Text- Based Data Visualization

**INTRODUCTION**

***WordCloud in Python***

WordCloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites. For generating word cloud in Python, modules needed are – matplotlib, pandas and WordCloud.

***Advantages of WordCloud***

1. Analyzing customer and employee feedback.
2. Identifying new SEO keywords to target.

***NLTK, Stemming and Lemmatization***

Natural Language Processing (NLP) is a field that focuses on making natural human language usable by computer programs. **NLTK** is a Python package that you can use for NLP.

A lot of the data that you could be analyzing is unstructured data and contains human-readable text. Before analyzing that data programmatically, we first need to preprocess it.

**Stemming** is a text processing task in which you reduce words to their root, which is the core part of a word. For example, the words “helping” and “helper” share the root “help.” Stemming allows you to zero in on the basic meaning of a word rather than all the details of how it’s being used. The stemmer used here is Porter Stemmer.  Like stemming, **lemmatizing** reduces words to their core meaning, but it will give you a complete English word that makes sense on its own instead of just a fragment of a word.

**ABOUT THE PROJECT**

This project has two parts:

1. A text data visual using WordCloud. Wordcloud is useful for quickly perceiving the most prominent terms and for locating a term alphabetically to determine its relative prominence. It is widely used in media and well understood by the public. A small text sample is inputted and then code is written to produce a data visual.
2. The next part involves using several NLP libraries to explore and analyze the various word strings that were found in a given sample of text that was derived from Twitter and several news feeds in German, English, Russian and Finnish. NLTK libraries were implemented in order to visualize and plot the frequency of words with their usage in the given samples.

The program reads a file that contains text samples from Twitter and other such sources and tries to split the whole document into multiple strings, counts it and displays the frequencies of the words so far that has occurred.

Using stemming and lemmatization in order respectively, it later splits the words according to the principles mentioned and visualizes the data accordingly. This helps us understand the correlation of text and visualization and the concept of text classification and visualization.

Files of the text visualization of all the 4 languages are attached with this document.

