**Background Detail:**

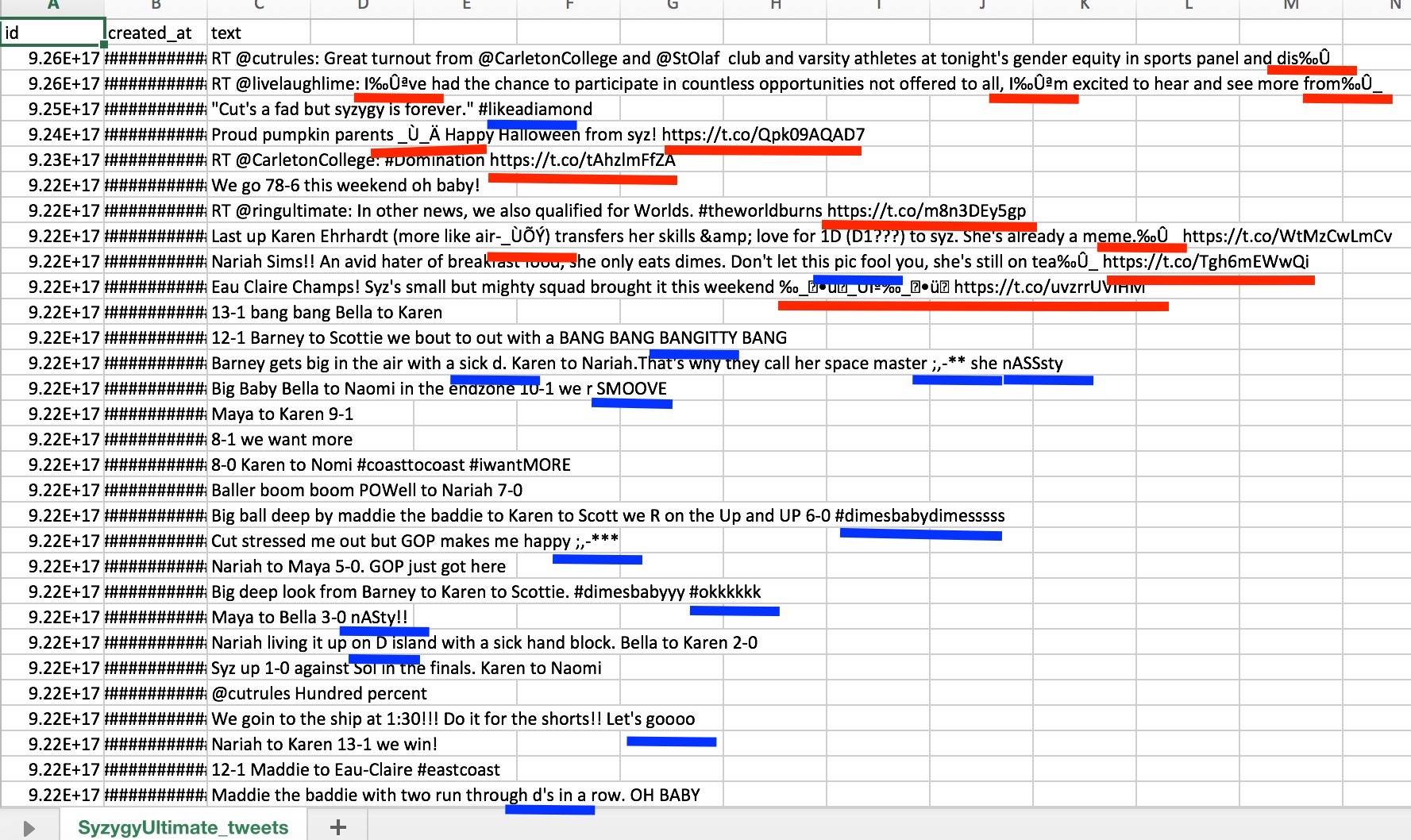
Twitter plays a huge role in the world of Ultimate frisbee. Through Twitter, teams can update fans on everything from scores to the latest team social event. Because tournaments are often far away, parents and friends usually can’t come to every game. Through Twitter, fans can get real time updates on a team’s performance. These updates tend to contain more information than just the score. For instance, a Tweet might say something like, “so and so gets a d”, “so and so to so and so for the score” and then state the score.

**Research Question/Statement:**

Given Twitter’s importance in the world of ultimate frisbee, I was interested in analyzing my frisbee team’s, Syzygy’s, Tweets. When looking at the Tweets, I was interested in creating word frequency charts. I made a chart for unigrams, bigrams and trigrams. I was also interested in using sentiment analysis to look into whether there were positive or negative trends in the Tweets

**Methodology of Approach:**

I began my research by downloading all syzygy Tweets from march 30th 2014 at 12:42 am to November 3rd 2017) at 5 am. I quickly noticed that there were some strange texts showing up in the Tweets. Below this paragraph is a picture of some of the downloaded Tweets. The red lines represent unrecognizable text such as http addresses and weird symbols that appeared randomly. The blue lines represent text that I understand the meaning of but, would not find the word in a dictionary, such as #likeadiamond, which is not a real word but, most native English speakers would understand the meaning. As you can see, almost every Tweet has at least one kind of underline. I ultimately decided to only lightly clean the text. Most of the weird symbols results from emojis, pictures, videos or links. It would be silly to not download any Tweets that contained one of those things as there is still valuable information in those. That being said, the red underlined Tweets did reek some havoc when trying to create word frequency charts. A surprising amount of the red underlined text has spaces in it. This caused my parser to divide up the sections into several more sections, doubling, if not tripling, their presence in the frequency charts.



The blue underlined text caused trouble when trying to perform semantic analyses on the Tweets. The program I used for semantic analyses, TextBlob, assigned an entire sentence a score of 0 if the sentence contained any unrecognizable worlds. Thus, a significant amount of my sentences received scores of 0. Also, in this context, changes in spelling are often used to add emphasis. Therefore, a score of 0 is even more inadequate for Tweets containing eccentric spellings.

**Unigram Frequency Table:**

Of the 3,229 Tweets downloaded there were a total of 25,841 words (including numbers) and 3,879 unique words (including numbers). I find 25,841 words (including numbers) to be a reasonable number. However, 3,879 unique words is a bit low. 3,879 Unique words implies that basically every Tweet had about one unique word in it. This low number indicates that the Tweets consist of a rather small vocabulary which makes sense considering the unique and specific context of Syzygy ultimate.

Below are the most the first few results from my unigram frequency chart. For the entire chart please refer to the unigram frequency section in the “NLPData” file. Please note that anything highlighted in yellow is a name of a Syzygy player.

to: 1203 – makes sense as I believe one of the most common Tweets is of the score which is typically presented as the assistor to the scorer (obviously assistor would be the name of the person who threw the assist and scorer is the name of the person who caught the disc in the end zone).

the: 915

a: 636

we: 430

and: 371

for: 370

with: 364

d: 318

on: 292

it: 291

in: 273

turn: 270

syz: 266 – short for syzygy

score: 242

katie: 233

they: 226

but: 220

of: 173

game: 171

thallon: 162

break: 157 – term used to refer to when the team that begins the point on defense and scores

up: 154

is: 150

half: 133

point: 128

out: 111

huck: 108 – a big throw

rt: 107 – stands for re-Tweet

ahna: 106

from: 105

zone: 100 – a type of defense

@cutrules: 98

9: 95 – I wonder why the number 9 appears so much

em: 90

gets: 88

lucia: 87

by: 87

go: 85

emma: 85

scores: 84

endzone: 82

against: 76

then: 74

buckner: 74

back: 74

our: 72

you: 71

that: 71

at: 70

great: 69

are: 69

throw: 66

soter: 65

nice: 65

this: 62

get: 62

big: 62

o: 61 – stands for offense

mar: 61 – stands for March, likely is related to time stamps on certain Tweets

so: 60

rosty: 60

one: 58

i: 58

it's: 57

an: 56

2: 56

maddie: 55

some: 51

naomi: 51

j: 51

again: 51

turns: 50

thall: 50

good: 50

@goprocks: 49

feb: 48 - stands for February, likely is related to time stamps on certain Tweets

ellen: 48

another: 48

zoe: 47

their: 46

kirstie: 46

down: 46

day: 46

3: 45

1-0: 45

all: 44

line: 43

ne: 42

win: 41

ubc: 41 – stands for University of British Columbia

take: 41

long: 41

field: 41

@eclipsethedisc: 41

5: 41

end: 40

beautiful: 40

was: 39

maya: 39

loosh: 39

leah: 39

claire: 39

It is important to point out that I removed the following from my frequency table,

',', '.', '(', ')', ':', ';','!','\\', "'s", 'https', "-",'"', !='\_', "?", "d9", “d9\_", "d9\_\_\_",“d9\_\_", “9\_", "/", "s", aa", st", “c8", “b4", '8", u143",“uc0", ce", “d5", “db”, fc", 95", d3". Each thing was removed because it had little to no significant meaning.

There are defiantly some interesting things happening in the unigram frequency chart. Many names have a significant number of Tweets. However, when considering Tweet quantities, especially those regarding names, it is important to take a couple of things into consideration. First of all, the time span in which these Tweets were downloaded has a significant impact on whose names occurs the most. Katie, the most frequently Tweeted name in my corpora occurs 233 times. Unlike other player’s names who occur less frequently, Katie’s entire four years on Syzygy is captured within the time span of the Tweets analyzed. Additionally, the maximum amount of years one can play on Syzygy is four (join freshman year and play each following year until graduation). However, some people do not join their freshman year and/or do not play each following year. Examples of this include Ahna weeks, who joined her sophomore year and Emma Nicosia who took her sophomore year off. Ahna had 106 Tweets within her 3 years and Emma occurs 85 times. Secondly, there are many people who share the same name on Syzygy. Last year alone there were 3 Emmas and 2 Claires on our team! Therefor some names in the frequency table do not necessarily refer to just one person. Thirdly, while some people might be on Syzygy all four years, that does not necessarily mean that they were physically able to play that entire time. Like any sport, frisbee is riddled with injured players. At any given time, it is fair to assume that about 15% of a Frisbee is out of commission due to sometype of injury or illness. As you can see, there are multiple factors that effect the amount of opportunities a player has to be tweeted about. Additionally, many people on the team have multiple ways they are referred to. For example, both “Thall” and “Thallon” refer to the same person, Claire Thallon. Thall occurs 50 times, Thallon occurs 162 times therefor, Claire Thallon is referred to at least 212 times (Claire also occurs 39 times however, because there have been so many Claires on Syzygy it is impossible to know how many of these refer to Claire Thallon). One interesting observation I have is that the quantities associated with each unigram become exponentially smaller as you go down the table. This makes sense given the small vocabulary the Tweets consist of.

**Bigram Frequency Table:**

\*Please note that unfortunately I was unable to remove erroneous results from my Bigram and Trigram tables, however, for the purpose of my analysis, I pulled out “noisy” tweets from the data below. The full, untouched data is provided in the NLPData document. Please note the “s” and “e” in the bigram and trigram frequency tables refer to the beginning and ends of Tweets. Here are the most prominent bigrams from Syzygy’s Tweets:

., e: 431 – Makes sense as every period is followed by an e. Thus, there must be about 431 Tweets that have periods at the ends of them. This is a surprisingly low number considering there are 3,229 total Tweets in the corpus.

for, the: 201

!, !: 189

with, a: 185

s, we: 176

!, e: 157

s, syz: 137 – almost exactly half (51% to be exact ☺) as much as its corresponding unigram

we, turn: 127

s, katie: 117 – Katie Ciaglo, once again appearing at the top of the frequency chart

the, score: 115

katie, to: 114

\_, ': 114

s, they: 110

in, the: 105

s, rt: 99

turn, .: 93

score, .: 93

to, thallon: 78

d, e: 78

with, the: 74

on, the: 62

d, .: 62

mar, e: 61

., we: 61

-, mar: 61

they, turn: 52

they, score: 50

to, ahna: 49

but, we: 49

a, d: 49

to, emma: 48

lucia, to: 48

feb, e: 48

-, feb: 48

of, the: 46

s, lucia: 44

turn, e: 43

huck, to: 43

s, thallon: 42

the, endzone: 41

on, d: 41

,, but: 41

to, the: 39

s, 2: 39

on, a: 39

gets, a: 39

3, -: 39

s, 3: 38

s, ": 38

to, em: 37

the, d: 36

s, 4: 36

2, -: 36

s, 6: 35

s, 5: 35

end, zone: 35

apr, e: 35

6, -: 35

., they: 35

-, apr: 35

5, -: 34

4, -: 34

thallon, to: 33

s, @cutrules: 33

it, up: 33

", e: 33

to, soter: 32

the, break: 32

rt, @cutrules: 32

out, of: 32

jan, e: 32

em, to: 32

d5, ': 32

@cutrules, :: 32

7, -: 32

-, jan: 32

1-0, e: 31 – 70% of its corresponding unigram

scores, .: 30

s, ne: 30

s, em: 30

s, 7: 30

may, e: 30

for, a: 30

-, may: 30

to, a: 29

s, maddie: 29

The respective names occur in relatively the same order as they did in the unigram frequency chart. As expected, the highest bigram frequencies are lower than the highest unigram frequencies. This makes sense as it is less likely for two consecutive words to occur than just one. Order wise, everything is relatively consistent with the unigram ranking.

The highest bigram ranking is about 35% percent of the highest unigram ranking. For the most part, this ratio is relatively consistent throughout the rest of the bigram quantities (+/- about 10%).

Many of the most frequent bigrams were of scores (highlighted in blue). While the numbers are concentrated in the lower portion of the chart, they definitely form a significant presence. There are far fewer numbers in the unigram frequency chart. This makes sense as numbers are usually reported in regards to a score and thus in the following format, # - #. The high quantity of numbers in the bigram chart speaks to the significance of one of Syzygy’s ultimate’s Twitters’ main purpose, to provide real time updates for followers during games.

**Trigram Frequency Table:**

for, the, score: 94

!, !, !: 74

s, we, turn: 65

s, katie, to: 62

the, score, .: 61

-, mar, e: 61

we, turn, .: 48

-, feb, e: 48

turn, ., e: 45

s, 3, -: 38

s, 2, -: 36

s, 6, -: 35

-, apr, e: 35

!, !, e: 35

s, 5, -: 34

s, 4, -: 34

rt, @cutrules, :: 32

-, jan, e: 32

s, 7, -: 30

-, may, e: 30

s, they, score: 29

s, rt, @cutrules: 29

s, 8, -: 29

in, the, endzone: 29

-, jun, e: 29

s, lucia, to: 28

-, jul, e: 27

but, we, turn: 26

the, end, zone: 25

s, 1, -: 23

-, oct, e: 23

d, ., e: 22

we, turn, e: 21

s, they, turn: 21

s, 9, -: 21

gets, a, d: 20

with, a, big: 19

starting, on, d: 19

s, thallon, to: 19

s, rt, @ultiworldlive: 19

s, katie, with: 19

rt, @ultiworldlive, :: 19

for, the, break: 19

-, sep, e: 19

run, through, d: 17

in, the, end: 17

., we, turn: 17

with, the, d: 16

with, a, nice: 16

we, turn, ,: 16

the, endzone, .: 16

s, em, to: 16

s, 10, -: 16

in, the, ez: 16

@ultiworldlive, :, w: 16

s, they, break: 15

katie, with, a: 15

d, in, the: 15

-, aug, e: 15

works, it, up: 14

with, a, d: 14

they, score, .: 14

s, maya, to: 14

s, emma, to: 14

on, a, huck: 14

of, the, day: 14

katie, to, emma: 14

for, the, win: 14

we, take, half: 13

the, d, e: 13

s, loosh, to: 13

s, 11, -: 13

-, nov, e: 13

', db\_, e: 13

they, turn, .: 12

the, ez, .: 12

s, rosty, to: 12

s, maddie, to: 12

The ratio between each individual trigram quantity and its corresponding bigram quantity is quite similar, perhaps about 5% higher, to the bigram to unigram ratio. Both “syz” and “1-0”, the two things I was tracking in the unigram and bigram charts, fell off the table in the trigram chart! I understand why 1-0 might not appear in the top trigrams however, I am very surprised that “syz” is no where to be found in the top trigram quantities! The overall presence of numbers is about the same as in the bigram chart, however, they appear higher up in the trigram chart. This makes sense as scores are usually reported in a trigram format (# - #). I am surprised how many negative Tweets there are (highlighted in red). I guess in some sense it speaks to our honesty in trying to keep followers up to date in the happenings of each game. I would love to look into how the number of negative Tweets Tweeted during a game relates to the score of the game.

**SENTIMENT ANALYSIS(positive):**

\*Used TextBlob to assist with sentiment analysis

1.0: AND WE WIN!! 15-11\

1.0: Great block by Emma N! Katie to Thallon 4-4\

1.0: Thallon to Ani. Beautiful! 9-2\

1.0: Thallon plays excellent dump defense and gets a D then dishes it for score to Soter 6-10\

1.0: Good luck @eclipsethedisc !! Bring it home to carleton! #carletoncollegeisthebest\

1.0: Emily is Lebon james. Puts it to Claire Thallon for the score. We win! 16-15\

1.0: Beautiful throw Karen! 12-0\

1.0: Perfect cup d and endzone offense 2-1\

1.0: Em to Megs for the win! Syz beats Virginia 12-10\

1.0: Awesome throw from Katie to Thall leaping over the line for the score 2-3\

1.0: Best battle scar of the weekend goes to Emma Nicosia #nobig https://t.co/lE17ZXNJZ7\

1.0: Great d by maddie sets up Ahna to Elaine for the score! Halftime 8-2\

1.0: Snyder to Ari-Kari for the win!! 11-9\

1.0: Reed is the best team tent ever! http://t.co/S6v5BvoWa3\

1.0: Rushing the field was #worththetmf as the ref said "that was an AWESOME catch"\

1.0: Beautiful day to be playing frisbee with the heavenly bodies! Catch teams X and Y flinging discs in all their celes\'89\'db\_ https://t.co/7n8skoBysy\

1.0: Kirstie to Chav for the win! 10-4\

1.0: beautiful hammer from kirstie! 1-0 syzzlebeans\

1.0: EXCELLENT TEAM D\

1.0: @cutrules let's go CUTe!!!!!!! \_\'d9\'d4\uc0\u143 \_\'d9\u143 \'c8\

1.0: Good luck to @eclipsethedisc and @goprocks!! #carletoncollegeisthebest\

1.0: Katie to sarah for the win! 11-2\

1.0: @eclipsethedisc @goprocks Good luck today!!!! We're all cheering for you guys\

1.0: We win! 13-2!!! 4-0 for the day #upandup\

There are a surprisingly high number of Tweets with a score of 1. So much so that I could not even include them all in this paper! It seems like the sentiment analysis program is particularly sensitive to specific words. I highlighted words in yellow that might have pulled the sentiment analyses score up. It makes sense that there are so many extremely positive Tweets given the context of the Tweets. Obviously, some Tweets read significantly more positive than others, despite all sharing the same score. This is simply the nature of most sentiment analysis as it struggles to take context into account.

**SENTIMENT ANALYSIS(negative):**

-0.6: 6-5 Dartmouth break. So cold. So rainy. Go Syz!!!\

-0.6: Chilly endzone and then Loosh to Em 4s #fearlessleaders\

-0.6: Em with a dirty break to em 12s\

-0.625: Oh yeah, and you can watch this game on Nexgen if you want! Huddled up now! http://t.co/LPfY8YStDR\

-0.625: Syz got that mad flow. 3-2\

-0.625: S-Y-Z-Y-G-Y GAME TIME!! http://t.co/clXSYbGX8G\

-0.7: bad ladies score 4-5\

-0.7: Bad guys score.  5-2\

-0.7: @goprocks \_\'d9\'f7\_\_\'d9\'f7\_\_\'d9\'f7\_ our bad, Ellen Degeneres tomorrow???\

-0.7: Bad guys get one off a floaty huck. 10-2.\

-0.7: NE with a run through D. Then a huck. But then bad throw to try to score and a turn.\

-0.714285714286: Ellen with a sick d\

-0.714285714286: Syz got two sick ds on that point that Iowa state caught again\

-0.714285714286: Naomi sick help d\

-0.714285714286: Maddie gets a sick catch d, maybe a Callahan, called not in and then endzone turn\

-0.714285714286: Kirstie's making people fall over with her sick fakes\

-0.714285714286: Ari-kari with the sick grab and self officiating. 1-0\

-0.714285714286: Lucia-bean with that sick D\

-0.714285714286: Naomi gets a sick d on a huck\

-0.714285714286: Zo with sick cutter D\

-0.714285714286: Soter with a sick layout D\

-0.714285714286: Nariah living it up on D island with a sick hand block. Bella to Karen 2-0\

-0.75: We catch it OB :-(\

-0.75: "RT @kraynolds90: (ctd.)\

-0.78125: Game one finished! Up and up Syz!! \_\'d9\'ce\'dd\_\'d9\'ce\'aa\'89\'f7\'e3\_\'d9\'ce\_\

-0.892857142857: Katie to Ani with a sick layout! 3-1 #quickguestTweeterreturn\

-0.9375: Some chilly O and another freshman scores! Beam us up Scotty! 2-0\

-0.9765625: RT @eclipsethedisc: NATIONAL CHAMPIONSHIP game at 10:30 EST!!!! \_\'d9\_\'e4\_\'d9\_\'e4\_\'d9\'d3\'b4\_\'d9\'d3\'b4\'89\_\'c1\'95\'fc\uc0\u143 \'89\u157 \_\'95\'fc\u143 \'89\u157 \_\_\'d9\'f7\u141 \_\'d9\'ce\'a5\_\'d9\'ce\'d0\_\'d9\'ce\'d1\_\'d9\'ce\'f7\_\'d9\'ce\'d4\_\'d9\'ce\'d5\_\'d9\'ce\'d2\_\'d9\'ce\'d3\

-1.0: 6-2. I apologize for the boring Tweets. Tryna be an #engagedsideline\

-1.0: Maya's is on her way to nASty! 8-0 \_\'d9\'d4\'a9\'89\'db\uc0\u141 \_\'d9\_\_\

There a way more scores of 1 than scores of -1. There are also way more positive Tweets (33 pages) than negative Tweets (19 pages). Also, the “negative” Tweets, for the most part, are not at all negative. They simply contain a word that is stereotypically considered a negative word, however, in this specific context is quite positive. An example of this is the use of the word “sick”, as in a “sick d”. Here the term “sick” is being used akin to the word cool rather than its stereotypical meaning of being under the weather. There are also some Tweets that I simply have no idea why they are considered negative (highlighted in green). These Tweets tend to contain random letters, numbers and symbols that likely are the result of downloading a link, picture, emoji or video.

**Conclusion:**

There are a lot of other cool things I would like to have been able to look into. One thing that could have been interesting would be analyzing the timing of the Tweets. From the timing, I could have looked into typical time intervals/patterns. Maybe I could have researched whether Syzygy Tweets more or less when losing. It also would have been interesting to see how Syzygy’s Tweets compares to other Frisbee team’s Tweets. I could have looked into this by comparing frequency charts of multiple team’s tweets. On a similar note, I could have seen how women’s team’s Twitters compare to men’s and how Frisbee team’s Twitters compare to other sport team’s Twitters. It also would have interesting to analyze Ultiworld’s Tweets (Ultiworld is one of the only news site for ultimate Frisbee, they also have a Twitter) or USA ultimate’s Tweets. I also would have like to looked into tracking how mnay times a player’s name appears during specified times. This might be able to provide insight to certain players trajectories. If I had had more time I would have liked to implement different waits to account for time span covered by Tweets, how many years each person played on Syzygy, and whether the player was unable to play for significant periods of time due to injuries.

**Ethical Statement:**

Frisbee Tweets often point out success of certain players while not giving credit to the teammates who helped enable their success (from the teammate(s) who helped get the d that point to the teammate(s) who pushed them at practice and made them a better player etc..). Also, people usually only Tweet about assist, goals, and ds. There is a lot more that happens on the field that is not as visually apparent (like shut down defense or making cuts that open up the field etc.). My code further exaggerates this by creating word frequencies tables. In these tables some player’s names pop up in the top twenty while others are hidden beneath layers of other words, which is not necessarily at all representative of each player’s frisbee abilities.

Looking at segments of Tweets out of context can enable misunderstanding. For semantic analysis, Tweets are removed from their sequential ordering and reordered based on their scores. This also happens with unigrams, bigram and trigrams in my frequency tables. Also, it is important to note that there is a lot of ultimate/Syzygy specific lingo that comes up in Tweets (nicknames, inside jokes). If you are not familiar with my team’s culture or Frisbee lingo then you might misunderstand a Tweet. An example of this would be Syzygy’s frequent use of beans. Unless on Syzygy you’re not goanna have any idea what talking about when we say “bean”.

At tournaments, an injured player might be making 100 hundred Tweets in a day. Sometimes scores and other things can be confusing causing incorrect information to be Tweeted. Thus, my program might inadvertently highlight misinformation.

There are also benefits to my code. It enables teams to see if they are Tweeting a couple people way more than the rest of the team. It also might highlight if a team is over positive or negative. It can also help enable a team to realize how many weird words they use that only their teammates/people in the ultimate community might know.

I don’t think I really need to implement any safeguards or restrictions for my code. I can’t come up with any scenario where my program could cause any serious harm. I guess I would just want to warn people about how my semantic analysis code will automatically cast a Tweet to 0 if it does not recognize any of the words in the Tweet. I wish my semantic analysis code casted a Tweet to a number not between -1 and 1(the range for my sematic analysis code) because it is misleading when it casts it 0 as some sentences genuinely earn a score of 0.