Natural Language Processing Case:
Problem Clustering for Codekicker.de

Codekicker.de, the German copy of stackoverflow, wants to automatically suggest related tickets based on the problem description in the ticket body. For this they ask you to apply a clustering algorithm to find patterns in their database.

Functional Requirements:

- Clustering of attached dataset ("unclustered_input.txt") into 5 clusters
- Computation of "recall" and "precision" for every cluster with comparison to the 5 original clusters ("cluster_x.txt")
- Analysis of the results in written form with the following content:
    - short summary of the method used and why it was chosen
    - short evaluation of successes and failures of the method, assumption why results are this way

Non-Functional Requirements:

- Only open-source libraries/technologies are used
- Code is documented and „clean“ in terms of common best practices for structure and naming conventions (doesn't matter which convention, as long as consistency is evident)
- Input data is read from input txt files, which can be modified (e.g. adding lines) without breaking the algorithm
- The algorithm should not be "overtrained", i.e. it should still deliver similar results after slight variations of the input dataset
- Output data is displayed in a readable way
- All results are provided as zip-package to „jobs@frag-aaron.de“

Additional Information:

- Any programming language is fine, e.g. Python
- We recommend the simple approach of using frequency counting of the provided tags (marked yellow in the overview excel) to assign the input sentences to a certain cluster. Methods like stemming or lemmatization will ensure independence of declination/conjugation. A python library which implements them for German language is ParZu (https://github.com/rsennrich/parzu)
- Any other NLP method or even machine learning method (e.g. k-means) is acceptable, as long as it is explained why it was chosen over the method suggested above
- If you have any questions concerning this code, please don’t hesitate to send them to „jobs@frag-aaron.de“