What is the DEAL with

HORROR

These Days?

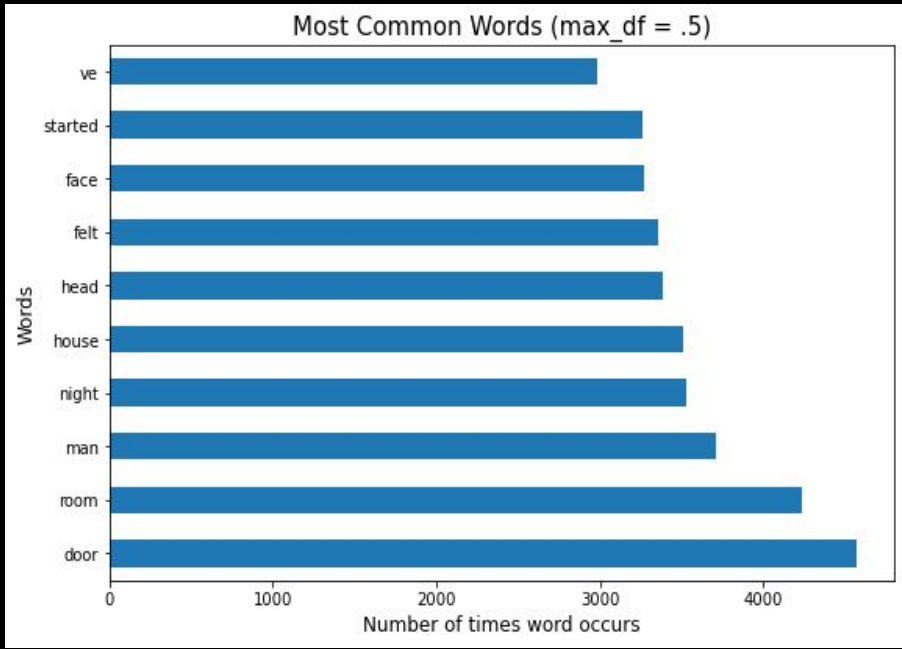
Finding the hot topics in Reddit's most popular horror subreddits

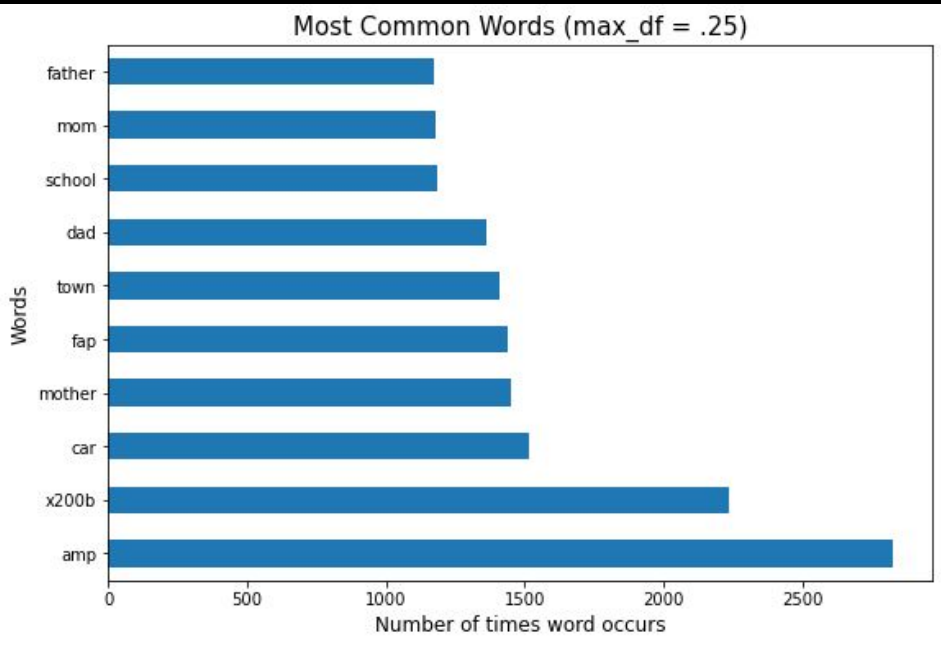
HALLOWEEN

- A horror novel publisher was looking for the hottest trending topics in new, interesting horror stories, so they hired a data scientist to scrape r/nosleep and r/creepypasta, two subreddits in which people publish original horror. stories, and find the most common themes present in those stories.
- The data scientist used PushShift API and CountVectorizer to find the most common themes between the two subreddits. Let’s see what they are...

HALLOWEEN II

- First pass with CountVectorizer - no hyperparameters, no results.
- Second and third passes (out of 10):





HALLOWEEN III: SEASON OF THE WITCH

- After Multiple passes (adjustments of CV hyperparameters), the data scientist came up with a list of “themes” to present to the publisher.

HALLOWEEN 4: THE RETURN OF MICHAEL MYERS

- During her research, the data scientist discovered that r/nosleep contains more and better original horror than r/creepypasta.
- Unfortunately, the heads of the horror publishing company have loved creepypasta since its internet infancy in the early 2000s (predating Reddit - some argue it started even earlier, in the chain emails of the 1990s). The data scientist had to convince them that creepypasta is not what it used to be, or at least, the subreddit is not as good as nosleep.
- The data scientist vowed, shaking her fists at the sky, to run some classification models that could determine whether a post is from r/nosleep or r/creepypasta, in order to prove her point that the content is different and nosleep now contains better, more original horror.

HALLOWEEN 5: THE REVENGE OF MICHAEL MYERS

The data scientist used two different vectorizers, CountVectorizer and TF-IDF, in combination with 3 classification models: Multinomial Naïve Bayes, Random Forest, and Logistic Regression. She optimized for both accuracy and sensitivity.

The baseline accuracy score was 51%.

Vectorizer/Model	Accuracy Score	Sensitivity Score
CV/MNB	69%	66%
TF-IDF/MNB	75%	84%
TF-IDF/RF	79%	91%
TF-IDF/LR	78%	86%

HALLOWEEN: THE CURSE OF MICHAEL MYERS

- Random Forest with the TF-IDF vectorizer gave the best scores for both accuracy and specificity.
- This is probably because with all of the features my data had available, Random Forest was able to handle the large dataset and overfitting very efficiently. TF-IDF was likely the best because it weighs word importance vs. word frequency.
- The data scientist was optimizing for accuracy and sensitivity: accuracy for obvious reasons (a good binary classifier is accurate), and sensitivity because r/nosleep was her positive class, and while there may be some decent original horror on r/creepypasta that could be misclassified as r/nosleep, she would not expect it to go the other way around.

... HALLOWEEN ENDS

So, the data scientist returned to the horror novel publisher with her list of trending horror topics and her strong beliefs about r/nosleep and r/creepypasta, and using her TF-IDF/Random Forest model, proved to them that the post content of the two subreddits was different, and that r/nosleep was a better place to find the trending topics in new original horror (and possibly source a few new writers for their imprint!).

But one of the publishers, the one who loved creepypasta the most, started to laugh, and ripped off his face to reveal… SLENDER MAN.

And then the lights went out.

HALLOWEEN: RESURRECTION

When the data scientist woke up the next day in a field, covered in dirt, surrounded by the bodies of the other publishers who had taken her advice, she had a moment to think about how she would continue her very important work:

- Bringing down even more posts from r/nosleep and r/creepypasta would likely improve the accuracy of a model, but it might also require a more powerful model to handle that amount of text data - Random Forest could do it, possibly through AWS on a computer with significant RAM.
- r/nosleep has 14.4 million members. Every day, there are thousands of submissions (although most are rejected). Staying on top of popular topics in horror would be as simple as pulling at least a thousand posts a week.

HALLOWEEN: QUESTIONS?

(That is not a real Halloween movie. But everything else was!)