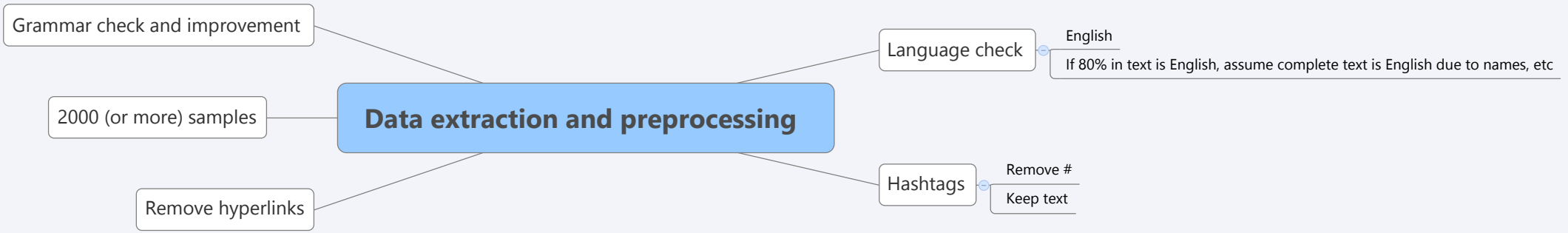
Grammar check and improvement
2000 (or more) samples
Remove hyperlinks
Data extraction and preprocessing
Language check
English
If 80% in text is English, assume complete text is English due to names, etc
Hashtags
Remove #
Keep text