Capstone 2 Final Report

The Texas Rangers are in the midst of a rebuilding season as such their fans, who have become accustomed to a well performing team, are now cheering for a team struggling to win 50% of the time. The team's public relation department realizes that the twitter interactions they have with the fans must be handled delicately so the fans focus on the future of the club and continue to invest in tickets and merchandise.

The Rangers have hired Sheneman Consulting to create a tweet bot to aid in their mission of creating positive interactions. The consulting firm will utilize deep learning to interact with fans. This will be done using Python to connect to twitter and download past interactions. Using this data a chat bot will be created in Tensorflow.

At the conclusion of the project Sheneman Consulting will provide the Rangers with the training code, a deep learning model to deploy and a summary report. The firm will also give a presentation on the project.

Data Set

In order to formulate a realistic model, we must first start with a real world data set. Here the data set will consist of a history of tweets where the official Texas Rangers user id (@Rangers) is mentioned. I have written a Python script that connects to the Twitter API and downloads a large number of these tweets. I then pickle the file so I can access it for further processing. The next step is processing the data. I again wrote a script to clean the text in Python. In this step all urls, user tags and emojis.

Before analysing the data I tokenized the text then removed the stopwords and numbers. Removing these tokens ensures that non-relevant tokens cause the model to misidentify the intent of the tweets. Then I performed stemming and lemmatization both of which extract the root of the word that is being used. This makes sure the plural, past tense, etc. of a word are counted as the same which is important when we are looking for the topics being discussed.

Exploring the Data

When performing natural language processing exploratory data anaylsis consist of trying to find the hot topics in the text. To do this I looked at a few different appoarches. First I extracted the ngrams from the text and got the following:

```
Number of 0
Most common ngrams:
('#',) 4492
('togetherwe',) 2742
```

('rangers',) 1083
('|',) 957
('home',) 896
('baseball',) 891
('game',) 861
('go',) 743
('never',) 725
('play',) 704

Number of 1
Most common ngrams:
('#',) 4492
('togetherwe',) 2742
('rangers',) 1083
('|',) 957
('home',) 896
('baseball',) 891
('game',) 861
('go',) 743
('never',) 725
('play',) 704

Number of 2
Most common ngrams:
('#', 'togetherwe') 2742
('josh', 'hamilton') 643
('ive', 'never') 590
('play', 'baseball') 585
('never', 'seen') 584
('seen', 'physically') 584
('physically', 'gifted') 584
('gifted', 'player') 584
('player', 'josh') 584
('hamilton', 'born') 584

Number of 3
Most common ngrams:
('ive', 'never', 'seen') 584
('never', 'seen', 'physically') 584
('seen', 'physically', 'gifted') 584
('physically', 'gifted', 'player') 584
('gifted', 'player', 'josh') 584
('player', 'josh', 'hamilton') 584
('josh', 'hamilton', 'born') 584
('hamilton', 'born', 'play') 584
('born', 'play', 'baseball') 584
('play', 'baseball', 'power') 584

Number of 4
Most common ngrams:

('ive', 'never', 'seen', 'physically') 584
('never', 'seen', 'physically', 'gifted') 584
('seen', 'physically', 'gifted', 'player') 584
('physically', 'gifted', 'player', 'josh') 584
('gifted', 'player', 'josh', 'hamilton') 584
('player', 'josh', 'hamilton', 'born') 584
('josh', 'hamilton', 'born', 'play') 584
('hamilton', 'born', 'play', 'baseball') 584
('born', 'play', 'baseball', 'power') 584
('play', 'baseball', 'power', 'speed') 584

Number of 5
Most common ngrams:
('ive', 'never', 'seen', 'physically', 'gifted') 584
('never', 'seen', 'physically', 'gifted', 'player') 584
('seen', 'physically', 'gifted', 'player', 'josh') 584
('physically', 'gifted', 'player', 'josh', 'hamilton') 584
('gifted', 'player', 'josh', 'hamilton', 'born') 584
('player', 'josh', 'hamilton', 'born', 'play') 584
('josh', 'hamilton', 'born', 'play', 'baseball') 584
('hamilton', 'born', 'play', 'baseball', 'power') 584
('born', 'play', 'baseball', 'power', 'speed') 584
('play', 'baseball', 'power', 'speed', 'instincts') 583

By simply analyzing the ngrams Josh Hamiliton's playing ability seems to be the main topic of conversation.

However, if we use the TextBlob sentiment analysis we see other tweets are also quite popular.

Number of tweets: 12006

Positive tweets percentage: 2.89%
Negative tweets percentage: 0.3 %
Neutral tweets percentage: 96.81 %

Positive tweets:

priceless time with my boy #unt #untalumni
#togetherwe vote for a chance to win two lower-level tickets couesy of rules
#togetherwe vote for a chance to win two lower-level tickets couesy of rules
its a great day when the gift shop stas to carry jerseys #togetherwe
its a great day when the gift shop stas to carry jerseys #togetherwe
#togetherwe vote for a chance to win two lower-level tickets couesy of rules
#togetherwe vote for a chance to win two lower-level tickets couesy of rules
radio option always when you get to listen to one of the best
#togetherwe vote for a chance to win two lower-level tickets couesy of rules
#togetherwe vote for a chance to win two lower-level tickets couesy of rules

Negative tweets:

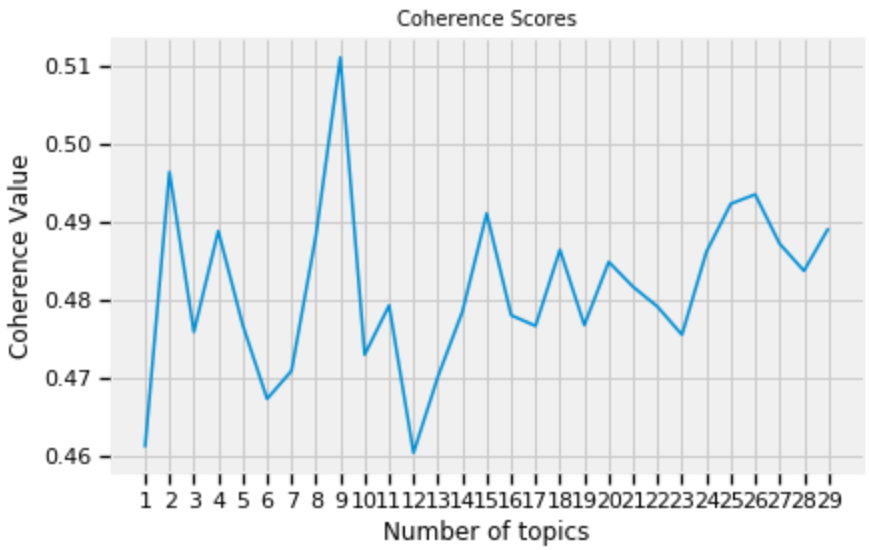
thank you for running my bihday gift from my wife no where on this stupid ad does it say you have to bu
another blown save coming up sickening
damnit smyly god this dude is awful
my god drew smyly is fucking terrible
thats a terrible estimate
over santanas head heard that before horrible trades
remember when unt went into fayetteville and kicked the living crap out of the h
maybe they should change name to doctor of base stealing #togetherwe |
maybe they should change name to doctor of base stealing #togetherwe |
these 9 pm stas are terrible texas should not be in the west div

Neutral tweets:

#togetherwe make old pop culture references
hey rangers is the reason we cant get a mike minor jersey in the stores because yall are trading him
a home game is coming
lets go rangers rangers team total over 55 240/200 #togetherwe #gamblingtwitter #mlb
back home live is back in arlington and we get you all set for the game on #togetherwe
this ones for you gerry
now radio net pregame join and me for chris woodward inside look and more
nice
nieces first game go #togetherwe #womeninbaseballday
grant i have a large inventory of rangers and i do trades of all teams

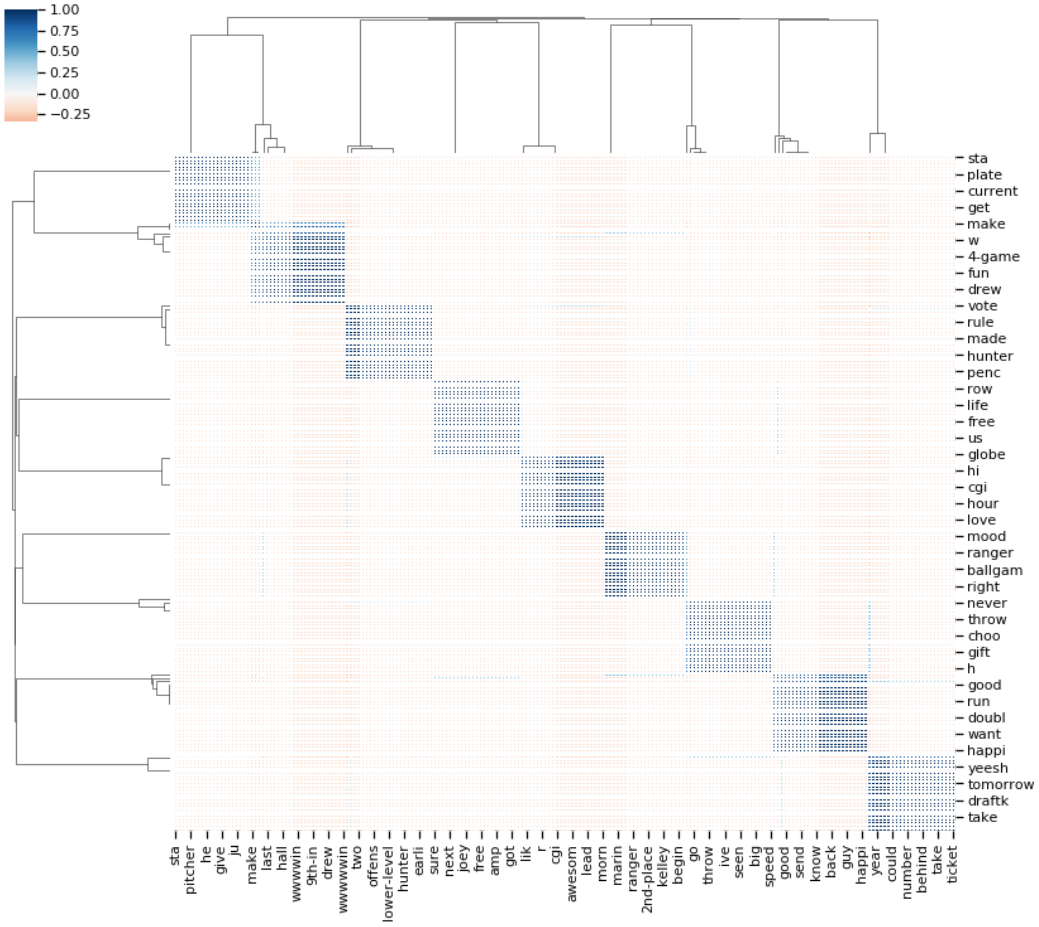
This gives us a good sense of the tweets and we can now see that there are many topics of conversation going. From here we can perform a statistical anaylsis of sorts called Latent Dirichlet Allocation, LDA. LDA transforms the bag-of-words counts into a topic space of lower dimensionality. LDA is a probabilistic extension of LSA (also called multinomial PCA), so LDA's topics can be interpreted as probability distributions over words.

One of the parameters that is given the LDA model is the number of topics. Since the dataset is extremely large and unlabeled a sweep of the data must be performed to find the ideal number of topics. The matrix used to determine this is how coherent the tweets are given N number of topics they can be classified under. The graph below shows the coherence scores given a topic range of 1 to 30 topics.



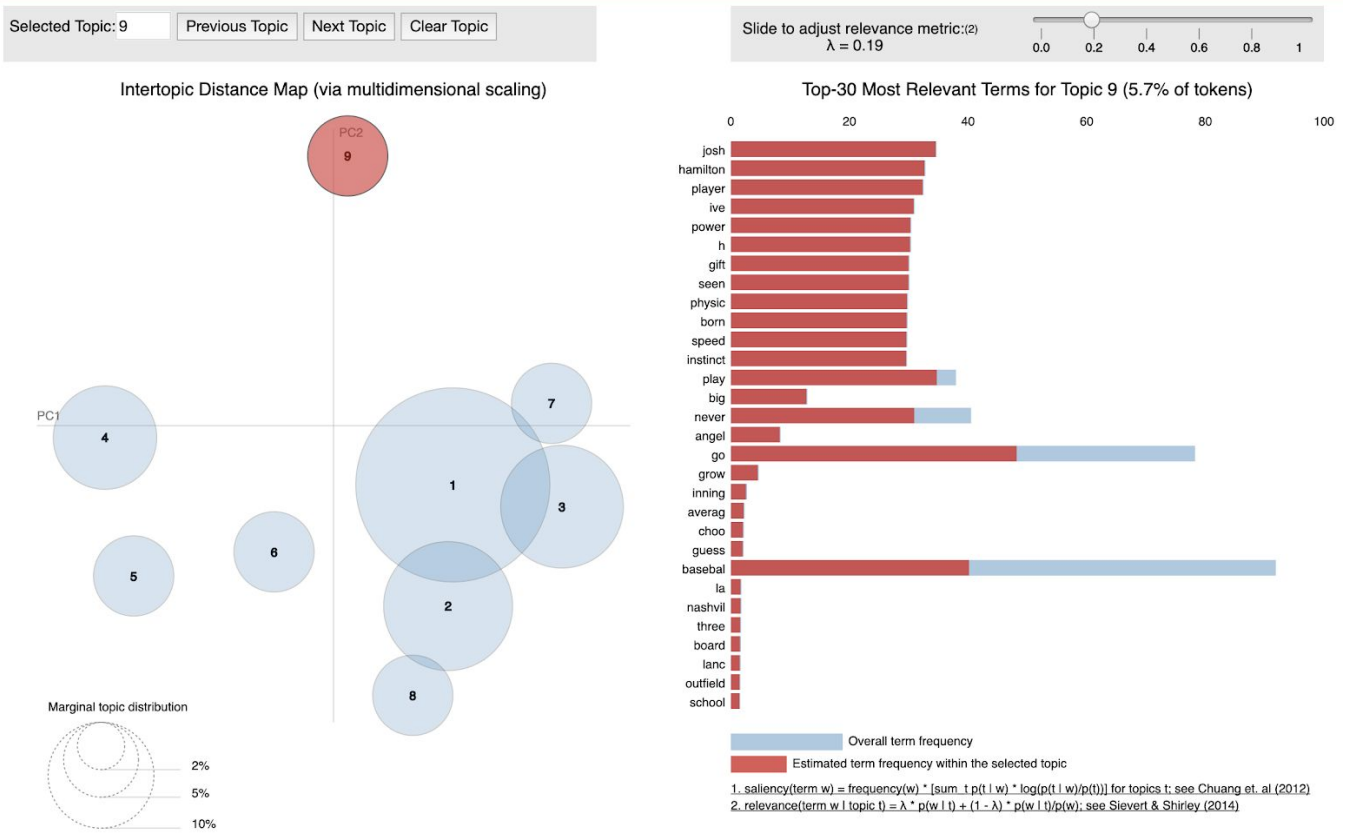
From this graph we see that 14 topics seems to be the ideal number of topics. We can the Ideal number of topics for our model is 7.

The next step is to visualize what distinguishes one topic for the next. This can be done using a heat map as shown below.



As you can see this is hard to quickly identify relationships when a large number of topics are present. Luckily a package called pyLDAvis has been written to interactively explore the topics. A screenshot is shown below.

39]:



Based on the findings we can extract the following topics:

Topic 1: undetermined/unuseable
Topic 2: Joey Gallo
Topic 3: Good game tonight
Topic 4: Ranger's are in 2nd place
Topic 5: 4-game winning streak
Topic 6: Hunter Pence
Topic 7: Joey Gallo on draftkings
Topic 8: undetermined/unuseable
Topic 9: Josh Hamilton

The next steps in the project is to plug the topic information into a neural network to create a state of the art PR machine.

The Model

To build a natural language processing neural network there are some preliminary steps that must be performed. To do this I mode use of two pieces of technology: 1) a free cloud server space known as Google Colaboratory and 2) Keras with a Tensorflow backend. The tokens then were prepped for the model using tokenizer from Keras.

The first model I tried was a simple recurrent neural network with a scaler loss function. The network had a 1024 node RNN layer and ran for 20 epochs for 51 steps.

```
Model: "sequential_6"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_6 (Embedding)      (1, None, 50)             91200
_________________________________________________________________
gru_18 (GRU)                 (1, None, 1024)           3302400
_________________________________________________________________
gru_19 (GRU)                 (1, None, 1024)           6294528
_________________________________________________________________
gru_20 (GRU)                 (1, None, 1024)           6294528
_________________________________________________________________
dense_6 (Dense)              (1, None, 1824)           1869600
=================================================================
Total params: 17,852,256
Trainable params: 17,852,256
Non-trainable params: 0
_________________________________________________________________
```

Then one of the topics can be given to the model and a suggested text will be returned. Here the model was given “good game tonight” and returned:

```
good game tonight give need listen jami reed ranger right injuri begin yo
good game tonight play sta player minor altuv vacat offici sta last summer
good game tonight loss mntwin 1/2 game ahead bring ticket upcom home game
good game tonight loss mntwin 1/2 game ahead stand date pa team univers
good game tonight loss mntwin 1/2 game ahead stand date pa team univers
good game tonight 8th name willson jose altuv congrat michel glenn height pair
good game tonight loss mntwin 1/2 game ahead stand date mani final frown
good game tonight finish seri angel watch game look remain undef rubber game
good game tonight loss mntwin 1/2 game ahead stand date pa team univers
good game tonight loss mntwin 1/2 game ahead stand date pa team univers
```

The model was given “joey gallo” and returned:

```
joey gallo slug total base ab pitcher behind count sinc kiddo wait
joey gallo slug total base ab pitcher behind count sinc kiddo wait
joey gallo slug total base ab pitcher behind count sinc kiddo wait
joey gallo slug total base ab pitcher behind count sinc kiddo wait
joey gallo slug hint togetherw need excus stay past bed time basebal
joey gallo slug total base ab pitcher behind count sinc kiddo call
joey gallo slug total base ab pitcher behind count sinc kiddo wait
joey gallo slug total base ab pitcher behind count sinc kiddo wait
joey gallo ugli five scoreless appear togetherw blow game embarrass fashion stir
joey gallo slug total base ab pitcher behind count sinc kiddo wait
```

The model given “josh hamilton” and returned:

```
josh hamilton turn today mvp season recent sea look keep bat ticket
josh hamilton good time basebal wouldnt bad idea let throw inaugu basebal
josh hamilton turn today mvp season recent sea hot get rememb career
josh hamilton turn today mvp season recent sea look keep bat great
josh hamilton turn today mvp season recent sea look keep bat shin-soo
josh hamilton turn today mvp season recent sea look keep bat bat
josh hamilton good time basebal wouldnt bad idea let throw inaugu throw
josh hamilton good time basebal wouldnt bad idea let throw inaugu player
josh hamilton turn today mvp season recent sea look keep bat offens
josh hamilton good time basebal wouldnt bad idea let throw inaugu player
```

Next I tried to increase the model complexity by adding dropout layers but stayed with the simple RNN with a scaler loss function. The network had a 1024 node RNN layer and ran for 200 epochs for 50 steps. The topology was:

```
Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_3 (Embedding)      (128, None, 500)          912000
_________________________________________________________________
cu_dnngru_9 (CuDNNGRU)       (128, None, 3072)         32937984
_________________________________________________________________
cu_dnngru_10 (CuDNNGRU)      (128, None, 3072)         56641536
_________________________________________________________________
cu_dnngru_11 (CuDNNGRU)      (128, None, 3072)         56641536
_________________________________________________________________
dense_3 (Dense)              (128, None, 1824)         5605152
=================================================================
Total params: 152,738,208
Trainable params: 152,738,208
Non-trainable params: 0
_________________________________________________________________
```

Then one of the topics can be given to the model and a suggested text will be returned. Here the model was given “good game tonight” and returned:

```
good game tonight ang paid among among among habit habit ip best hour
good game tonight ang best hour among among habit best hour ok ok
good game tonight ang paid among among among habit habit habit ip best
good game tonight ang best hour among among habit best hour wast far
good game tonight ang best hour among among habit best hour wast far
good game tonight ang paid american far hour habit hour habit state far
good game tonight ang best hour among among habit best hour ok ok
good game tonight ang paid among among among habit habit ip best hour
good game tonight ang paid american far hour habit hour habit habit habit
good game tonight ang best hour among among habit best hour wast far
```

The model was given “joey gallo” and returned:

```
joey gallo teammat teammat best among improv sho habit best best wast
joey gallo ip ip ip best habit best habit lost wife ip
joey gallo ip gerri ip best best habit habit best habit ip
joey gallo lost gerri among wan habit best state habit surpris state
joey gallo teammat ip opener best best ok ip receiv ball among
joey gallo ip teammat ip best ip ip fearlessli ball land ok
joey gallo ip best wast among best suppo hour best bishop hour
joey gallo sho habit among far best wan habit surpris state best
joey gallo ip ip ray habit receiv best among receiv horribl habit
joey gallo ip ip ip arizona best habit gerri best among wast
```

The model given “josh hamilton” and returned:

```
josh hamilton supervis far go late make longestlistofnighttimewinnersinnohtexa op born cocain allstar
josh hamilton late best otd anaheim best gl late go wear embarrassingli
josh hamilton among star longestlistofnighttimewinnersinnohtexa 31 otd paid tog pitch late even
josh hamilton unfoun game pitch continu realli bobbi 5 late make pitch
josh hamilton safe home lot basketbal basebal loss pitch make terribl r
josh hamilton nice far houston offenc much woodi allstar bas youth royal
josh hamilton among anaheim ray exclus super park understand height definit lead
josh hamilton among entir best level pitch akin night someth bet take
josh hamilton among best aug late make best even home yr pitch
josh hamilton helen two see pla touch want houston who competit allstar
```

Lastly I replaced the RNN layers with long- short term memory gates for the following topology:

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_3 (Embedding)      (1, None, 500)            912000
_________________________________________________________________
lstm_9 (LSTM)                (1, None, 1024)           6246400
_________________________________________________________________
lstm_10 (LSTM)               (1, None, 1024)           8392704
_________________________________________________________________
lstm_11 (LSTM)               (1, None, 1024)           8392704
_________________________________________________________________
dense_3 (Dense)              (1, None, 1824)           1869600
=================================================================
Total params: 25,813,408
Trainable params: 25,813,408
Non-trainable params: 0
_________________________________________________________________
```

Here the model was given “good game tonight” and returned:

```
good game tonight wolf hope dalla advic tra flythew game game game game
good game tonight poor save smh situat 40th congratul editor game game game
good game tonight differ protect domin tog hear oriol game yout strong shit
good game tonight pla suppo despit arod 56min yout somehow critic game game
good game tonight spi game game game game game nolan let physic nolan
good game tonight scoreless 56min game game game game game togetherw player arlington
good game tonight advic poor scoreless arm sposwrit lanc save protect video game
good game tonight spi game game game game sta closet r tonight game
good game tonight sister obvious fast away guyer togetherw hunti era day togetherw
good game tonight 56min advic game game game nolan septemb ball togetherw game
```

The model was given “joey gallo” and returned:

```
joey gallo game great star pick take littl tiger expect tog poor
joey gallo soon bummer chanc keller that booth exam level veloc degre
joey gallo game continu suspend enrout top ace strikeout level 4 jerri
joey gallo forc amaz way star cloth throw wolf k 2 odubel
joey gallo land valuabl bullshit that make lead tweet back josh top
joey gallo game walsh futur see angel fan highlight unt n still
joey gallo beltr big triple pa warn famili perform far lynn attend
joey gallo pt game could wise dont way remot beach e wwe
joey gallo game la that dream smile set g entir whoever someon
joey gallo bullshit hall run queen behind championship offic casino hit alarm
```

The model given “josh hamilton” and returned:

```
josh hamilton game game make section top good wrote fals player 9th
josh hamilton game game amp back yall instead train ariel icymi appear
josh hamilton game game one togetherw better win 31 train doesnt score
josh hamilton game game game hit yall jedi finish whi skipper despit
josh hamilton game game yet via fan fan finish 2 beat score
josh hamilton game game favorit seam sometim cours power sometim 1st score
josh hamilton game game men get avg toss back spot fra theyr
josh hamilton game game amp get overcom warn name kid tough intercontinent
josh hamilton game game thank prankster univers staff josh game game game
josh hamilton game game keller board owner reach 31 repres mo allow
```

Conclusion

The models I built provide a good basis of tweets for the public relations department to consider though it is clear modifications need to be made. The network is set up but lacks a robust training data. The RNN with 1024 nodes produces the best results. Further work will be done to improve the model, however, the team has access to the sentiment analysis and topic modeler while the code is being refined.

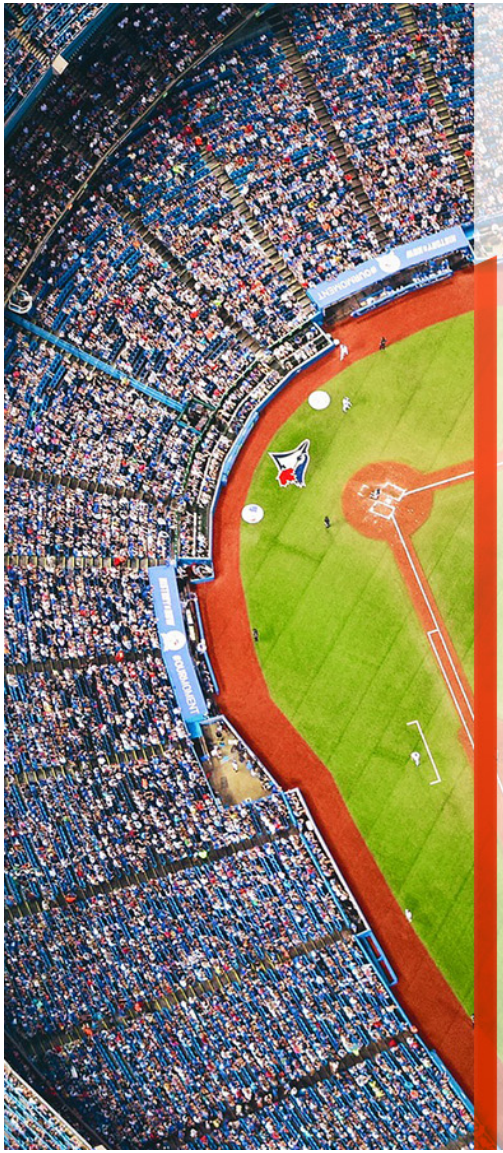
Code for this project can be found at: https://github.com/LSheneman/texas_rangers_modeler


Texas Rangers:
Twitter Relations
FPPT.com

Twitter: PR Machine

Driving Force

- The Rangers are in the middle of a rebuilding season.
- The team has a record of around .500 and fans are getting angsty.
- Twitter is a place where fans can voice there opinions in real time.



How should the team respond?

- The Rangers are seeking a product that has three features:
  - Gage the overall sentiment
  - Find the topics users are currently discussing
  - Creates suggested responses.

Download data directly from Twitter

OBTAIN THE DATA

Twitter Data

- Setup an app on Twitter.
- Downloaded the maximum number of tweets allowed.
- All tweets mentioned contained @Rangers.

Clean data

- Removed stopwords.
- Removed numbers.
- Removed plurals and matched word tenses using
  - Stemming
  - Lemmatization

Topic modeling

EXPLORE USER TWEETS

Sentiment Analysis

Number of tweets: 12006

Positive tweets percentage: 2.89%

Negative tweets percentage: 0.3 %

Neutral tweets percentage: 96.81 %

Positive Tweets

- priceless time with my boy #unt #untalumni
- #togetherwe vote for a chance to win two lower-level tickets couesy of rules
- #togetherwe vote for a chance to win two lower-level tickets couesy of rules
- its a great day when the gift shop stas to carry jerseys #togetherwe
- its a great day when the gift shop stas to carry jerseys #togetherwe
- #togetherwe vote for a chance to win two lower-level tickets couesy of rules

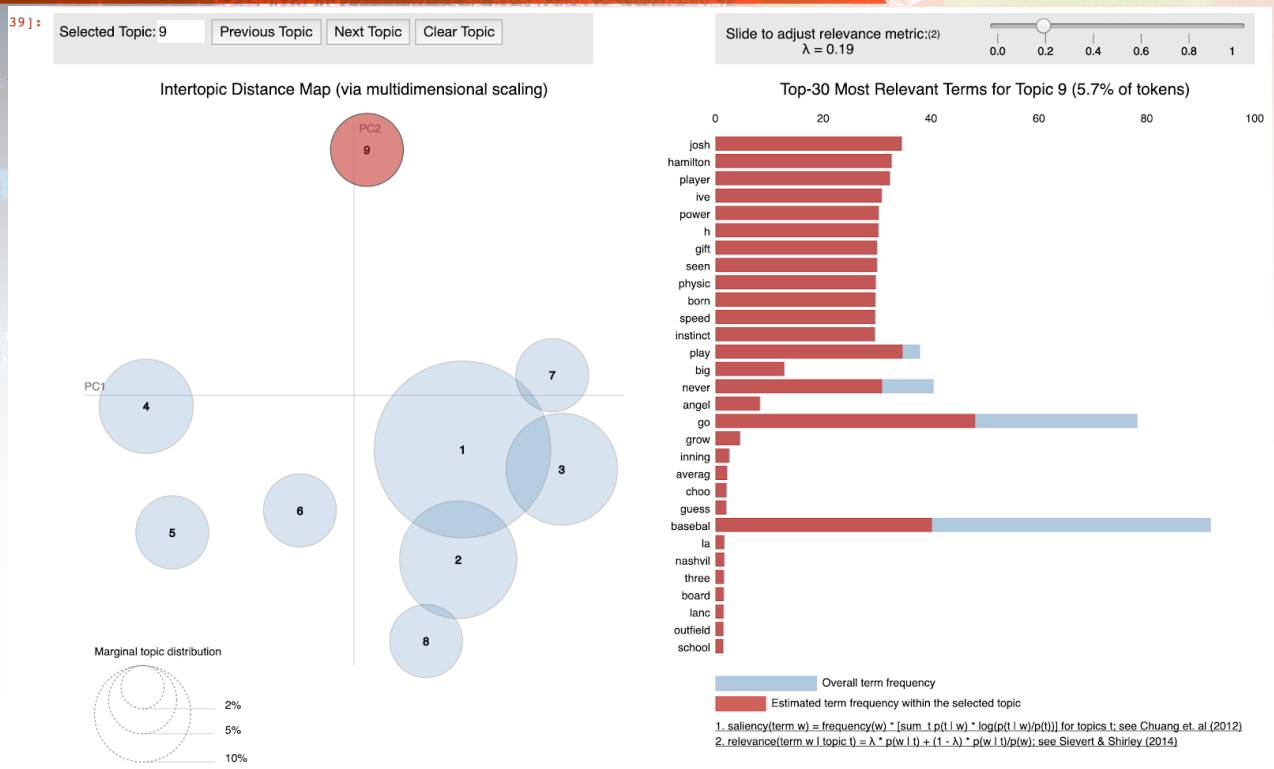
Negative Tweets

- thank you for running my bihday gift from my wife no where on this stupid ad does it say you have to bu
- another blown save coming up sickening
- damnit smyly god this dude is awful
- my god drew smyly is fucking terrible
- thats a terrible estimate
- over santanas head heard that before horrible trades

Neutral Tweets

- hey rangers is the reason we cant get a mike minor jersey in the stores because yall are trading him
- a home game is coming
- lets go rangers rangers team total over 55 240/200 #togetherwe #gamblingtwitter #mlb
- back home live is back in arlington and we get you all set for the game on #togetherwe
- this ones for you gerry
- #togetherwe make old pop culture references

Topic modeling



Top Topics

- Joey Gallo
- Good game tonight
- Ranger’s are in 2nd place
- 4-game winning streak
- Hunter Pence
- Joey Gallo on draftkings
- Josh Hamilton

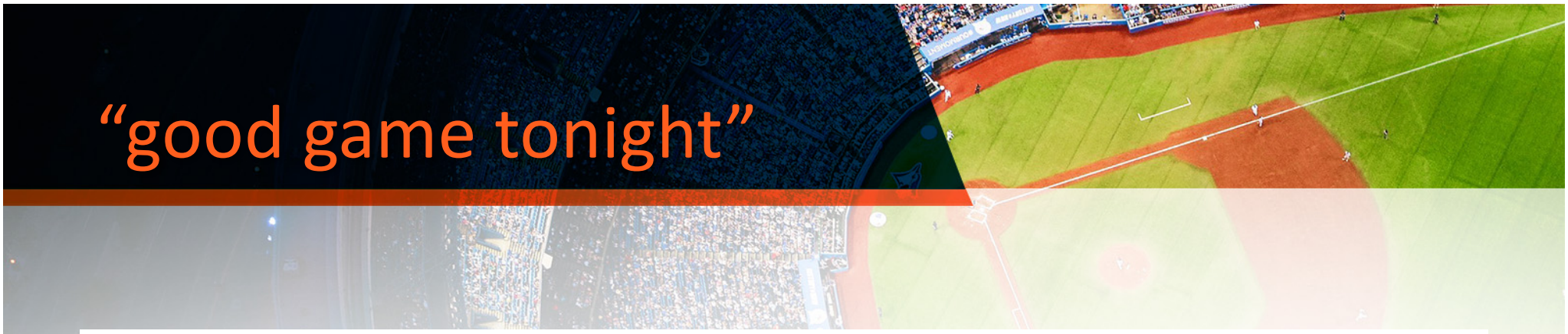
Implement a neural network

BUILD A RESPONSE SYSTEM

Recursive Neural Network

```
Model: "sequential_6"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_6 (Embedding)      (1, None, 50)             91200
_________________________________________________________________
gru_18 (GRU)                 (1, None, 1024)           3302400
_________________________________________________________________
gru_19 (GRU)                 (1, None, 1024)           6294528
_________________________________________________________________
gru_20 (GRU)                 (1, None, 1024)           6294528
_________________________________________________________________
dense_6 (Dense)              (1, None, 1824)           1869600
=================================================================
Total params: 17,852,256
Trainable params: 17,852,256
Non-trainable params: 0
_________________________________________________________________
```

“good game tonight”



```
good game tonight give need listen jami reed ranger right injuri begin yo
good game tonight play sta player minor altuv vacat offici sta last summer
good game tonight loss mntwin 1/2 game ahead bring ticket upcom home game
good game tonight loss mntwin 1/2 game ahead stand date pa team univers
good game tonight loss mntwin 1/2 game ahead stand date pa team univers
good game tonight 8th name willson jose altuv congrat michel glenn height pair
good game tonight loss mntwin 1/2 game ahead stand date mani final frown
good game tonight finish seri angel watch game look remain undef rubber game
good game tonight loss mntwin 1/2 game ahead stand date pa team univers
good game tonight loss mntwin 1/2 game ahead stand date pa team univers
```

”joey gallo”

```
joey gallo slug total base ab pitcher behind count sinc kiddo wait
joey gallo slug total base ab pitcher behind count sinc kiddo wait
joey gallo slug total base ab pitcher behind count sinc kiddo wait
joey gallo slug total base ab pitcher behind count sinc kiddo wait
joey gallo slug hint togetherw need excus stay past bed time basebal
joey gallo slug total base ab pitcher behind count sinc kiddo call
joey gallo slug total base ab pitcher behind count sinc kiddo wait
joey gallo slug total base ab pitcher behind count sinc kiddo wait
joey gallo ugli five scoreless appear togetherw blow game embarrass fashion stir
joey gallo slug total base ab pitcher behind count sinc kiddo wait
```

“josh Hamilton”

```
josh hamilton turn today mvp season recent sea look keep bat ticket
josh hamilton good time basebal wouldnt bad idea let throw inaugu basebal
josh hamilton turn today mvp season recent sea hot get rememb career
josh hamilton turn today mvp season recent sea look keep bat great
josh hamilton turn today mvp season recent sea look keep bat shin-soo
josh hamilton turn today mvp season recent sea look keep bat bat
josh hamilton good time basebal wouldnt bad idea let throw inaugu throw
josh hamilton good time basebal wouldnt bad idea let throw inaugu player
josh hamilton turn today mvp season recent sea look keep bat offens
josh hamilton good time basebal wouldnt bad idea let throw inaugu player
```

FINAL SCORE?

It’s a win!

- Without human input the final model can:
  - Analyze the sentiment of fans
  - Generate a list of the top topics
- The model can suggest tweets that the PR team can make small edits to before posting.