

IST-736 Final Project

Building a *Marvel Search Engine*



Marvel Cinematic Universe is getting too big

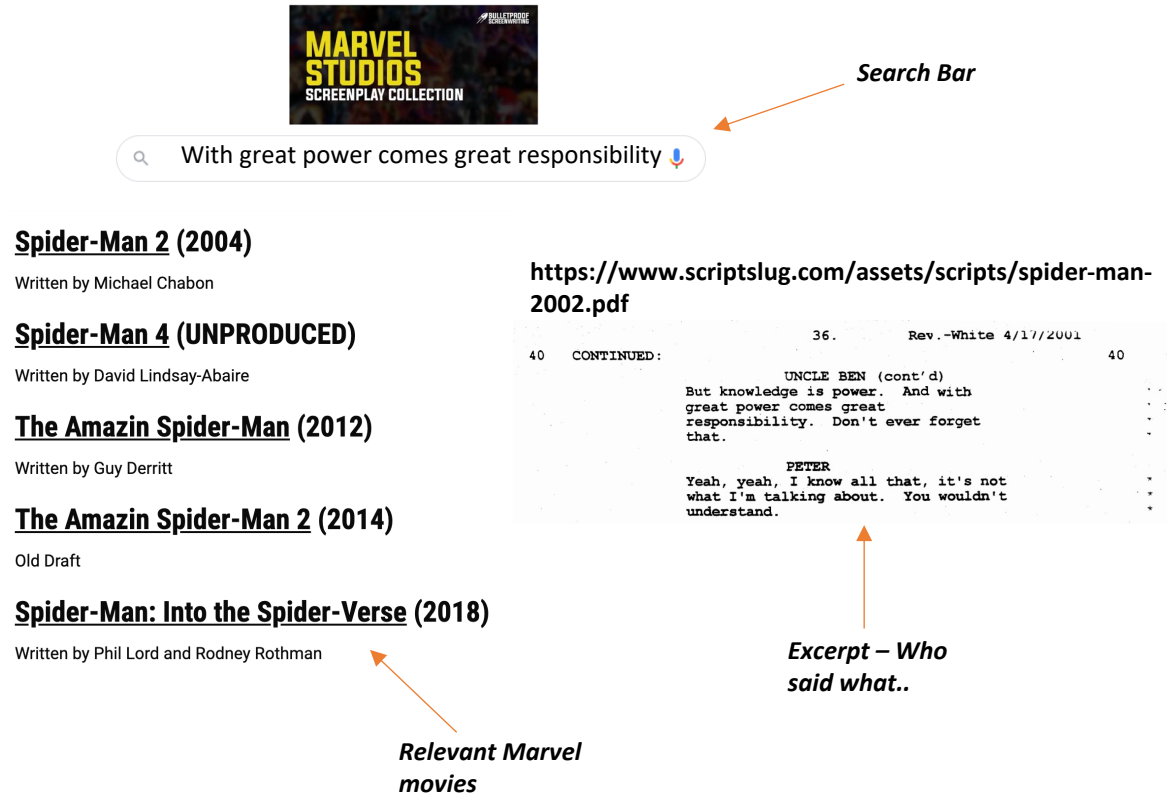
Problem Statement

The Infinity Saga (Phases 1-3) of the MCU, consisted of **23 movies** which is would take more than 3000min to watch it all.

As Phase 4's 25 TV shows and movies approach, how do we know which movies to pre-watch to get relevant information??



Finding Relevant MCU information through Movie Scripts



MARVEL STUDIOS SCREENPLAY COLLECTION

Search Bar: With great power comes great responsibility

Spider-Man 2 (2004)
Written by Michael Chabon

Spider-Man 4 (UNPRODUCED)
Written by David Lindsay-Abaire

The Amazin Spider-Man (2012)
Written by Guy Derritt

The Amazin Spider-Man 2 (2014)
Old Draft

Spider-Man: Into the Spider-Verse (2018)
Written by Phil Lord and Rodney Rothman

Relevant Marvel movies

Excerpt – Who said what..

<https://www.scriptslug.com/assets/scripts/spider-man-2002.pdf>

40 CONTINUED: 36. Rev.-White 4/17/2001 40

UNCLE BEN (cont'd)
But knowledge is power. And with great power comes great responsibility. Don't ever forget that.

PETER
Yeah, yeah, I know all that, it's not what I'm talking about. You wouldn't understand.

Proposed Solution

Build a model for a large text corpus of transcripts of Marvel movie screenplay – then using a *search engine*, given a query (like a dialogue, monologue, location or character) retrieve corresponding documents ranked in order of relevance.

The search engine will then help MCU fans, find a place in the MCU where they can learn more about a specific topic rather than rewatching all the movies.

Finding Relevant MCU information through Movie Scripts

Text Mining Concepts

Scripts are not standardized. Scripts can come as pdfs or text files where each one might follow a different patter. Using text scraping and regex, the text should be standardized.

Raw Script PDF

BLACK PANTHER

Written by
Ryan Coogler & Joe Robert Cole

EXT. DEEP SPACE

A dark screen is lit up by twinkling stars .

SON

Baba?

FATHER

Yes, my son?

SON

Tell me a story .

FATHER

Which one?

SON

The story of home .
A meteorite drifts into frame , heading towards tiny Earth off in the distance.

FATHER

Millions of years ago , a meteorite made of vibranium, the strongest substance in the universe struck the continent of Africa affecting the plant life around it.

The meteorite hits Africa and we see plant life and animals affected by vibranium.

Processing Script

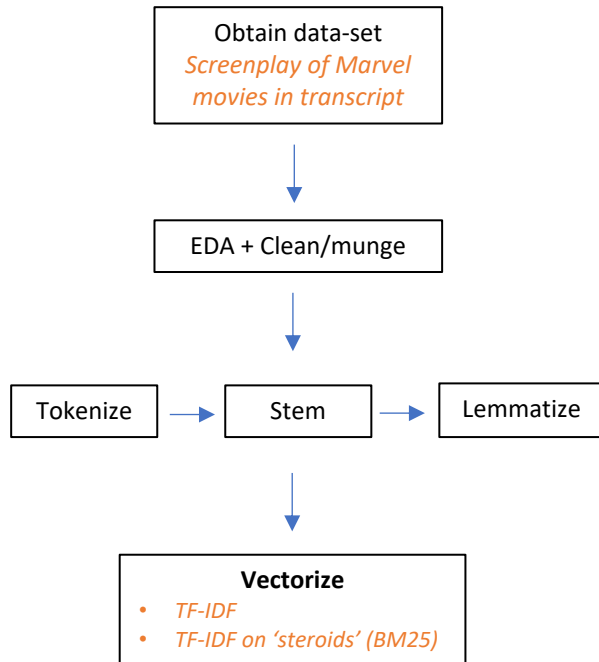
- Get important characters and their dialog.
- Get setting and narration of script.
- Remove un-important comments.
- Clean up Text for modeling.

Vectorized Script

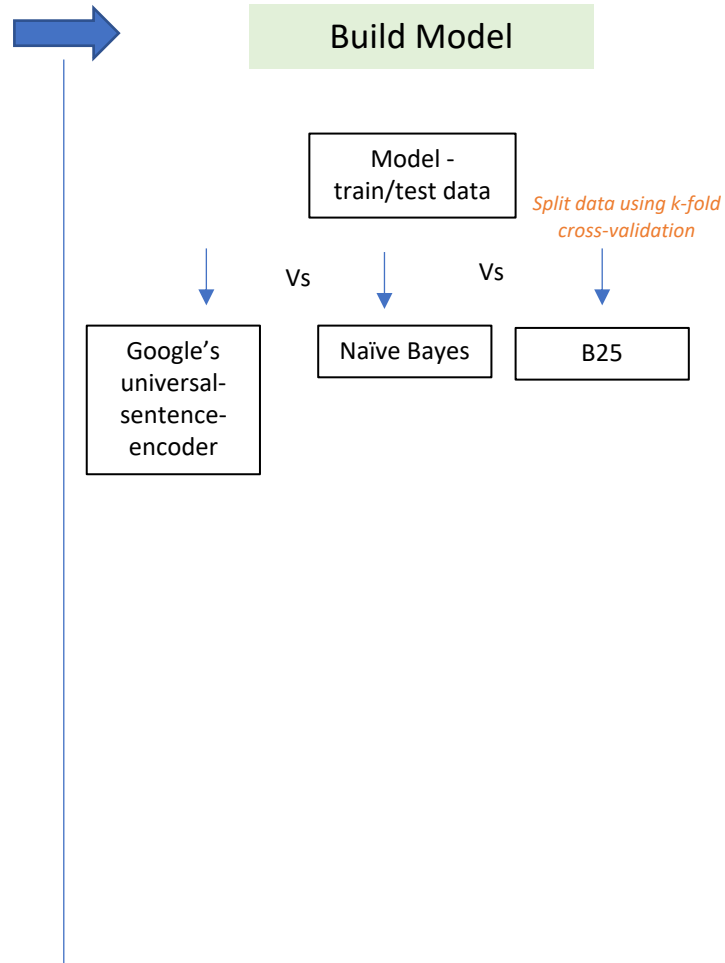
character	line	movie	year
TONY STARK	Uh...mind leaving that on?	Iron Man 3	2013
NATASHA ROMANOFF	Ow. Those really do sting.	Captain America: The Winter Soldier	2014
TONY STARK	How bout that?	The Avengers	2012
STEVE ROGERS	What targets?	Captain America: The Winter Soldier	2014
PEPPER POTTS	Thank you.	Iron Man 2	2010
TONY STARK	Where are we going?	Iron Man 2	2010
PETER PARKER	Well, what if Mr. Stark needs me or something....	Spider-Man: Homecoming	2017
PETER PARKER	Ah, you know, it's boring. Got better things t...	Spider-Man: Homecoming	2017
PEPPER POTTS	Tony? Oh, my God. Are you all right? What's go...	Avengers: Infinity War	2018
JAMES RHODES	Yeah, thanks. Tony, look, I'm sorry, okay?	Iron Man 2	2010
TONY STARK	Three, two, one. Hey, May. My gosh, uh, I want...	Spider-Man: Homecoming	2017
LOKI	Hm. There's not many people that can sneak up ...	The Avengers	2012
BRUCE BANNER	Oh.	Avengers: Age of Ultron	2015
BRUCE BANNER	Why? Shouldn't we be shooting back or something?	Thor: Ragnarok	2017
THOR	This is how my father explained it to me... Yo...	Thor	2011
PETER PARKER	They're not stopping.	Captain America: Civil War	2016
LOKI	No, don't hit it, just press it gently.	Thor: The Dark World	2013
THOR	Know this, son of Coul. You and I, we fight fo...	Thor	2011
TONY STARK	And he didn't invite me!	The Avengers	2012
TONY STARK	Then leave it urgently. Security breach. That...	The Avengers	2012

Search Pipeline

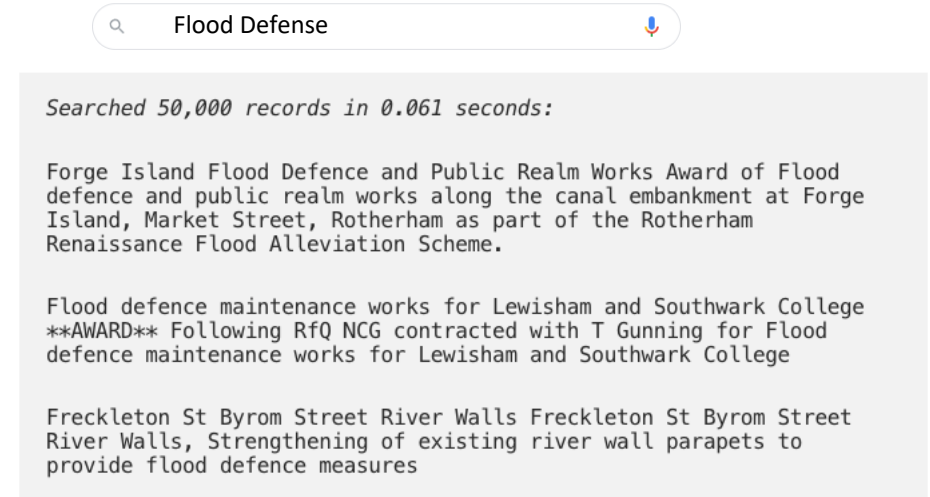
Data preparation



Build Model



Search in action!



Current Status

Text Mining Concepts

Scripts are not standardized. Scripts can come as pdfs or text files where each one might follow a different pattern. Using text scraping and regex, the text should be standardized.

References