Words have meaning: initial descriptive language choice and startup success

Zachary Hayes, Justin Liu, Mike McCormick

2022-04-18

Introduction

text text text

Motivation

text text text

Background

text text text

Hypothesis

text text text

Data

Y Combinator startup company descriptions

text text text

Figure 1 - histogram of words per company description, pre-processing

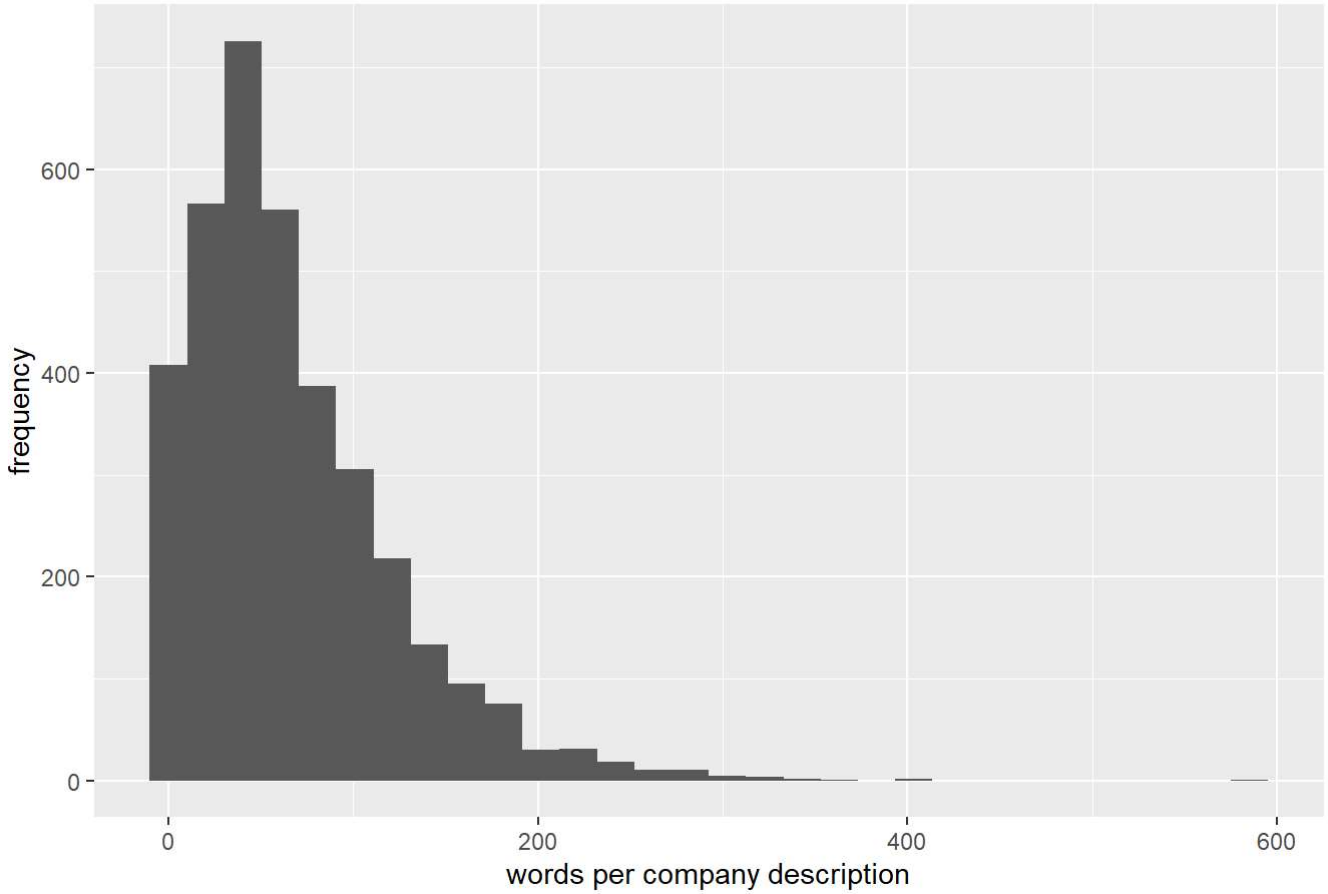
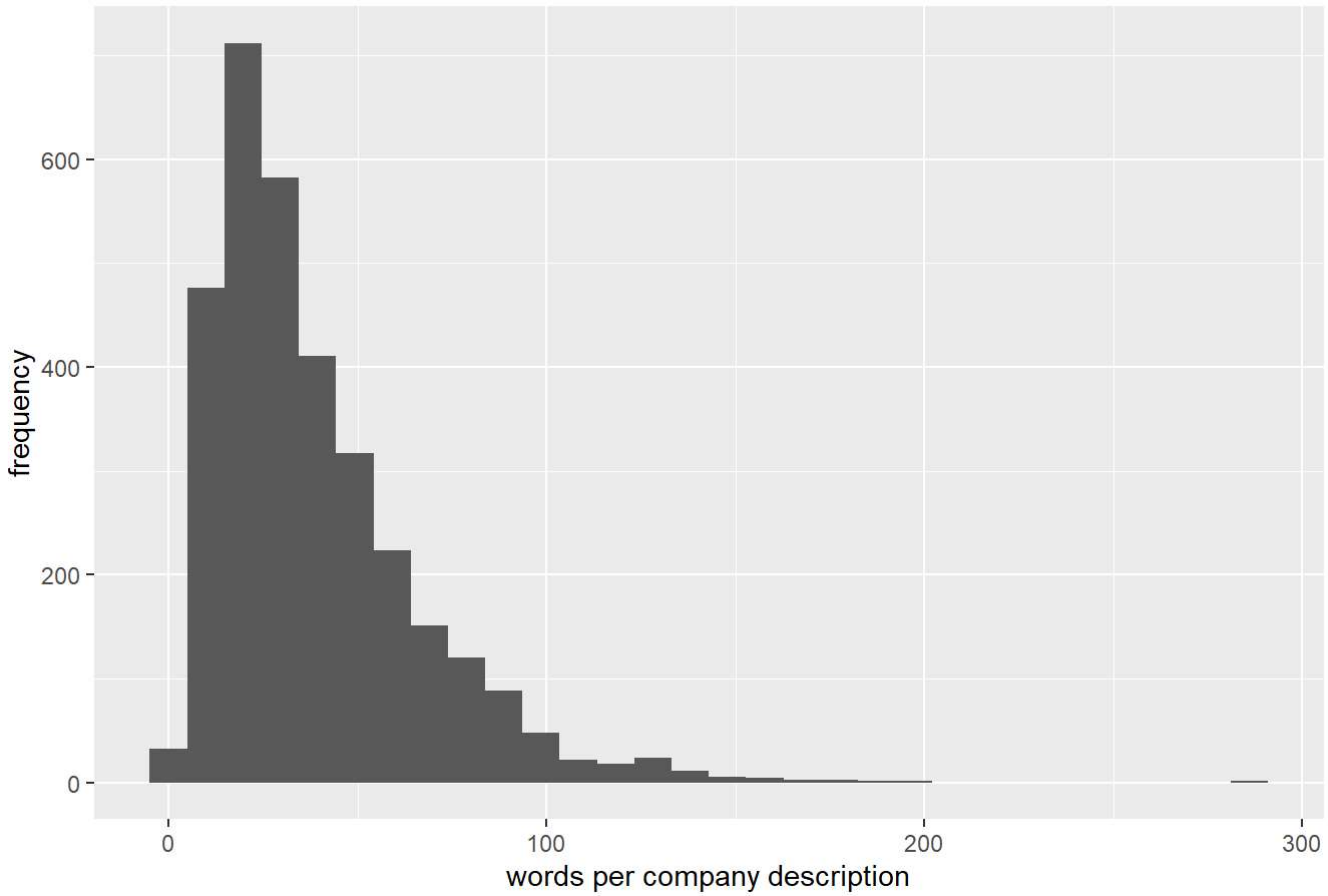


Figure 2 - histogram of words per company description, post-processing



Open source textbooks

text text text

Table 1 - textual data summary statistics

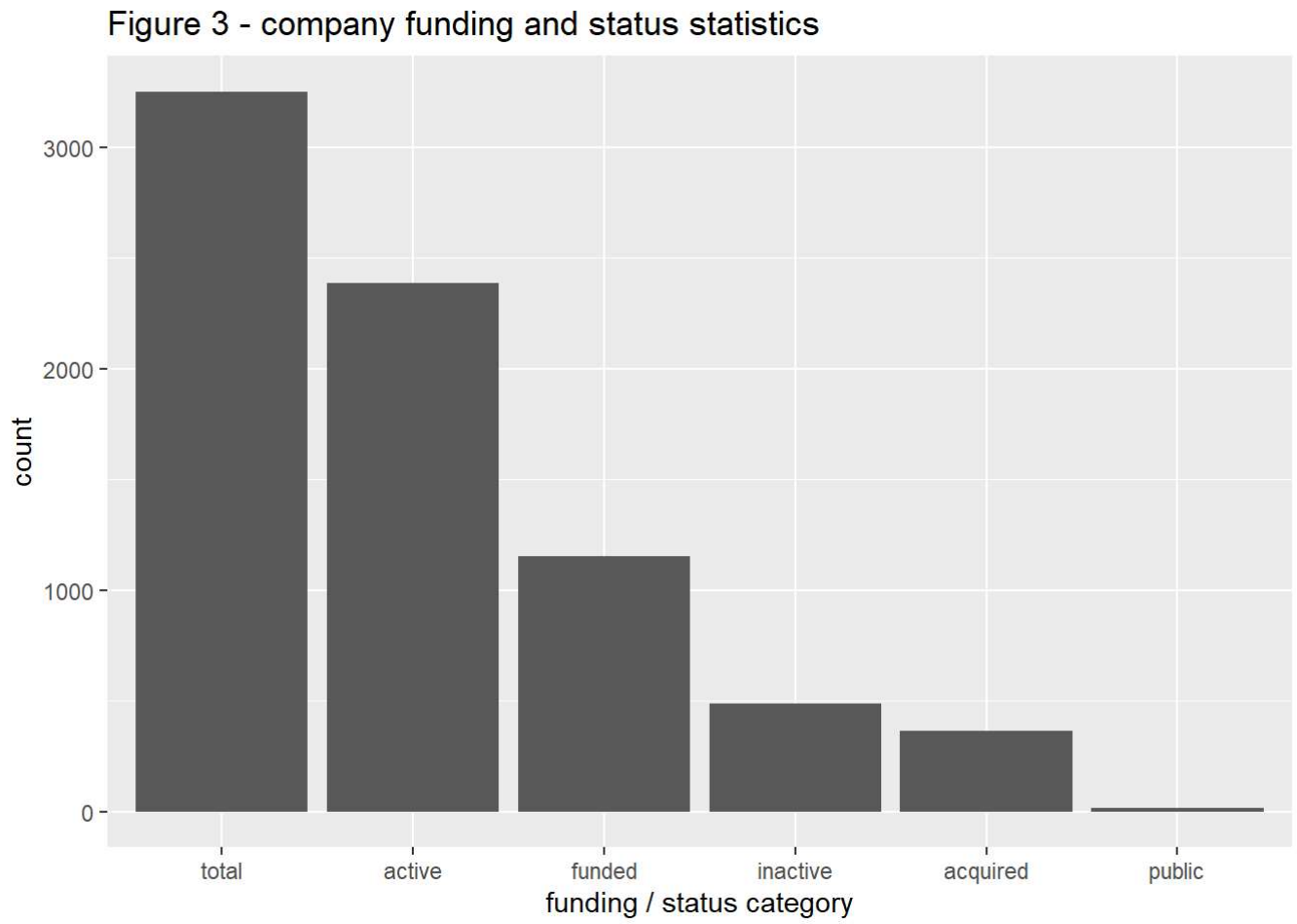
data	document count	words per document						
		minimum	lower quartile	mean	median	upper quartile	maximum	standard deviation
companies, pre-processing	3586	1	28	66.45901	53	93	586	55.06245
companies, post-processing	3251	1	19	38.47155	31	52	287	27.17250
textbooks, pre-processing	5	344	2039	4883.00000	6241	7499	8292	3499.99136
textbooks, post-processing	5	181	1067	2435.00000	3153	3724	4050	1712.98059

Crunchbase startup funding data

text text text

Table 2 - funding and exit data for Y Combinator companies

status	count
total	3251
active	2386
funded	1153
inactive	488
acquired	362
public	15



Methodology and results

Methodology

text text text

Data processing

text text text

Similarity scores

text text text

Table 3 - Mean cosine and similarity scores by company status

		Mean cosine similarity scores				
status	count	entrepreneurship	finance	leadership	marketing	strategy
Acquired	362	0.0738666	0.0390857	0.0335410	0.0272614	0.0498762
Active	2386	0.0784647	0.0460266	0.0317776	0.0335066	0.0526428
Inactive	488	0.0651645	0.0323107	0.0277279	0.0272202	0.0461308
Public	15	0.1074728	0.0429126	0.0502199	0.0352015	0.0725202

		Mean Jaccard similarity scores				
status	count	entrepreneurship	finance	leadership	marketing	strategy
Acquired	362	0.0001355	0.0001355	0.0001355	0.0001304	0.0001354

		Mean Jaccard similarity scores				
status	count	entrepreneurship	finance	leadership	marketing	strategy
Active	2386	0.0001301	0.0001301	0.0001301	0.0001265	0.0001301
Inactive	488	0.0001275	0.0001275	0.0001275	0.0001231	0.0001275
Public	15	0.0001492	0.0001492	0.0001492	0.0001420	0.0001492

Data integration and regression analysis

Results

Regression analysis table

```
## Warning: Model matrix is rank deficient. Parameters fin_j, ldr_j were not
## estimable.

## Warning: Model matrix is rank deficient. Parameters fin_j, ldr_j were not
## estimable.
```

		active			**funded**	
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	2.19	1.65 – 2.93	**<0.001**	0.53	0.40 – 0.69	**<0.001**
ent c	6.17	0.33 – 115.48	0.223	0.14	0.01 – 1.90	0.139
fin c	2290.84	125.16 – 46566.83	**<0.001**	0.03	0.00 – 0.31	**0.004**
ldr c	3.36	0.18 – 68.71	0.426	28.08	1.95 – 401.77	**0.014**
mkt c	1.07	0.15 – 8.32	0.949	0.16	0.03 – 0.98	**0.049**
str c	0.07	0.00 – 3.27	0.169	48.55	1.43 – 1642.02	**0.031**
ent j	0.00	0.00 – Inf	0.472	Inf	0.00 – Inf	0.662
mkt j	Inf	Inf – Inf	**0.029**	0.00	0.00 – Inf	0.221
str j	Inf	0.00 – Inf	0.626	0.00	0.00 – Inf	0.757
Observations	3251			3251		
R^2 Tjur	0.017			0.011		

```
##
## Call:
## glm(formula = active ~ ent_c + fin_c + ldr_c + mkt_c + str_c +
##     ent_j + fin_j + ldr_j + mkt_j + str_j, family = binomial(link = "logit"),
##     data = data_joined)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2553  -1.4558   0.7458   0.8316   1.2328
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.859e-01  1.467e-01   5.358 8.40e-08 ***
## ent_c        1.819e+00  1.492e+00   1.220   0.2226
## fin_c        7.737e+00  1.509e+00   5.126 2.96e-07 ***
## ldr_c        1.212e+00  1.521e+00   0.797   0.4256
## mkt_c        6.629e-02  1.031e+00   0.064   0.9487
## str_c       -2.718e+00  1.977e+00  -1.375   0.1692
## ent_j       -2.880e+04  4.001e+04  -0.720   0.4716
## fin_j               NA         NA      NA       NA
## ldr_j               NA         NA      NA       NA
## mkt_j        8.670e+03  3.961e+03   2.189   0.0286 *
## str_j        1.949e+04  4.005e+04   0.487   0.6265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3766.7  on 3250  degrees of freedom
## Residual deviance: 3706.3  on 3242  degrees of freedom
## AIC: 3724.3
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## glm(formula = funded ~ ent_c + fin_c + ldr_c + mkt_c + str_c +
##     ent_j + fin_j + ldr_j + mkt_j + str_j, family = binomial(link = "logit"),
##     data = data_joined)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2841  -0.9546  -0.8793   1.3906   1.9305
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.388e-01  1.358e-01  -4.704 2.56e-06 ***
## ent_c       -1.989e+00  1.344e+00  -1.480  0.13900
## fin_c       -3.620e+00  1.255e+00  -2.885  0.00392 **
## ldr_c        3.335e+00  1.358e+00   2.455  0.01409 *
## mkt_c       -1.803e+00  9.175e-01  -1.965  0.04944 *
## str_c        3.882e+00  1.796e+00   2.161  0.03068 *
## ent_j        1.748e+04  3.994e+04   0.438  0.66165
## fin_j               NA         NA      NA       NA
## ldr_j               NA         NA      NA       NA
## mkt_j       -4.547e+03  3.719e+03  -1.223  0.22150
## str_j       -1.236e+04  3.998e+04  -0.309  0.75720
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4228.1  on 3250  degrees of freedom
## Residual deviance: 4191.1  on 3242  degrees of freedom
## AIC: 4209.1
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## glm(formula = inactive ~ ent_c + fin_c + ldr_c + mkt_c + str_c +
##     ent_j + fin_j + ldr_j + mkt_j + str_j, family = binomial(link = "logit"),
##     data = data_joined)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8663  -0.6236  -0.5447  -0.4264   2.6329
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.105e+00  1.799e-01  -6.140 8.26e-10 ***
## ent_c       -3.475e+00  1.885e+00  -1.843   0.0653 .
## fin_c       -9.912e+00  2.048e+00  -4.840 1.30e-06 ***
## ldr_c       -4.911e+00  2.034e+00  -2.414   0.0158 *
## mkt_c        1.475e+00  1.302e+00   1.133   0.2571
## str_c        4.474e+00  2.435e+00   1.837   0.0662 .
## ent_j       -3.001e+05  6.435e+06  -0.047   0.9628
## fin_j               NA         NA      NA       NA
## ldr_j               NA         NA      NA       NA
## mkt_j       -1.035e+04  4.934e+03  -2.098   0.0359 *
## str_j        3.092e+05  6.435e+06   0.048   0.9617
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2749.7  on 3250  degrees of freedom
## Residual deviance: 2684.3  on 3242  degrees of freedom
## AIC: 2702.3
##
## Number of Fisher Scoring iterations: 11
```

```
##
## Call:
## glm(formula = acquired ~ ent_c + fin_c + ldr_c + mkt_c + str_c +
##     ent_j + fin_j + ldr_j + mkt_j + str_j, family = binomial(link = "logit"),
##     data = data_joined)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1707  -0.5033  -0.4761  -0.4414   2.4464
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.360e+00  2.073e-01 -11.384   <2e-16 ***
## ent_c        2.489e-01  2.075e+00   0.120    0.905
## fin_c       -2.911e+00  1.977e+00  -1.473    0.141
## ldr_c        2.539e+00  1.995e+00   1.273    0.203
## mkt_c       -1.824e+00  1.481e+00  -1.232    0.218
## str_c       -4.095e-01  2.790e+00  -0.147    0.883
## ent_j        5.328e+04  4.003e+04   1.331    0.183
## fin_j               NA         NA      NA       NA
## ldr_j               NA         NA      NA       NA
## mkt_j       -3.208e+03  5.282e+03  -0.607    0.544
## str_j       -4.735e+04  4.009e+04  -1.181    0.238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2271.3  on 3250  degrees of freedom
## Residual deviance: 2254.3  on 3242  degrees of freedom
## AIC: 2272.3
##
## Number of Fisher Scoring iterations: 5
```

```
##
## Call:
## glm(formula = public ~ ent_c + fin_c + ldr_c + mkt_c + str_c +
##     ent_j + fin_j + ldr_j + mkt_j + str_j, family = binomial(link = "logit"),
##     data = data_joined)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4201  -0.1017  -0.0792  -0.0634   3.6359
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.089e+00  1.030e+00  -6.886 5.72e-12 ***
## ent_c        7.981e+00  8.159e+00   0.978   0.3280
## fin_c       -1.277e+01  1.039e+01  -1.230   0.2187
## ldr_c        1.205e+01  6.484e+00   1.859   0.0631 .
## mkt_c       -5.819e-01  6.241e+00  -0.093   0.9257
## str_c        3.986e+00  1.189e+01   0.335   0.7375
## ent_j       -2.834e+05  2.897e+07  -0.010   0.9922
## fin_j               NA         NA      NA       NA
## ldr_j               NA         NA      NA       NA
## mkt_j       -2.133e+03  2.128e+04  -0.100   0.9202
## str_j        2.915e+05  2.897e+07   0.010   0.9920
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 191.29  on 3250  degrees of freedom
## Residual deviance: 180.33  on 3242  degrees of freedom
## AIC: 198.33
##
## Number of Fisher Scoring iterations: 14
```

Conclusion

References

R libraries

Data sources

Company descriptions

Textbooks

Entrepeneurship: openstax, Entrepeneurship Finance: Robert C. Higgins, Analysis for Financial Management, 10th Edition Leadership: Richard L. Daft, The Leadership Experience Marketing: Introducing Marketing, Open Textbook Library Strategy: Strategic Management, VA Tech

Funding data

Crunchbase funding data provided by Professor Katie Moon