Trends and Patterns of TED Talks

Jingya Huang, Yuezhou Sun, Wenlong Zhao, Zhaokai Xu
Univerisity of California, San Diego
{jih201, yus174, wez094, zhx121}@ucsd.edu

Abstract

This project is a final project for the Fall 2018 course COGS 109 Modeling and Data Analysis at UC San Diego. In this project, we are interested in study the factors that contribute to the popular culture. In recent decades, TED talks become a popular platform to share ideas across fields, so we explored the TED Talks dataset[1]. The specific tasks we conducted are the following: (a) apply linear regression with L1 and L2 regularization to predict the number of views with the popular tags and duration of the videos. (b) We apply logistic regression and support vector machines with cross-validation to classify video themes with the ratings. (c) We apply word2vec to embed transcripts as vectors and apply logistic regression to predict ratings. Logistic regression demonstrates strong ability on all the three tasks; in particular, regularization methods further improve the results. SVMs in the second task and the embedding method in the third task fail to show meaningful results.

[1] Dataset is available on Kaggle https://www.kaggle.com/rounakbanik/ted-talks

1 Introduction

This project is an attempt of our group to give a comprehensive report on our command of the knowledge that we have learned from the Fall 2018 course COGS 109 Modeling and Data Analysis at UC San Diego. During the course, we have studied a number of classical techniques for data modeling and analysis, which extends from linear and non-linear regression to classification methods such as logistic regression and SVMs. We have also learned model selection techniques, including cross-validation, as well as feature selection techniques which allow us to find the best predictors and conduct the most accurate prediction in regression and classification tasks. Further, we have learned about how model bias and variance need to be balanced, and how regularization methods can help us avoid the models that overfit the training data due to capturing noise and thus perform bad on testing data. In addition, we have learned feature dimension reduction methods such as PCA.

We applied all the methods mentioned above in our project to investigate an interesting dataset: the TED Talks dataset. This dataset contains a number of features for 2550 TED Talks. While the features does not demonstrate any obvious patterns, they certainly provide a great amount of information which might be correlated. We are thus interested in applying data modeling and analysis techniques to figure out whether certain features may have correlations or predictive relations. We conducted regression and classification tasks which all aim at using certain features to predict other features. Worth mentioning is that, since we find that the transcripts of the TED Talks --- the actual contents of the TED Talks --- have a great chance to be correlated with the features of TED Talks, we experimented with some text embedding techniques to turn the transcript into computable data.

We list below the detailed scientific questions that we have investigated and the corresponding hypotheses that we have examined. (1) We apply linear regression with L1 and L2 regularization to predict the number of views with the popular tags and duration of the videos. (2) We apply logistic regression and support vector machines with cross-validation to classify video themes with the ratings. (3) We apply word2vec to embed transcripts as vectors and apply logistic regression to predict ratings.

The first two tasks are fairly straight forward. The third task uses vector embeddings of each talk's transcript to predict whether the talk is rated positive or negative by its viewers. We assume that transcripts carry most of the information within each talk such as its topics and sentiments, and that viewers' reaction to the talk can be predicted from the talk's information. We make the hypothesis that transcript-based models can predict a talk's rating, at least better than using random guess.

2 Materials and Methods

2.1 Dataset

We experiment with the TED Talks dataset[2] available from Kaggle. The TED Talks dataset contains 2550 data, each with 17 attributes, including numerical attributes such as date posted, video duration and ratings, nominal data such as tags, and textual data such as name and transcripts. See description of features that we used below:

[2] https://www.kaggle.com/rounakbanik/ted-talks

The “duration” attribute originally contains the duration of the talk in seconds.
The “tags” attribute originally contains the themes associated with the talk.
The “views” attribute originally contains the number of views on the talk.

The “ratings” attribute originally contains 14 tags, each with a count. For example, a video may have 4000 counts under the rating tag “beautiful”, which indicates that 4000 viewers have rated the video as beautiful. In this task, we modify the ratings into a binary target. Talks which have “obnoxious”, “confusing”, “unconvincing”, or “longwinded” in its 7 most rated tags are classified as “negative” and are labeled by 1. Those with only positive ratings in its top 7 tags are classified as “positive” and are labeled 0. After the preprocessing, we have 850 out of 2464 data points tagged 1 as negative.

The “transcripts” attribute originally is a string of text with average length of XX. All transcripts together contain XX unique words, disregard of cases. Here, we tokenize each talk’s transcript with the tokenizing function implemented by nltk. Further compression of this atomic data into numerical is a subject of this task, and will be explained in the Analysis Approach section.

In our regression experiments, we use the most common “ratings” on a TED talk and the “duration” of the TED talk as predictors to predict the number of “views” received by videos. In the classification experiments, we start with using “ratings” to predict “tags”, and go on to use “transcripts” and “ratings”.

2.2 Methods

2.2.1 Linear Regression with Regularization

Linear regression is a basic linear statistical model to fit the relationship between a dependent variable and other independent variables. The model with one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. A naive solution to model the data is using linear regression so that we can see a clear pattern if the two variable has a strong relationship.

2.2.2

2.2.2.1 Logistic Regression

Logistic regression provide relatively robust prediction for binary classification problems. In our study, we are interested in two classification related questions. First, weather theme tag could be predicted by ratings. Second, whether the positivity of the rating could be predicted by the transcript. In order to answer those two classification questions, we pre-process the given feature and obtain relatively well prediction result. Indeed, we perform specific analysis through statistics of the model to interpret the relevance of features.

2.2.2.2 Support Vector Machine

Similar to Logistic Regression, Support Vector Machine is a common practice of classification task which usually have state-of-art performance. In this study, we use it as a comparative model to the Logistic regression.

2.2.2.3 Cross-Validation

By applying cross-validation in the training process, we run our model both in training and validation dataset. If both the training and validation accuracy are high, the model we construct is valid such that it will neither overfit nor underfit.

2.2.3

In this task, we experiment with different embeddings of the transcript, which is to be used as the predictor. We fit a simple logistic regression model implemented by nltk with each embedding and compare their behaviors.

The embedding methods we attempted are as following:

2.2.3.1 Statistical model

In this model we embed each transcript with a 2464 * XX matrix, with each row corresponding to a datapoint and each column corresponding to a word in the vocabulary. The entry on the i-th row, j-th column denotes the j-th word's number of occurrences in the i-th talk. For dimension reduction, we attempt four methods:

a) Removing most frequent words

The most frequent words, such as prepositions and pronouns, may not convey useful meaning. We start with a full-span model and remove columns by vocabulary frequency in descending order to observe changes in train and test losses.

b) Removing least frequent words

The least frequent words, which only appear one or twice in the entire corpus, are not generalizable onto unseen data. We start with a full-span model and remove columns by vocabulary frequency in ascending order to observe change in train and test losses.

c) Removing words with extreme frequencies from both ends

We combine (a) and (b) to observe if words of medium frequency are more informative. The frequency thresholds are determined by vocabulary distribution.

d) Reducing dimension by fitting a PCA model

We directly run a PCA on the statistical matrix to extract information.

2.2.3.2 Word2vec semantic model

Word2vec is a neural network that automatically learns a vocabulary's semantic features from its corpus, and thus projects each word into a vector space.

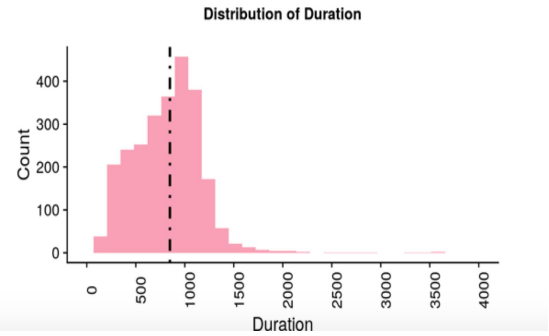
Here, we select embedding dimensions to be 10 and 100. We sum over the vector representation of every word in a transcript and take their average to be the transcript's embedding. We also experiment on leaving out the most or least frequent words.

3 Results

3.1 Regression: predict the number of views to watch a TED Talk

Scientific Question: How can we predict the number of viewers when a new TED talk is published?

Hypothesis: People are interested in watching a TED talk that has very popular tag and the duration of the TED talk is just fit.



Based on the plots of duration distribution, it is seen that the median value is less than 1000 seconds (16 minutes), and the data seem to be evenly distributed. So we are interesting in exploring the correlation between the duration of a TED talk and the number of views. By assumption, if a TED talk has a proper time duration, the number of viewers will grow, and neither a long talk nor a short talk is good for users. Besides, the tag assigned to a talk is another important feature that can be used to predict the number of views because popular tags should attract more audient.

By applying several regression models (linear, rigid, lasso) learned in Cogs 109, we feed the duration and the most popular tag as two features into the models, and predict the number of viewers of a TED talk. Before starting the experiment, we shuffle the data and separate the 2,550 sized dataset into 2,000 training dataset and 550 testing dataset. Since there is no requirement of turning parameters for the above regression models, we do not create validation set. Since the predicted view number will not be exactly same as the actual view number, to fairly evaluate the models, we define the prediction to true if the predicted value is in the range of **± 50%** of the actual view number, otherwise, we predict false.

3.1 Linear regression model

We obtained the accuracy of 84.72% with 466 correct prediction and 84 incorrect prediction.

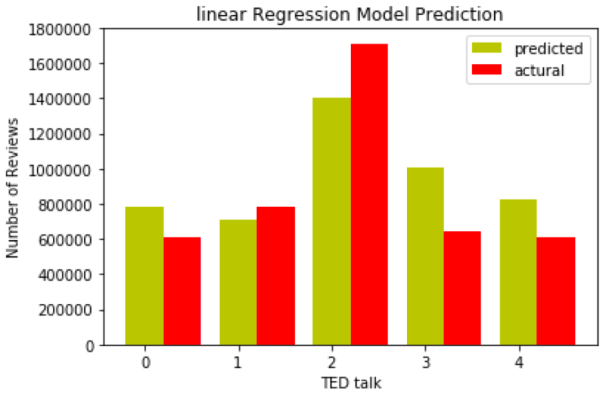
To visualized the prediction result, we randomly select five data from the test dataset, and then plot the actual view numbers and the predicted view numbers as below.

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.247
Model:	OLS	**Adj. R-squared:**	0.246
Method:	Least Squares	**F-statistic:**	327.1
Date:	Tue, 04 Dec 2018	**Prob (F-statistic):**	1.35e-123
Time:	22:30:17	**Log-Likelihood:**	-31965.
No. Observations:	2000	**AIC:**	6.394e+04
Df Residuals:	1997	**BIC:**	6.395e+04
Df Model:	2		
Covariance Type:	nonrobust		

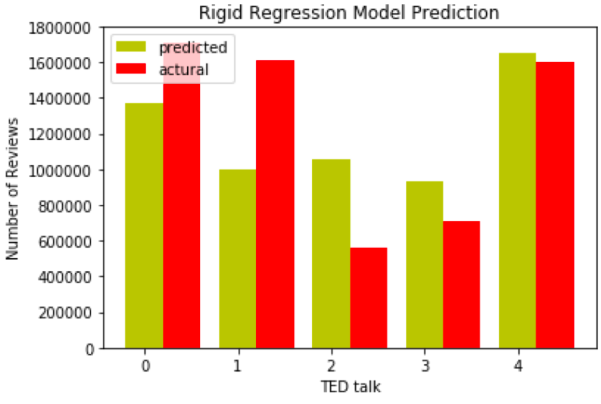
	coef	std err	t	P>|t|	[0.025	0.975]
const	1.064e+06	1.15e+05	9.278	0.000	8.39e+05	1.29e+06
x1	-221.4032	124.869	-1.773	0.076	-466.290	23.484
x2	4158.1224	162.888	25.528	0.000	3838.675	4477.570

Omnibus:	2015.736	**Durbin-Watson:**	2.010
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	269049.970
Skew:	4.506	**Prob(JB):**	0.00
Kurtosis:	59.101	**Cond. No.**	2.27e+03



3.2 Rigid regression model

We obtained the accuracy of 87.81% with 483 correct prediction and 67 incorrect prediction, which is a



slightly improvement compared to the linear regression model.

Similarly, we randomly select five data from the test dataset and plot the prediction result.

3.3 Lasso regression model

In the experiment, we are surprised to see a very closed result obtained from lasso regression model compared to rigid regression model, which shows an accuracy of 87.81% as well.

3.4 Linear model comparison

From the above results, we can conclude that rigid/lasso regression is comparably better than linear regression to predict the number of viewers. As rigid regression and lasso regression applied the L1 and L2 distance for regularization, the model is more optimized to fit the train data.

3.2 Classification: perception of Tech and Culture Talk

Scientific Question: Do people perceive talks related to each theme differently?

Hypothesis: The theme tag of a talk may be predicted by the ratings given by viewers .

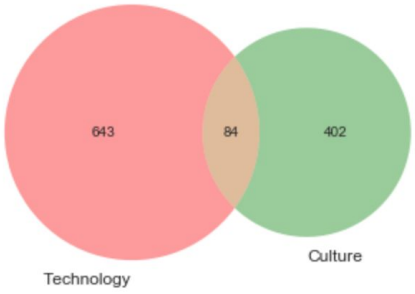
Ted talk with its slogan "ideas worth spreading" is known to inspire people across fields. Here we are interested in how strong those difference are reflected in viewer's perception. In particular, we pick two most popular and intuitively disjoint fields: Technology and Culture (Among 2550 talks in the data set, there are 727 tagged as Technology and 486 tagged as Culture).

Theme	Technology	Science	Global Issue	Culture	TEDx	Design
Occurrence	727	567	501	486	450	418

Table: number of occurrences for top 6 most popular themes

There are wide-range of themes (461 themes in total) covered by Ted Talks and a lot of them have multiple theme tags. Therefore, we subsample the data to only include talks either tagged as 'Technology' or 'Culture' and split them into Training (60%), Validation (30%) and Testing (10%) Set. For each classification task, we compare two models (Logistic regression and Support Vector Machine) and tune the regularization parameter C with 10-fold cross-validation.

We study that the number of tags for each of the 14 perception ratings[3] in terms of their ability to predict whether the talk is technology or culture related. As shown in the Venn diagram to the right, the talks under Technology and Culture are not separable that some talk



[3] ['Longwinded', 'Persuasive', 'Inspiring', 'OK', 'Beautiful', 'Unconvincing', 'Confusing', 'Funny', 'Jaw-dropping', 'Ingenious', 'Courageous', 'Informative']

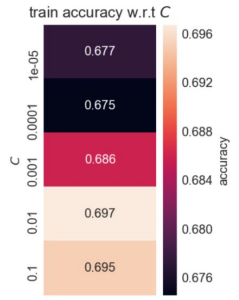
are tagged as both Technology and Culture. In order to handle the intersection between the two categories, we separate the classification to two one label classification.

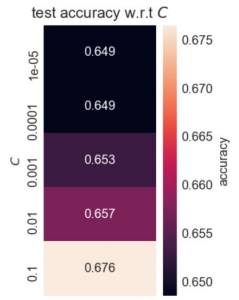
3.2.1 Technology vs. Non-Technology

3.2.1.1 Logistic Regression

Optimal regularizer parameter: C = 0.1
Testing Accuracy: 0.867256637168; True positive: 98 / 133, True negative: 0 / 133

Significant Predictors (p < 0.05)
['Funny', 'Courageous', 'Ingenious', 'Obnoxious', 'Unconvincing', 'Ok', 'Persuasive', 'Longwinded']





Logit Regression Results

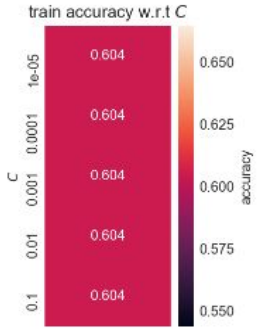
Dep. Variable:	y	No. Observations:	677
Model:	Logit	Df Residuals:	662
Method:	MLE	Df Model:	14
Date:	Tue, 04 Dec 2018	Pseudo R-squ.:	0.1285
Time:	20:52:13	Log-Likelihood:	-396.06
converged:	True	LL-Null:	-454.47
		LLR p-value:	2.637e-18

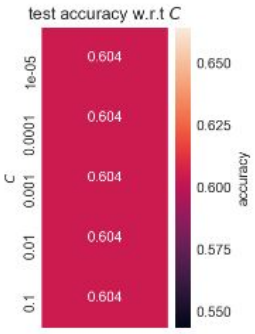
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.9055	0.140	6.462	0.000	0.631	1.180
Longwinded	-0.0013	0.000	-3.355	0.001	-0.002	-0.001
Persuasive	0.0025	0.003	0.807	0.420	-0.004	0.009
Inspiring	-8.718e-07	0.000	-0.003	0.997	-0.001	0.001
OK	-0.0008	0.000	-2.312	0.021	-0.001	-0.000
Beautiful	0.0026	0.001	4.697	0.000	0.002	0.004
Fascinating	0.0021	0.001	2.416	0.016	0.000	0.004
Obnoxious	-0.0010	0.000	-2.119	0.034	-0.002	-7.73e-05
Unconvincing	-0.0039	0.001	-3.261	0.001	-0.006	-0.002
Confusing	-0.0009	0.000	-2.061	0.039	-0.002	-4.33e-05
Funny	-0.0001	0.000	-0.650	0.516	-0.001	0.000
Jaw-dropping	-0.0020	0.001	-3.944	0.000	-0.003	-0.001
Ingenious	-0.0002	0.000	-1.103	0.270	-0.000	0.000
Courageous	0.0003	0.000	0.868	0.386	-0.000	0.001
Informative	0.0004	0.000	1.122	0.262	-0.000	0.001

3.2.1.2 Support Vector Machine (Linear Kernel)

<u>Result of Support Vector Machine</u>
(10 - folds cross validation)

Optimal regularizer parameter: C
* C does not have any effect on the training and validation accuracy
Testing Accuracy: 0.902654867257
True positive: 102 / 133, True negative: 0 / 133



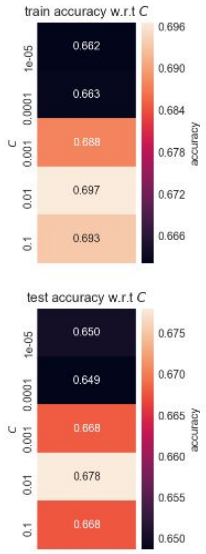


3.2.1.3 Analysis on Logistic Regression

According to the result abolve, we are able to select a set of predictors that is signfiicant in this classification task, we hypothesize that using thsoe predictors might improve the classfication accuracy by reducing the variance introduced through extra parameters that is not significnat (eg. Confusing).

Optimal regularizer parameter: C = 0.01
Testing Accuracy: 0.867256637168; True positive: 98 / 133, True negative: 0 / 133

*Most predictors remain significant while 'Conconvincing' are not signifcant anymore



Logit Regression Results

Dep. Variable:	y	No. Observations:	677
Model:	Logit	Df Residuals:	668
Method:	MLE	Df Model:	8
Date:	Tue, 04 Dec 2018	Pseudo R-squ.:	0.1236
Time:	20:52:18	Log-Likelihood:	-398.30
converged:	True	LL-Null:	-454.47
		LLR p-value:	1.253e-20

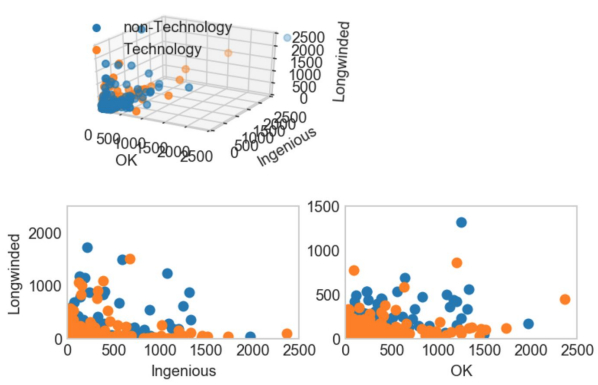
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.9333	0.136	6.867	0.000	0.667	1.200
Funny	-0.0013	0.000	-3.578	0.000	-0.002	-0.001
Courageous	-0.0007	0.000	-2.234	0.025	-0.001	-8.61e-05
Ingenious	0.0027	0.001	5.044	0.000	0.002	0.004
Obnoxious	0.0023	0.001	2.690	0.007	0.001	0.004
Unconvincing	-0.0008	0.000	-1.784	0.074	-0.002	7.75e-05
OK	-0.0033	0.001	-3.220	0.001	-0.005	-0.001
Persuasive	-0.0010	0.000	-2.758	0.006	-0.002	-0.000
Longwinded	-0.0020	0.000	-3.957	0.000	-0.003	-0.001

i) Sequential Feature Selction

The result that some predictors are no longer significnat in the 2nd Logistic Regression (with significant predictors only) may happen due to the correlation among those predictors (eg. persvasive and unconvising are negatively correlated). In order to exclude such effect, we use the forward and backward feature selection. Specifically, we started with begin with all the predictors and iterate through all of them. If the predictor's p-value < threshold 0.01, it is included in the forward process. Then we iterate all the predictors in the included list. Through this backward selection process, we remove all predictors whose p-value > threshold 0.05. Through this process, 5 predictors are identified.

Perdictors	OK	Persuasive	Ingenious	Longwinded
P-value	1.66561e-07	0.00199577	0.000942413	0.000120561

Here is the visualization of data according to the first three predictors



Those features allow the Logistic regression to predict the class label 'Technology' with relatively high accuracy 0.867. However, the data visualization displayed showed that the data are not binary separable, therefore, the classification result may highly

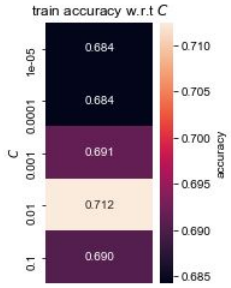
depended on the distribution of the data studied. A larger dataset would be needed.

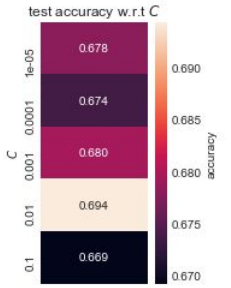
3.2.2 Culture vs. Non-Culture

3.2.1.1 Logistic Regression

Result of Logsitic Regression
(10 - folds cross validation)

Optimal regularizer 3.2.1.2 Support Vector Machine (Linear Kernel)



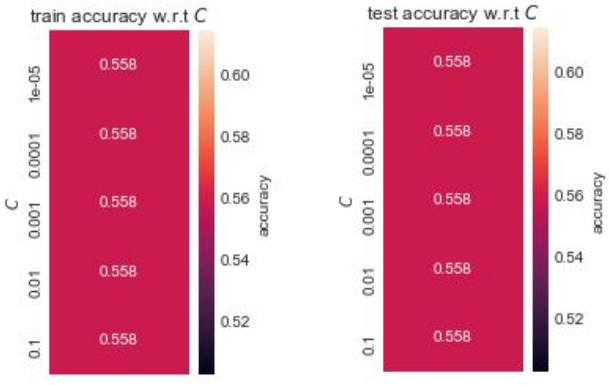


Logit Regression Results

Dep. Variable:		y	No. Observations:			677
Model:		Logit	Df Residuals:			662
Method:		MLE	Df Model:			14
Date:		Tue, 04 Dec 2018	Pseudo R-squ.:			0.1496
Time:		20:47:29	Log-Likelihood:			-395.13
converged:		True	LL-Null:			-464.64
			LLR p-value:			1.115e-22
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8438	0.144	-5.842	0.000	-1.127	-0.561
Longwinded	0.0009	0.000	2.303	0.021	0.000	0.002
Persuasive	0.0037	0.003	1.092	0.275	-0.003	0.010
Inspiring	-0.0002	0.000	-0.787	0.431	-0.001	0.000
OK	0.0018	0.001	3.120	0.002	0.001	0.003
Beautiful	-0.0039	0.001	-5.389	0.000	-0.005	-0.002
Fascinating	-0.0020	0.001	-2.470	0.014	-0.004	-0.000
Obnoxious	0.0005	0.000	1.185	0.236	-0.000	0.001
Unconvincing	0.0056	0.001	3.849	0.000	0.003	0.008
Confusing	-0.0002	0.000	-0.408	0.683	-0.001	0.001
Funny	0.0010	0.000	3.030	0.002	0.000	0.002
Jaw-dropping	0.0013	0.000	2.944	0.003	0.000	0.002
Ingenious	-9.324e-05	0.000	-0.350	0.727	-0.001	0.000
Courageous	0.0002	0.000	0.434	0.665	-0.001	0.001
Informative	-6.68e-05	0.000	-0.185	0.853	-0.001	0.001

Result of Support Vector Machine
(10 - folds cross validation)

Optimal regularizer parameter: C
* C does not have any effect on the training and validation accuracy
Testing Accuracy: 0.601769911504
True positive: 0 / 133, True negative: 68 / 133



3.2.1.3 Analysis on Logistic Regression

According to the result abolve, we are able to select a set of predictors that is signfiicant in this classification task, we hypothesize that using thsoe predictors might improve the classfication accuracy by reducing the variance introduced through extra parameters that is not significnat (eg. Confusing).

Result of Logsitic Regression
(with ‘Significant’ Predictors)

Optimal regularizer parameter: C = 0.01
Testing Accuracy: 0.699115044248
True positive: 52 / 133
True negative: 27 / 133

Among all those predictors, [‘OK’, ‘Courages’, ‘Ingenious’, ‘Inspiring’] are found to be significant through the forward and backward feature section process described above.

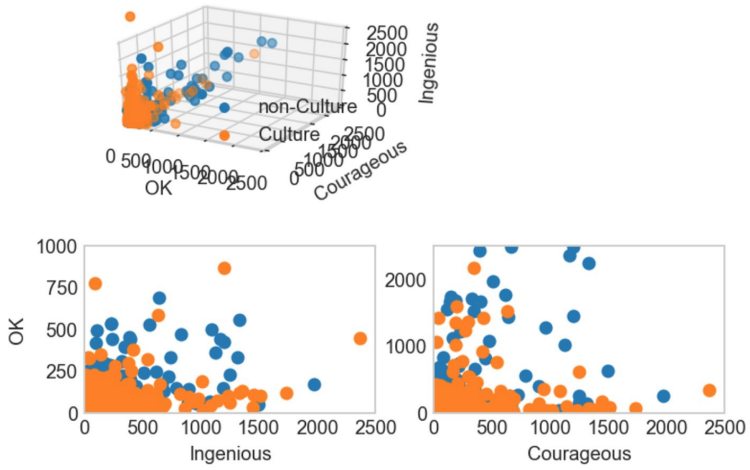
Logit Regression Results

Dep. Variable:	y	No. Observations:	677
Model:	Logit	Df Residuals:	669
Method:	MLE	Df Model:	7
Date:	Tue, 04 Dec 2018	Pseudo R-squ.:	0.1452
Time:	20:47:35	Log-Likelihood:	-397.16
converged:	True	LL-Null:	-464.64
		LLR p-value:	5.755e-26

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8077	0.142	-5.698	0.000	-1.085	-0.530
Funny	0.0008	0.000	2.080	0.037	4.69e-05	0.002
Courageous	0.0021	0.001	3.908	0.000	0.001	0.003
Ingenious	-0.0039	0.001	-5.771	0.000	-0.005	-0.003
Obnoxious	-0.0018	0.001	-2.280	0.023	-0.003	-0.000
OK	0.0062	0.001	4.719	0.000	0.004	0.009
Inspiring	0.0007	0.000	2.926	0.003	0.000	0.001
Longwinded	0.0013	0.000	3.070	0.002	0.000	0.002

Perdictors	OK	Courages	Ingenious	Inspiring
P-value	4.34023e-11	8.75865e-05	2.69335e-06	0.00582796

Here is the visualization of data according to the first three predictors



Compare the significant predictors (selected through forward and backward feature selection) for ‘Technology’ and ‘Culture’ tagged talks, they share [‘OK’, ‘Ingenious’, ‘Courageous’].

This result could motive the future work in using those features as predictor to perform multi-label classification for talks tagged as ‘Technology’ and ‘Culture’.

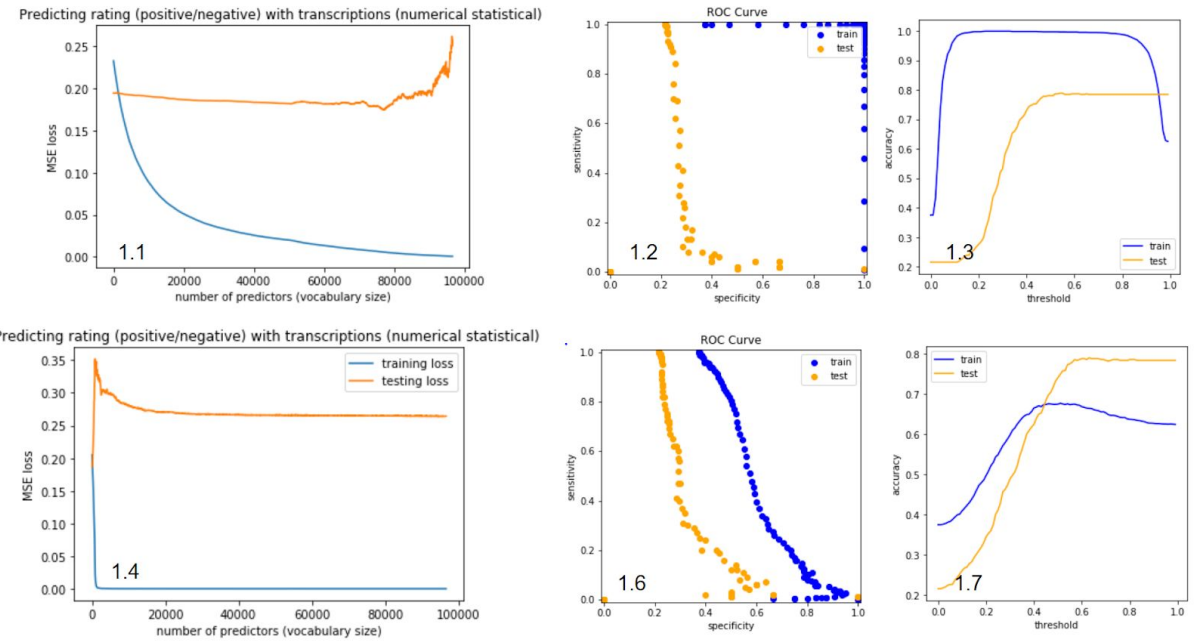
3.3

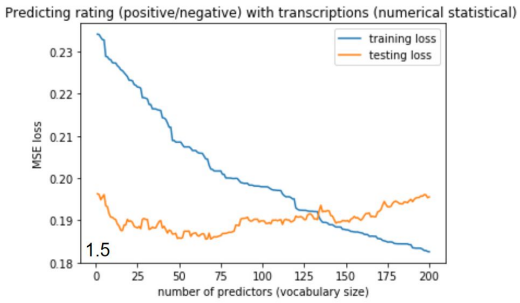
Our models do not show meaningful classification results under this task. However, we can still analyze the failure's pattern and gain useful information.

Image 1 corresponds to model (i)(a) and (i)(b), where a transcript is embedded statistically by each vocab's number of occurrences. Thereby, each vocabulary is a predictor. Subplot 1.1 demonstrates the loss curve given by Mean Squared Error when we add these predictors starting from the lowest frequency words; 1.4 demonstrates that when we start from the highest frequency ones.

As shown in 1.1, always having the most frequent words as predictors yield a smooth and reasonable training curve. Training and test loss both decrease as the number of predictors increase, until when number of predictors build to 76900, overfitting kicks in, that training loss continues to decrease yet test loss rises significantly. The ROC curve in 1.2 and the accuracy versus threshold plot in 1.3 characterize this best model with 76900 most frequent words as predictors. The almost perfect training accuracy and the 78% test accuracy slightly better than random chance (80% of the test data and 65% of all data are labeled 0) indicate that we have enough model complexity. The insatiable validation performance may be evidence that statistical embedding is not a best choice.

In comparison, building a model starting from the least frequent word predictors results in worse classification performances. As shown in 1.4, the model immediately overfits onto the training data, while the test loss gradually drops back to normal rate as the number of predictors increase. This pattern is different from most training curves we have seen in this class but is easily interpretable. When a model only has the least frequent words as predictors, most of these words may only appear in one talk, that that model can easily learn to classify the training data by hard memorization. Note that the test loss curve starts from a relatively low position, so we zoom in on it in Image 1.5, and analyze the model with the lowest test loss. Incorporating 70 least frequent words as its predictors, this model performs even worse than random chance (Image 1.6, 1.7).





From this case comparison between (i)(a) and (i)(b), we conclude that using more common features as predictors can give a model stronger modeling capacity and generalizability, even though the features can be noisy or seemingly lack of information.

In attempt to seek a middle ground, we run several trials excluding both the highest and lowest ends of our data, and a PCA compression. All these trials fail without exception, and their ROC and accuracy curves are plotted in Image 2 for reference.

It is interesting to notice how the model consistently overfits on the training data, again evidencing the sufficient model complexity. Meanwhile, the test data cannot even plot its ROC curve, with its specificity (1 - false negative rate) clustered between 0.2 to 0.4. Indeed, 20% of our test data are tagged 1, so blindly classifying every datapoint as 0 can result in the 0.2 specificity and 0.8 accuracy. In this background, our models using statistically embedded transcript matrices to predict TED talk's ratings have completely failed.

Image 3 demonstrates two models trained on word2vec embedded transcripts, one of dimension 10 and the other of dimension 100. This model cannot even overfit onto the training data. We further furnished our data by leaving out words of extreme frequencies and attempting other dimensions, but none of the efforts produced meaningful results. Overall, brute-force averaging the word vectors in a transcript is not ideal for our rating prediction task.

In general, both our statistical and word2vec models fail to predict whether people give a TED talk positive or negative ratings based on its transcripts. The various training and testing accuracies we observe including 100% training accuracies indicate that logistic regression itself is sufficient for this task. The problem is within the predictors, and there are three potential and parallel reasons behind the failure. First, the transcript may not be so relevant to a talk's rating. Second, the transcript is actually relevant, but our embeddings methods are to be polished. Third, the transcript is relevant to the rating, but the fact that we only have two thousand datapoints makes the predictors too sparse for the model to acquire any generalizability.

4 Discussion and Conclusions

In conclusion, logistic regression demonstrates strong ability on all the three tasks; in particular, regularization methods further improve the results. SVMs in the second task and the embedding method in the third task fail to show meaningful results.

There are a lot of analysis that can be attempted in future work. For example, in our classification experiments, we have only used binary classification about one label. Further experiments can be done with multilabel multiclass classifications. Also, the text embedding methods that we adopt do not lead to very meaningful results, and thus more advanced text embedding techniques might be attempted.

Appendix

Image 1

Image 2