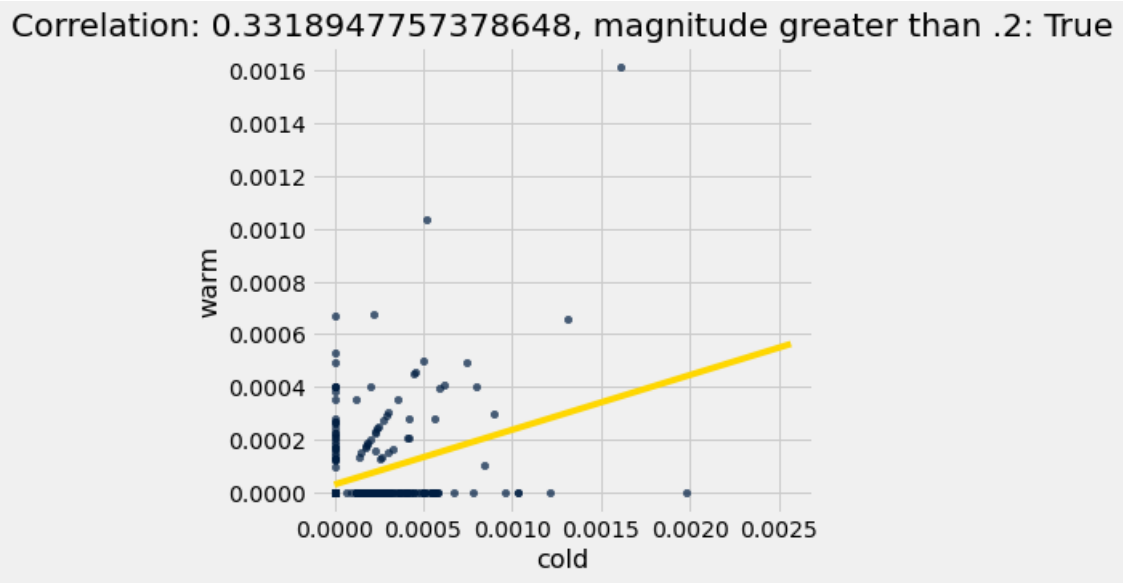
Question 1.2.2 Choose two *different* words in the dataset with a magnitude (absolute value) of correlation higher than 0.2 and plot a scatter plot with a line of best fit for them. Please do not pick "outer" and "space" or "san" and "francisco". The code to plot the scatter plot and line of best fit is given for you, you just need to calculate the correct values to `r`, `slope` and `intercept`.

Hint 1: It's easier to think of words with a positive correlation, i.e. words that are often mentioned together*. Try to think of common phrases or idioms.

Hint 2: Refer to Section 15.2 of the textbook for the formulas.

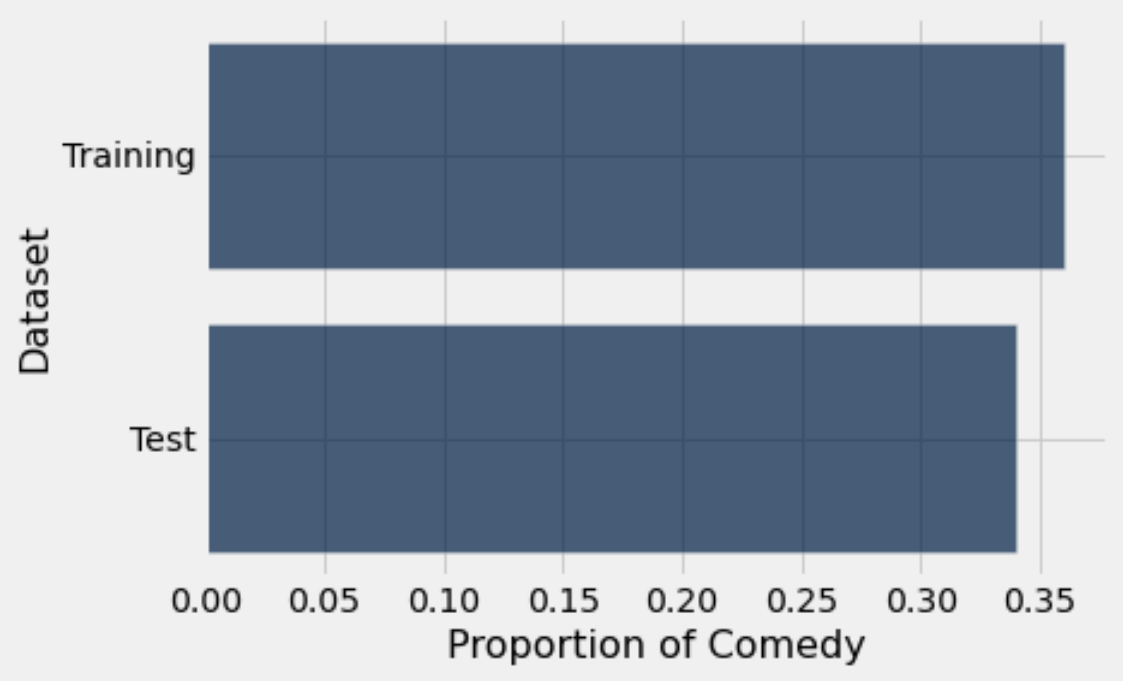
In [28]:
```
word_x = "cold" # SOLUTION
word_y = "warm" # SOLUTION

# These arrays should make your code cleaner!
arr_x = movies.column(word_x)
arr_y = movies.column(word_y)

x_su = (arr_x - np.mean(arr_x)) / np.std(arr_x) # SOLUTION
y_su = (arr_y - np.mean(arr_y)) / np.std(arr_y) # SOLUTION

r = np.mean(x_su * y_su) # SOLUTION

slope = np.std(arr_y) / np.std(arr_x) * r # SOLUTION
intercept = np.mean(arr_y) - (np.mean(arr_x) * slope) # SOLUTION

# DON'T CHANGE THESE LINES OF CODE
movies.scatter(word_x, word_y)
max_x = max(movies.column(word_x))
plots.title(f"Correlation: {r}, magnitude greater than .2: {abs(r) >= 0.2}")
plots.plot([0, max_x * 1.3], [intercept, intercept + slope * (max_x*1.3)], color='gold');
```



Question 1.3.1 Draw a horizontal bar chart with two bars that show the proportion of Comedy movies in each dataset (`train_movies` and `test_movies`). The two bars should be labeled "Training" and "Test". Complete the function `comedy_proportion` first; it should help you create the bar chart.

Hint: Refer to Section 7.1 of the textbook if you need a refresher on bar charts.

```
In [33]: # BEGIN SOLUTION NO PROMPT
         def comedy_proportion(table):
             """Return the proportion of movies in a table that have the comedy genre."""
             return table.where('Genre', are.equal_to('comedy')).num_rows / table.num_rows

         datasets = make_array('Training', 'Test')
         prop_comedy = make_array(comedy_proportion(train_movies), comedy_proportion(test_movies))
         Table().with_columns(
                 'Dataset', datasets,
                 'Proportion of Comedy', prop_comedy)\
                .barh('Dataset')
         # END SOLUTION
         """ # BEGIN PROMPT
         def comedy_proportion(table):
             # Return the proportion of movies in a table that have the comedy genre.
             ...
             return ...

         # The staff solution took multiple lines.  Start by creating a table.
         # If you get stuck, think about what sort of table you need for barh to work
         ...
         """; # END PROMPT
```



Question 3.1.7 In two sentences or less, describe how you selected your features.

Type your answer here, replacing this text.

SOLUTION: The staff features don't work very well. It's a good idea to pick words that are common in thriller movies but uncommon in comedy movies and vice-versa. These are points that are far away from the diagonal in the above plot.

Question 3.3.3

Do you see a pattern in the types of movies your classifier misclassifies? In two sentences or less, describe any patterns you see in the results or any other interesting findings from the table above. If you need some help, try looking up the movies that your classifier got wrong on Wikipedia.

Type your answer here, replacing this text.

SOLUTION: The classifier tends to misclassify movies that have both comedy and thriller elements in them.

Question 4.2

Do you see a pattern in the mistakes your new classifier makes? How good an accuracy were you able to get with your limited classifier? Did you notice an improvement from your first classifier to the second one? Describe in two sentences or less.

Hint: You may not be able to see a pattern.

Type your answer here, replacing this text.

SOLUTION: Any reasonable student solution is fine.

Question 4.3

Given the constraint of five words, how did you select those five? Describe in two sentences or less.

Type your answer here, replacing this text.

SOLUTION: You were meant to put in at least some minimal effort.