GENDERED INTERACTION ONLINE

By Katie Thomas

Motivation

- Sociolinguistics
- *Women, Men, and Language*, by Jennifer Coates
  - *Women use minimal responses and back channels more*
    - Only really applies in speech, not online interaction
  - *Women use more hedges*
    - Expressing uncertainty
    - Socialized to believe asserting themselves isn't ladylike
  - *Women give and receive more compliments than men*
  - *Women prefer collaborative speech style; men prefer competitive speech style*
  - *Women use questions to avoid the role of expert*
    - Ex) right? Isn't it? Don't you? Etc.
  - *Men avoid self-disclosure and talk about more impersonal topics*

Plan for analysis

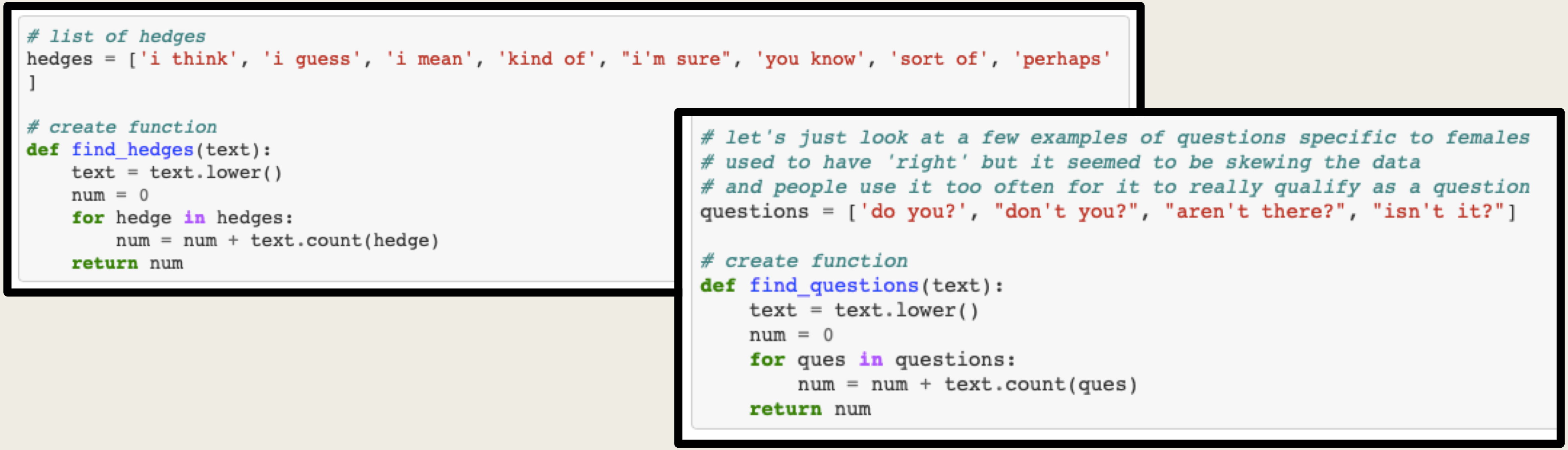
- Questions:
  - *How do women and men present themselves differently online?*
  - *Do people respond differently to male vs. female posters?*
  - *Do male responders respond differently to male vs. female posters? Do female responders respond differently to male vs. female posters?*
- Hypothesis:
  - *Female responders “favor” female posters, male responders “favor” male posters*
    - Unsure of specifics, but thought there would be a difference
  - *Women use more hedges than men*
  - *Women use more questions that “avoid the role of expert” than men*
- What did I look at?
  - *Post length, response length, average sentence length, Google k-band*
    - Never found anything significant with Google k-band
  - *Hedges and questions (maybe compliments?)*
  - *t-tests to determine significance by gender*

Original data

- Original data from a study at Stanford University called RtGender (https://nlp.stanford.edu/robvoigt/rtgender/)
- Format:
  - *Facebook Congress: know gender of poster*
  - *Facebook Wiki: know gender of poster*
  - *Fitocracy: know gender of poster and responder*
  - *Reddit: know gender of poster and responder*
  - *TED: know gender of “poster” (speaker)*

Modifying data

- Merged posts and responses into a single data frame for each source
- Tokenized and found post/response length, post/response sentence length, and average Google k-band
- Hedges and questions:

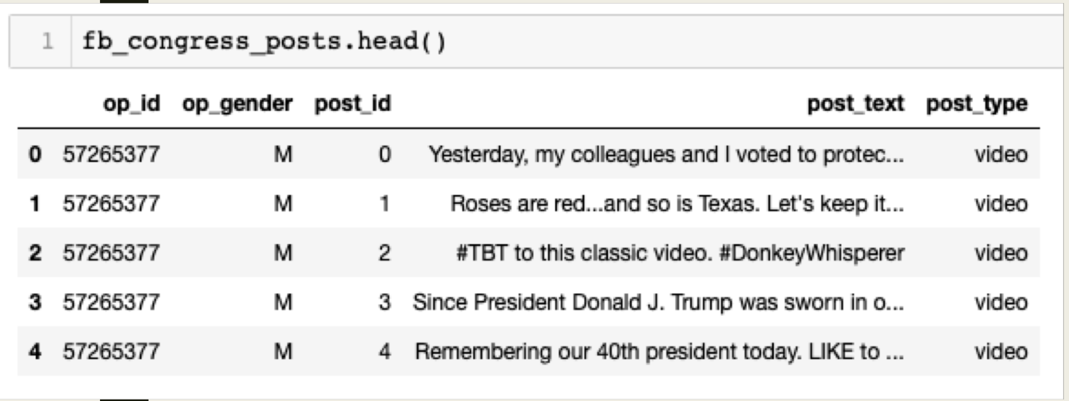


```python
# list of hedges
hedges = ['i think', 'i guess', 'i mean', 'kind of', "i'm sure", 'you know', 'sort of', 'perhaps'
]

# create function
def find_hedges(text):
    text = text.lower()
    num = 0
    for hedge in hedges:
        num = num + text.count(hedge)
    return num
```

```python
# let's just look at a few examples of questions specific to females
# used to have 'right' but it seemed to be skewing the data
# and people use it too often for it to really qualify as a question
questions = ['do you?', "don't you?", "aren't there?", "isn't it?"]

# create function
def find_questions(text):
    text = text.lower()
    num = 0
    for ques in questions:
        num = num + text.count(ques)
    return num
```

- Split into smaller samples for analysis and machine learning

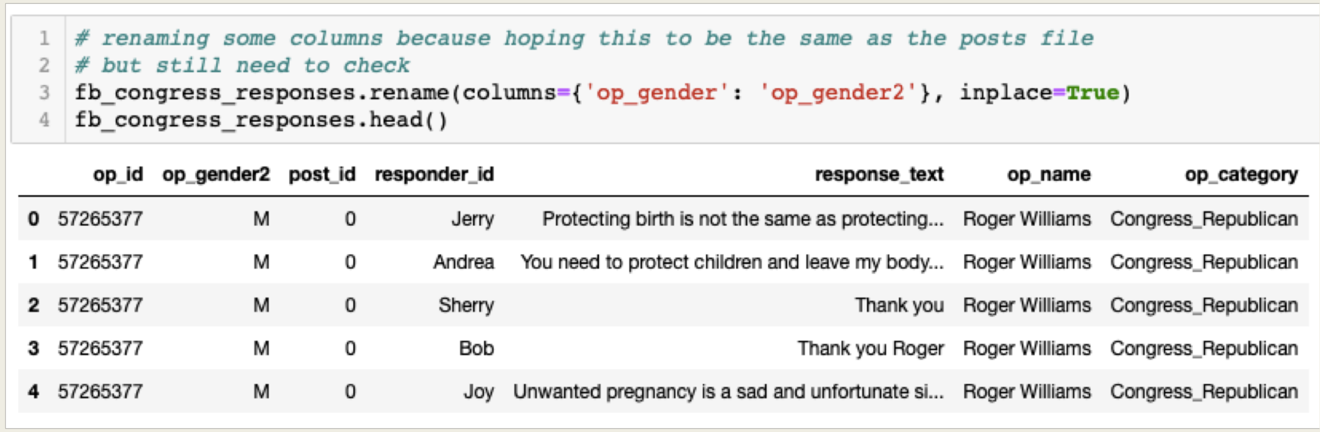
Example: Facebook Congress



```python
fb_congress_posts.head()
```

	op_id	op_gender	post_id	post_text	post_type
0	57265377	M	0	Yesterday, my colleagues and I voted to protec...	video
1	57265377	M	1	Roses are red...and so is Texas. Let's keep it...	video
2	57265377	M	2	#TBT to this classic video. #DonkeyWhisperer	video
3	57265377	M	3	Since President Donald J. Trump was sworn in o...	video
4	57265377	M	4	Remembering our 40th president today. LIKE to ...	video

+

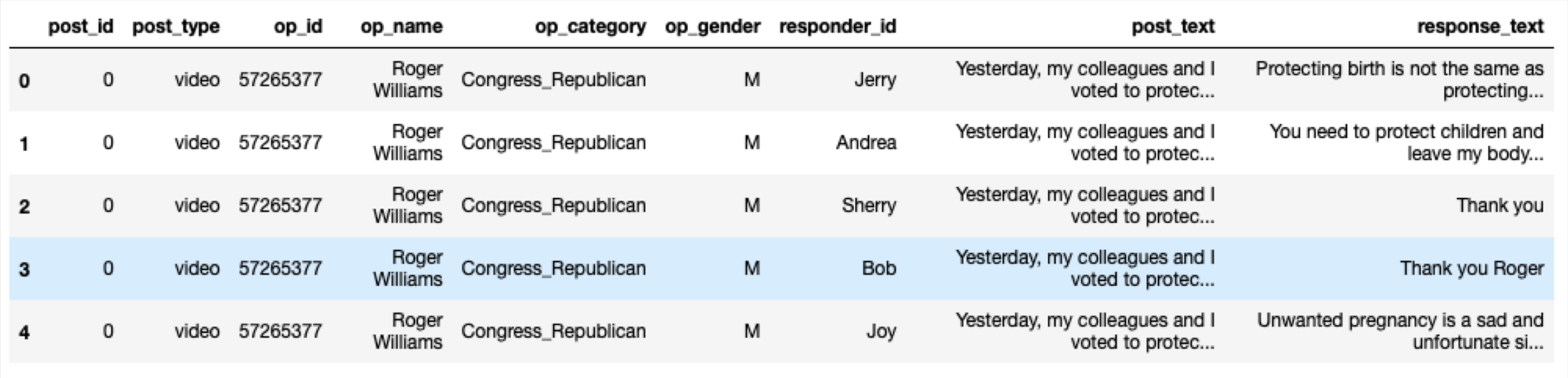


```python
# renaming some columns because hoping this to be the same as the posts file
# but still need to check
fb_congress_responses.rename(columns={'op_gender': 'op_gender2'}, inplace=True)
fb_congress_responses.head()
```

	op_id	op_gender2	post_id	responder_id	response_text	op_name	op_category
0	57265377	M	0	Jerry	Protecting birth is not the same as protecting...	Roger Williams	Congress_Republican
1	57265377	M	0	Andrea	You need to protect children and leave my body...	Roger Williams	Congress_Republican
2	57265377	M	0	Sherry	Thank you	Roger Williams	Congress_Republican
3	57265377	M	0	Bob	Thank you Roger	Roger Williams	Congress_Republican
4	57265377	M	0	Joy	Unwanted pregnancy is a sad and unfortunate si...	Roger Williams	Congress_Republican

Merged on post ID

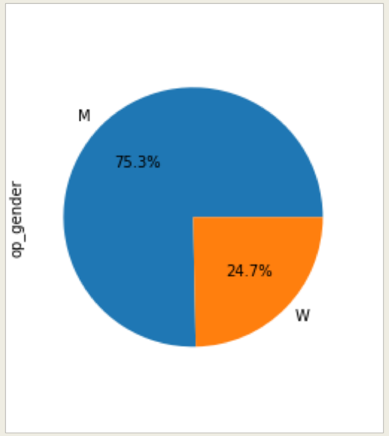
=



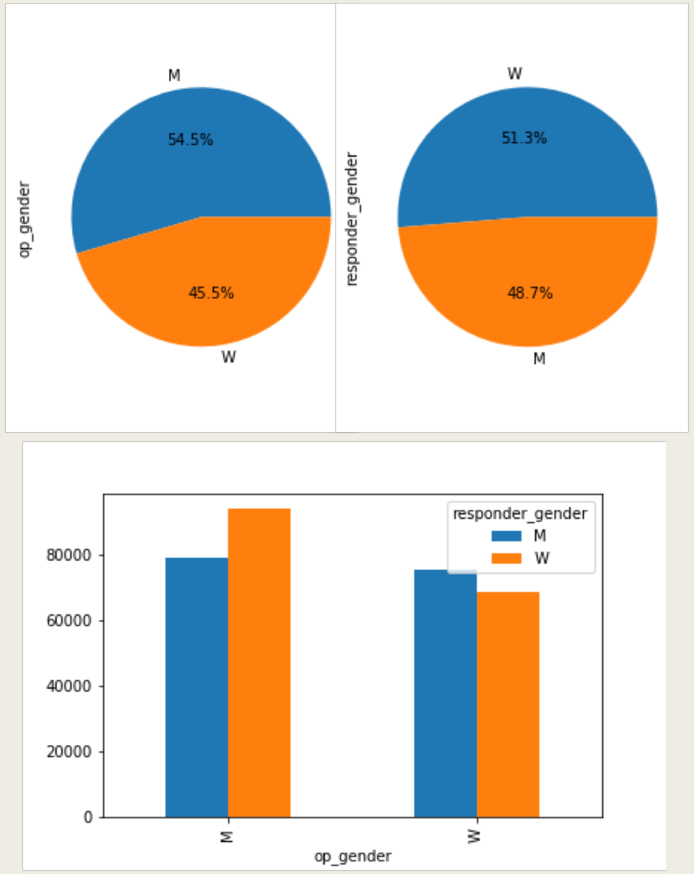
	post_id	post_type	op_id	op_name	op_category	op_gender	responder_id	post_text	response_text
0	0	video	57265377	Roger Williams	Congress_Republican	M	Jerry	Yesterday, my colleagues and I voted to protec...	Protecting birth is not the same as protecting...
1	0	video	57265377	Roger Williams	Congress_Republican	M	Andrea	Yesterday, my colleagues and I voted to protec...	You need to protect children and leave my body...
2	0	video	57265377	Roger Williams	Congress_Republican	M	Sherry	Yesterday, my colleagues and I voted to protec...	Thank you
3	0	video	57265377	Roger Williams	Congress_Republican	M	Bob	Yesterday, my colleagues and I voted to protec...	Thank you Roger
4	0	video	57265377	Roger Williams	Congress_Republican	M	Joy	Yesterday, my colleagues and I voted to protec...	Unwanted pregnancy is a sad and unfortunate si...

Gender distributions

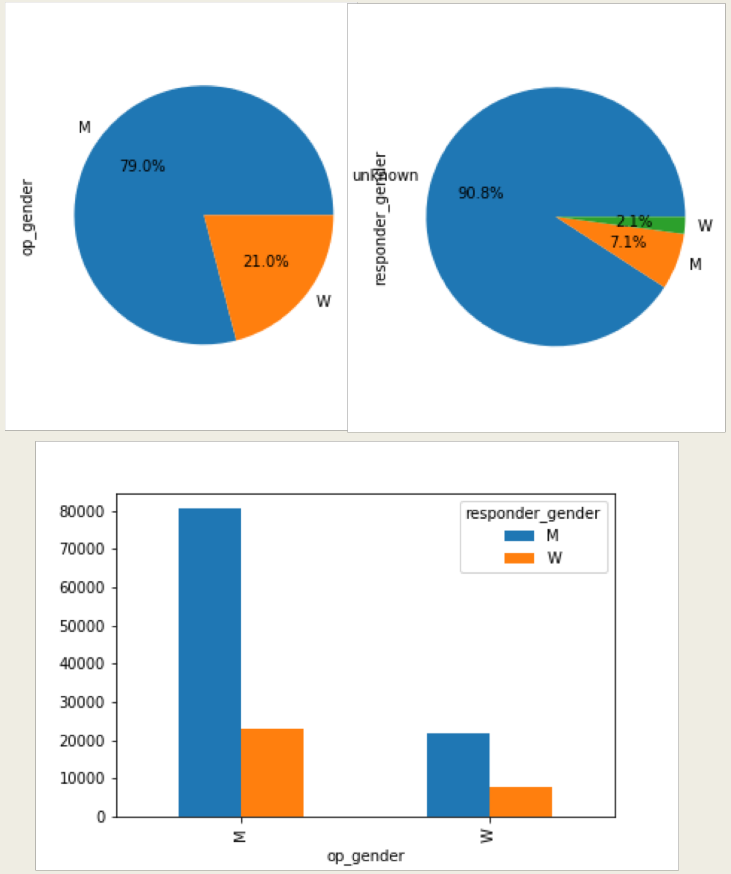
Facebook Congress



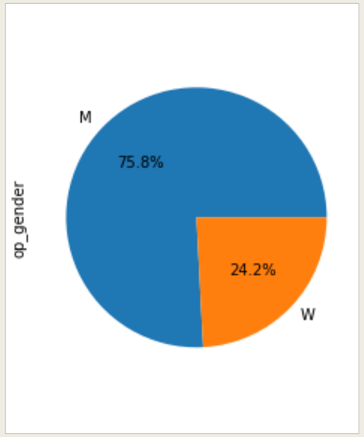
Fitocracy



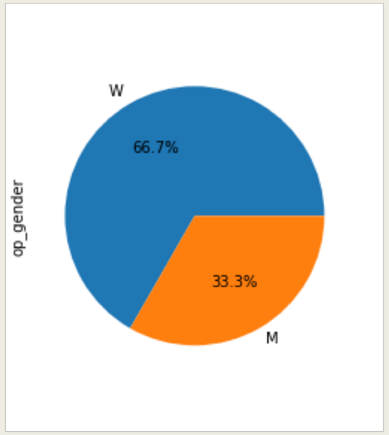
Reddit



TED

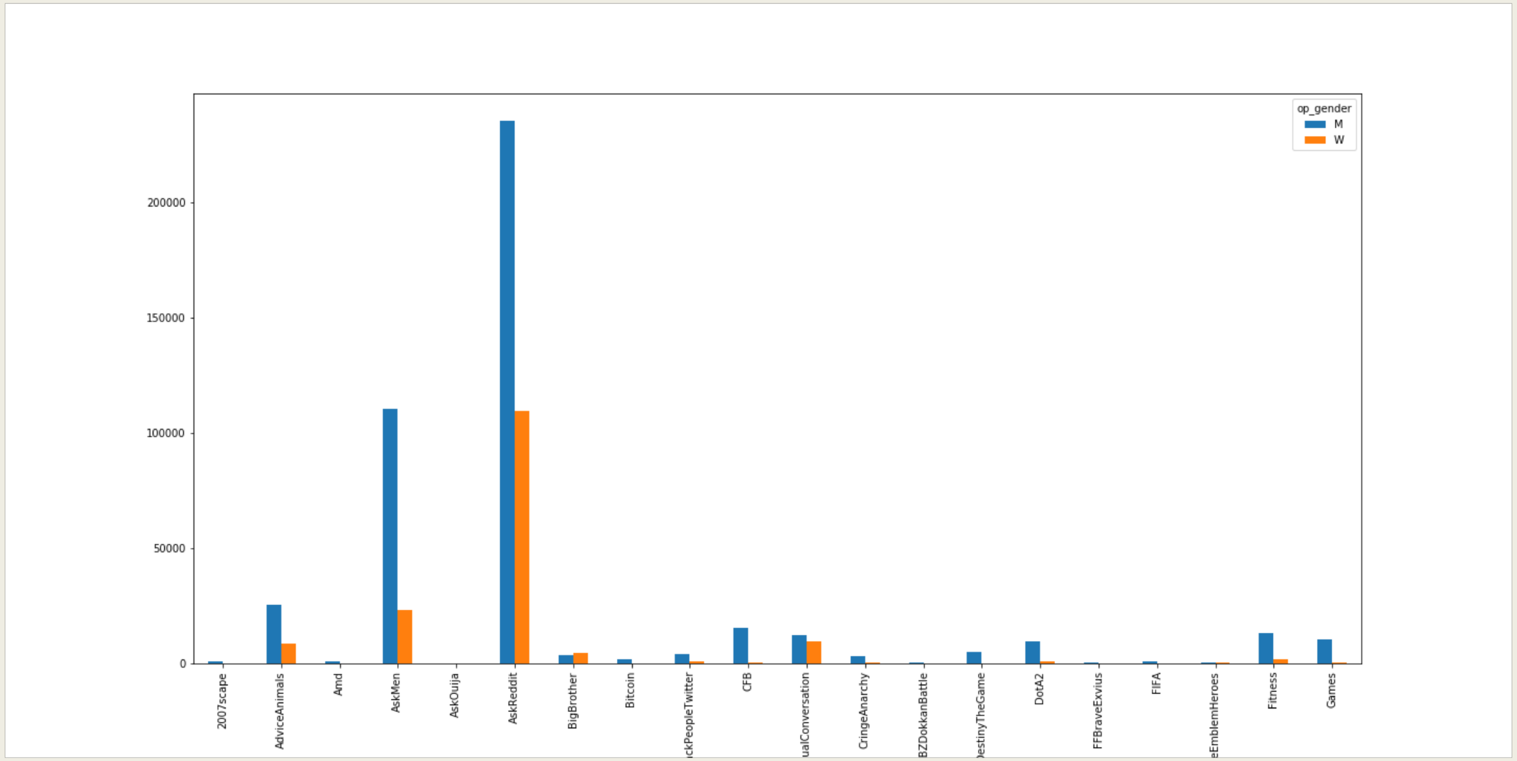


Facebook Wiki



Reddit: more in depth

- 75.8% male: very male dominated
- Is this because of the specific subreddits?



- Out of 98 subreddits, only 5 have more female posters than male posters
  - *BigBrother, awww, counting, relationships, and rupaulsdragrace*

Findings: Gender

Post/response length seems to be correlated with sentence length

	Facebook Congress	Facebook Wiki	Fitocracy	Reddit
Post length	• Female posters have longer posts	• Male posters have longer posts	• Female posters have longer posts	• Female posters have longer posts
Sentence length	• Female posters use longer sentences	• Male posters use longer sentences	• Responses to female posters use longer sentences • Female responders use longer sentences	No significance
Response length	No info about responder gender	No info about responder gender	• Responses to female posters are longer • Female responders have longer responses	• Responses to female posters are longer • Female responders have longer responses
Hedges	• Female posters use more hedges	• Male posters use more hedges	No significance	• Female posters use more hedges • Female responders use more hedges
Questions	No significance	No significance	No significance	No significance

Findings: Gender x Gender

	Male responder	Female responder
Male poster	Fitocracy • Longer responses • Longer sentences in response Reddit • Shorter responses	Fitocracy • Shorter responses • Shorter sentences in response
Female poster	Fitocracy • Shorter responses • Shorter sentences in response Reddit • Longer responses	Fitocracy • Longer responses • Longer sentences in response

Note:

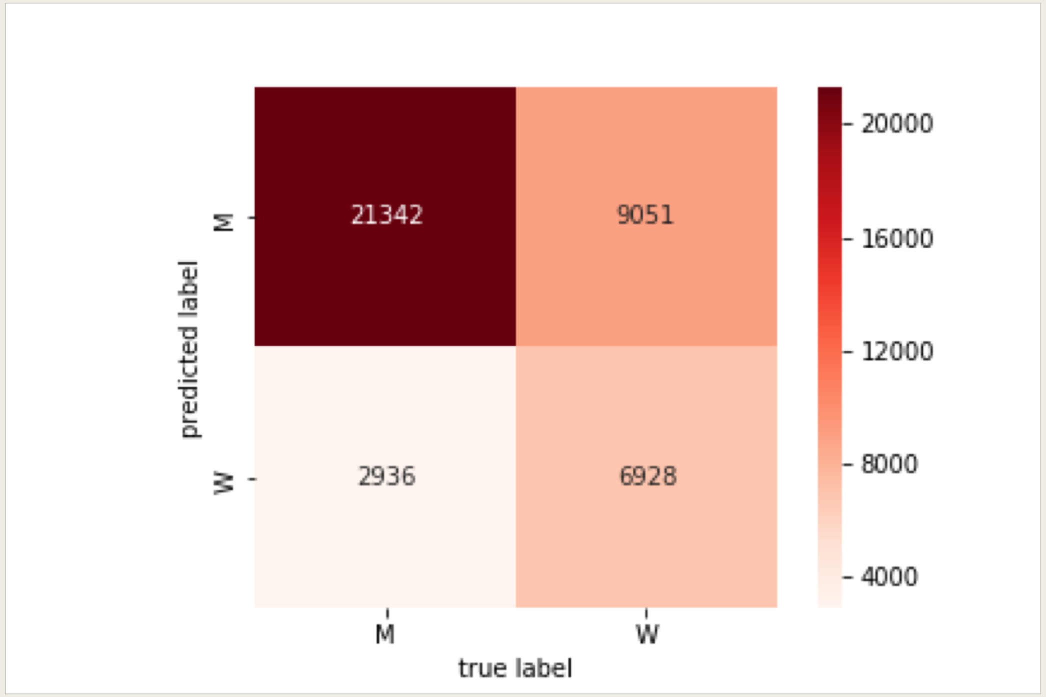
- when saying “longer”, “shorter” - this refers to the difference between the rows in the specific responder column
- Reddit is opposite of Fitocracy
  - Could this be because female posters are so much more rare?

Machine Learning

- Goals:
  - *1. Identify gender by looking at text, regardless of if poster or responder*
    - Merged all sample files of posters (always know gender) and some responders (sometimes know gender)
  - *2. Identify gender of poster and responder by looking at response text*
    - Merged Fitocracy and Reddit files when gender of both poster and responder was known and visible

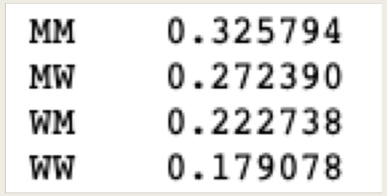
Goal 1: Simply identify gender

- Baseline: 60% male
- Used train test split, TfidfVectorizer, and MultinomialNB
  - *Using nltk's tokenizer improved accuracy*
  - *Punctuation is important?*
- Accuracy score: 70.2%



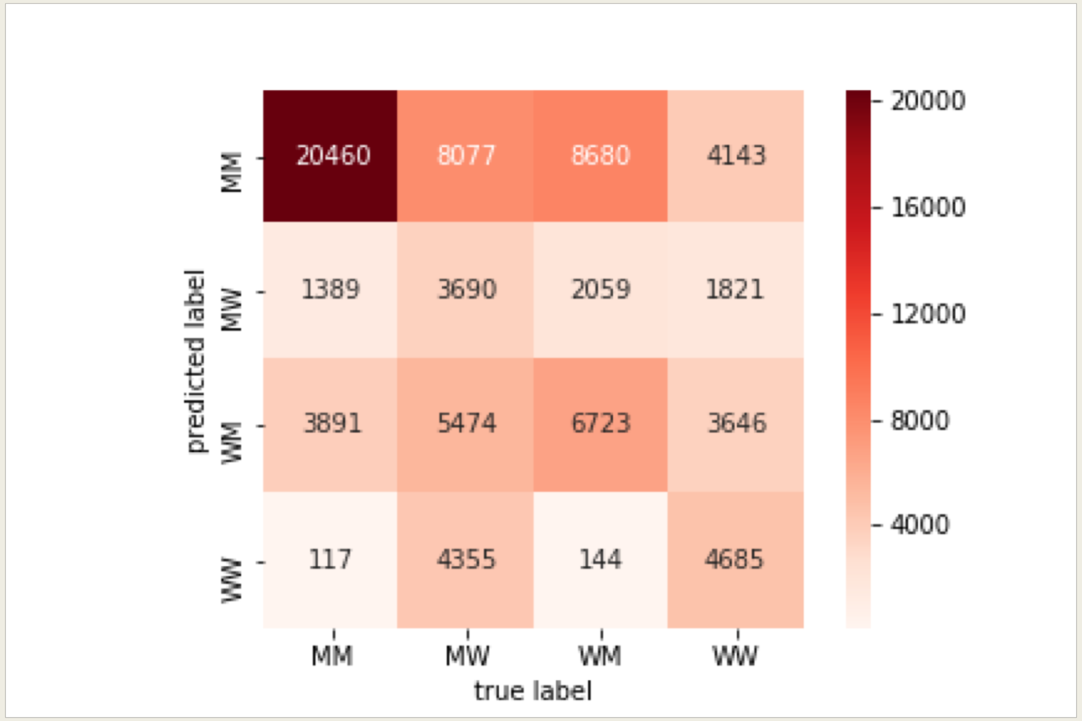
Goal 2: Identify both genders

- Created new column
  - *First letter: gender of poster*
  - *Second letter: gender of response*



```
MM     0.325794
MW     0.272390
WM     0.222738
WW     0.179078
```

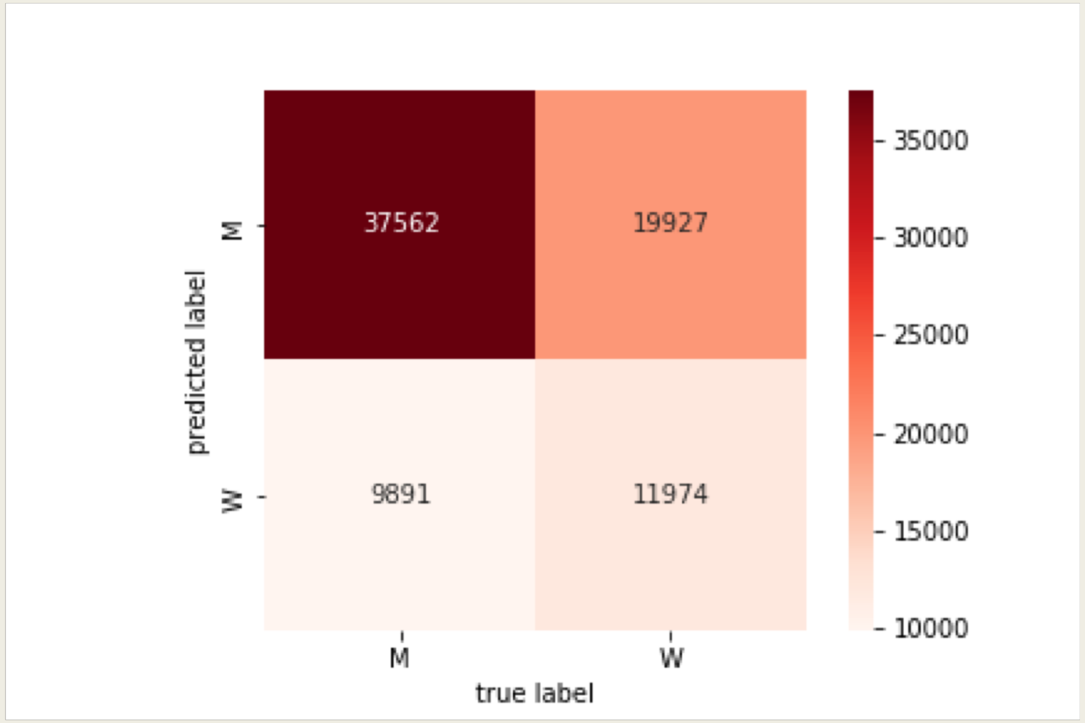
- Baseline: 32.6% male poster/male responder
- Used train test split, TfidfVectorizer, and MultinomialNB
- Accuracy score: 44.8%
- Confusion:
  - *When true label is WM (female poster, male responder):*
    - Predicted as both MM and WM
  - *When true label is WW (female poster, female responder):*
    - Predicted as both MM and WW
  - *Least accurate is MW*



	MM	MW	WM	WW
MM	20460	8077	8680	4143
MW	1389	3690	2059	1821
WM	3891	5474	6723	3646
WW	117	4355	144	4685

Goal 2.5: identify gender of poster given response

- Attempting to simplify the last task
- Baseline: 59.8% male
- Used train test split, TfidfVectorizer, and MultinomialNB
- Accuracy score: 62.4%



Improvements for the future

- Go deeper into hedge/compliment/question analysis
- Look at most informative features – why are things being classified the way that they are?
- FeatureUnion????
- Annotation for compliments and questions

THANK YOU!