TEXT MINING & SENTIMENT ANALYSIS PROJECT

Analysis of Circular Economy Terms in Scientific Literature

Doina Vasilev
Università degli Studi di Milano

TABLE OF CONTENT

THE PROJECT

The aim of the following project is to introduce the methods that can be used for tracking changes in word meanings over time, with the purpose of performing ***semantic shift analysis***.

Across the many possible alternatives, whose use depend on the specific problem at hand, I'll be focusing on techniques that enable the examination of the terms surrounding '**circular economy**' in scientific literature, as a way to detect changes on its meaning.

THE DATASET

I was able to retrieve the dataset from the paper:
"***Conceptualizing the circular economy: An analysis of 114 definitions***", *by Julian Kirchherr , Denise Reike, Marko Hekkert (2017)*.

This dataset compiles 114 definitions of the 'circular economy' concept, each extracted from a distinct scientific publication, spanning the years 2006 to 2017. The concise nature of these definitions, typically just a few lines each, makes the dataset especially suitable for my project, given the constraints related to the capacity of the methods I plan to use.

DATA PREPROCESSING & *CONTEXTUALIZED* EMBEDDINGS EXTRACTION

While models like **BERT** are robust at handling a wide range of text data, certain pre-processing steps are crucial to ensure the text is efficiently interpreted by the model. These steps help minimize noise and focus the model's attention on salient content.

These include:

- **Removing punctuation;**
- **Standardizing the text format by converting to lowercase and removing special characters;**
- **Removing stopwords to reduce the dataset to meaningful content only.**

Understanding BERT & Contextualized Word Embeddings

What is BERT?

Developed by Google, BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model, that, unlike traditional models that process text in one direction, interprets text bidirectionally, allowing for a nuanced understanding of context.

The backbone of BERT is the Transformer architecture, which uses attention mechanisms to discern the relevance and relationship between words in a sentence. This approach allows BERT to generate contextually rich embeddings that reflect the varied meanings words can have in different linguistic environments.

Understanding BERT & Contextualized Word Embeddings

What are *contextual word embeddings?*

Traditional word embeddings represent words in a high-dimensional space, where each word is assigned a fixed vector.
These embeddings capture some semantic meanings but fail to account for context.
The typical example is the word 'bank', which could mean both '*bank*' as the financial institution, but also '*river bank*'.

In such case, a uncontextualized model would return the same vector representation regardless of the different meanings.
Contextualized word embeddings, instead, such as those generated by BERT, overcome this limitation by providing embeddings that take into account the word's context, by 'looking' at the text bidirectionally.

Analysis of Circular Economy Terminology

Vector represention of the words enables the evaluation of the *terminology* surrounding the target expression: ‘*circular economy*‘, by exploiting the *cosine similarity* between the vectors, which can be translated into the ‘*closeness*‘ of the words in time.

In this analysis, I am not only able to extract the terminology characterising *circular economy*, but I can also measure how such closeness varies across time, and thus possibly discovering emerging trends, or declyining concepts.

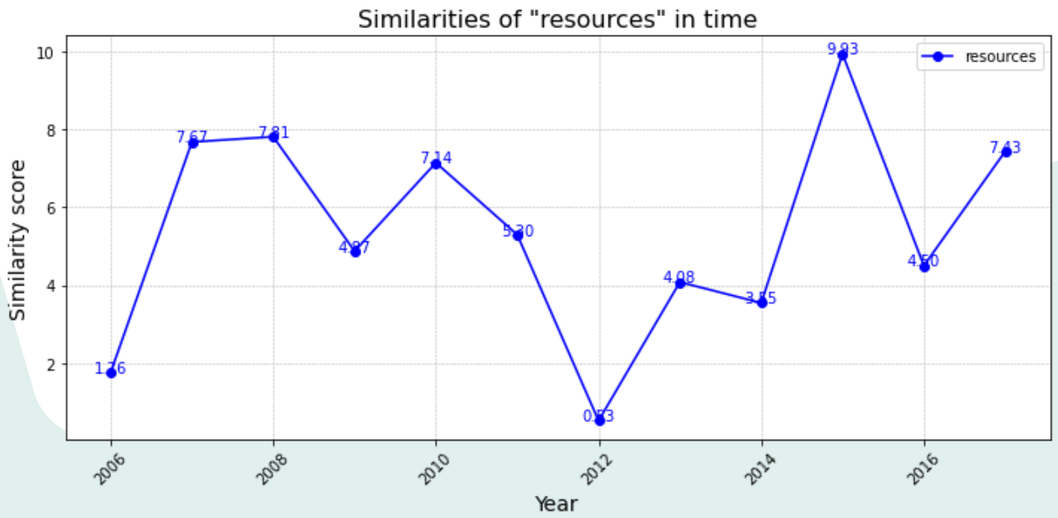
In the following slides, we can see that, while the majority of terms maintain a steady association with '*circular economy*' across the timeline, a few exhibit upward trends in recent years. Notably, terms like '*regeneration*' and '*value*' are on the rise. These trends could potentially signal shifts in focus and emerging priorities within the circular economy discourse, highlighting an increasing emphasis on sustainable practices and value creation as central themes.

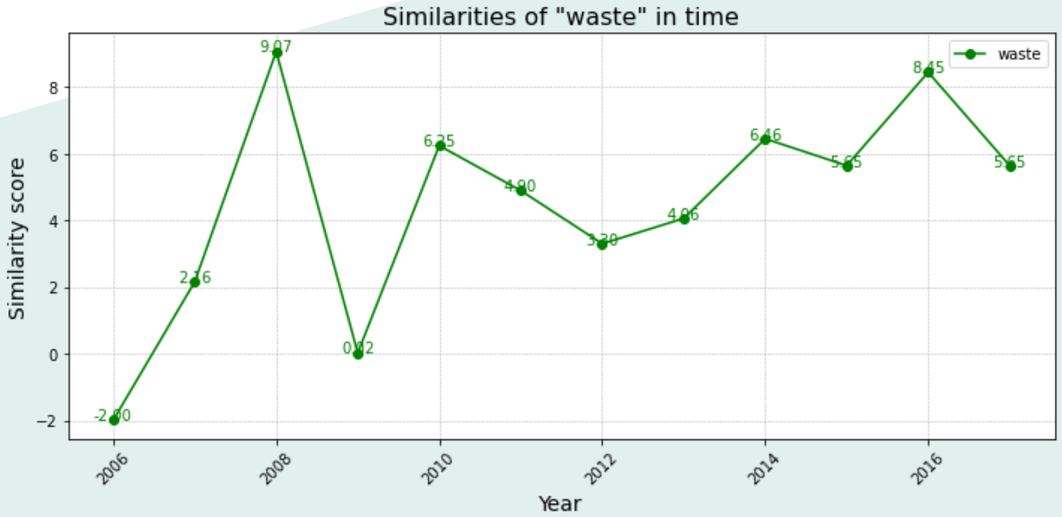
Analysis of Circular Economy Terminology

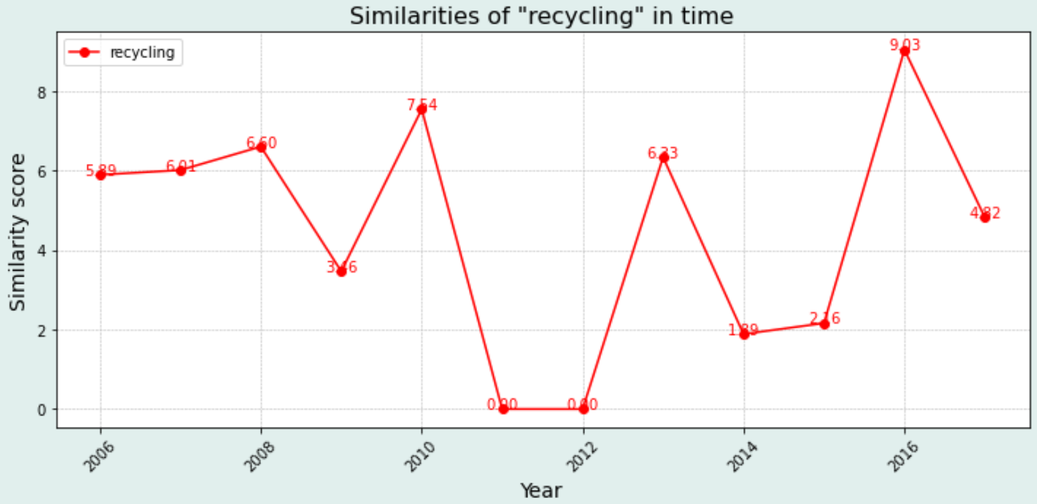
The following matrix showcases the similarity scores over time for the 12 *closest* words to *circular economy*.

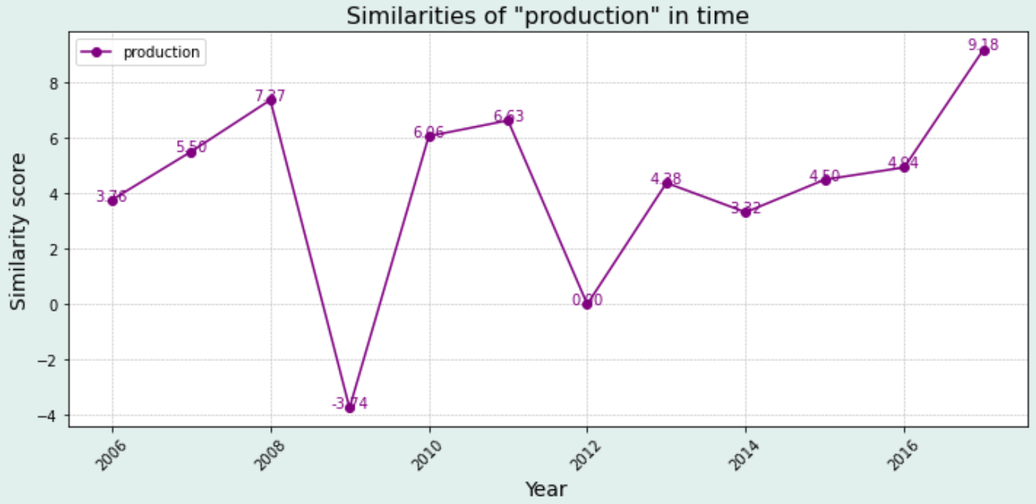
	resources	waste	recycling	production	sustainable	development	reuse	product	value	environmental	growth	regenerative
2006	1.763066	-1.995882	5.890133	3.764491	0.000000	-2.298856	1.973638	0.000000	0.000000	0.000000	0.000000	0.000000
2007	7.673624	2.158309	6.007074	5.504724	6.610572	5.872868	0.000000	6.708714	0.000000	0.000000	0.000000	1.416992
2008	7.806597	9.071983	6.598020	7.367278	2.044785	9.434771	3.316276	6.445321	4.066265	10.029106	0.000000	0.000000
2009	4.874933	0.021294	3.463244	-3.739317	-4.512886	5.070293	0.000000	0.000000	0.000000	6.574088	8.900323	0.000000
2010	7.141171	6.247017	7.537005	6.064473	0.000000	3.611096	8.078314	7.999768	6.179660	4.940399	1.882968	0.000000
2011	5.297078	4.898800	0.000000	6.629486	2.469191	3.394648	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2012	0.531274	3.301647	0.000000	0.000000	0.000000	0.000000	3.816591	0.000000	3.072336	0.000000	0.000000	4.753189
2013	4.082935	4.061091	6.331325	4.377513	0.000000	3.364060	6.364116	4.346654	4.463513	3.859568	0.000000	6.035515
2014	3.549820	6.455097	1.885553	3.318938	7.711700	6.379561	7.488108	5.612489	4.818056	5.845633	6.542039	6.371398
2015	9.927141	5.647614	2.158209	4.499228	5.629954	5.141438	6.179089	5.445374	6.630569	2.991633	10.186809	6.605929
2016	4.502395	8.451423	9.030802	4.935603	6.933575	6.425570	4.421515	4.524833	9.939872	5.744277	6.922936	4.232020
2017	7.432587	5.646369	4.816237	9.182620	5.740952	4.127337	8.545987	8.651990	9.123248	6.842910	4.087599	8.678449

Analysis of Circular Economy Terminology



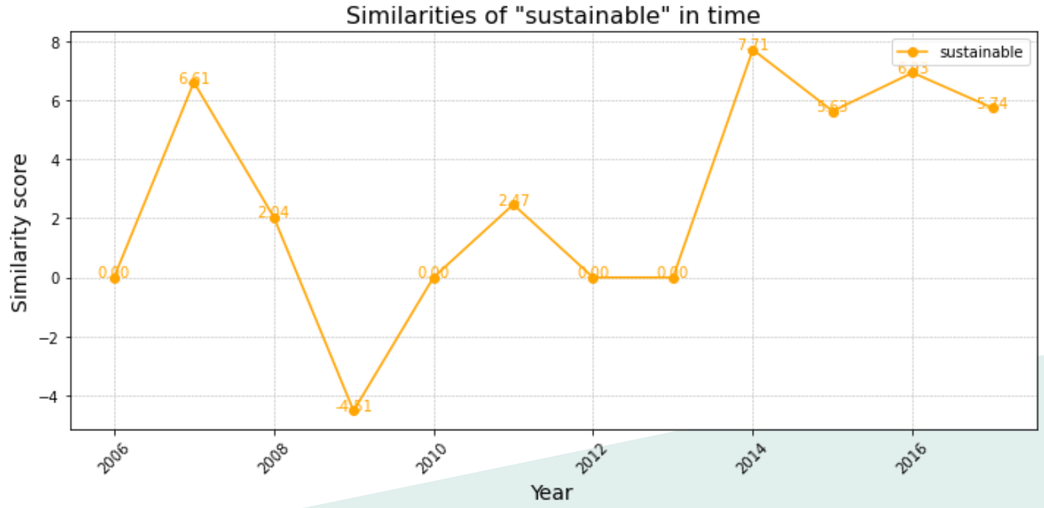


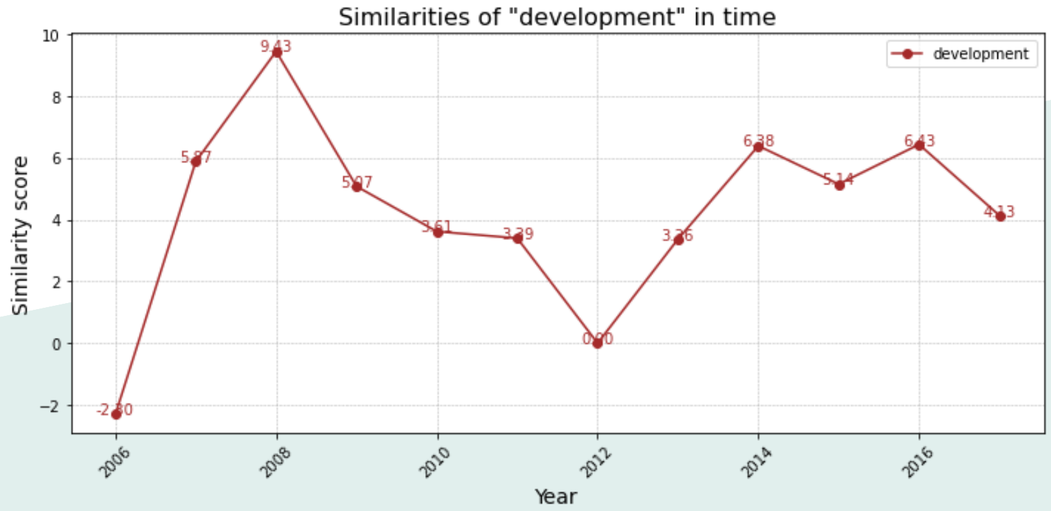


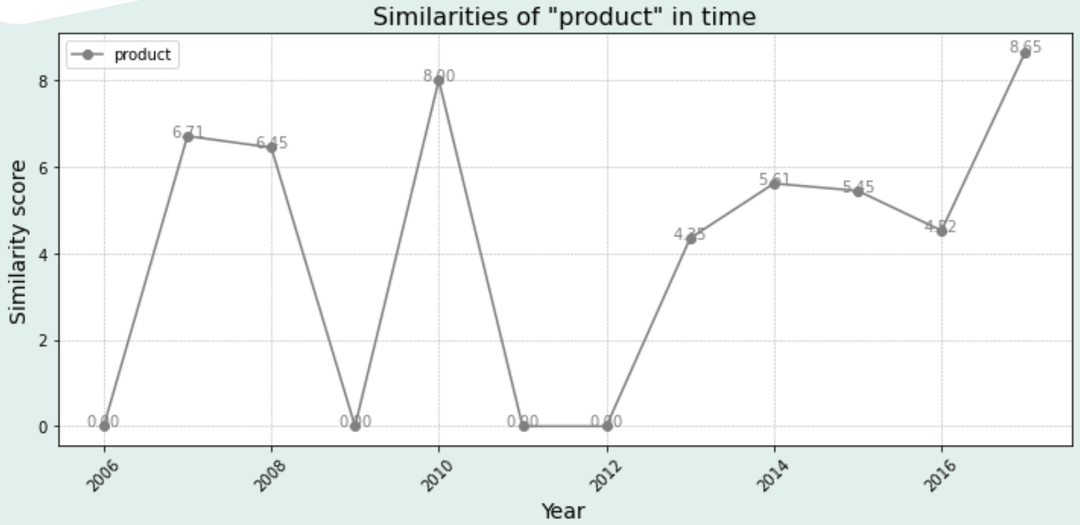


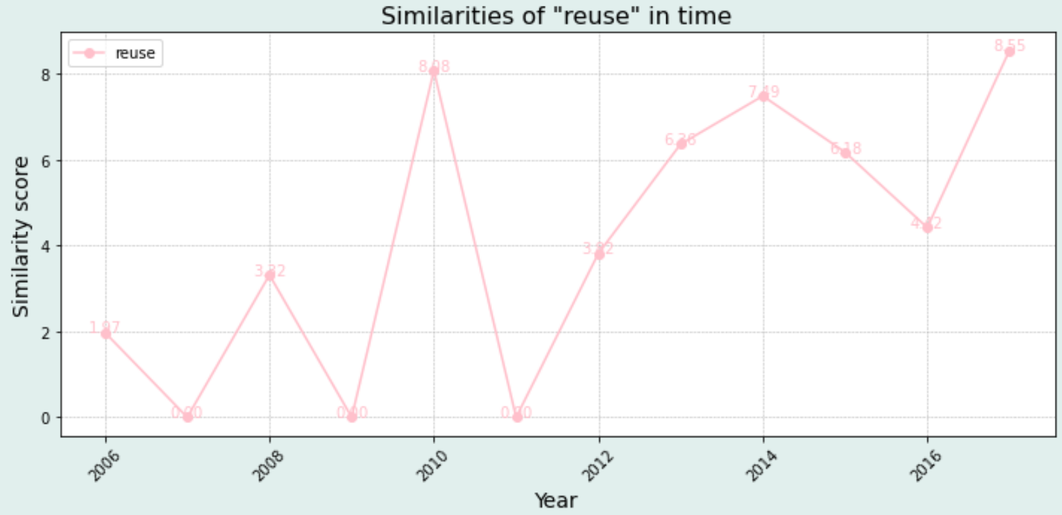
Graphical representations of the variations in similarity scores for the 12 closest words.

Analysis of Circular Economy Terminology



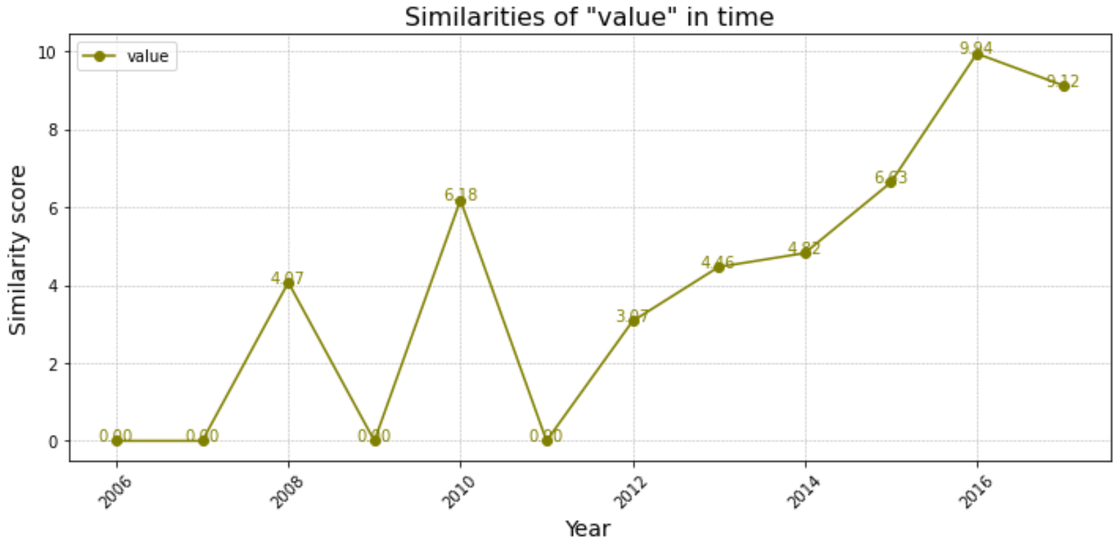


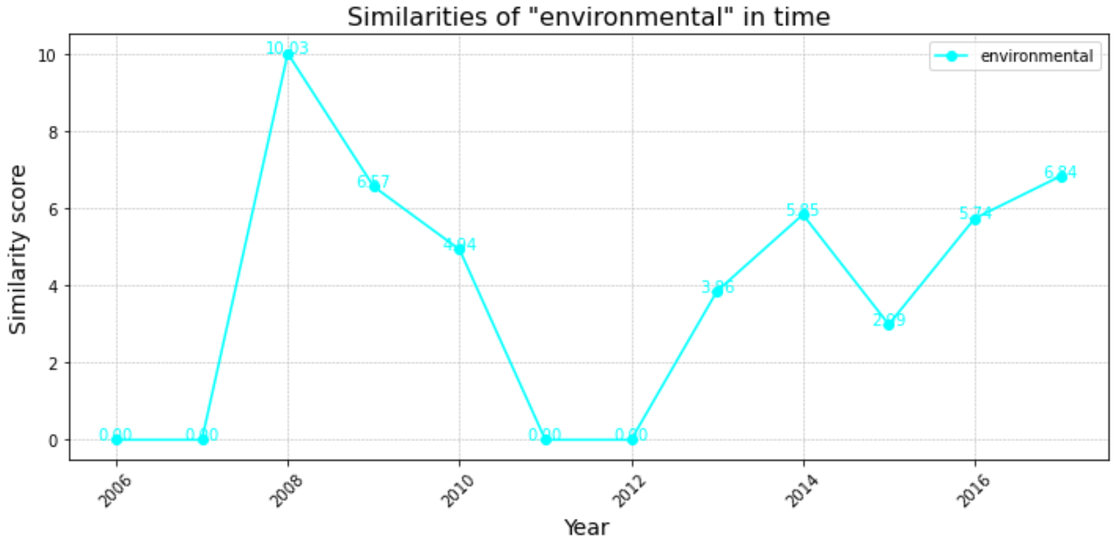


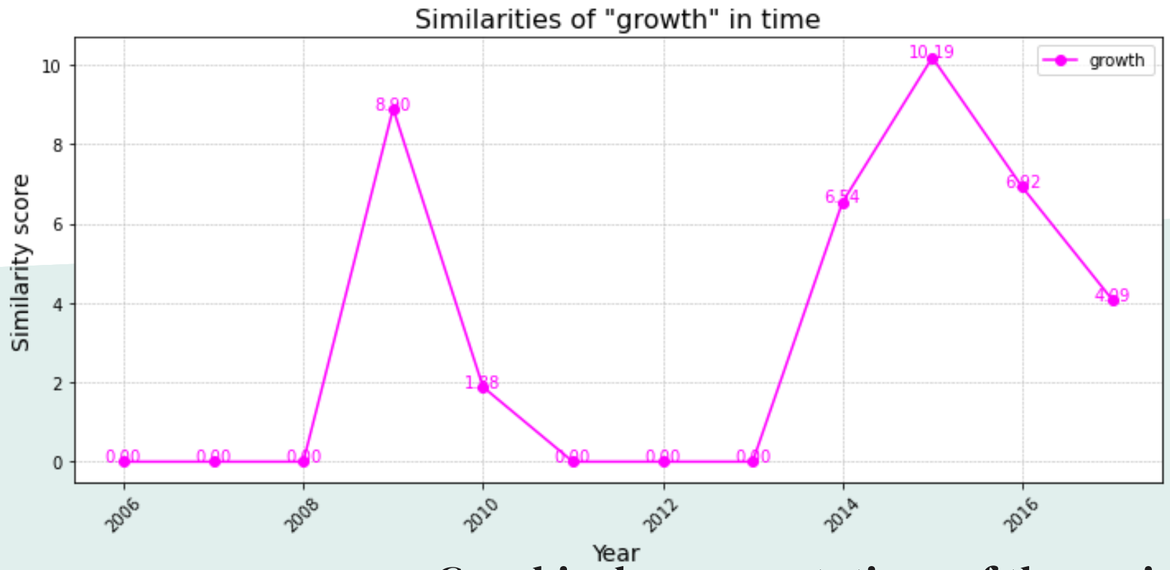


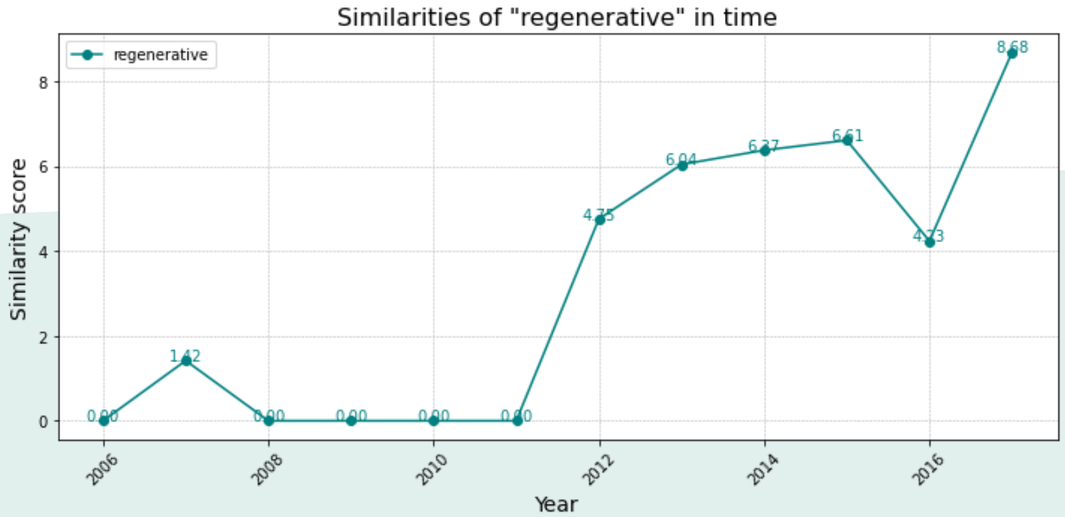
Graphical representations of the variations in similarity scores for the 12 closest words.

Analysis of Circular Economy Terminology









Graphical representations of the variations in similarity scores for the 12 closest words.

CONCLUDING REMARKS

The preceding pages showcased how we can exploit natural language processing for gathering insightful information on the evolution of words through time.
In the case study, not only can we gather information on the terminology used in the context of circular economy, but also observe how this terminology has evolved over time.

FURTHER WORK:

While the current project has the scope of showcasing some methodologies for evaluating semantic shift, further work must be done to extract more insightful information.

- *Extending the Time Frame*: Incorporating more recent studies could provide insights into the latest trends and shifts in the circular economy discourse.

- *Expanding the Dataset*: Analyzing entire papers, rather than selected excerpts, offers a more comprehensive view of the terminology used.

- *Training Models on the Dataset*: This approach is not possible using BERT, but can be attempted with uncontextualized models such as Word2Vec. Training the model on the dataset could lead us to extract even more precise word representations, and thus gather more nuanced information of the texts.