**Assignment I – Data Type Portability: Discretization**

* Screenshot of the complete workflow:

**A computer screen shot of a diagram

Description automatically generated**

* Discretization step

**A screenshot of a computer

Description automatically generated**

* Resulting data

**A screenshot of a computer

Description automatically generated**

* Top 10 countries

A screenshot of a computer

Description automatically generated

**Assignment II – Pre-Processing of Text Data**

1. a)  First, remove all preprocessing steps except “Transformation” and transform the corpus into lowercase. This avoids to have multiple versions (“Large” and “large”) of the same word. If you checked the box “Apply Automatically” on the lower left, the changes are directly ap- plied. You can use the “Word Cloud” widget to observe the effect of the transformation. Please note that also punctuation symbols are also counted. Save the word cloud clicking on the disk symbol.

A close-up of a text

Description automatically generated

Fig: Word Cloud – Before lower case conversion

We can see that one of the most frequent words are ‘the’ and ‘THE’.

A close-up of words

Description automatically generated

Fig: Word Cloud – After lower case conversion

Now we don’t have multiple versions of same word.

1. b)  Now apply “Tokenization”. This is used to break the text into smaller pieces like sentences, words, ... Please have a look at the aforementioned documentation to understand the different tokenization methods. Use Regular Expressions to split the text by words without keeping punctuation. This is quite a common way to break down text. What are the top 3 words? Save the word cloud clicking on the disk symbol.

A close-up of words

Description automatically generated

Fig: Tokenization without keeping punctuations

The top 3 words are ‘the’, ‘of’ and ‘and’.

1.c) Now filter the stopwords. What are the top 3 words now? Save the word cloud clicking on the disk symbol. A close-up of a text

Description automatically generated Fig: Word Cloud – Stop words filtered

Now the top 3 words are ‘government’, ‘states’ and ‘congress’.

1. d) Apply the standard normalization (Porter Stemmer). This does, to put it simple, convert words to their base form, like e.g. “the boy's cars are different colors” à “the boy car be differ color”. What do you observe?

A close-up of words

Description automatically generated

Fig: Word Cloud – After Normalization

Now we observe that word are converted to their base forms for ex: ‘government’ is converted to ‘govern’.

A screenshot of a computer

Description automatically generated

Fig: Complete Workflow