Assignment 5

Ishaq Khan

INFS 6353

11/28/2023

1. Collect Twitter retweet data based on one or more keyword(s) of your choice. Then perform the following on the data:

Keyword #1: Formula 1

The keyword that was selected for the first set was “Formula 1”.

This keyword will be collected from X(Twitter) to perform the following:

1. This is the retweet network with the Formula 1 Retweet Network

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

This is the retweet network with the default Fruchterman Reingold Layout.

A diagram of a circle with arrows and points

Description automatically generated with medium confidence

This is the retweet network but with the default layout and node labels.

A black and white circle with blue text

Description automatically generated with medium confidence

1. This is the retweet network using the Random Layout. A network of lines and dots

   Description automatically generated

This is the retweet network using the Circular Layout.

A blue circle with black lines

Description automatically generated

1. This is the in-degree, out-degree, betweenness, closeness, and Egenvector centralities of each vertex calculated.

A screen shot of a computer

Description automatically generated

A screen shot of a computer

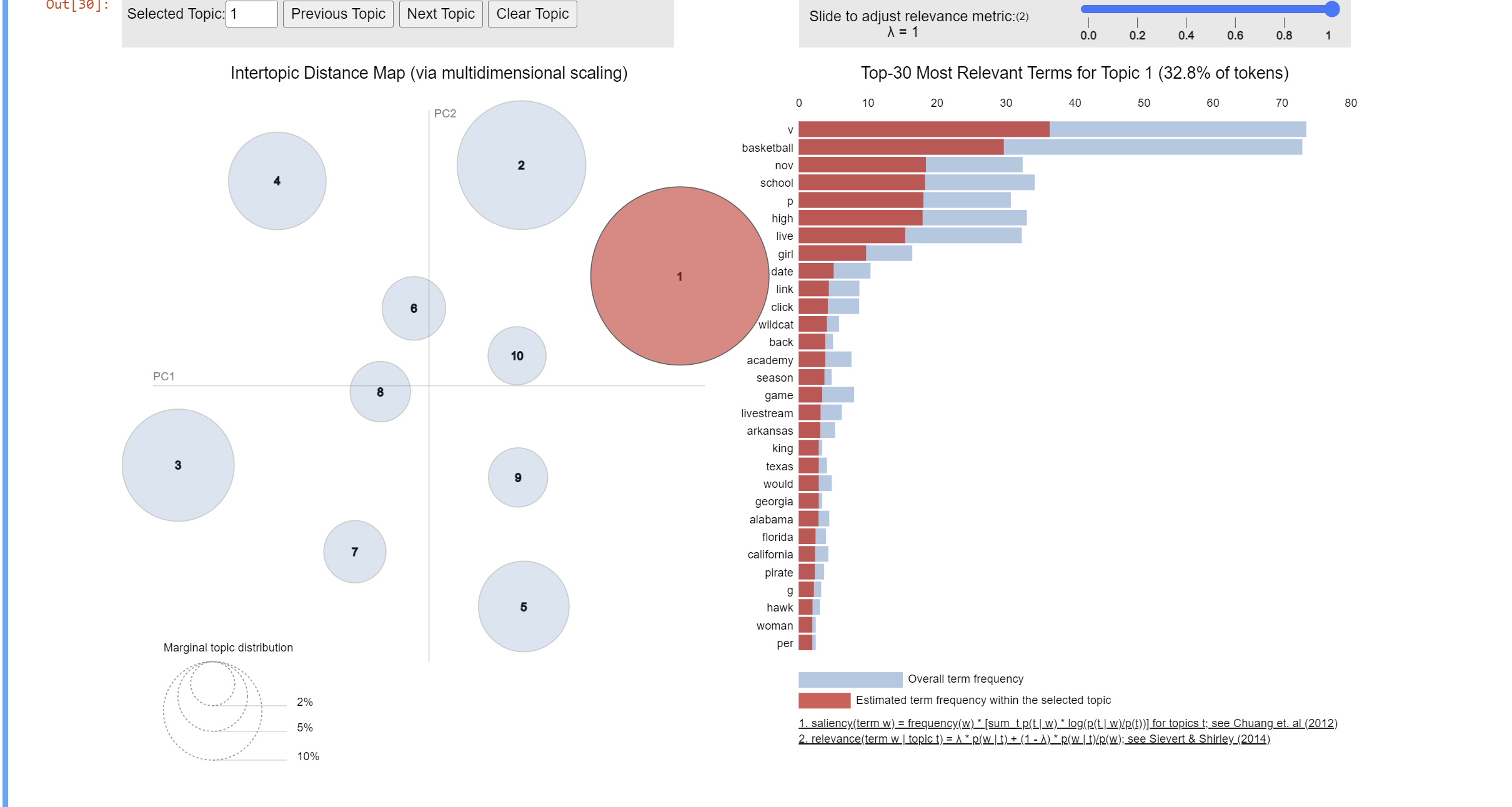
Description automatically generated



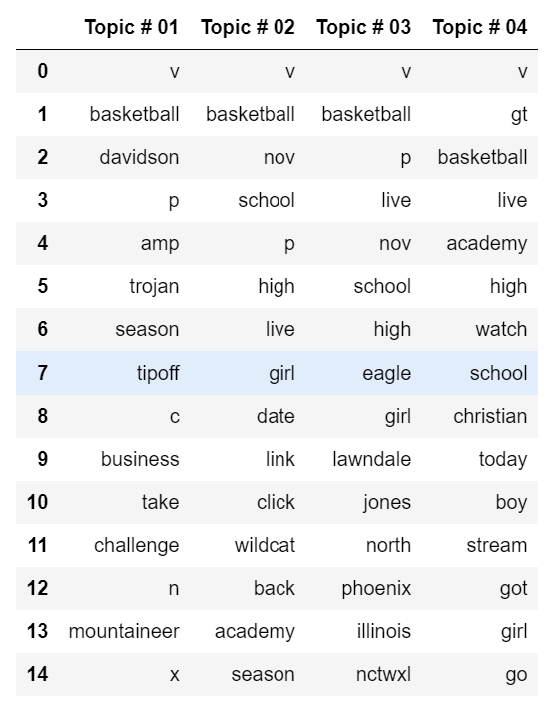
1. Collect original tweet data based on one or more keyword(s) of your choice. Then perform the following on the data:
   1. Case 1: The keyword that was used was “Basketball.” The topic number that was selected was 10 and the 10 top words for each topic.



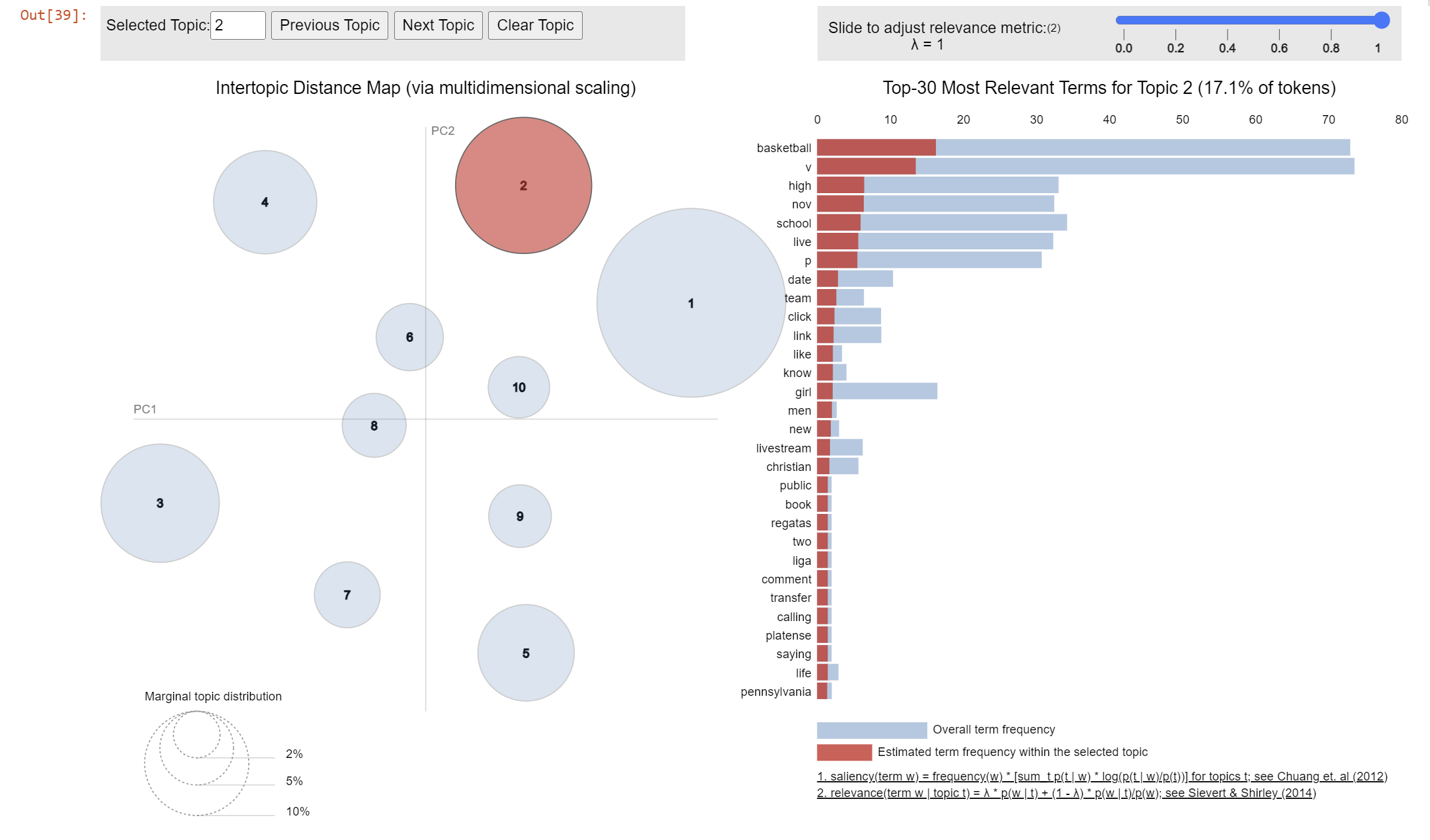
* 1. This is the topic modeling on the tweets and we used LDA and visualized the results down below. Topic 1’s bubble is the largest since it represents how many tweets in our dataset discuss the topic. In topic 1, “v” had the highest frequency but basketball was the second highest.



* 1. Case 2: The keyword that was used was “Basketball.” The topic number that was selected was 4 and the 15 top words for each topic.



* 1. This is the topic modeling on the tweets and we used LDA and visualized the results down below. Topic 2’s bubble is the 2nd largest since it represents how many tweets in our dataset discuss the topic. In topic 2, “basketball” had the highest frequency but “v” was the second highest, followed by “high”.



* 1. I think less topic numbers but higher top words are the better fit for the dataset, at least for the keywork “Basketball”. If you have a larger number of topics, then its possible you get repeated words and singular letters that you cannot make an inference on. But if you minimize the number of topics and increase the top words, it allows for better range of data to work with without potentially having repeated words and singular letters, as show in case 2.