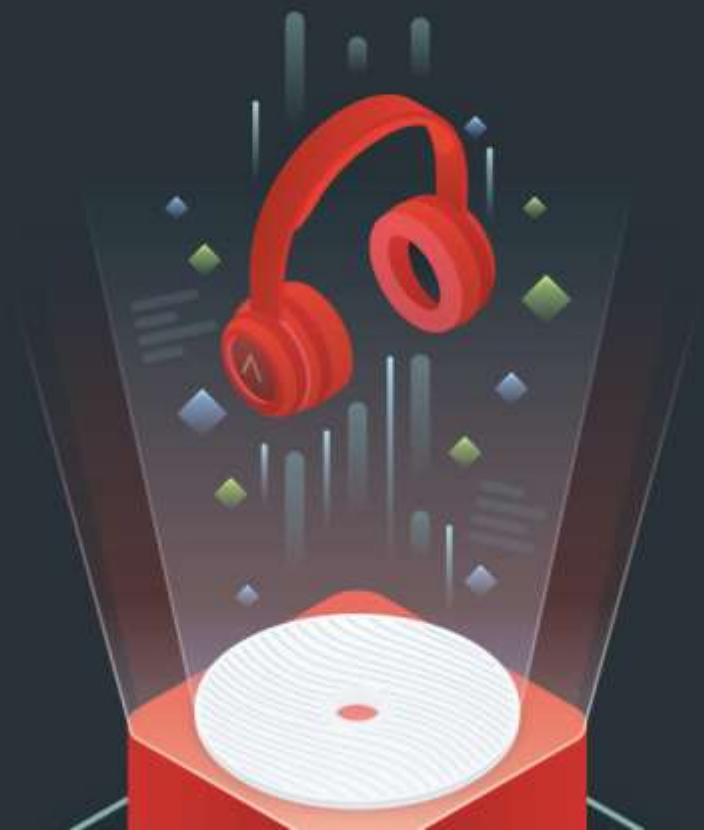
https://www.sparktech.ro/textract/

SPARKTECH

HOME ABOUT SCHOLARSHIPS INTERNSHIP CAREERS TEXTRACT CONTACT

2018 TEXTRACT MACHINE LEARNING HACKATHON

8DEC



HACKATUNE IN

So no one told you it was gonna be this way,
A new Textract edition is proudly underway!
Let's meet: 8th of December on a Saturday,
24 hour ride for you to plug and play.

Assemble your crew, make sure you come prepared,
You have to show them that you're really not scared.
You'll tackle 99 problems, but datasets ain't one,
We know developers don't just wanna have fun.

To help you figure out exactly what this means,
It involves prediction, song lyrics and algorithms.
Whatever happens, don't leave it all to chance,
Find creative ways to address this circumstance.

Let the drums roll, everything's gonna be alright!
At midnight get up! Stand up! Don't give up the fight!
If you hit a bump, you know the show must go on!
The coolest ideas may win this hackathon!

You may say we are dreamers, but we're not the only ones,
Of course, a Machine Learning competition is not for everyone,
But if we're alike from many points of view
All we want for Textract is you!

Task 1

The file “Lyrics-Genre-Training.csv” is a dataset of ~18000 songs and the file “Lyrics-Genre-Test-GroundTruth.csv” is a test dataset with ~8000 songs. For each song, the dataset contains its title, artist, year, complete lyrics and genre. Your task is **to build and compare models which predicts the song’s genre using ONLY its lyrics**.

The files “Lyrics-Genre-Training.csv” and “Lyrics-Genre-Test-GroundTruth.csv” are on the Moodle

The Data

Song	Song year	Artist	Genre	Lyrics	Track_id
Forest-enthroned	2007	catamenia	Metal	I am a night in to the darkness, only soul lost with me, …	18096

The Task

- Try different features: content words (BOW, word embeddings), stylistic markers (stop words), rhyme, rhythm, etc.
- Try different learning algorithm
- Find good combinations
- Don't overfit (report results on (cross)validation and test)

Deliverables

- **Research Report**: document everything you tried and results obtained
- Code

Custom tasks are possible, but you must discuss with me in advance

Hard Deadline: January 23, 2022 23:59