**Assignment 2**

**Task 4**

4.1

My expectation is that removing stop words will not drastically change the result of the KNN prediction.

The reason is that the implementation in task 2 uses TF-IDF values – which actually gives lesser weightage to most common words anyways (like stop words) – so removing stop words altogether might not change the results significantly.

4.2

Stemming might improve / change the results of the KNN prediction.

Currently the words such as [”game”,”gaming”,”gamed”,”games”], each would have its own tf-idf value and weightage. But in terms of categorization they represent the same context. So, calculating a tf-idf value for a stemmed word such as game will more accurately represent the frequency of the words and hence improve KNN classification.

**Task 3**

Here is the output from task 3. It took 51 minutes.

Count of categories: 1763328

\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Average number of categories is: [Row(avg(count)=79.943160886687)]

\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Minimum number of categories is: [Row(min(count)=1)]

\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Maximum number of categories is: [Row(max(count)=2586023)]

\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Top Categories are:

[Row(cat='Noindexed\_pages', count=2586023), Row(cat='All\_stub\_articles', count=2243344), Row(cat='WikiProject\_Biography\_articles', count=1650535), Row(cat='Articles\_with\_short\_description', count=1516225), Row(cat='Redirects\_from\_moves', count=1489008), Row(cat='Unprintworthy\_redirects', count=1412217), Row(cat='Coordinates\_on\_Wikidata', count=1048235), Row(cat='Biography\_articles\_of\_living\_people', count=970853), Row(cat='Stub-Class\_biography\_articles', count=939454), Row(cat='Living\_people', count=938708)]

Spark History

Graphical user interface, table, Excel

Description automatically generated

**Task 2**

**allDocsAsNumpyArrays.take(3)**

[('44405232', array([0.10621242, 0.02004008, 0.03006012, ..., 0. , 0. ,

0. ])), ('44132170', array([0.05460751, 0.02540766, 0.02313235, ..., 0. , 0. ,

0.00037922])), ('44007727', array([0.11340206, 0.0257732 , 0.03092784, ..., 0. , 0. ,

0. ]))]

**allDocsAsNumpyArraysTFidf.take(2)**

[('23543157', array([0.01212293, 0.0107774 , 0.00408303, ..., 0. , 0. ,

0. ])), ('23508196', array([0.01025712, 0.00967064, 0.00388791, ..., 0. , 0. ,

0. ]))]

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

[('Articles\_using\_small\_message\_boxes', 1), ('EuroLeague\_team\_past\_rosters', 1), ('1927\_in\_basketball', 1), ('1996\_in\_basketball', 1), ('Articles\_to\_be\_expanded\_from\_August\_2014', 1), ('All\_articles\_with\_empty\_sections', 1), ('Maccabi\_Tel\_Aviv\_B.C.\_players', 1), ('All\_articles\_to\_be\_expanded', 1), ('1939\_in\_basketball', 1), ('Articles\_with\_empty\_sections\_from\_August\_2014', 1)]

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

[('Articles\_using\_small\_message\_boxes', 1), ('EuroLeague\_team\_past\_rosters', 1), ('1927\_in\_basketball', 1), ('All\_articles\_with\_empty\_sections', 1), ('Articles\_to\_be\_expanded\_from\_August\_2014', 1), ('1996\_in\_basketball', 1), ('Maccabi\_Tel\_Aviv\_B.C.\_players', 1), ('All\_articles\_to\_be\_expanded', 1), ('1939\_in\_basketball', 1), ('Articles\_with\_empty\_sections\_from\_August\_2014', 1)]

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

[('All\_disambiguation\_pages', 2), ('Disambiguation\_pages\_with\_short\_description', 2), ('Disambiguation\_pages', 2), ('All\_article\_disambiguation\_pages', 2), ('2004–05\_in\_Central\_American\_football\_leagues', 1), ('2004–05\_in\_Honduran\_football', 1)]

**Spark History**

Table

Description automatically generated