Capstone 2 Milestone 1 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:
```

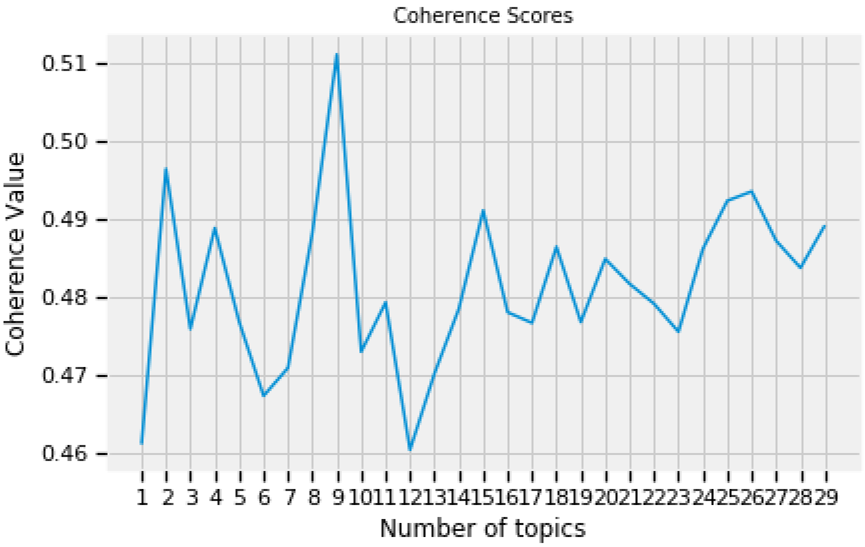
thank you for running my bihday gift from my wife no where on this stupid ad does it say you have to bu
another blown save coming up sickening
damnit smyly god this dude is awful
my god drew smyly is fucking terrible
thats a terrible estimate
over santanas head heard that before horrible trades
remember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

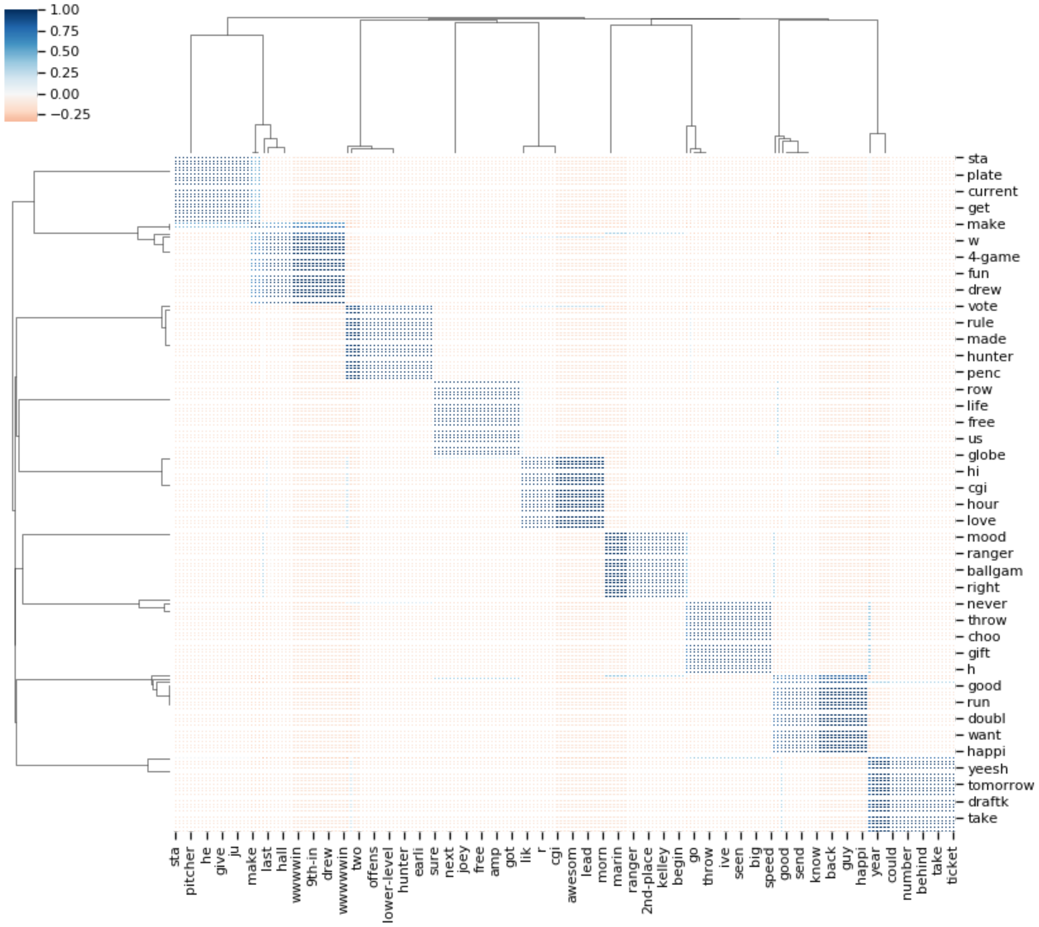
This gives us a good sense of the tweets and we can now see that there are many topics of conversation going. From here we can perform a statistical anaylsis of sorts called Latent Dirichlet Allocation, LDA. LDA transforms the bag-of-words counts into a topic space of lower dimensionality. LDA is a probabilistic 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.

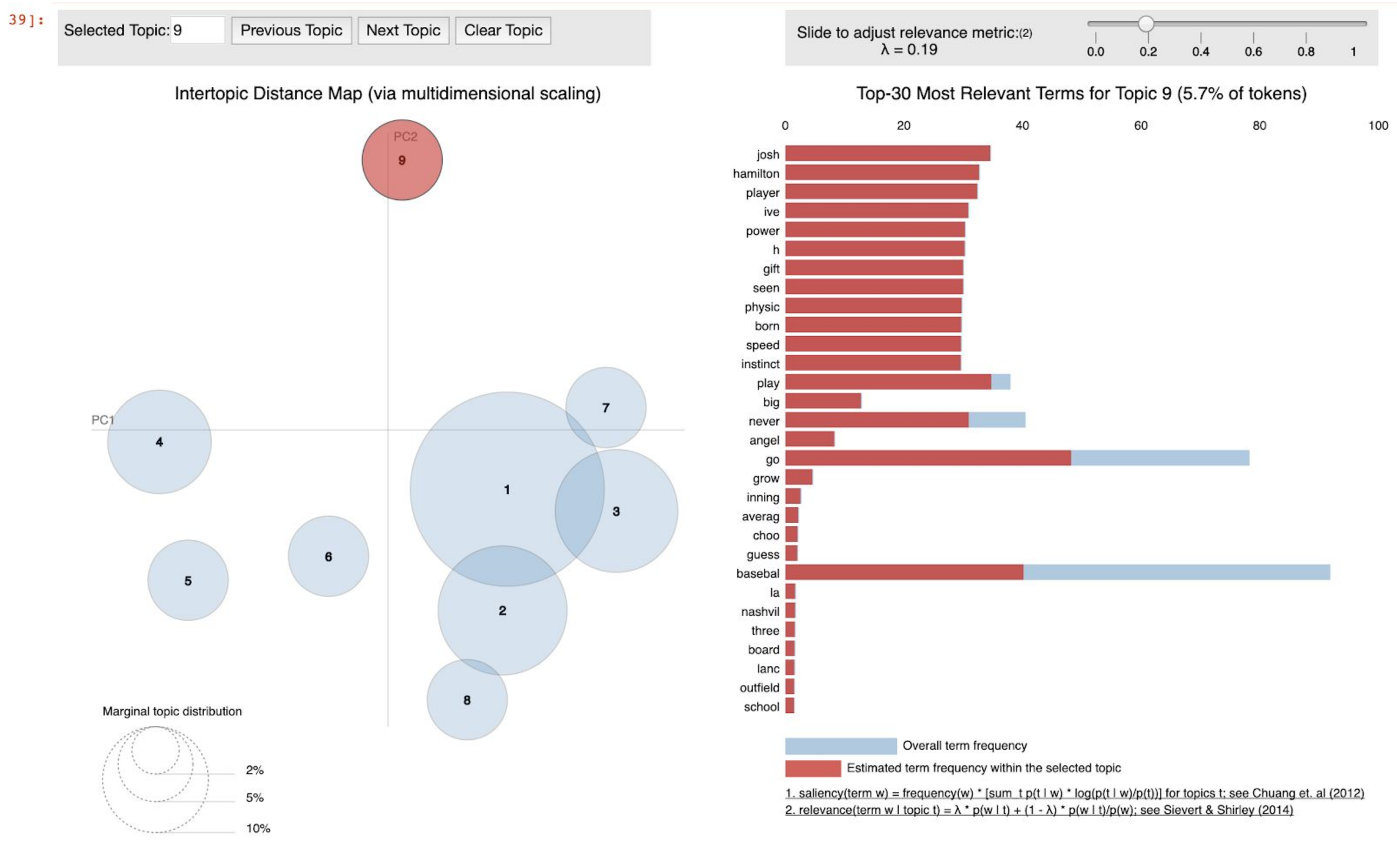


From this graph we see that 14 topics seems to be the ideal number of topics. We can the Ideal number of topics for our model is 7.

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.



Selected Topic: 9

Slide to adjust relevance metric:(2)
λ = 0.19

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

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 act PR machine.

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