Анализ аспектов оценки коротких историй

Данные

Данные представляют собой результаты оценки коротких историй по 9 критериям. Оставим в данных только сами критерии, удалив ненужные на данный момент колонки. Критерии оценивались по разным шкалам, что отображено в саммари.

Шкала 0 - 4

overall_quality_rating – главный критерий, общая оценка качества истории.
coherency_score – последовательность и логичность событий.
fluency_score – естественность и грамматическая правильность языка.
interest_score – насколько история интересна.

Шкала 0 - 1

narrative_arc_answer – наличие нарративной арки.
common_sense_answer – соответствие истории здравому смыслу.
plot_twist_answer – наличие неожиданной смены событий.
humanlike_score – насколько история похожа на написанную человеком.

Шкала 0 - 2

share_score – желание поделиться историей с другими (даже уверенней).

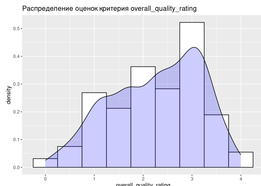
```
data <- read_csv('https://raw.githubusercontent.com/yashkens/AnDan_CompLing_Project/main/story_evaluation_scores.csv', show_col_types = FALSE)

to_delete = c('story', 'Human', 'OLL', 'Emo', 'EmoArc', 'ATD')
data = data[, !(names(data) %in% to_delete)]
summary(data, digits = 2)
```

```
##  overall_quality_rating narrative_arc_answer common_sense_answer coherency_score
##  Min.   :0.0            Min.   :0.00         Min.   :0.00         Min.   :0.0
##  1st Qu.:1.6            1st Qu.:0.20         1st Qu.:0.40         1st Qu.:2.0
##  Median :2.4            Median :0.60         Median :0.60         Median :2.8
##  Mean   :2.3            Mean   :0.58         Mean   :0.59         Mean   :2.7
##  3rd Qu.:3.0            3rd Qu.:0.80         3rd Qu.:1.00         3rd Qu.:3.4
##  Max.   :4.0            Max.   :1.00         Max.   :1.00         Max.   :4.0
##  fluency_score interest_score plot_twist_answer  share_score    humanlike_score
##  Min.   :0.0   Min.   :0.0    Min.   :0.0       Min.   :0.00   Min.   :0.00
##  1st Qu.:2.0   1st Qu.:1.6    1st Qu.:0.2       1st Qu.:0.40   1st Qu.:0.40
##  Median :2.6   Median :2.2    Median :0.4       Median :0.80   Median :0.60
##  Mean   :2.6   Mean   :2.2    Mean   :0.5       Mean   :0.89   Mean   :0.59
##  3rd Qu.:3.2   3rd Qu.:2.8    3rd Qu.:0.8       3rd Qu.:1.20   3rd Qu.:0.80
##  Max.   :4.0   Max.   :4.0    Max.   :1.0       Max.   :2.00   Max.   :1.00
```

Распределение интересующей нас переменной.

```
ggplot(data, aes(x=overall_quality_rating)) +
  geom_histogram(aes(y=after_stat(density)), binwidth=.5, colour="black", fill="white") +
  geom_density(alpha=.2, fill="blue") +
  ggtitle("Распределение оценок критерия overall_quality_rating")
```



Гипотеза

Создателям историй (как сгенерированных, так и написанных) важно понимать, какие истории нравятся читателям и почему. В этом исследовании изучим факторы, составляющие общую оценку и выясним, какой вклад вносит каждый из них.

Гипотеза – *общий критерий overall_quality_rating независим от остальных критериев, которые рассматриваются в этой работе.*
Альтернативная гипотеза – *общая оценка истории зависит от остальных критериев.*

Корреляция

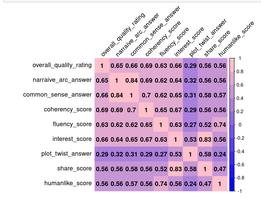
Сначала посчитаем корреляцию Пирсона между всеми критериями. *Примечательные заметки по графику:*

- *Наиболее высокая корреляция для критерия overall_quality_rating наблюдается с критерием последовательности, а также с критериями интересности и соответствия здравому смыслу.*
- *Критерий plot_twist_answer имеет самую низкую корреляцию с остальными факторами, но при этом заметно коррелирует с фактором interest_score.*

Также выделяются следующие факты:

- *Критерий share_score положительно коррелирует с interest_score.*
- *Критерий fluency_score положительно коррелирует с humanlike_score.*
- *Критерий common_sense_answer положительно коррелирует с narrative_arc_answer, что довольно сложно объяснить.*

```
corrplot(cor(data), method="color", addCoef.col = "black", tl.srt = 45, tl.col = "black", col= colorRampPalette(c("blue","pink"))(10))
```



OLS Regression

Далее изучим линейную зависимость используемого нас главного критерия overall_quality_rating со всеми остальными.

P-value < 0.05 для всех критериев, кроме двух – plot_twist_answer и share_score. Также p-value > 0.01 для критерия humanlike_score, что могло бы быть важно при более строгом порог значимости.

Гипотеза подтвердилась для plot_twist_answer и share_score – общее качество истории не зависит от этих критериев, они не имеют значительного вклада в общую оценку.

Для остальных критериев нулевая гипотеза была отвергнута и принята альтернативная. То есть общая оценка истории статистически значимо зависит от оставшихся факторов. Все обнаруженные зависимости положительные, значит, при улучшении отдельных факторов ожидается рост общего качества.

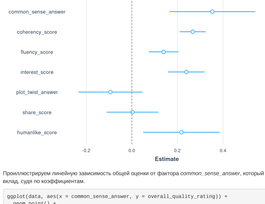
```
model <- lm(overall_quality_rating ~ ., data=data)
summary(model)
```

```
##
## Call:
## lm(formula = overall_quality_rating ~ ., data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.80390 -0.39683  0.02351  0.40799  2.09843
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.191804   0.064198   2.988 0.002865 **
## narrative_arc_answer 0.316955   0.094774   3.344 0.000849 ***
## common_sense_answer  0.353263   0.095683   3.694 0.000239 ***
## coherency_score      0.267321   0.029138   9.174 < 2e-16 ***
## fluency_score        0.139476   0.033340   4.183 3.07e-05 ***
## interest_score       0.230326   0.040846   5.639 5.92e-09 ***
## plot_twist_answer   -0.094055   0.071573  -1.314 0.189063
## share_score          0.002824   0.057906   0.049 0.961109
## humanlike_score      0.217575   0.085692   2.539 0.011234 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5882 on 1270 degrees of freedom
## Multiple R-squared:  0.5586, Adjusted R-squared:  0.556
## F-statistic: 239.7 on 8 and 1270 DF,  p-value: < 2.2e-16
```

Построим график, на котором наглядно показаны коэффициенты при каждом факторе в линейной зависимости.

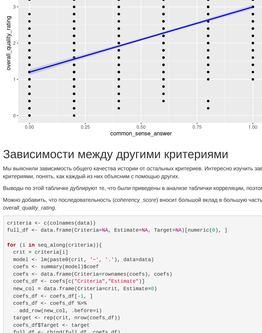
Хотя мы выяснили, что общее качество не зависит от наличия неожиданного перехода (plot twist), интересно, что коэффициент этого фактора отрицательный.

```
plot_summs(model)
```



Проиллюстрируем линейную зависимость общей оценки от фактора *common_sense_answer*, который вносит наибольший положительных вклад, судя по коэффициентам.

```
ggplot(data, aes(x = common_sense_answer, y = overall_quality_rating)) +
  geom_point() +
  stat_smooth(method = "lm", formula = y ~ x, col = "blue") +
  ggtitle("Linear Regression")
```



Зависимости между другими критериями

Мы выяснили зависимость общего качества истории от остальных критериев. Интересно изучить зависимости между другими критериями, понять, как каждый из них объясним с помощью других.

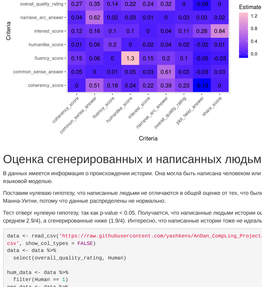
Выводы по этой таблице дублируют те, что были приведены в анализе таблички корреляции, поэтому я их опущу.

Можно добавить, что последовательность (coherency_score) вносит большой вклад в большую часть критериев, как и overall_quality_rating.

```
criteria <- c(colnames(data))
full_df <- data.frame(Criteria=NA, Estimate=NA, Target=NA)[numeric(0), ]

for (i in seq_along(criteria)){
  crit = criteria[i]
  model <- lm(paste0(crit, '~', '.'), data=data)
  coefs <- summary(model)$coef
  coefs <- data.frame(Criteria=rownames(coefs), coefs)
  coefs_df <- coefs[c("Criteria","Estimate")]
  new_col = data.frame(Criteria=crit, Estimate=0)
  coefs_df <- coefs_df[1,]
  coefs_df <- coefs_df %>%
    add_row(new_col, .before=1)
  target <- rep(crit, nrow(coefs_df))
  coefs_df$Target <- target
  full_df <- rbind(full_df, coefs_df)
}

ggplot(full_df, aes(Criteria, Target)) +
  geom_tile(aes(fill = Estimate)) +
  geom_text(aes(label = round(Estimate, 2))) +
  scale_fill_gradient(low = "blue", high = "pink") +
  theme(axis.text.x = element_text(angle = 45, hjust=1)) +
  ggtitle("Regression Coefficients") +
  xlab("Criteria") + ylab("Criteria")
```



Оценка сгенерированных и написанных людьми историй

В данных имеется информация о происхождении истории. Она могла быть написана человеком или могла быть сгенерированная языковой моделью.

Поставим нулевую гипотезу, что написанные людьми не отличаются в общей оценке от тех, что были сгенерированы. Проведем тест Манна-Уитни, потому что данные распределены не нормально.

Тест отверг нулевую гипотезу, так как p-value < 0.05. Получается, что написанные людьми истории оцениваются значительно выше (в среднем 2.9/4), а сгенерированные ниже (1.9/4). Интересно, что написанные истории тоже не идеальны.

```
data <- read_csv('https://raw.githubusercontent.com/yashkens/AnDan_CompLing_Project/main/story_evaluation_scores.csv', show_col_types = FALSE)
data <- data %>%
  select(overall_quality_rating, Human)

hum_data <- data %>%
  filter(Human == 1)
gen_data <- data %>%
  filter(Human == 0)

wilcox.test(hum_data$overall_quality_rating, gen_data$overall_quality_rating, paired = FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  hum_data$overall_quality_rating and gen_data$overall_quality_rating
## W = 339496, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
mean(hum_data$overall_quality_rating)
```

```
## [1] 2.870056
```

```
mean(gen_data$overall_quality_rating)
```

```
## [1] 1.82754
```