Track 3: Clustering Technology for Policy and Finance

Team: P. Clusterers

Team Members: Elsy Milan, Matthew Arran, Victor Rosenberg, Nouha Samlani
Challenge Organizer: David Dekker

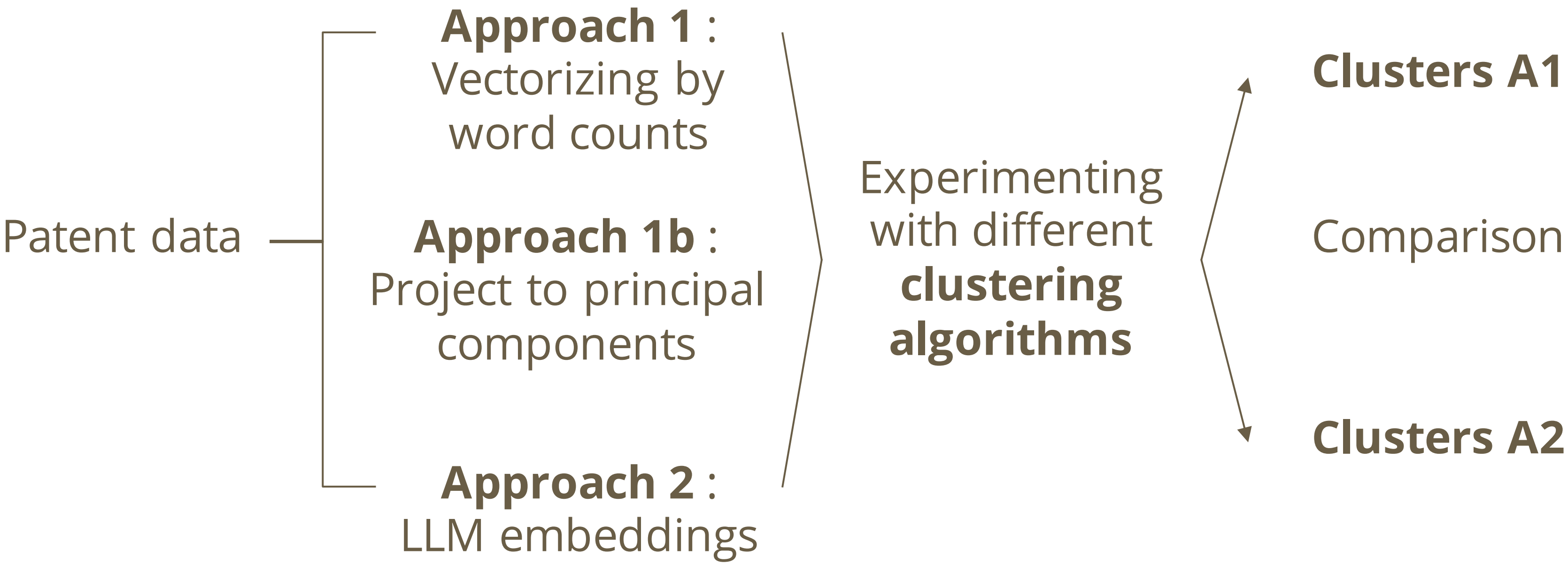
Goals:

⇒ Find technology clusters from patent data
⇒ Compare between a simple approach and LLM
⇒ Compare different clustering algorithms
⇒ Assess the quality of these clusters and Interpret the clusters

Challenges:

⇒ Learn about word vectorizing and embeddings
⇒ LLM computational time is long which limit the options could try

Methodology



Patent data

Approach 1 :
Vectorizing by
word counts

Approach 1b :
Project to principal
components

Approach 2 :
LLM embeddings

Experimenting
with different
clustering
algorithms

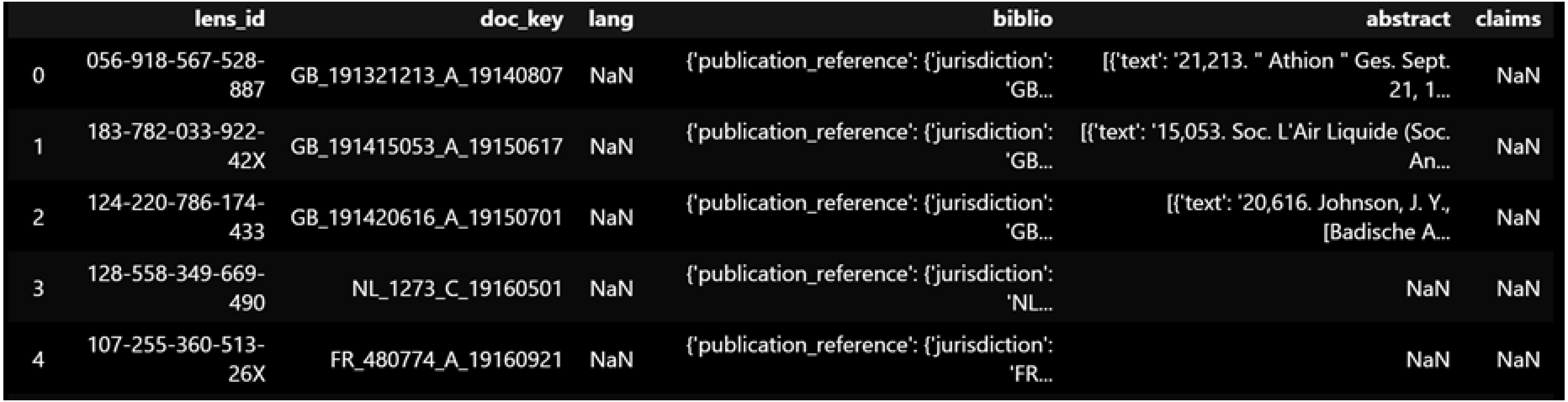
Clusters A1

Comparison

Clusters A2

⇒ Do we need LLMS ?

Data Exploration - Original Dataset

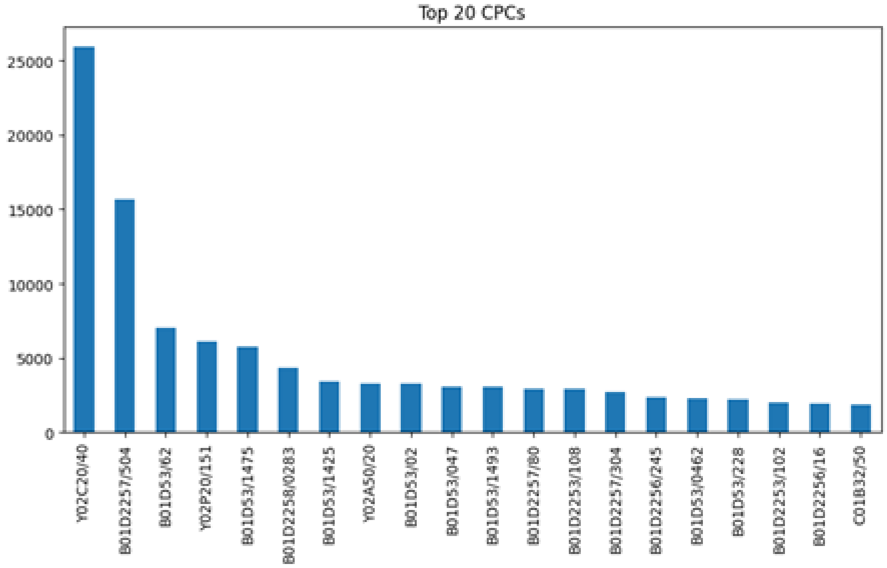


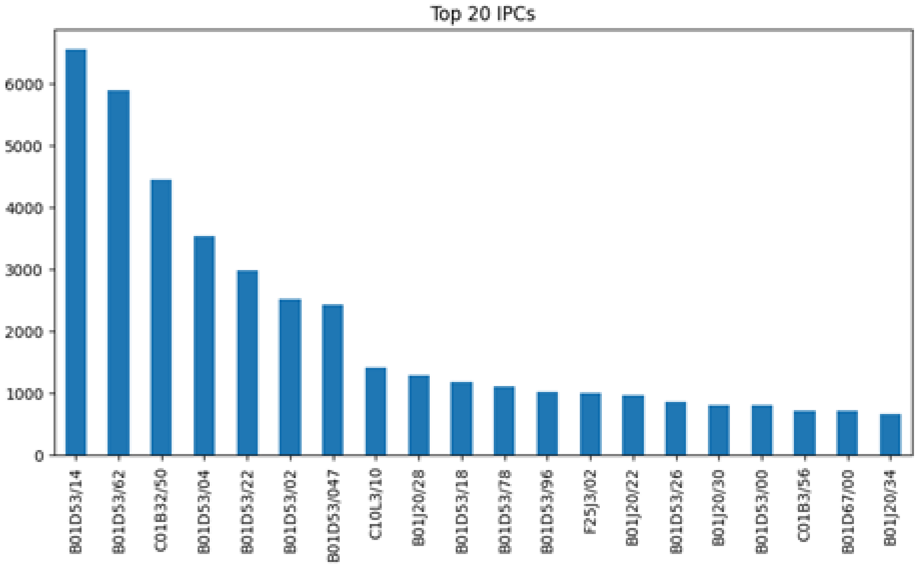
	lens_id	doc_key	lang	biblio	abstract	claims
0	056-918-567-528-887	GB_191321213_A_19140807	NaN	{'publication_reference': {'jurisdiction': 'GB...	[{'text': '21,213. " Athion " Ges. Sept. 21, 1...	NaN
1	183-782-033-922-42X	GB_191415053_A_19150617	NaN	{'publication_reference': {'jurisdiction': 'GB...	[{'text': '15,053. Soc. L'Air Liquide (Soc. An...	NaN
2	124-220-786-174-433	GB_191420616_A_19150701	NaN	{'publication_reference': {'jurisdiction': 'GB...	[{'text': '20,616. Johnson, J. Y., [Badische A...	NaN
3	128-558-349-669-490	NL_1273_C_19160501	NaN	{'publication_reference': {'jurisdiction': 'NL...	NaN	NaN
4	107-255-360-513-26X	FR_480774_A_19160921	NaN	{'publication_reference': {'jurisdiction': 'FR...	NaN	NaN

⇒ 70 525 patent
⇒ 52 510 with abstract
⇒ 25 945 with abstract in English

Features of the Patents

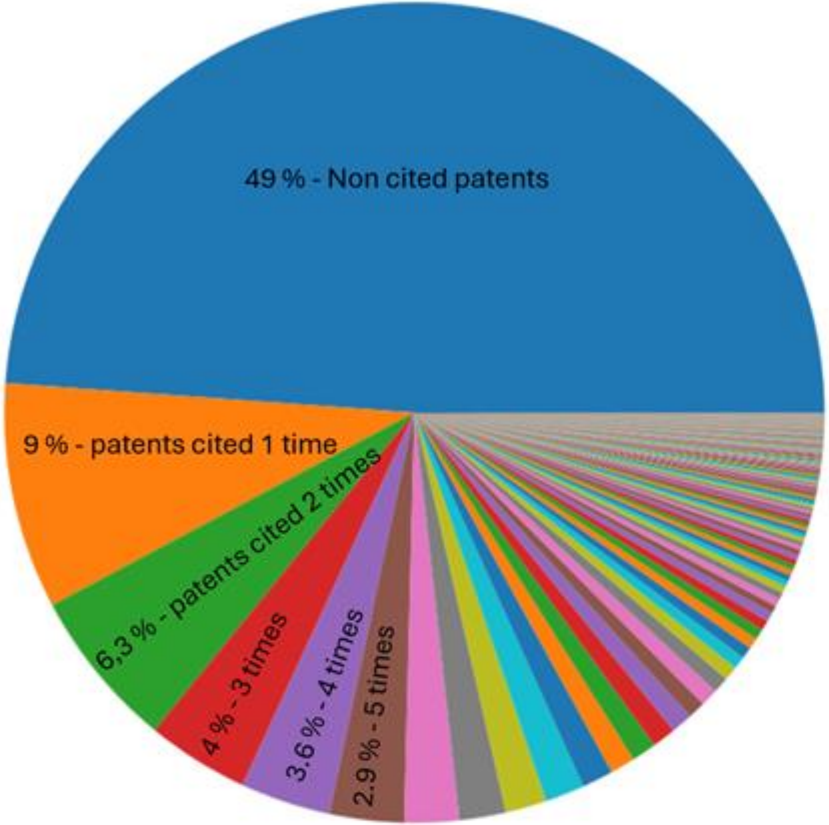
⇒ Classifications of Patents


Top 20 CPCs

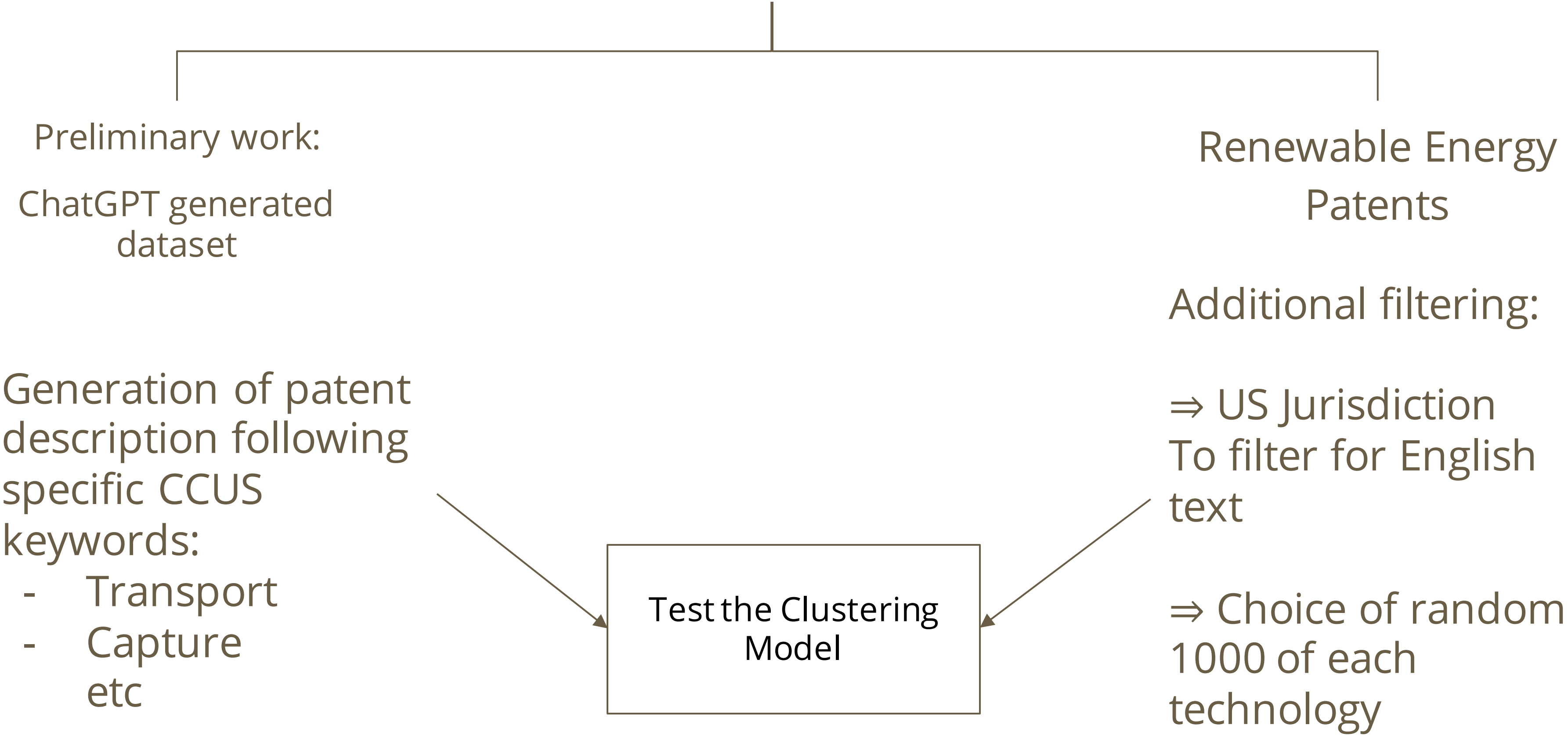

Top 20 IPCs

Features of the Patents

⇒ Citation of patents



Validation Datasets



Preliminary work:

ChatGPT generated dataset

Generation of patent description following specific CCUS keywords:

- Transport
- Capture

etc

Test the Clustering Model

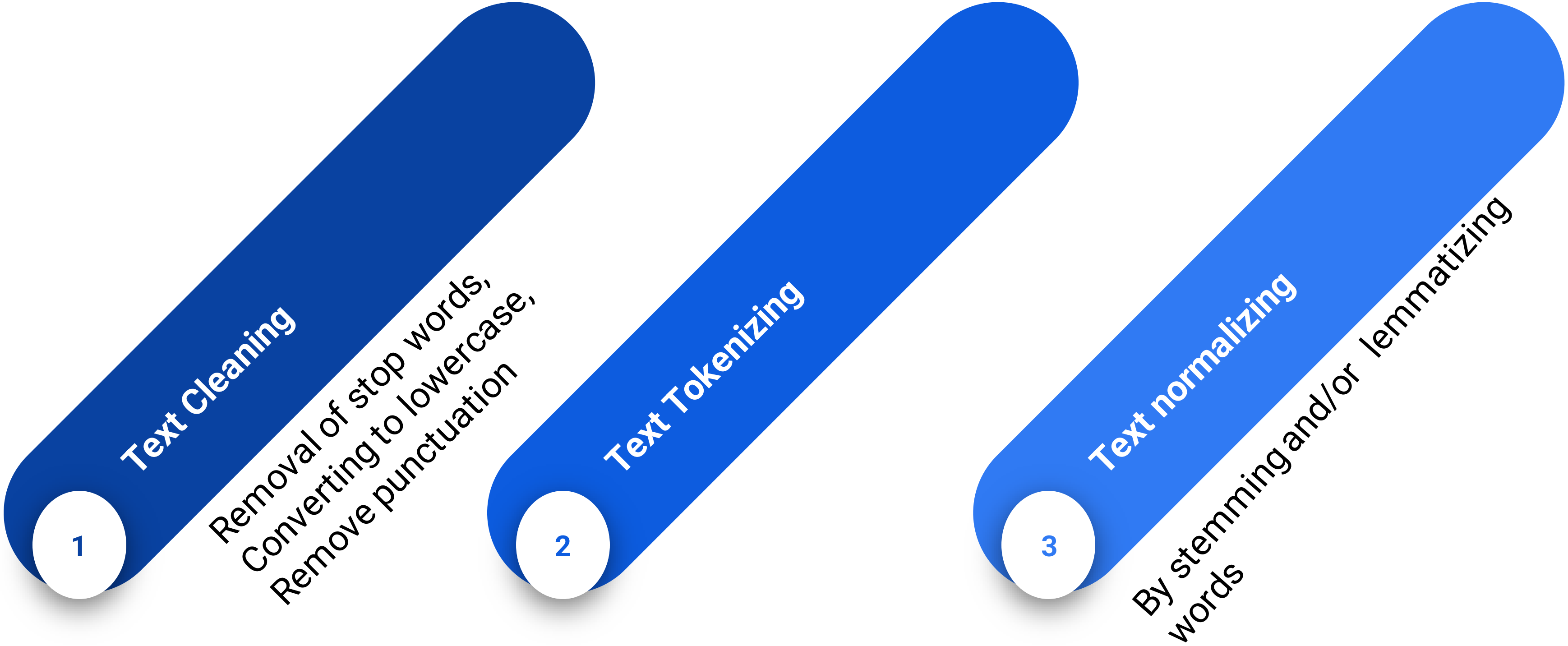
Renewable Energy Patents

Additional filtering:

⇒ US Jurisdiction
To filter for English text

⇒ Choice of random 1000 of each technology

Approach 1: Pre-processing



1 Text Cleaning

Removal of stop words,
Converting to lowercase,
Remove punctuation

2 Text Tokenizing

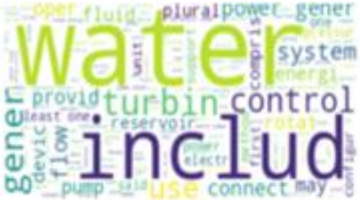
3 Text normalizing

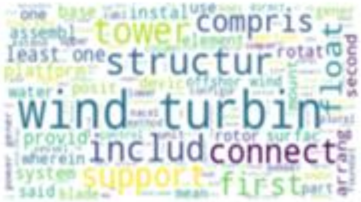
By stemming and/or lemmatizing words

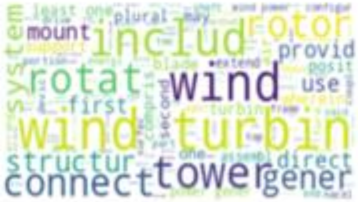
Preprocessed validation data

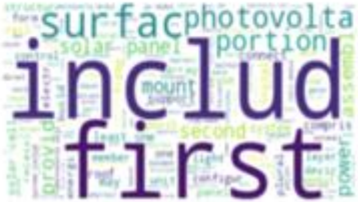
Hydroelectric: Offshore wind: Onshore wind: Solar PV:

Stems:

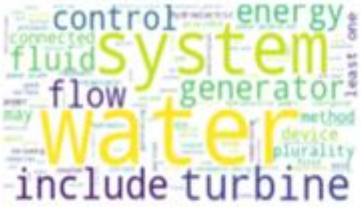


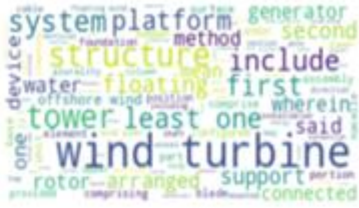


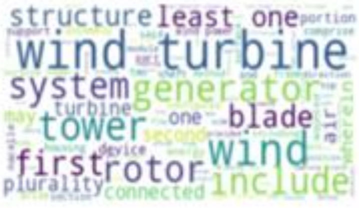


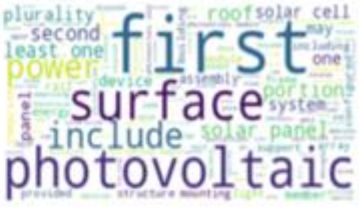


Lemmas:



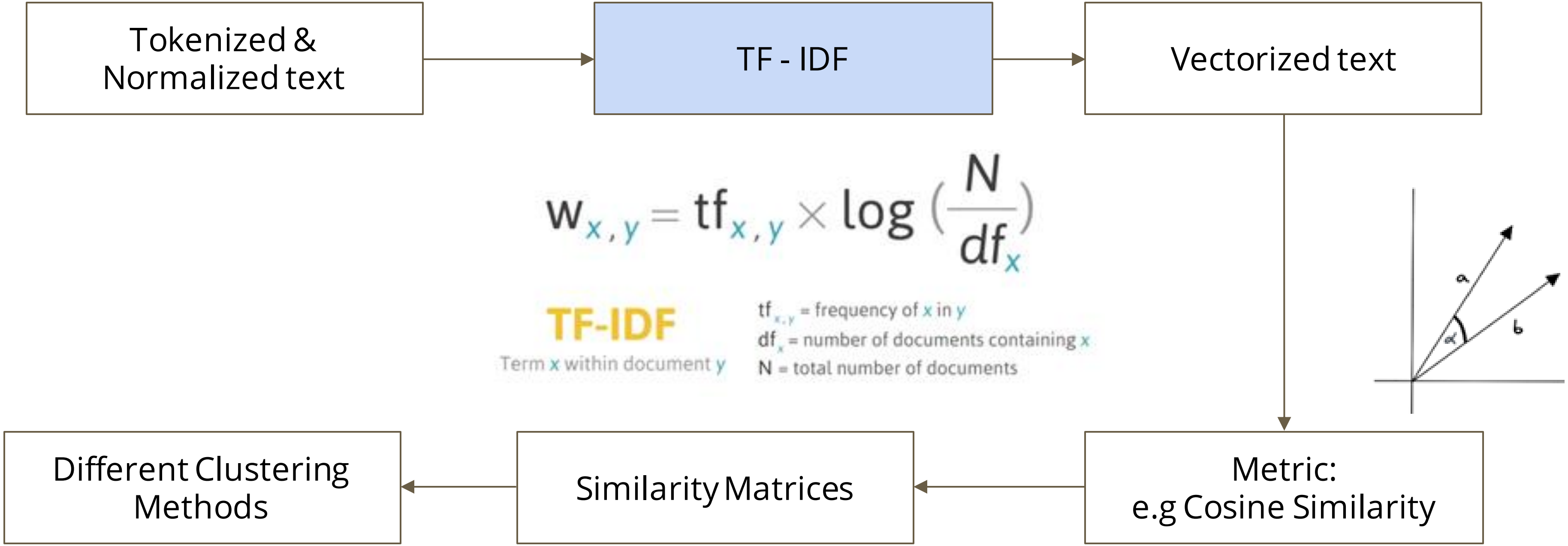






Approach 1: TF-IDF Vectorizer

⇒ Term frequency - Inverse Document Frequency (TF - IDF): Convert a collection of raw documents to a matrix of TF-IDF features.



Tokenized & Normalized text

TF - IDF

Vectorized text

Metric: e.g Cosine Similarity

Similarity Matrices

Different Clustering Methods

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

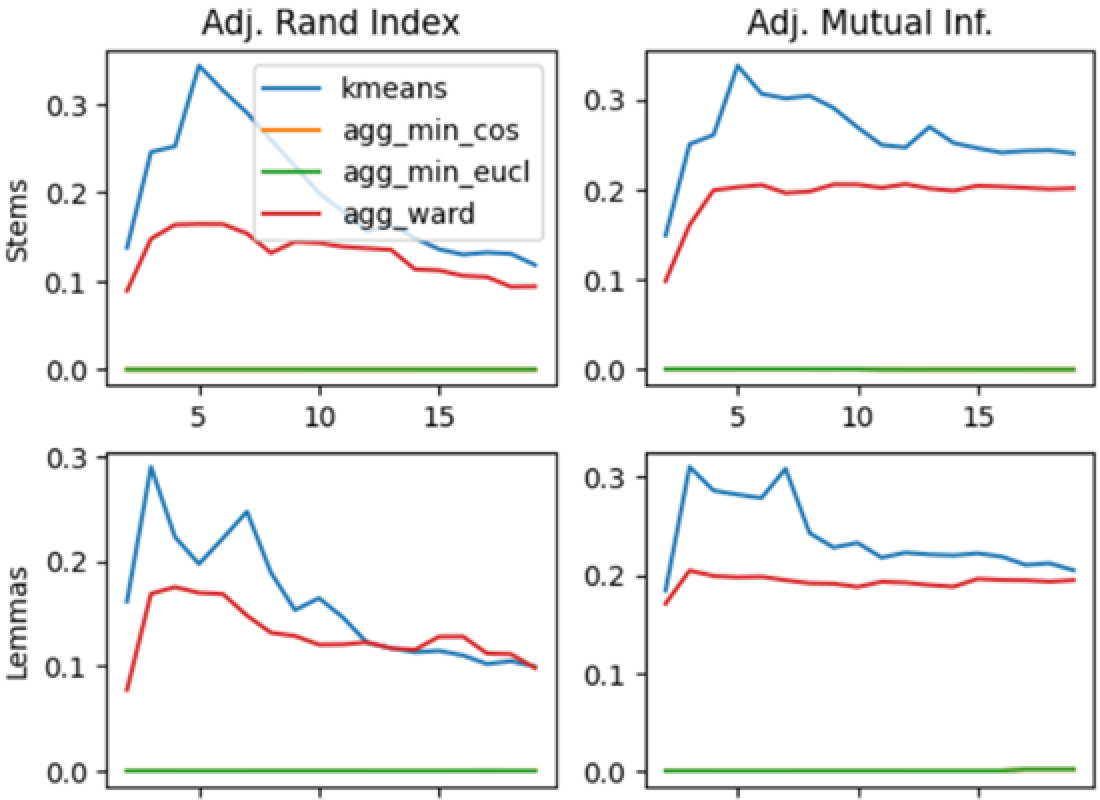
Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

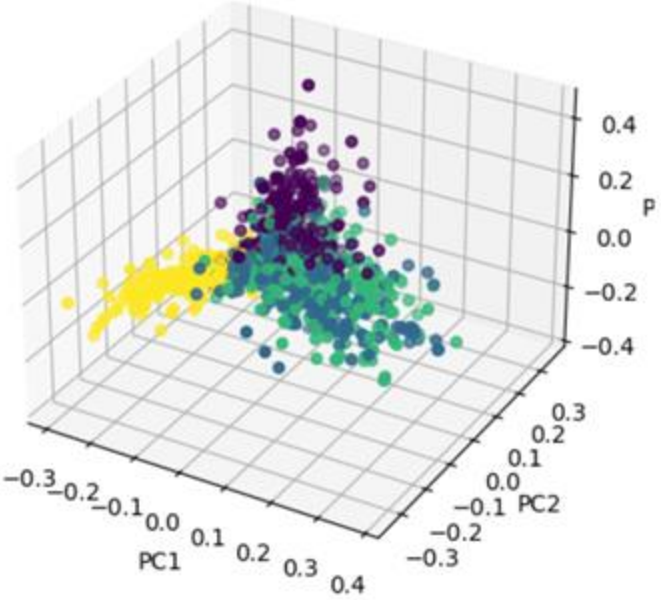
N = total number of documents

Approach 1a: Clustering results

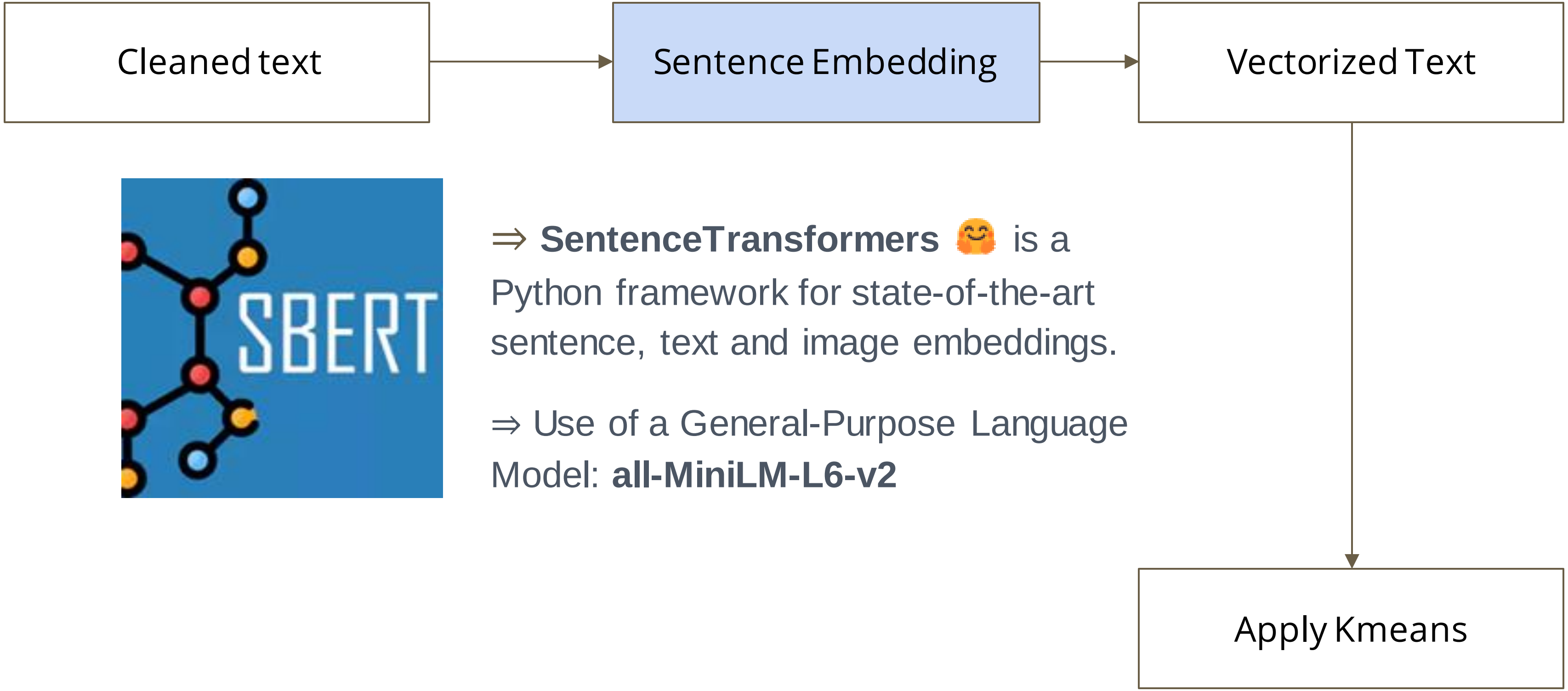


Motivation for Principal Component Analysis

- TF-IDF vectors sparse and high-dimensional
- Distance metrics more meaningful in lower dimension



Approach 2 - LLM



Cleaned text

Sentence Embedding

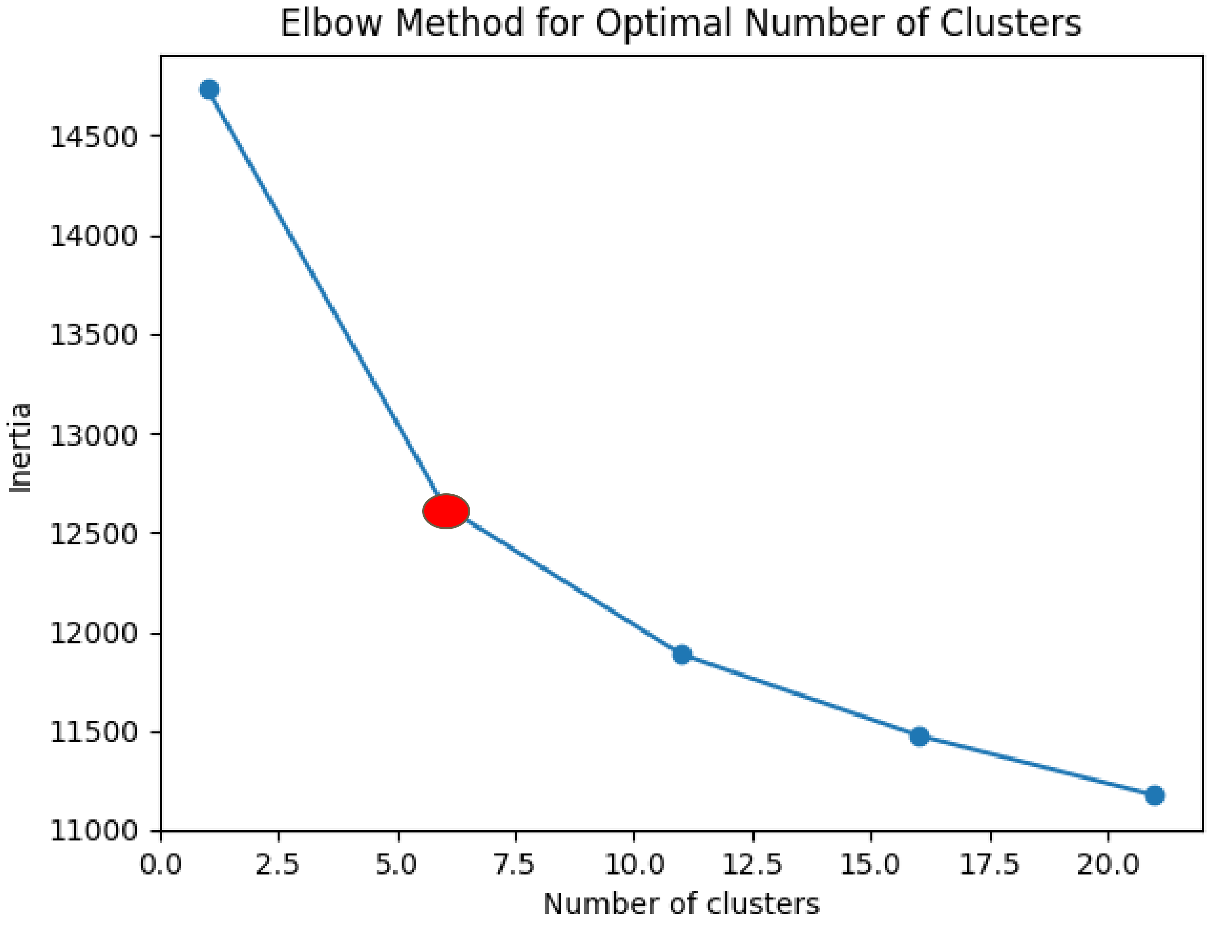
Vectorized Text

Apply Kmeans

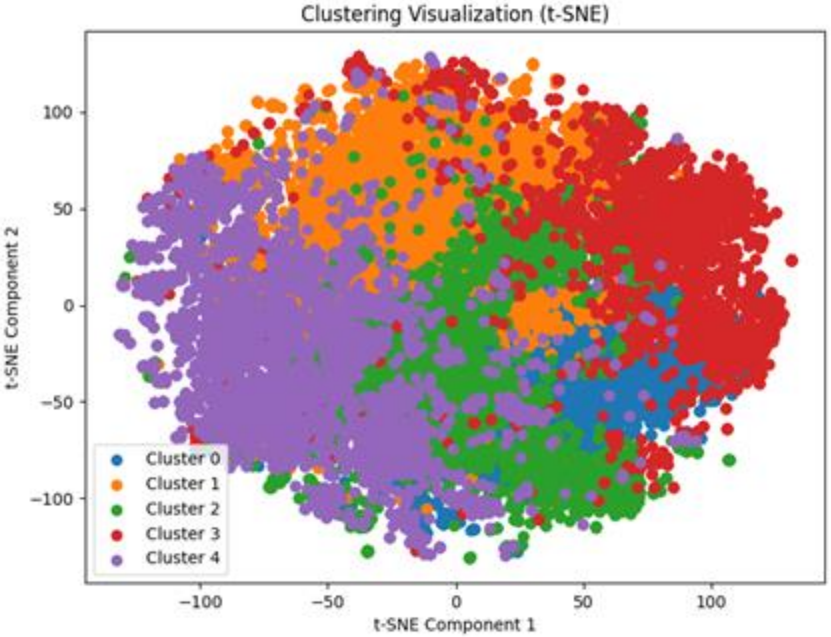
⇒ **SentenceTransformers** 🤗 is a Python framework for state-of-the-art sentence, text and image embeddings.

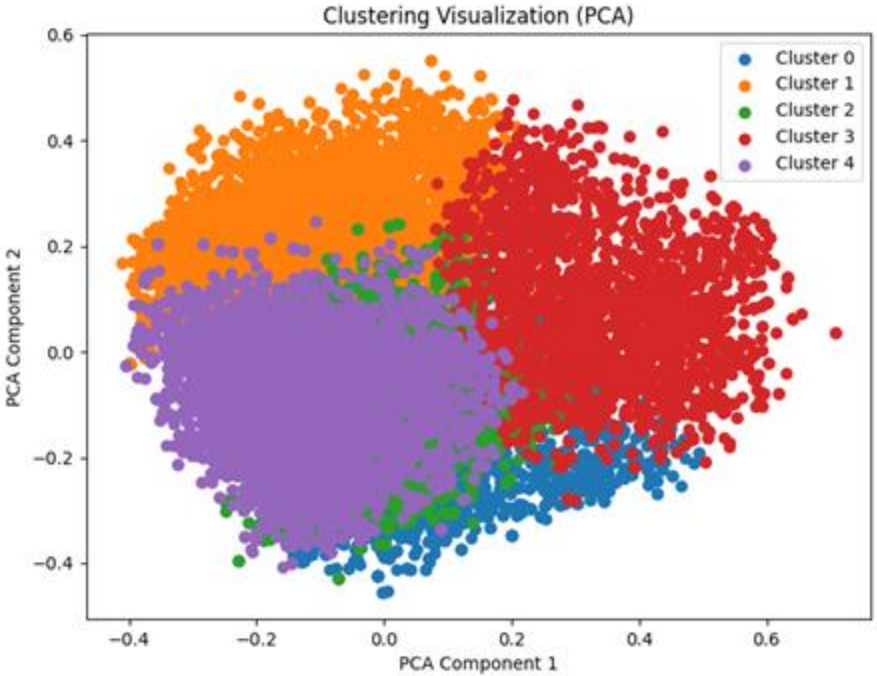
⇒ Use of a General-Purpose Language Model: **all-MiniLM-L6-v2**

Determining optimal cluster number (K)

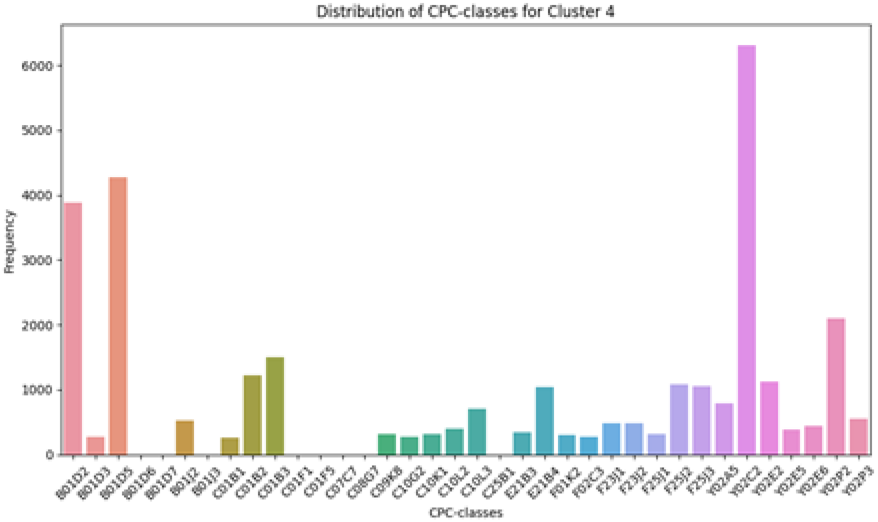


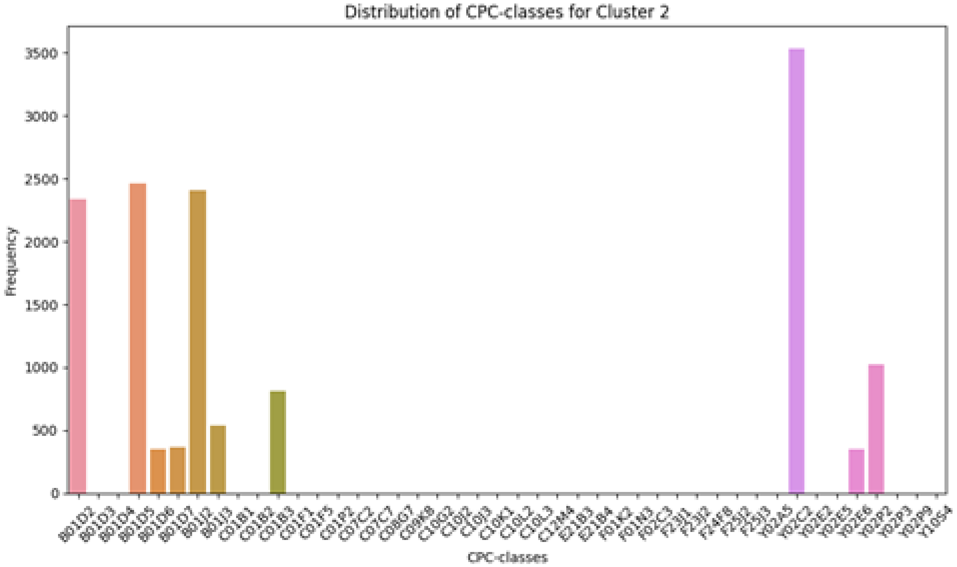
Results


Clustering Visualization (t-SNE)
t-SNE Component 2
100
50
0
−50
−100
−100
−50
0
50
100
t-SNE Component 1
● Cluster 0
● Cluster 1
● Cluster 2
● Cluster 3
● Cluster 4


Clustering Visualization (PCA)
PCA Component 2
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4
0.6
PCA Component 1
● Cluster 0
● Cluster 1
● Cluster 2
● Cluster 3
● Cluster 4

Distributions of CPC classes

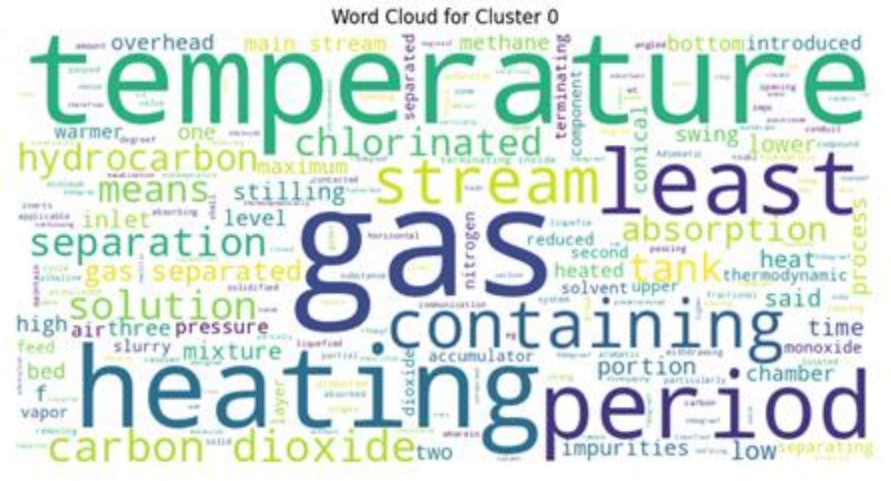


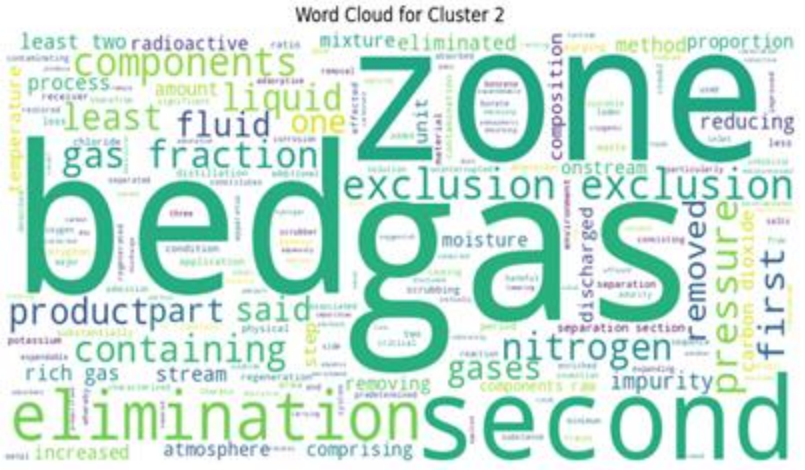


Assessing & Interpreting clusters

Approach 1: Create a summary out of all patents' abstract of each cluster

Approach 2: Looking at the abstract of the closest patents to each cluster's centroid


Word Cloud for Cluster 0


Word Cloud for Cluster 2

Opportunities for future research

- Address multilingual abstract texts
- Include VC investment behavior data
- Fine tune the language model on policy text and/or abstracts within the resulting clusters
- Building a predictive model to forecast Carbon Capture, Utilization, and Storage (CCUS) technological evolution and innovation pace