1 Spark Streaming with Real Time Data and Kafka

The data source used is NewsAPI. Once you create an account on NewsAPI, you get your own API key. This API key can be used to fetch news from the NewsAPI. I chose to get the top headlines in the US and send the title of each headline to a Kafka topic.

The output mode used is “complete” so that as new headlines come into the NewsAPI, these named entities are written to Kafka. If “update” is used, then only those headlines that originally came from NewsAPI will be updated and any new headlines will be ignored.

The results indicate that over a smaller amount of time, the named entities in the news headlines tend to stay the same. As more time passes, the dominating named entities in the news headlines tend to change. After 30 minutes, the same 3 named entities were the most frequent (Israel, Hamas, and NBC News) while the rest of the named entities were all tied. Thereafter, there was more variation in the named entity frequencies. After 45 minutes, more headlines with US and Week 12 started showing up. After 75 minutes and thereafter, the rest of the top named entities overtook the frequency of NBC News. This shows that maybe the headlines containing NBC News became “old” news as more “new” news began to populate the NewsAPI. Throughout the time, the other 9 named entities continued to increase in headline frequency. An interesting note is that television/sports broadcasting companies tend to occur frequently in headlines (e.g. NBC News, Reuters, and NFL appeared in the top 10 but others like CNBC and CNN appeared further down in the count too).

2 Analyzing Social Networks using GraphX/GraphFrame

The chosen dataset is a who-trust-whom online social network of a general consumer review site Epinions.com (https://snap.stanford.edu/data/soc-Epinions1.html). Members of the site can decide whether to "trust" each other. All the trust relationships interact and form the social network graph.

The degrees operation returns the count of all edges of each vertex, so inDegrees counts only incoming edges, and outDegrees counts only outgoing edges. Node 645 has the highest outdegree so that person trusts the most people. Node 18 has the highest indegree so that person is the most trusted.

The idea of PageRank is that each incoming edge represents an endorsement and makes the vertex more relevant in the given graph. In this social network, if a person is trusted by various people, he or she will be ranked highly. The parameter maxIter is set to 3. Even though a higher number is recommended, the algorithm takes too long to finish if set any higher and the results make sense anyways. Correspondingly, node 18 has the highest indegree and the highest PagePank, making this node the most relevant in the graph. Nodes 737, 790, and 136 also have both indegrees and PageRank values in the top 5.

The connected components algorithm computes the connected component membership of each vertex and finds isolated clusters or isolated sub-graphs. These clusters are sets of connected vertices in a graph where each vertex is reachable from any other vertex in the same set. The algorithm returns each vertex and the component (as an auto-generated id) to which it is connected. After grouping by component and sorting by count, the top 5 components with the largest number of nodes are found. Component 0 contains almost all the nodes in the graph, but there are other components that connect a small number of nodes as well (less than 8 nodes). This social network has one huge cluster that connects most of the nodes and many other smaller clusters, which makes sense for a social network to have multiple mutual connections that cover a large span of people.

Triangle count computes the number of triangles passing through each vertex and is commonly used as community detection in a social network graph. A triangle is a set of three vertices, where each vertex has a relationship to the other two vertices in the triangle. In a social network community, it is easy to find a considerable number of triangles connected to each other. The nodes with the highest triangle count correspond to nodes with highest outdegree (645, 634) or highest indegree (18, 737), which makes sense since high degrees means those nodes are connected to many other nodes.