




# **SUPERHEROES** WEB-SCRAPING & NLP APPLIED TO PREDICTIVE MODELING IN PYTHON

Daniel Maiorano, Justin Moczadlo, Grace Seaton,  
Aaron St. John, Reilly Williams



To try and predict the creator  
(Marvel, DC, ect.) of a  
superhero from only the history  
text about a hero

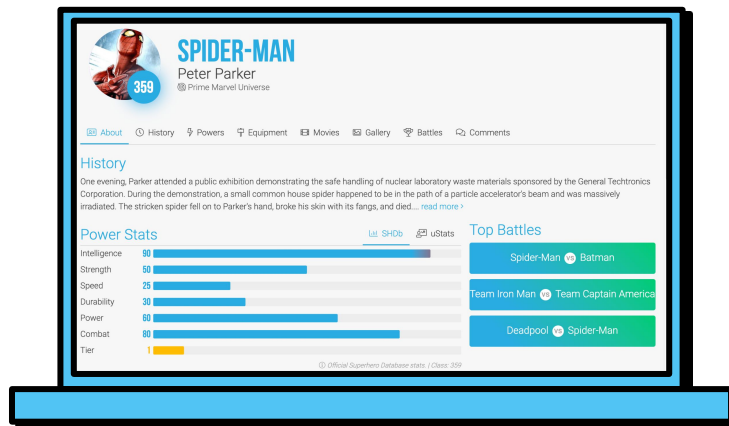
**—PROJECT GOAL**

# SUPERHERO DATABASE

Data was compiled by collecting all unique superheroes from the “Characters” page. For each hero, the following webpages were collected:

- About
- History
- Powers

These raw .html files were then cleaned to create our dataset





# DATA CLEANING & FEATURE ENGINEERING



The following features were extracted from the webpages for each superhero:

- Superhero Name
- Real Name
- Overall score
- Superpowers
- Creator
- Universe
- Alter-egos
- Place of birth
- First appearance
- Alignment
- Base
- Teams
- Relatives
- Gender
- Species
- Height
- Weight
- Eye Color
- Hair color
- History text
- Binary columns for the top 50 most often occurring superpowers
  - Telekinesis
  - Telepathy
  - Magic
  - Immortality
  - Stealth
  - Regeneration
  - Flight
  - etc.



# NATURAL LANGUAGE PROCESSING (NLP)



## FILL NA'S

Fill text column NAs with blanks, numeric columns with 0

## LOWERCASE

Make all text lowercase

## STOPWORDS

Remove English stop words using nltk stopwords list

## LEMMATIZE

Truncate each word to its root ('running' becomes 'run')

## TOKENIZE

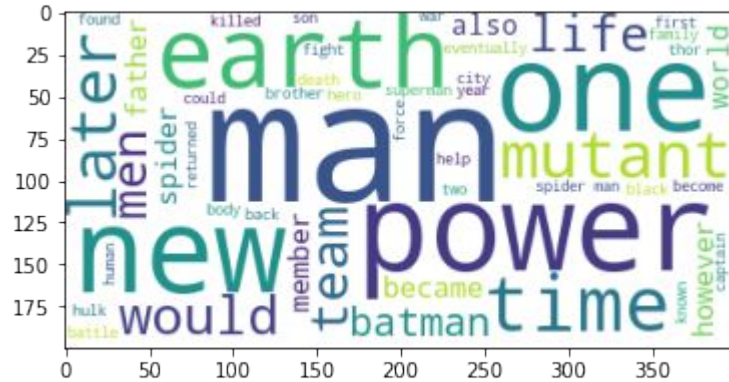
Convert string sentences to lists of individual words

# FINAL DATASET EXCERPT

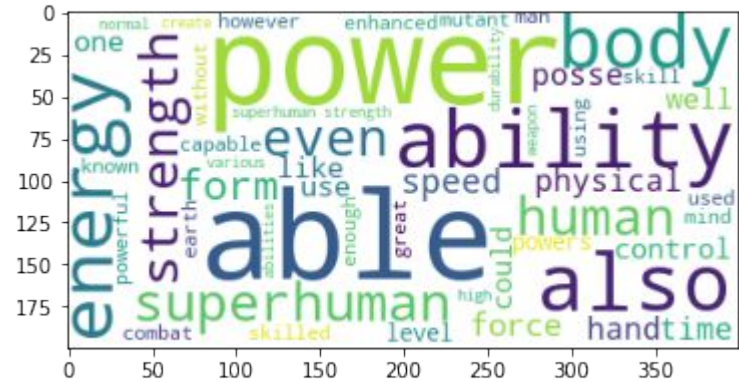
	name	real_name	full_name	history_text	powers_text	creator
0	3-d man	delroy garrett, jr.	delroy garrett, jr.	[delroy garrett, jr., grew, become, track, star, competed, olympic, games, tested, positive, steroids, lost, three, gold, medals, despair, turned, religion, specifically, trine, understanding, group's, founder, jonathan, tremont, found, one, three, fragment, mysterious, object, used, give, garrett, power, 3-d, man, garrett, assumed, power, newfound, spiritual, enlightenment, tremont, never, disabused, notion.]	[]	marvel comics
1	514a (gotham)	bruce wayne		[one, many, prisoner, indian, hill, transferred, another, facility, upstate, order, court, however, fish, mooney, hijack, bus, drive, gotham, city, bus, crash, fired, upon, butcher, glazean, gang, mobster, flee, sight, resurrected, fish, leaf, scene, elderly, hobo, lady, hears, cry, prisoner, release, them, horrified, monstrous, appearance, monster, depart, prison, van, make, way, gotham, city, bruce, wayne, look-alike, departs, bus, thanks, elderly, woman, entering, city, look-a-like, later, saw, selina, kyle, ivy, pepper, give, money, street, gang, selina, kyle, leaf, go, meet, fish, mooney, gang, ivy, visited, bruce, wayne, look-a-like, frantically, asks, bruce, wayne, is, confused, look, almost, identical, him, confused, terrified, ivy, run, ...]	[]	dc comics
2	a-bomb	richard milhouse jones	richard milhouse jones	[richard, "rick", jones, orphaned, young, age, expelled, several, orphanage, disciplinary, reason, placed, state, institution, called, tempest, town, troubled, rebellious, youth, jones, soon, came, attention, institution's, chief, administrator, smashed, guitar, gift, late, father, severely, thrashed, soon, afterward, jones, ran, away, institution, spent, first, half, teen, drifting, town, town, throughout, southwest, trying, avoid, juvenile, authorities, menial, work, could, get, it, age, 16, got, driver's, license, managed, save, enough, money, buy, used, car, overhearing, teenager, dare, friend, ride, desert, rumored, atomic, bomb, going, tested, jones, offered, take, upon, challenge, drove, car, test, site, discover, challenger, timid, show, up, dr., robert, bruce, banner, designer, ...]	[rare, occasions, unusual, circumstances, jones, able, tap, mysterious, near-limitless, energy, source, known, destiny, force, destiny, force, believed, inherent, humanity, jones, used, power, alter, reality, past, bringing, figure, imagination, life, even, figure, different, time, existence, proven, able, render, thousand, kree, skrull, warrior, immobile, thought, single-handedly, overcome, atlantean, army, augment, physical, attributes, heal, sustaining, life, threatening, energy, levitate, full, limit, destiny, force, overall, nature, jones, able, harness, certain, time, unknown, exposed, weapon, powered, gamma, radiation, designed, use, hulk, jones, mutated, radiation, result, jones, transform, superhuman, form, resembles, abomination, jones, demonstrated, ability, transform, and, so, gain, additional, 4, foot, 3, inch, height, 1,835, ...]	marvel comics
3	aa	aa		[aa, one, passive, member, pumice, people, race, stoneworld, hal, jordan, attempted, restart, green, lantern, corps, sent, one, rookies, briq, draw, new, recruits, briq, selected, one, member, stoneworld's, two, dominant, races, aa, pumice-people, kworri, obsidian-folk, horribly, recruited, aa, kworri, briq, captured, flicker, agent, pan-galactic, placement, service, corporation, "cosmic, headhunter", flicker, intended, sell, up-and-coming, green, lantern, race, known, quanhooga, already, succeeded, capturing, hal, jordan, attempted, brainwash, giving, information, concerning, earthing, carol, ferris, aka, star, sapphire, hal, broke, free, control, rescued, aa, others, prepared, eventuality, flicker, left, hero, believe, escaped, four, flew, outer, space, as, left, pan-galactic, company, ship, aa, took, liberty, capturing, one, ...]	[]	dc comics
4	aaron cash	aaron cash	aaron cash	[aaron, cash, head, security, arkham, asylum, hook, hand, real, hand, eaten, killer, croc.]	[]	dc comics
...	...	...	...	...	...	...
1445	zatanna	zatanna zatara	zatanna zatara	[zatanna, daughter, adventurer, john, zatara, wife, sindella, member, mystic, tribe, sorcerer, called, hidden, ones, homo, magi, zatanna, inherited, mother's, ability, manipulate, magic, father's, penchant, heroism, sindella, later, faked, death, return, hidden, one's, sanctum, turkey, leaving, daughter, john, zatara's, care, zatara, traveled, world, donna, later, taught, harness, magical, abilities, zatanna, later, raised, strangers, however, evil, witch, allura, cursed, zatanna, prevented, seeing, father, zatanna, left, search, fruitlessly, natural, parents, zatanna, discovered, father's, diary, created, stage, persona, herself, quest, find, father, led, brief, affair, john, constantine, later, help, justice, league, america, zatanna, able, lift, allura's, curse, reunite, father, later, mother, tragically, sindella, died, rescuing, ...]	[zatanna, genetically, talented, magic, abilities, part, homo, magi, race, such, cast, incredible, number, spells, usually, speaking, backwards, tribute, father, this, however, necessary, cast, spells, zatanna, shown, great, control, magic, ability, even, been, shown, control, elements, element, include, fire, generation, air, control, earth, control, liquid, control, ice, control, fire, control, weather, control, zatanna, also, able, use, magic, telepathic, purpose, reading, minds, remove, specific, memories, wipe, mind, completely, shown, erased, dr. light's, memory, also, tremendous, amount, ability, plant, control, sense, magic, telekinesis, probably, manipulation, much, more, upper, limit, spellcasting, ability, unknown, amongst, powerful, magic, user, universe.]	dc comics
1446	zero	dwn--ec: zero	dwn--ec: zero	[zero, created, late, dr., albert, wily, sometime, early, twenty-first, century, wily, alluded, bass's, ending, mega, man, power, battle, mentioned, developing, robot, blow, away, mega, man, bass, schematic, blueprint, body, seen, bass's, ending, mega, man, 2, power, fighters, learning, past, mistakes, including, accidental, creation, bassium, construction, bass, king, wily, constructed, zero, far, advanced, robot, anything, ever, built, before, power, level, far, superior, bass, mega, man, using, proto, man, reference, zero's, design, also, presumably, began, create, maverick, virus, around, time, wily, even, dubbed, zero, "greatest, masterpiece", zero, contained, flaw, cognitive, program, made, violent, unwilling, obey, instructions, this, wily, decided, seal, capsule, decade, ...]	[]	capcom
1447	zoom (new 52)	hunter zolomon		[hunter, zolomon, better, known, zoom, speedster, super-villain, enemy, flash, became, third, reverse-flash, originally, working, metahuman, profiler, kcpd, zolomon, crippled, gorilla, gmod, gained, power, accident, cosmic, tremor, zolomon, lost, mind, dedicated, life, making, former, friend, wally, west, better, hero, tragedy.]	[tricking, barry, allen, wally, west, breaking, force, barrier, zoom, gained, access, speed, force, -, mysterious, cosmic, force, push, time, space, forward, speed, force, conduit, radically, accelerates, aspect, hunter's, being.]	dc comics



## Tf-IDf & WordClouds



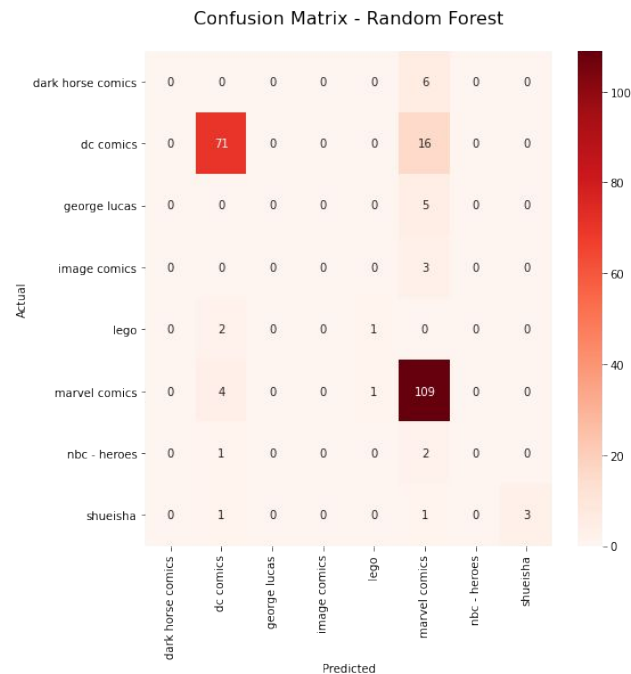
## history\_text



**powers\_text**

# MODELING: RANDOM FOREST

- RandomForest from sklearn.ensemble
- Integer encoded labels (`creator` variable) & limited to `count(creator) > 10`
- Tf-Idf on (`history\_text` variable) for features
- Tuned `n_estimators`, `max_features`, and criterion with GridSearchCV
- Best Result of 81.4% accuracy on test data





# MODELING: XGBOOST

- XGBoost utilizes gradient boosted decision trees and excels at speed and performance for structured or tabular data.
- One-hot encoded labels (`creator` variable)
- Tokenized and vectorized feature (`history\_text` variable)
- Passed xgboost's `XGBClassifier` through sci-kit learn's `MultiOutputClassifier`
- Result of 79.3% accuracy on test data

# CONCLUSIONS

## XGBOOST

- Test accuracy of 79.3%
- Consider data augmentation to increase dataset
- Look into tuning (grid search, genetic algorithm, etc)

## RANDOM FOREST

- Test accuracy of 81.4%
- Further tuning using GridSearchCV
- Calculate variable importances



The background features a comic book aesthetic with various polka dot patterns in red, yellow, and blue. A large, jagged green starburst is positioned on the right side. A thick black border frames the central content area.

# THANKS!

## DO YOU HAVE ANY QUESTIONS?

**CREDITS:** This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**