Who am I?

Tech Lead at GeoPhy

Formerly @ OED & other OUP dictionaries

Hip-hop head
Who are you?

Lexicographers?

Technologists?

Hip-hop music lovers?
Agenda

What is the Right Rhymes?

Peeking under the hood

Writing the dictionary

My latest stab at a lyrics corpus
What is The Right Rhymes?
Some motivation
A Dictionary

Defines word senses used in rap

Cites lyrics as evidence

Heavily influenced by OED, HDAS, etc.
A Hobby

Started as a way to learn about building corpora

Dictionary grew out of that

Under continuous development since

Progresses (or not) as my free time allows 😞
A Passion Project

Elmattic
@thisiselmattic

No doubt. Dude who wrote that is doing it for the love:
therightrhymes.com

The Right Rhymes
A Hip-Hop Dictionary
therightrhymes.com

1 9:59 PM - May 10, 2018
A Web Application

Speaking of ... demo?
What’s under the hood?
An old-school web application

Written in the Django Web Framework

Persistence in PostgreSQL database

Deployed on a DigitalOcean droplet (1 vCPU, 2GB RAM)
A RESTish API

Declared with Django Rest Framework

Endpoints service the front-end

Not really designed for public consumption...
API Demo?
Lessons learned

Django is great

JavaScript is powerful, but a whole different skill-set

Design is hard, interview your users

I do not scale 😞
Let’s talk data
Dictionary Data Overview

One big XML document 😲

Validated with a custom XML Schema

Versioned in Git
Editing

Draft entries in oXygen XML Editor

Search Corpus UI for evidence

Paste examples into XML

Edit as needed
Publishing

Denormalize XML with XSLT pipeline

Push to GitLab

Pull from GitLab & load into Django

Post-process for optimization
Shall we look at some XML?
Benefits

Control over the data model

I am comfortable with XML

Low-cost tooling
Pitfalls

Open-source software 🤣 XML

Yet another technology stack

oXygen has no web interface

Zero integration with my other tools
My latest stab at a lyrics corpus
But before that, some history
2012

My first attempt ...

IF YOU'RE HAVIN' PERL PROBLEMS I FEEL BAD FOR YOU, SON—

I GOT 99 PROBLEMS,

SO I USED REGULAR EXPRESSIONS.

NOW I HAVE 100 PROBLEMS.
Lessons learned

Downloading whole sites doesn’t scale

Adding structure is not trivial

Perl...
My second attempt...

The Right Rhymes
Lessons learned

I prefer Python to Perl; SQLite is great

Sourcing reliable metadata is hard

Data integration, also hard

Never throw anything away
2019

Query Spotify API for music metadata

Scrape Genius.com lyrics

Incorporate information from DBPedia, MusicBrainz, etc.
Design

Flask application written in Python

API endpoints to trigger data pipelines

Everything sinks into Postgres

Basic UI
Demo?
Lessons Learned

I ❤ Spotify, but their metadata...

Genius.com lyrics are pretty good, but not perfect

Should I write this in a different language?
Post-Elex Roadmap

Draft some entries 😜

Start integrating Seed 2.0 with The Right Rhymes

Feed corpus data into Sketch Engine?
Thank you so much for listening!

Any questions?