Names: Jakob Sig Dransfeldt, Endika Mitxelena, Alpha Mbarushimana Ntakiyimana and Mikkel Rune Jørgensen

Executive report

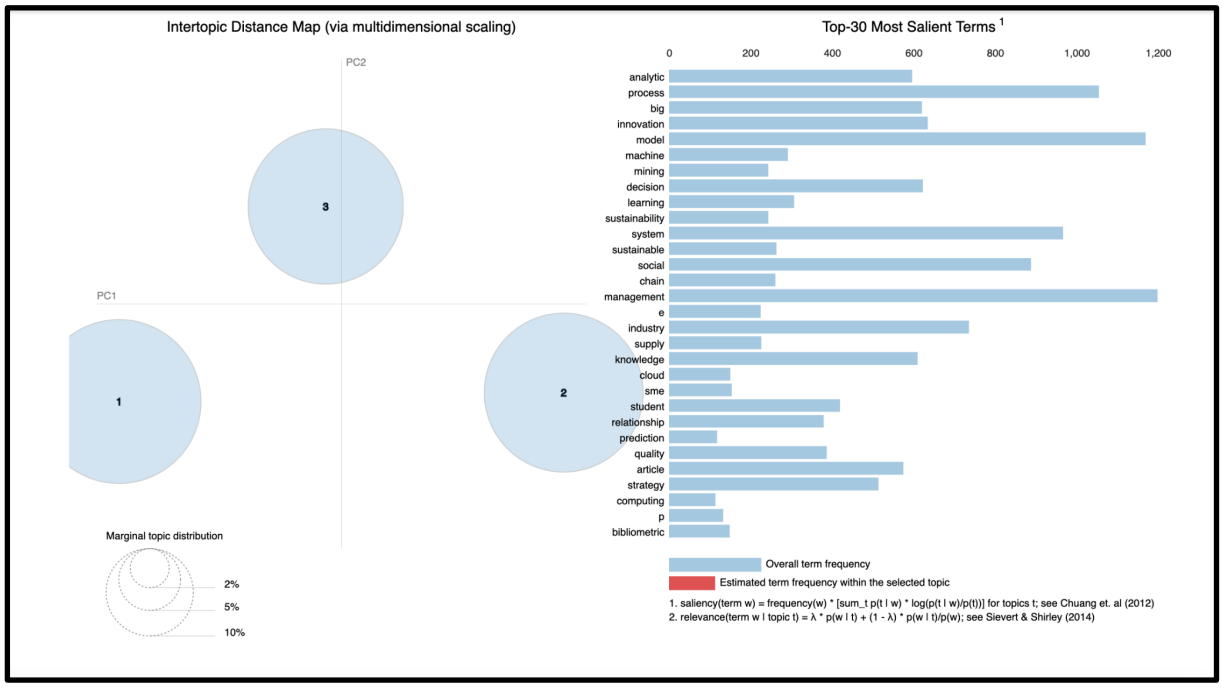
## Introduction

Within this small paper, Natural Processing (NLP) has been used to see the application of the scientific literature that has used Business data science. To analyze the literature on this topic, the website Scopus has been used to extract that data. From Scopus, one can get the Authors and their ID's, the title of their papers related to the topic. Furthermore, Scopus gives the data for the Abstract, year, source title, and key Author.

## Stakeholders executive report

The first thing there is visualized in the notebook comes from the topic modeling. It is the method of extracting needed attributes from a bag of words as in this case comes from articles from Scopus. The reason it is used is because each word is treated as a feature in the NLP-analysis. This allows us to use feature reduction to focus on the relevant material rather than wasting time going through all of the data's text.

## Topic Modeling




Illustrated above is our intertopic distance map which shows the distribution of topics from our data collected from Scopus. It is an interactive graph if you look at it in the notebook where the

most salient terms will shift when you hover over the different clusters. For choosing the amount of clusters we used LDA which is an unsupervised clustering technique that is commonly used for text analysis. It is a type of topic modeling in which words are represented as topics, and documents are represented as a collection of these word topics which reveals co-occurrences among words, as well as long-span latent topic information. In this particular case 3 clusters were chosen which could have been based on statistical analysis but in this case we simply went forward with simple manual testing and ended up with 3. When you hover over the clusters it will show blue and red in the *Top-30 Most Salient Terms*-graph, which indicates the overall term frequency with the blue bars and the red indicates the estimated term frequency within the selected topic. The 3 clusters top topics are as followed with focus on the estimated term frequency:

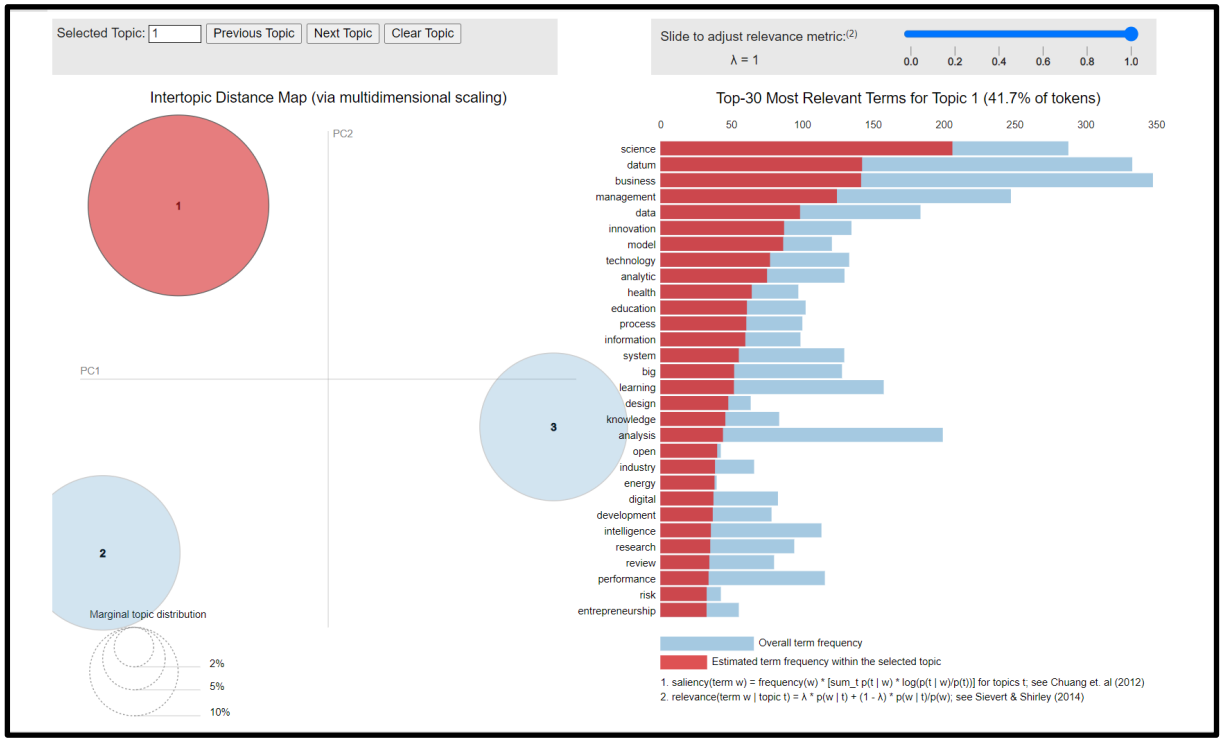
1) Within this topic nr.1 in the business science field, involves everything about process, systems and technology. Meaning the articles and reports in this topic, are about how the data is processed and analyzed.

2) The articles in this topic mostly describe business management by using data science and analysis. The word industry is also strongly presented at the top, meaning that the application of data science within management is strongly dependent.

3) The third topic is based on innovation belonging to the data science field. Here we see the articles describe data science in some different fields, such as social, medicine and healthcare.

Names: Jakob Sig Dransfeldt, Endika Mitxelena, Alpha Mbarushimana Ntakiyimana and Mikkel Rune Jørgensen

Cluster




Where we in the topic modeling used the article abstract to find patterns in the words used. We have in the clustering part used the author keywords to find the most used words in the clusters.

There are 3 clusters, where the clusters arrange the words in mostly used to less used. We use the clusters to train the model. The clusters have specific words that later will be able to explain the labels.

Names: Jakob Sig Dransfeldt, Endika Mitxelena, Alpha Mbarushimana Ntakiyimana and Mikkel Rune Jørgensen

Model training

| y=0 top features | | y=1 top features | | y=2 top features | |
|---|---|---|---|---|---|
| Weight? | Feature | Weight? | Feature | Weight? | Feature |
| +3.150 | nan | +2.794 | science | +1.752 | chain |
| +1.520 | social | +1.568 | health | +1.463 | supply |
| +1.103 | research | +1.470 | model | +1.301 | artificial |
| +1.098 | medium | +1.446 | energy | +1.284 | corporate |
| +1.088 | analysis | +1.406 | open | +1.208 | application |
| +1.055 | media | +1.342 | innovation | +1.153 | tourism |
| *… 744 more positive …* | | +1.225 | information | *… 1042 more positive …* | |
| *… 1743 more negative …* | | *… 1068 more positive …* | | *… 1445 more negative …* | |
| -1.083 | technology | *… 1419 more negative …* | | -1.390 | research |
| -1.135 | health | -1.210 | analysis | -1.682 | nan |
| -1.354 | education | -1.468 | nan | -1.718 | data |
| -1.631 | management | -1.492 | social | -1.942 | science |

The model made is predicting the labels extracted from the keywords, and then when the supervised model is given an abstract, the model will be able to categorize in the label it is supposed to be in. In the tabel each label has a set of words that either has a positive or negative weight. This means that the words with a positive weight will have an impact on where a new article will be put in based on which words the abstract holds.

Network analysis

The purpose of making a network analysis is to look how the authors are connected, and to whom. The approach was first to make an edge list, which can be used to create a graph to visualize and conduct centrality measures. It's pretty similar to the UML where it also takes and looks at the connections and patterns. Based on the eigenvector centrality we get an idea on how far each author is based on the edges. As you can see on the visualization there are a lot of people who cite each other, while some are not cited so much. It also looks like there is different groups within the citing, which can relate to each topic as we did in topic modeling.

Names: Jakob Sig Dransfeldt, Endika Mitxelena, Alpha Mbarushimana Ntakiyimana and Mikkel Rune Jørgensen