Predicting Age + Gender from Movie Dialogue

starring Max Harlynking and Arnav Luthra

INT. LIBRARY - DAY

Max and Arnav sit in a library cluster, pondering how to predict age and gender from a dataset of 2,000 uncategorized movie scripts. Sweat drips off Arnav’s brow.

MAX
Eureka!

ARNAV
(flabbergasted)
What happened?

MAX
I think we just cracked the code- If we find a way to separate each characters dialogue, we can use text analysis to try to see which words are most indicative of gender and age.

ARNAV
But… That’s impossible.

MAX
Nothing is impossible. This is CMU.

Goals

From dialogue, predict the speaker's gender and age

Use this data on speech to determine if rule is generalizable

Dataset

Started with 3 files:

A list of 2,000 movie script links

Metadata for about 6,000 movies

A list of characters in the 2,000 movie scripts

Dataset

Step 1: Reduce the number of scripts

Step 2: Download the scripts

Step 3: Find a way to parse script text

Step 4: Match script characters to characters in data

Step 5: Create a CSV of all dialogue linked to character and film data

Step 6: Profit

Dataset

A CSV with 163,413 lines. 18.4 megabytes.

Features:

- Script_id
- Movie title
- Year
- Gross
- Character_name
- Age (nominal)
- Gender
- Dialogue

Methods

Placed Data in Buckets

Correlations

Naive Bayes

~~Decision Trees~~

~~Support Vector Machines~~

Results - Gender

Female

42,151 instances

Male

121,262 instances

Results - Gender

Based on correlation:

Men	Women
the, a, of	me, i, so
hey	oh
man, his, dude, guys	her, she
fight, problem	husband, honey, married
hell	god
jesus	love, loved
america, mission, squad	daddy
friend	please
fucking, shit, fuck	slut
gandalf, frodo	brigette, ginger, margaret, ted, richard, esher, mark, max, toto…

Results - Gender

Naive Bayes, 10-fold cross validation

Accuracy: 74% of data

Most accuracy came from predicting male speech

Many mistakes came from predicting male on female dialogue- mostly stop words

Confusing words:
her, she, please
want, mean, sorry

Results - Age

0-18

5,792 instances

18-36

65,899 instances

36-54

60,649 instances

54-72

13,714 instances

72+

4,145 instances

Results - Age

Naive Bayes, 10-fold cross validation

Accuracy: 43% of data

Most confusion between buckets 36-54 and 54-72

Was able to predict 18-36 most accurately-
Probably because it’s the most quantity

Stop words caused the most confusion

Results - Age

0-18	18-36	36-54	54-72	72+
toto	huh	stump	hellboy	vulcan
daddy	really	the	?	starfleet
dad	dude	a		warp
hhhhelp	shit	catwoman		enterprise
mom	yo			klingon
mommy	cool			thrusters
auntie	kumar			jedi
uncle	names…			kenobi
treehouse				donatello
extraterrestrials				leonardo
				aye
				kafka
				chekov

And now...

INT. WEAN CLASSROOM - DAY

Bored students barely pay attention to the presentation that is about to end. Raja raises his hand.

RAJA
I'll stump you guys with the Catwoman squad.

MAX
Shit.

ARNAV
Dude, yo, really? Jesus.

RAJA
You have a f***ing problem?

MAX
... Cool.

ARNAV
America.

Conclusions

Gender stereotypes are very present

Large difference in dialogue quantity between genders

Less difference in age- but only present over big differences in age

Are movies at fault or gender?

Thank You!