**NFL Event Detection Project – Plan Book**

**The Overall Process**

* Reddit/Twitter data gathering 🡺 ETL - Data filtering/transform 🡺 Twitter/Reddit Segmentation 🡺 Bursty Segment Extraction 🡺 Bursty Segment Clustering 🡺 Event Summarization 🡺 Visualization

**Details of the Respective Steps**

1. **Reddit/Twitter Data Gathering**

* How it can be achieved?
* We can use Twitter API to gather twitter tweets, and latest Reddit data set from 732 for Reddit posts/comments.
* Application/Technology used?
* Twitter API
* Things to consider.
* What will be an exact time window that we will use for data gathering? From (game kick-off time) to (game end time + alpha)?
* How are we going to handle re-tweets? Is there a way for us to know it is a re-tweet? Are we going to ignore them as the are duplicate?

1. **ETL - Data Filtering/Trasform**

* How it can be achieved?
* We will use specific key words to filter out all the tweets and Reddit posts unrelated to the game and output the filtered dataset as a parquet.
* Application/Technology used?
* Spark DataFrame? RDDs?
* Things to consider.
* What are the keywords that we should use for filtering?
* What are the fields (information) that we will keep after ETL process? Ordering of the transformed data is required?

1. **Reddit/Tweet Segmentation**

* How it can be achieved?
* After we ensure we have all the tweets/posts relevant to the game, we will split them into non-overlapping meaningful segments. (e.g. “goal scored”, “nasty foul”, “timeout”)
* From the text, we only keep meaningful segments, removing all the distracting words/phrases or memes. (e.g. “I can do better than him”)
* We will make great use of hashtags in our event detection as it usually represent the most important word in respective posts.
* Application/Technology used?
* Spark? AWS?
* Things to consider.
* How are we going to figure out what to keep and what to remove? (e.g. using Wikipedia titles)
* How are we going to store the extracted segments? (e.g. tuple of [time, segment])
* Are we going to give more weight to those segments extracted from #hashtag?
* How are we going to segment hashtag keyword since there is no spacing? (e.g. LeaveMeAlone 🡺 [leave me alone])
* #hashtags might not be available in Reddit. Is there anything that is used just like hashtag in Reddit?

1. **Bursty Segment Extraction**

* How it can be achieved?
* Since some segments will not necessarily be useful for event detection, we will focus on “bursty” segments by filtering out those segments that were NOT clustered together.
* We will use normal distribution probability model to detect those segments that may be meaningful for our event detection. (called “Busty segment”)
* Application/Technology used?
* Spark?
* Things to consider.
* How are we going to set the threshold to determine if a certain segment is “bursty”?
* Are we going to incorporate complicated math equations defined by us? Or maybe using the existing libraries?
* Are we going to take into account a factor of “user popularity” by counting the number of their retweets since they are more trusted source of information? (e.g. segment follower count)

1. **Bursty Segment Clustering**

* How it can be achieved?
* A time window is split into multiple sub-windows (5 mins interval?) and see what are the segments most tweeted/posted on Reddit.
* There might be some segments that are closely related (e.g. “goal scored”, “(some player) rocks”), and they should be grouped as one event (to be indexed).
* Those components can be considered as candidate event clusters.
* Application/Technology used?
* Spark ML clustering?
* Things to consider.
* How do we ignore some of frequently tweeted/posted segments that are not relevant to those events that we are interested in?

1. **Event Summarization**

* How it can be achieved?
* Based on the index of the event given to each segment during step 5, we will be able to summarize a series of major events that happened during the game.
* Output all the summarized events as csv(?).
* Application/Technology used?
* Spark?
* Things to consider.
* How are we going to come up with a “predicted” time of respective events? Are we going to use MLE to make prediction?

1. **Visualization**

* How it can be achieved?
* With the data obtained from step 6, we will visualize major events with the time stamp, and along with probability of the event detected.
* Also displays a summary of events detected. (# of tweets/posts classified to be related to the event. What keywords/segments considered to be relevant?)
* Application/Technology used?
* Power BI
* Things to consider.
* What are the things we want to display? Do we want to display our results on a time line?