**Podcast Recommender System**

# Introduction

As a project I am building a podcast recommender system. I’ve been wanting to do a data-science project for a while now and this type of project can combine my interest in podcasts too! I run my [own podcast](https://linktr.ee/howyanowpodcast) with a friend so the recommender system may be biased when I’m finished!

# First Steps

I have no experience with recommender systems so, like all great programmers, I first went to google to see if anyone else had done something similar. I found [this](https://www.linkedin.com/pulse/building-recommender-system-podcasts-stephen-witkowski/) well-documented tutorial by Stephen Witkowski which I decided to use as my base to understand the approach to recommender systems.

Stephen references a CSV file of data which he is going to be working from provided by a website called ListenNotes. I couldn’t find a succinct CSV like him but instead, access to an API. It was time for a detour to build something to pull the necessary data from this API.

# Getting Something Working

As it turned out, the dataset from ListenNotes was freely available after some Googling so I decided to use this. The method Stephen uses to generate recommendations is something called “tf-idf” which is short hand for term frequency-inverse document frequency.

The methodology is quite simple. First the text, in my case it’s the podcast’s description, is scanned and any stop words are removed. Then words are weighted based on their frequency and relative to how common they are within a document. These scores are weighted and then similarities can be drawn based on other podcasts with scores as high on these words.

I have implemented this but it isn’t very effective. This is because it usually clings to one or two unusual words (e.g. pop-culture, year-old) and then bases recommendations on this. While this is better than nothing it certainly doesn’t feel very effective.