**Assignment 3 Sheet**

**[include this sheet in your submission package]**

Name: Miguel Curiel

**Due 11:59 PM 11/02 (Wed.) - start working on it now!**

**Important points**

* This is an independent assignment. But you can work in groups to troubleshoot;
* The deadline can be extended. But you must contact me in advance to request an extension. The deadline will NOT be extended unless you show progress in completing the assignment.
* All codes you need to complete this assignment can be found in tutorials.

**How and what to submit**

Submission through Moodle. Your submission package should include this answer sheet.

**Task 1: Collect tweets based on a hashtag of your choice (set n = 2000. It's okay if the final data frame contains fewer than 1,000 tweets).**

Enter the Twitter hashtag used in your collection: #kanyewest

**Task 2: Follow the tutorial to show the top words ranked by TF-IDF. Copy and paste the output below.**

**Graphical user interface, text, application

Description automatically generated**

**Task 3: Based on the top words ranked by TF-IDF, what can you learn about the general themes in the collected tweets?**

|  |
| --- |
| **Purely from the TF-IDF, I can infer the hashtag #kanyewest seems to be involved in US sociopolitics (#usa, #news, #court). There also seems to be a relationship with other celebrities and social themes/organizations (#steve, #harvey, #movie, #nfl).** |

**Task 4: Follow the tutorial to create a semantic network based on hashtag co-occurrence. The semantic network should include top 50 hashtags and relationships among the 50 hashtags. Copy and paste the output below.**

**Diagram, radar chart

Description automatically generated**

**Task 5: Please try to interpret the semantic network. What does it reveal about the general themes in the collected tweets?**

|  |
| --- |
| **The hashtag #kanywest seems to be associated to three clusters: One regarding the migration/humanitarian crisis in Central America, the other regarding the artist’s antisemitic comment, and the last regarding his association to other pop-culture celebrities and organizations.** |

**Task 6: Please create topic models based on the collected tweets. Try 3 different versions of *k*. Copy and paste below the top words from each topic from each model.**

**Graphical user interface

Description automatically generated with low confidence**

**Table

Description automatically generated with low confidence**

**Table

Description automatically generated**

**Task 7: From the previous step, you will end up with three different topic models (based on the 3 different k), pick one of the topic models, inspect the result, and explain why the model is a fit or a misfit.**

|  |
| --- |
| **Using k = 10 as the point of reference, this model seems to be a misfit. While most groupings make sense, not all of them provide value as the categories are too broad or are missing context for them to be properly understood. For example, while words like "Belize", "latinos" and "maya" are grouped together, they make sense, and they can be analyzed, there is another group which seems to correspond to monosyllabic Spanish words ("de", "la", "en", "que"), hence not providing much value or context. Or while, on one hand, there are a couple of groupings that refer to Kanye's antisemitic comments, there are also a couple that relate the artist to other public figures (such as Elon Musk or Steve Harvey).** |