LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Creating Clean Multilingual Corpora For Low-resource Languages

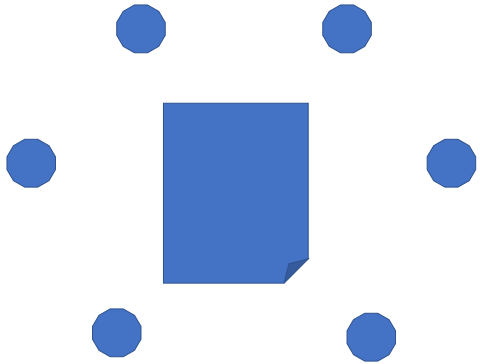
Ongoing

Amir Hossein Kargaran
amir@cis.lmu.de

October 7, 2023

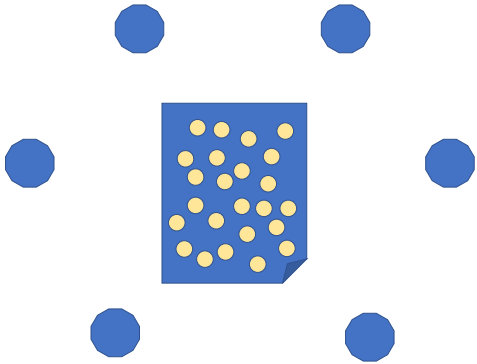
2) How to check the Homogeneity of corpus?

Iterative Clustering



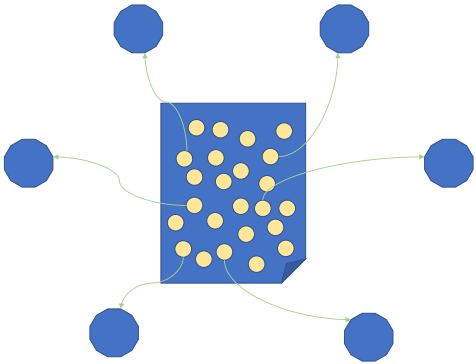
2) How to check the Homogeneity of corpus?

Iterative Clustering



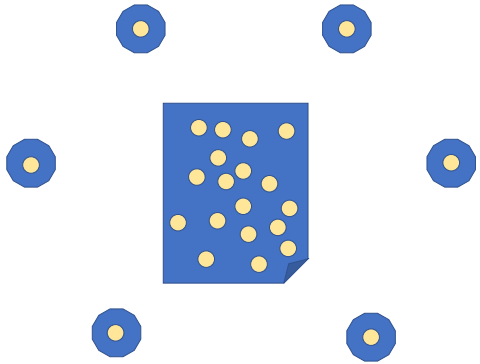
2) How to check the Homogeneity of corpus?

Iterative Clustering



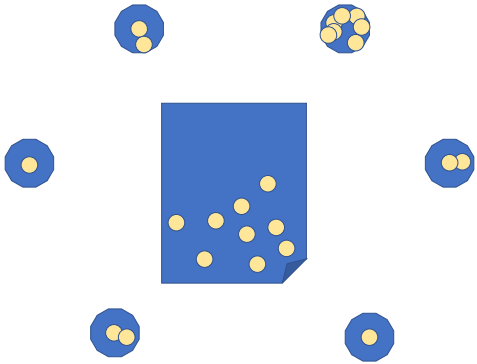
2) How to check the Homogeneity of corpus?

Iterative Clustering



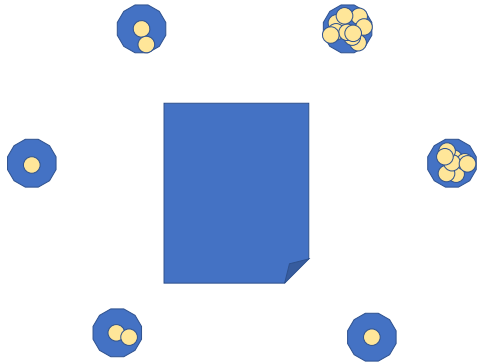
2) How to check the Homogeneity of corpus?

Iterative Clustering



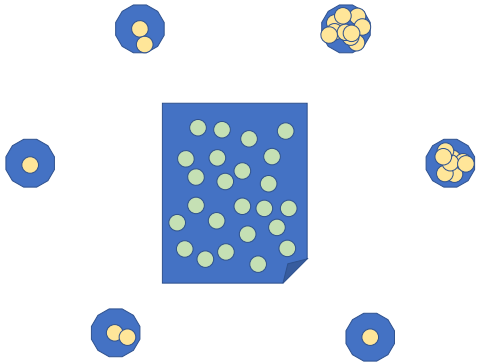
2) How to check the Homogeneity of corpus?

Iterative Clustering



2) How to check the Homogeneity of corpus?

Iterative Clustering



2) How to check the Homogeneity of corpus?

Iterative Clustering Limitations

- Poor seed selection $\rightarrow$ Poor result
- May not work on close languages
- An incorrectly selected sentence will bring incorrect sentences in the next step