CS777 – Week 2 Homework Submission Template

**Task 1 – Generate the Top 20K dictionary and Create the TF-IDF Array (4 Points)**

Get the top 20,000 words in a local array and sort them based on the frequency of words. In the end, produce an RDD that includes the docID as key and a NumPy array for the position of each word in the top 20K dictionary:

(docID, [dictionaryPos1, dictionaryPos2, dictionaryPos3...])

* In your code print out print(allDocsAsNumpyArrays.take(3)).

| *Your output should go here.* |
| --- |

* In your code print out print(allDocsAsNumpyArraysTFidf.take(2)):

| *Your output should go here.* |
| --- |

* Include the relevant code excerpt that you used for creating an RDD that satisfies the above conditions

| *The code excerpt should go here.* |
| --- |

**Task 2 – Implement the getPrediction function (8 Points)**

Print out the results for the following queries:

* print(getPrediction('Sport Basketball Volleyball Soccer', 10))

| *Your output should go here.* |
| --- |

* print(getPrediction('What is the capital city of Australia?', 10))

| *Your output should go here.* |
| --- |

* print(getPrediction('How many goals Vancouver score last year?', 10))

| *Your output should go here.* |
| --- |

* Include the relevant code excerpt that shows how you implemented the getPrediction() function

| *The code excerpt should go here.* |
| --- |

**Task 3 – Using Dataframes (6 points)**

**Task 3.1**

Use Spark Dataframe to provide summary statistics (max, average, median, std) about the

number of Wikipedia categories that are used for Wikipedia pages. Print the results on the output

console, or store them on the cloud storage.

| *Your output should go here.* |
| --- |

**Task 3.2**

Use Spark Dataframe to find the top 10 most used Wikipedia categories. Print the results on the output console, or store them on the cloud storage.

| *Your output should go here.* |
| --- |

**Task 4 – Removing Stop Words, do Stemming and redo task 2 (2 points)**

**Task 4.1 – Remove Stop Words (1 point)**

Describe if removing the English Stop words (most common words like ”a, the, is, are, i, you, ...”) would change the final kNN results here.

You do not need to implement this task, only discuss your expected outcome results.

| *Discuss your reasoning here.* |
| --- |

**Task 4.2 – Do English word stemming (1 point)**

We can stem the words [”game”,”gaming”,”gamed”,”games”] to their root word ”game”. Read more about stemming here <https://en.wikipedia.org/wiki/Stemming>

You do not need to implement this task, only discuss your expected outcome results.

| *Discuss your reasoning here.* |
| --- |

**Spark History Output:**

**Task 1:**

| *Your screen capture should go here.* |
| --- |

**Task 2:**

| *Your screen capture should go here.* |
| --- |

**Task 3:**

| *Your screen capture should go here.* |
| --- |