XML

XML steht für eXtensible Markup Language und ist eine beschreibende Auszeichnungssprache, mit der Bedeutungen innerhalb der Texte codiert werden können. XML wird verwendet, um Dateien in einer strukturierten, aber auch menschenlesbaren Form zu speichern. Weitere Formate um Text- und Metadaten strukturiert zu speichern sind zum Beispiel CSV und JSON.

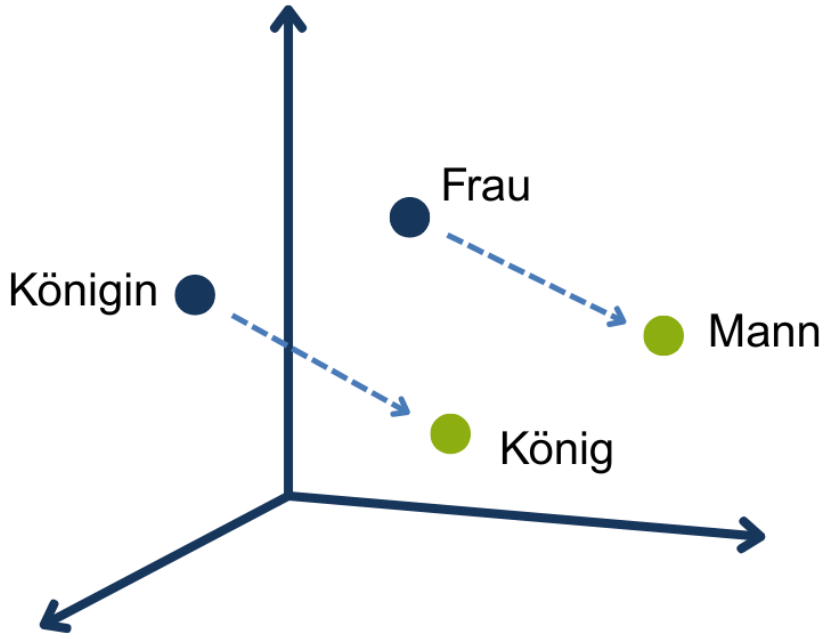
Im Rahmen der Linie „Freiraum 2023“ gefördert durch die Stiftung Innovation in der Hochschullehre

Taggen

Taggen bezeichnet das Kennzeichnen von Wörtern. Das Wort stammt vom englischen Wort *tag* ab, was so viel bedeutet wie Etikett. So annotiert, können die Texte schnell nach den gewollten Informationen, zum Beispiel alle Fremdwörter, durchsucht werden.

Im Rahmen der Linie „Freiraum 2023“ gefördert durch die Stiftung Innovation in der Hochschullehre

Embeddings



Word embeddings sind Vektorrepräsentationen von Wörtern, also pro Wort jeweils eine Reihe von Zahlen, die die Bedeutung repräsentieren sollen. Damit kann man das Wort in einem riesigen Koordinatensystem verorten und numerisch repräsentieren. Die Zahlen werden so berechnet, dass ähnliche Wörter im Koordinatensystem näher zusammenstehen als unähnliche. Dies wird natürlich nicht händisch gemacht, sondern mit Algorithmen, die Wörter danach bewerten, in welchen Kontexten sie auftreten.

Im Rahmen der Linie „Freiraum 2023“ gefördert durch die Stiftung Innovation in der Hochschullehre

Linked Open Data

Linked Open Data bezeichnet frei verfügbare, strukturierte Daten, die im Internet über eindeutige Identifikatoren (URIs) veröffentlicht, maschinenlesbar gemacht und so miteinander verknüpft werden, dass sie über Organisations- und Themengrenzen hinweg offen nutzbar sind und komplexe Zusammenhänge sichtbar machen (Paderta, 2024). Ein Beispiel ist Wikidata, auf dem beispielsweise Wikipedia aufbaut.

Im Rahmen der Linie „Freiraum 2023“ gefördert durch die Stiftung Innovation in der Hochschullehre

Textkorpora

Textkorpora (Plural von „das Korpus”) sind systematische Sammlungen von Texten, die auf verschiedene Weise miteinander verbunden sind. Entscheidend ist hierbei, was untersucht werden soll. Denkbar sind unter anderem Sammlungen nach Autor*in, Genre, Epoche oder auch sprachlichen Merkmalen, wie Dialekten und vieles mehr.
Textkorpora werden typischerweise für die computergestützte Analyse aufbereitet, d.h. die enthaltenen Texte sind maschinenlesbar.

Im Rahmen der Linie „Freiraum 2023“ gefördert durch die Stiftung Innovation in der Hochschullehre

Metadaten

Metadaten sind „Daten über Daten“: Sie enthalten Informationen über andere Daten.
Das können technische Infos wie der Dateityp, administrative Infos zur Lizenzierung oder beschreibende Infos wie Schlagworte, Autor*in und Titel sein.
Durch das Hinterlegen von Schlagworten und anderen beschreibenden Infos werden Daten auffindbar und durchsuchbar.

Im Rahmen der Linie „Freiraum 2023“ gefördert durch die Stiftung Innovation in der Hochschullehre