Applied Data Science with Python

Labwork #3

Text Mining:

- Create a random text (or copy any text paragraph) and import it as variable into a Python program
- Separate the text into a collection of words.
- Count vocabulary of words and display the word frequency. Display the most frequent words in the text
- Use functions available in Pandas and convert words into numerical vector. Is the data discret or continuous?
  i.e: try the function word2vec in *gensim* package
- How to calculate a distance between two words? Or propose an appropriate distance for a pair of words.
- Apply k-means on the numerical data.
- Visualize the obtained results and calculate the clustering quality criteria (if possible).
- Try to test all these steps for different texts