CRANFIELD UNIVERSITY

Théophile Cordiez

Clustering of Human Activity Data to Identify Smart Living Opportunities

Academic Year: 2020-2021

MSc: Applied Artificial Intelligence

Supervisors:

Dr. Saba Al-Rubaye

Dr. Weisi Guo

Sponsor:

Smart City Consultancy LTD

Acknowledgements

I would like to express my gratitude to my supervisors, Dr Saba Al-Rubaye and Dr Weisi Guo, and my professors who have provided me with all the essential knowledge in Artificial Intelligence and guided me in this work when I needed it. I am also grateful to my family and friends who gave me their support during this year impacted by the Covid-19 pandemic.

Abstract

This master's thesis focuses on the identification of 5G business opportunities related in the city of Milton Keynes. This geo-spatial analysis is based on the processing of textual data from company websites to identify business activities from which we derive an analysis of 5G opportunities for smart cities. This approach, specific to the scenario of 5G deployment in Milton Keynes, is explained and justified through its various successive steps, including, the collection and preprocessing of textual data, the knowledge extraction of these documents and then the clustering of the retrieved information to map the areas of the city corresponding to a need for 5G service deployment. The information retrieval phase is divided into two parts: the first one aims at identifying the company's activity sector based on the semantics of the terms used through their website. The second part is an attempt to extract knowledge about the opportunities related to 5G, based on the articles of a technology magazine. Each of these two parts are considered via different natural language processing techniques - including Latent Dirichlet Allocation and a transformer-based model - and are qualitatively compared in their ability to capture the semantics of the documents. Finally, the identification of Milton Keynes' business activities allows us to implement a geo-spatial analysis of the target areas of the city where the implementation of 5G will be commercially beneficial.

Contents

List of figures

List of abbreviations

EMBB	Enhanced Mobile Broadband
MMTC	Massive Machine-Type Communications
URLLC	Ultra-Reliable Low-Latency Communication
LPWA	Low-Power, Wide-Area
LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations from Transformers
KDE	Kernel Density Estimation
UMAP	Uniform Manifold Approximation & Projection
HDBSCAN	Hierarchical Density-Based Spatial Clustering
RNN	Recurrent Neural Network
IR	Information Retrieval
SVM	Support Vector Machine

I. Introduction

A. Background and motivation

The smart city concept is considered as an urban model based on innovative technologies to improve the quality of life of its inhabitants. This ideal city model encompasses the economic, societal, spatial, and organizational dimensions of urban spaces to create an optimal city form. The smart city concept integrates information and communication technologies and various physical devices connected to the IoT network to optimize the efficiency of municipal operations and services and connect to citizens. By deploying its 5G service around the city centre, Red Bull Racing and the stadium, Milton Keynes intends to know which innovative businesses are emerging that can potentially benefit from 5G.

5G enables much more efficient communications, in terms of security, speed, low latency and helps a wide range of sectors to grow, such as autonomous driving, virtual reality, remote healthcare and IoT. The business opportunities associated with 5G are still nascent and deserve to be targeted geographically and dynamically. Identifying and highlighting the businesses that will best reap the benefits of 5G is therefore critical to ensure the proper deployment of 5G. An analysis based on the geospatial distribution of businesses according to their type of activity should help to identify the most innovative and relevant areas of the city when developing 5G.

Furthermore, data mining and natural language processing techniques open a vast field of applications in the field of organisational management and decision making. The increasing availability of textual data through the internet and the media is a source of knowledge that touches fields such as economics, politics and sociology and offers many possibilities of data analysis, in real time. In this master thesis, we want to use data from corporate websites to get a more accurate and representative view of corporate activities. We will therefore aim to extract the semantics of these websites and put them in perspective with the commercial opportunities of 5G to target the most interesting areas of the city for the deployment of this communication technology.

B. Aims and objectives

This master thesis aims to produce a tangible proof-of-concept result for the targeting of opportunity areas for businesses in need of 5G, within the city of Milton Keynes. The identification of the business sector is the main objective of this master thesis. We want to build a business description for the city's businesses from an Industry 4.0 perspective. Digital technologies are often poorly captured by standard classifications [1]. The approach we take therefore aims at identifying these industrial sectors with more precision and relevance, in real time. A second major objective of this thesis is to perform a geo-spatial analysis of the areas where the most innovative industries are located. This approach can be seen as a proof of concept in the geospatial analysis of 5G opportunities and is intended to be replicable in other geographical areas. In this work, we mobilize state-of-the-art techniques in the field of Natural Language Processing (NLP) and Machine Learning concepts.

II. Methodology

A. Literature review

1. Information retrieval

Information retrieval (IR) is the field that studies how to find information in a corpus. The research focuses on various aspects of information retrieval systems to improve their efficiency and reliability [2]. This includes different models of information retrieval and important questions on document representation, similarity measurement and query expansion.

Many IR systems are based on data mining methods. Data mining deals with large amounts of data and aims to discover interesting knowledge from a large amount of data. Data mining techniques include classification, clustering, sentiment analysis. Clustering and classification are widely used to categorize textual data. Search engines often give a huge amount of information for a given query, which makes it difficult for users to navigate or identify relevant information. With the help of clustering methods [3] we can automatically group the retrieved information into a list of important categories. Online articles can be classified into various categories which include, politics, entertainment, news, travel. Text classification is also a method used to sort news into different classes. Generally, classification is based on a set of training documents that are labeled according to their class.

Document categorization is used to associate them with abstract themes. Some algorithms like K-mean, SVM (Support Vector Machines) and HMM (Hidden Markov Model) help to classify groups into subgroups. HMM for example is useful to filter the parts of the html code that correspond to tags. [4] The rest of the text can then be processed by SVM to perform the classification. SVM is a support vector machine which is a binary classifier. It is used for example to distinguish keywords. The k-means is an algorithm that allows to create clusters, by specifying the number of classes desired. These algorithms are most often based on a vector representation of words or sentences [5]. In this field, a lot of progress has been made with the use of RNNs, and then the arrival of transformers which have surpassed

them in speed and reliability. With the help of descriptor sets and large libraries on which these models are trained, it is possible to represent the document by a vector. It is also possible to use a priori knowledge about how terms are distributed in documents according to their importance. Part VI will give more details on these methods.

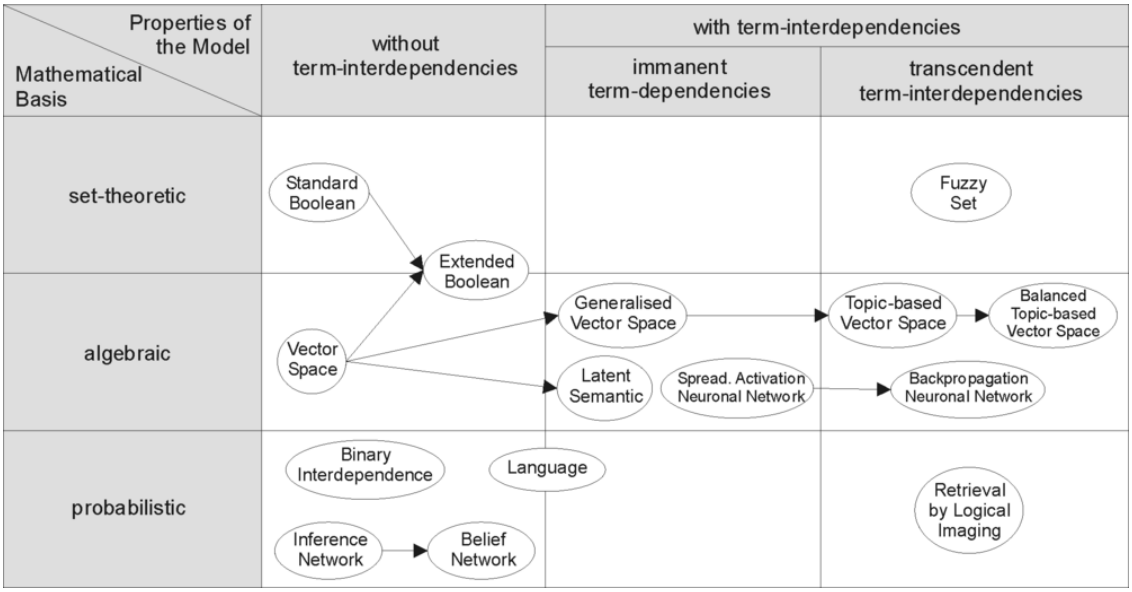


Fig 1- Categories of Information Retrieval models

2. Web data collection

There are several difficulties affecting the collection of data on the internet. Documents on the Web are distinguished from general text documents by the presence of hyperlinks. Documents may also contain heterogeneous data, i.e., in addition to text, they may contain other content such as audio, video and images [1]. Some of the documents are close or exact duplicates of other documents. Finally, another difficulty is the lack of stability, as the content of web pages is often changed. To address these difficulties, there are many libraries that can parse the text. For example, the use of regular expressions is a way to filter the tags of the html code. Getting these web pages can be done via a simple GET request. Python programming language has a vast quantity of documentation and libraries to perform web data collection.

B. Thesis structure

The thesis is divided into these steps:

- Review of 5G opportunities

A detailed review of 5G's features and implications in the business world. This chapter provides an overview of the industry sectors that will be affected by the deployment of 5G.

- Web-based data collection

An explained implementation of the website data collection system. This will be used to form corpora containing the text of businesses in Milton Keynes. We also detail how these businesses and sites were found.

- Industry-independent geospatial analysis

We conduct a first preliminary study of the density of companies in Milton Keynes to distinguish a priori, the most economically active areas. We use the Kernel Density Estimation, to realize this heatmap.

- Information retrieval methods

In this chapter, we review state-of-the-art methods in the field of topic modeling and document clustering.

- Latent dirichlet allocation

We explain here the details of the algorithm we will use to model the themes of the document corpora we have built. We detail the text processing precautions and present the results obtained with this first approach.

- Transformer-based clustering

To get a better understanding of the business activities, we use a model based on transformers. All the steps of this algorithm are explained. We then present the clustering results for the different corpora.

- Visualising clusters

Finally, we exploit the activity detection that was performed earlier, to perform the geo-spatial analysis that highlights the target areas for 5G in Milton Keynes. We use a density estimation method to highlight the locations where relevant activities are concentrated. To produce the final map of the areas of interest, which we will call the "target area", we will cover the following points (not necessarily in this order):

(A)

(1) Listing of companies and startups in Milton Keynes and their geographical location

(2) Collection of data from the websites of Milton Keynes companies to obtain a corpus of text to be analysed.

(3) Clustering of *CorpoWeb* documents via two NLP methods:

(a) Latent Dirichlet Allocation

(b) BERTopics

(B)

(1) Collection of articles on 5G from TechCrunch magazine, gathered in a corpus of texts called *Tech5G*.

(2) Topic modeling techniques applied to the *Tech5G* corpus

(C)

(1) Combining results from (A) and (B)

(2) Presentation of the heatmap of the target areas

III. Review of 5G opportunities

A. Overview of 5G features

5G is the fifth generation of mobile telephony standards. It succeeds the fourth generation, called 4G, by offering higher speeds, while avoiding the risk of network saturation due to the increase in digital uses [6]. Its deployment is being contested, particularly about the health effects of electromagnetic waves and the environmental impact of this technology [7]. The 5G technology provides access to speeds well beyond those of 4G, with very short latency times and high reliability, while increasing the number of simultaneous connections per area covered. This technology creates a new breakthrough in the evolution of mobile communications whose evolution since the 1980s created a wide range of innovations that revolutionized the way the world communicates.

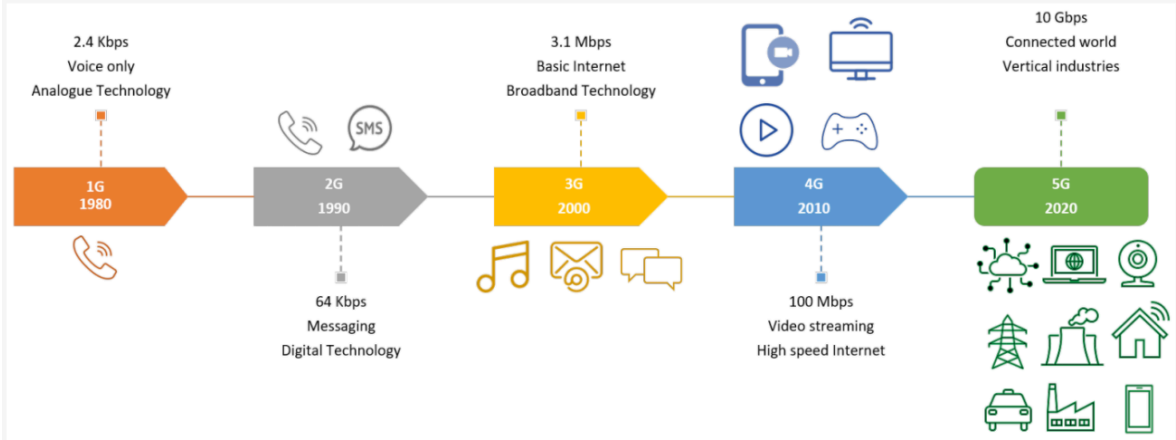


Fig 2- The evolution of mobile communications- source [8]

It aims to support up to one million mobiles per square kilometer (ten times more than 4G). [9] [10] Once deployed, it should enable mobile telecommunications speeds of several gigabits of data per second, up to 1,000 times faster than the mobile networks used in 2010 and up to 100 times faster than the original 4G.

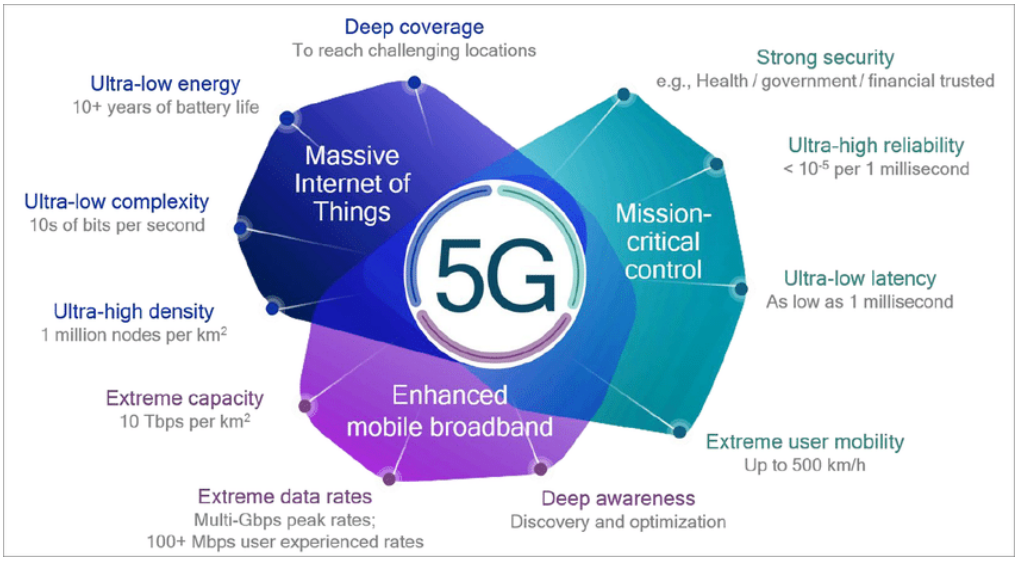


Fig 3- Main features within the target domains of 5G mobile technology - source: [11]

B. High density of devices in smart cities

A smart city is a place where networks and services are made more reliable with the use of digital technology. It can contribute to the development of an urban area and improve the quality of life for its inhabitants. Cities that use the Internet of Things (IoT) are equipped with sensors and data centers to collect real-time data to inform and respond to changing consumption patterns. The number of connected devices globally is expected to reach 75 billion by 2025 [10]. With the explosion of these devices, a smart city will able to collect and analyze massive amounts of data to make informed decisions for its residents and its companies.

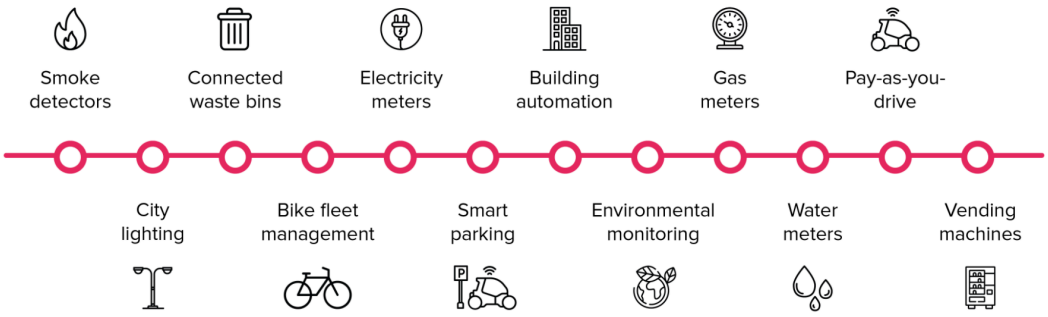


Fig 4- Variety of connected systems in smart cities - source [10]

Digital experiences and smart cities will be strongly impacted by the features of 5G. Not only will the speed of data exchange be higher, but also the network will be able to handle a larger number of mobile devices while having a very low latency. Less latency means less time between sending and receiving the signal. Also, with the new networks, speed and latency do not degrade, even with a large amount of connected devices. Therefore, 5G offers a higher density of devices within urban areas. Today, in crowded stadiums for instance, the connection degrade. With 5G, this will no longer be the case. It will enable communication between an ecosystem of connected devices and objects and offer its potential, in homes, on the streets and in factories and offices.

C. 5G IoT use cases

5G is expected to enhance the capabilities of various Internet of Things (IoT) applications. Some of these include video streaming and remote control of various devices. The widespread adoption of 5G will most likely happen in waves, with EMBB (Enhanced mobile broadband), MMTC (Massive Machine-Type Communications), and URLLC (Ultra-Reliable Low-Latency Communication). For each application, adoption will depend on various factors such as network speed and coverage, and the evolution of regulations [9].

- EMBB will be enabling faster download speeds in various coverage areas. It will mainly impact video streaming. EMBB applications are growing rapidly due to the emergence of 5G smartphones [7] and the increasing number of mobile operators that are investing in the infrastructure needed to support this technology.

MMTC is a 5g technology feature that will support extremely high connection density of online devices. Yet, even as 5G grows, 4G will still be the preferred technology for low-power, wide-area (LPWA) applications. This is because 4G's coverage will remain available at a lower cost than 5G.

- URLLC offers for a faster and better quality of service in critical function of mobile devices and include the control of autonomous vehicles [7]. To gain widespread acceptance of the 5G network, both 5G networks and chipsets must improve their

performance. This step will take a couple of years, and the development of these components will most likely take another four to five years [9].

D. Impact of 5g on vertical industries

5G is a "key technology" because its potential data rates meet the growing demand for data which will be driven by the rise of smartphones and networked communicating objects. It is expected to develop [12] cloud computing, integration, and interoperability of communicating objects and smart electrical networks in a home environment, contributing to the development of the "smart city" concept. It could also develop 3D or holographic image synthesis, data mining, big data and Internet of Things management, complex interactive and multi-player games, instantaneous automatic and assisted translation, or remote control in fields such as telemedicine, autonomous vehicles, and industrial automation [10]. Among the smart city industry verticals that will be transformed by 5G are energy, manufacturing, logistics, healthcare, manufacturing, entertainment, and media and will have a wide variety of specific applications at the core of each of these areas.

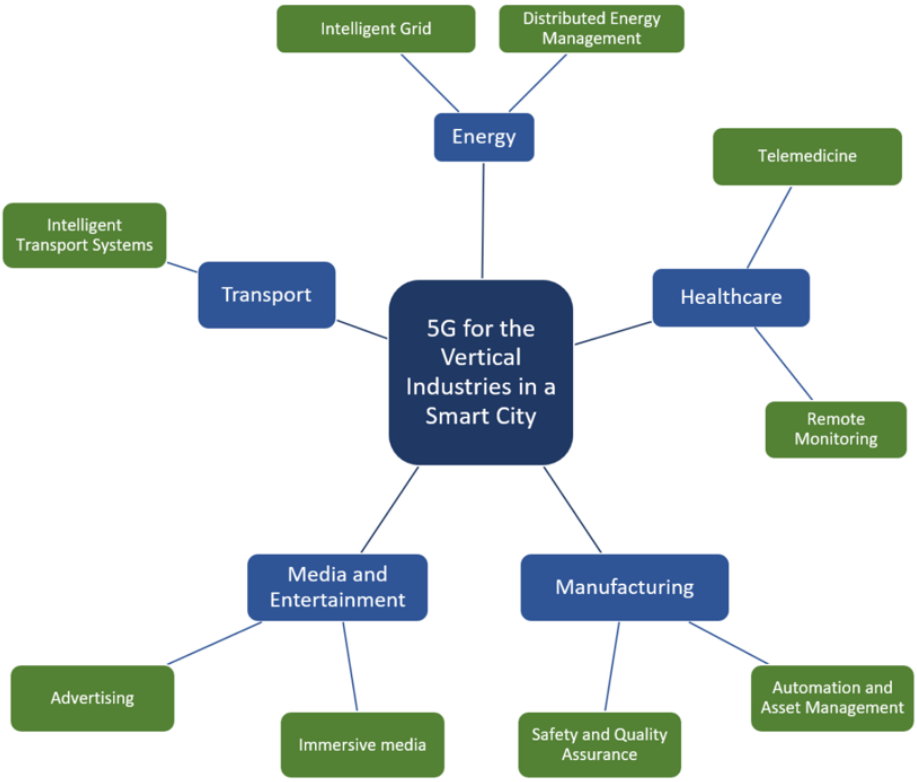


Fig 5- 5G for the vertical industries in a Smart City - source: [8]

a) Energy

Electricity generation could eventually exist through a smart grid network. It is a network that promotes the flow of information between suppliers and consumers to adjust the flow of electricity in real time and allow for more efficient management. This revolution could be enabled using a more powerful communication network such as 5G technology.

b) Healthcare

Aging Western populations and longer life expectancy will create a huge market for the healthcare sector. Scaling up the means of care will have to involve digital technologies and IoT. This transformation of the sector, if it is to keep pace with the increase in demand, will need to be aided and accelerated using the 5G network, both for research and for real-time diagnosis of patients.

c) Manufacturing

The increase in productivity brought about by the robotization of certain industries has been unprecedented in human history. The ability to automate, speed up production and delivery functions relies heavily on digital technologies. Among the additional applications that 5G will provide, we include autonomous guided vehicles, as well as the control and management of a larger number of robotic machines.

d) Media and Entertainment

The last two decades have been marked by the emergence and democratization of media and entertainment via the Internet. The colossal revenue generation of media and entertainment companies has been enabled by wider and more fluid access to content. For online gaming, for example, real-time access is critical. The same is true for video streaming. Accelerating streaming capabilities will open opportunities for augmented reality and greater consumer consumption of this content. Media could become immersive and provide users with a multi-dimensionnal experience [9].

e) Transport

According to the article [8] 5G could increase revenues in Europe's automotive and transportation sector by more than 200 billion euros. 5G by providing increased connectivity between vehicles. The efficiency of transportation is significantly linked to the speed and power of the information system that manages it. For smart cities, this is a considerable challenge. Integration through a system interconnected with infrastructures, vehicles, and users, and based on 5G could offer improvements in terms of decision making, resource saving and travel time reduction. The appearance of autonomous vehicles in urban space and the integration of urban data in real time, will serve the transport, logistics and e-commerce sector.

IV. Web-based data collection

A. Preliminary remarks

In this section, we explain the process and tools that helped to conduct the collection of basic data and spatial coordinates of businesses in Milton Keynes. This covers, among other things, the listing of companies in Milton Keynes and the collection of their basic information, including address and industrial sector, geolocation and website search, the listing of companies listed as startups and finally the extraction of textual data from the websites of these companies. This data collection could be done using different services detailed below and was implemented in Python language. We discuss the nature and origin of this data as well as the methods used to obtain it.

B. External sources and services for collecting business data

1. Companies House API

To build up a database we needed to collect data on businesses in Milton Keynes, including: their name, address, and sector. To do this we used the Companies House API, which is the UK's business register. This API allowed us to retrieve information on the 22,000 active businesses in the city, including their SIC code, which is the type of business they report. The use of this API is free, but the response time is limited to at least 0.5 seconds per request, so the data collection took more than 3 hours.

2. GoogleMaps API

Using the Companies House API, the addresses of businesses in Milton Keynes were extracted. These had to be mapped to a geographical position in a coordinate system (*latitude, longitude*) required for geospatial analysis. We therefore used the Google Places API to make the link. This API has the advantage of being free, thanks to the monthly credits offered by Google, and offers very good geolocation results overall. Moreover, it also gives

the company's website when it has been identified. With a rate of one request per second, the processing of the list will have taken more than 5 hours.

Several precautions were taken before undertaking the queries. Only active companies with a SIC code was kept. Companies that may have 'Limited' or 'Ltd' in their name were removed to obtain more reliable results. We also assessed for a sample of companies, different sets of features to qualify the address. The optimal query has the company name, followed by the street number and name, followed by 'Milton Keynes' and 'United Kingdom'. Using the postcode meant fewer responses from the API. In addition, after removing the outliers, there were still 20,065 businesses geo-located in Milton Keynes.

The use of the GoogleMaps API made it possible to obtain both the geographical coordinates and the websites of 5,210 companies (success rate of 26% compared to all the companies). In total, 2'722 different websites were retrieved for this set of companies. We note that a website can be associated with different addresses, as is the case for companies with several locations. We also noted the presence of several companies located at an identical address or very close. This leads to having the same website for companies located in shared offices. We will try to distinguish these places where there is a high concentration of companies, which are a priori considered as target areas for 5G.

3. Crunchbase service overview

Crunchbase is an online service used by investors for sourcing startups. We signed up for a free trial of this service to access a list of 488 startups in the city. The drawback of this data source is that it does not provide the reference number of the companies, but only the name of the service or product. We tried to identify the legal name of these companies, but it was not satisfactory. It is therefore impossible to discern these startups within the Companies House database. On the other hand, more than 95% of these startups had a website. This is not surprising, as startups need a public presence on the internet more than any other type of business. Many indicators are given by Crunchbase, such as fundraising dates, industry, or revenue range, but are not available for the majority of startups, so we chose to leave them aside.

4. Crunchbase data collection with Selenium

Selenium is an open source automation tool for emulating user interactions with a web browser. This tool allows us to work on the most common languages for web programming, including Python, and its global features cover simple and systematic scenarios in which we will seek to locate UI elements where coveted data is located. We used this tool to access the Crunchbase site and extract a table of data spread over several dozen pages. We first inspected the source code of the page, and then located the tags where the company information is located. With this script, the details of 488 companies, contained in 50 variables, were collected, and saved in a csv file.

Unfortunately, neither the address nor the legal name of the companies was provided by Crunchbase. However, a search on GoogleMaps or on Open Corporate, with a few well-chosen keywords, could provide the addresses of these companies. We then proceeded to manually search for the addresses of each of the 488 companies. In total, we found addresses for 386 of them. This work was very long and tedious but essential for the geospatial analysis that follows.

External service name	Price	Average rate	Company registration number	Number of hits	Number of collected websites	Registered address available	Registered industry	Technical collection
Companies House	free	2 requests/s	Yes	22'000	Not applicable	Yes	SIC Code	Python script, in-built API
Crunchbase	$29 / user / month, billed annually 1 week free trial	Not applicable	No	488	467	No	Industry focus	Python script, selenium
Google Places	20$/1000 request 300 credit	1 request/s	Not applicable	20'065	2'722	Not Applicable	Not Applicable	Python script, in-built API

Table 1- Information on external services used

C. Collection of text from company websites and online articles

1. Collection of text from websites

In our scenario, we collect only the main page of the website. It would have been possible to browse each site to store more text. But in general, the main page of a website gives a good indication of the nature of the service. [13] We used a Python script for this text extraction. We omitted all dynamic and visual aspects of the sites to extract only the raw text. With the urllib module, we extracted the html code of each main page, then we parsed it with the BeautifulSoup module. We also had to deal with errors due to the inaccessibility of some websites, or due to a very long access time. We made new access attempts to get the maximum of sites. Finally, we filtered the answers with sequences indicating an error, like '401 (Unauthorized)' or '404 (Not Found)'. But also, the answers containing less than 80 words were removed because they are considered as not giving relevant information about the company's service or correspond to errors. We finally obtained 2'080 websites out of the 2722 related to Companies House.

2. Getting articles from TechCrunch

To get content on 5G, we turned to the technology magazine TechCrunch. This choice seemed relevant to us because the magazine publishes content of good semantic quality and in sufficient quantity. This source, unlike social media like Twitter, provides data that is not affected by language errors or jargon, although TechCrunch may have some jargon related to technology or investment, with in some cases subtleties of language that remain a challenge in the NLP domain. This source will therefore serve as a lexical knowledge base. By searching the TechCrunch website, we obtained 970 articles including the keyword 5G. To retrieve these articles, we used Selenium, in a two-step approach. First, we went through the dozens of result pages for the word 5G. For each page, the algorithm identified the links to the articles, hidden in the hmtl code. Then in a second step, we accessed the links, as in the approach described in the previous paragraph.

V. Industry-independent geospatial analysis

A. Preliminary remarks

In this chapter we undertake the identification of the target area based solely on the density of businesses in Milton Keynes. This includes geo-located business listings from the Companies House and CrunchBase sources. This is based on the simplifying assumption that the communications capabilities enabled by 5G will benefit the area's most densely populated with businesses. Indeed, these business-dense locations indicate that the area is commercially attractive and that a spectrum of actors - consumers and businesses - will have greater a priori use of 5G than at the edge of these high-density areas. In particular, the density of startups will serve as an indicator of the attractiveness of an area. Incubators and business incubators, for example, are among the target areas. These places gather innovative companies, often with a technological focus, and have natural reasons to want to be given priority in the use of 5G service. This technology indeed offers a leverage for productivity, by facilitating access to information, coordination of certain systems and related services in the field of IoT for example. We will therefore use the two lists of traditional businesses and startups, in Milton Keynes to undertake this geospatial analysis.

B. Map of Milton Keynes' businesses

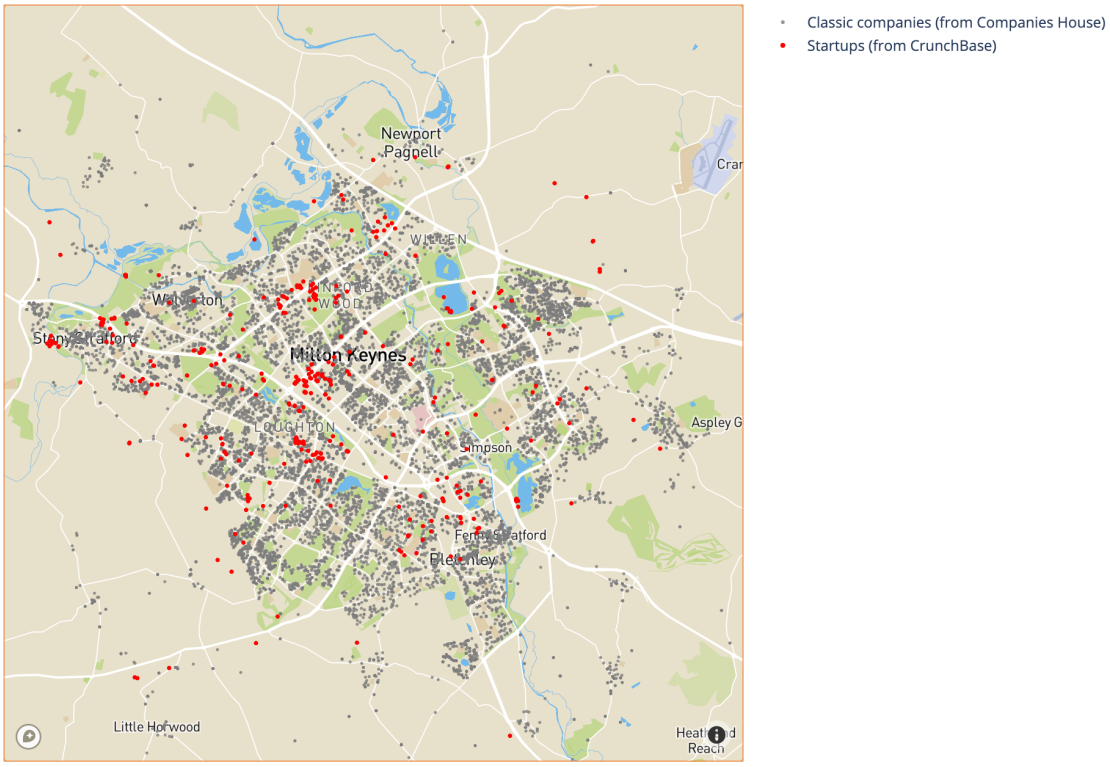


Fig 6- Geospatial distribution of companies and startups in Milton Keynes

C. Kernel density estimation

The kernel density estimation method is used in geospatial analysis to identify and estimate the concentration of points of interest. Kernel density estimation is a non-parametric method of estimating the probability density of a random variable. It is based on a sample of a statistical population and allows to estimate the density at any point of the support. In our case, the samples are startups or classical companies and follow a spatial distribution that we seek to estimate. It is a way to evaluate the concentration of companies at any point in the city.

The estimation is done by centering a function (obtained from a window, or kernel) on each observation and averaging the functions over all observations. A bandwidth h that is too

small would reveal undesirable details, and a bandwidth that is too large would smooth out relevant details. We therefore tried different bandwidths to obtain the following result.

D. Density map of Milton Keynes' businesses

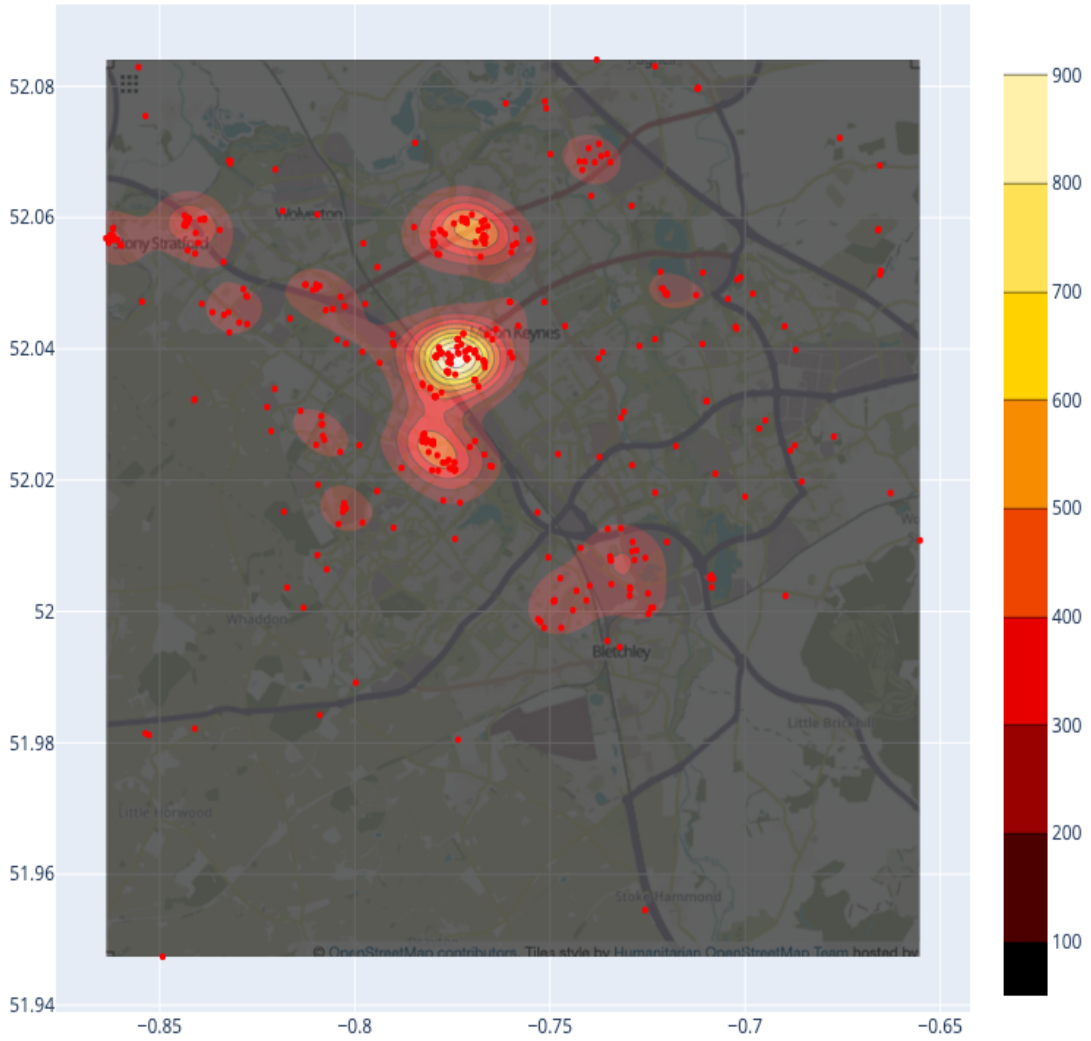


Fig 7- Density of startups in Milton Keynes - Gaussian Kernel Density Estimation, Bandwidth : 0.15

The representation of the startup density highlights a central group and several auxiliary groups. If the clusters are clearly distinguishable, it can be expected that this is due to specific infrastructures for startups, such as incubators, or to the phenomenon of business concentration, explained by Michael Porter [14]. This density obtained can be reused during the aggregation phase of the indicators.

E. Review of the approach

This simple approach is an effective way to see the important points of activity in the city and is useful information for the general implementation of a communications network. The shortcoming of this method is that it fails to distinguish between companies that will specifically need 5G for their business. Therefore, for the rest of this thesis, we will incorporate data on company activities, whether it is information from company websites or their known activity on Companies House or Crunchbase.

In geography, the Local Quotient is an indicator used to assess the strength and size of a particular industry in an area. It is a way of quantifying the concentration of a particular industry or even demographic group in a region relative to the country. Being part of a cluster increases the productivity of companies by facilitating the supply of inputs, access to necessary information, technology and institutions, coordination with related companies, and the measurement and stimulation of progress. We undertake a geospatial analysis which is based on companies' activity.

VI. Information retrieval methods

A. Remarks

The identification of business activities is a necessary step for many works in economics, finance, or for business development purposes. As previously mentioned, standard industrial classification is old and inadequate for describing emerging industrial sectors and its use to find target areas for 5G may be limited by its lack of granularity [1]. We therefore undertake an approach from a machine learning and NLP perspective that aims to analyse the *CorpoWeb* and *Tech5G* corpora to support target area identification, based on the themes extracted from these document sets. This is inspired by the article [1] which uses topic modeling to propose an alternative classification of industrial sectors. We would like to apply these techniques to the 5G-related texts grouped in the *Tech5G* set to help target the relevant companies. In this section, we propose a review of the literature on document embedding and topic analysis, which are essential for knowledge extraction.

B. Classical approaches

1. Latent Dirichlet Allocation

Topic modeling, in the field of machine learning, is a general approach to the problem of information retrieval in textual documents. It consists of a natural language processing technique for extracting topics from a set of documents. In 1999, T. Hofmann designed one of the first thematic models, Probabilistic Latent Semantic Analysis (PLSA), which inspired Latent Dirichlet Allocation (LDA) [15] which is now one of the most widely used models in the field of topic analysis. It is a type of analysis that is said to be unsupervised because one does not have to know the topics beforehand. They are given by the analysis. Another method that is related to LDA is its version called Hierarchical Latent Tree Analysis

(HLTA) [16]. This method makes it possible to restore the hierarchical characteristic of concepts in the analysis of topics.

The LDA model assumes that each document i is a mixture θ_i of a small number of topics or themes α. We therefore sample for each document i a topic distribution $\theta_i \sim Dir(\alpha)$ where Dir() denotes a Dirichlet distribution. This generates an initial "topic model": topics present in the documents and the words defining the topics. This topic model is highly implausible because it is randomly generated.

We seek to improve the randomly generated topic model in initialization. To do this, in each document, we take each word and update the topic to which it is related. This new topic is the one that would have the highest probability of generating it in this document. We assume that all themes are correct, except for the word in question. More precisely: for each word w in each document i we compute two quantities for each theme α :

- $p(\alpha \mid i)$ the probability that document i is assigned to topic α
- $p(w \mid \alpha)$ the probability that the topic α in the corpus is assigned to the word w

We then choose the new topic t with probability $p(\alpha \mid i) \times p(w \mid \alpha)$. This corresponds to the probability that topic α generates word w in document i.

By repeating the above steps, enough times, the assignments stabilize. The mix of topics present in each document is obtained by counting each representation of a topic (assigned to words in the document). The words associated with each topic are obtained by counting the words associated with it in the corpus.

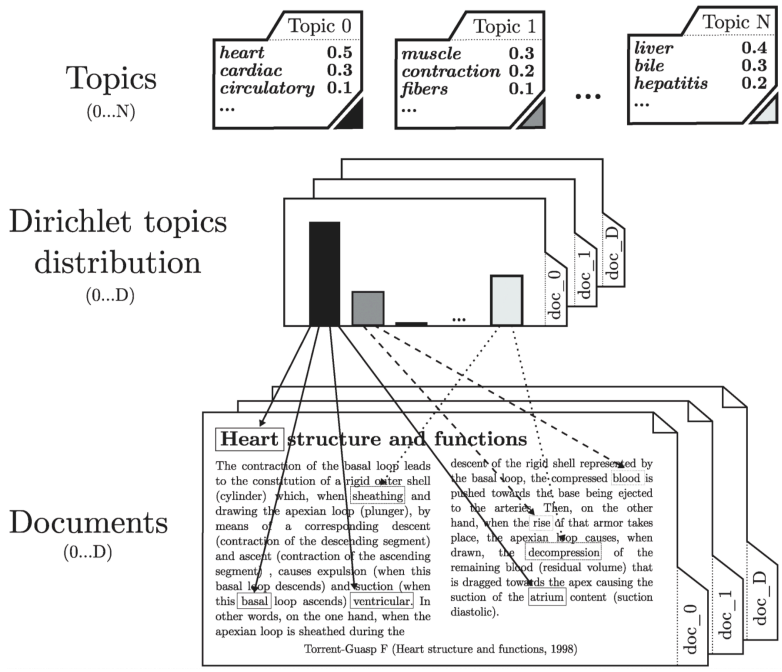


Fig 8- LDA underlying intuition. Generation of documents through topics following the Dirichlet distribution. Source: [17]

This is a dimensionality reduction method, as it allows us to reduce the texts to the main descriptive characteristics that are the topics. These topics extracted from the corpus are groups to which each text is related according to the frequency of each word it possesses. We say that topic modeling is a soft clustering method as opposed to hard clustering where each document is assigned to a single class. An intrinsic drawback of this method is that it does not consider the word order of the documents and fails to capture the semantic context of each word within a sentence or paragraph.

2. Word embeddings

Word embedding is a natural language processing technique that is used to represent words as high-dimensional vectors, aiming to preserve the semantic links of words in the vector space. This has allowed powerful applications for many NLP tasks and offers very interesting possibilities for document analysis and clustering. The technique named word2vec appeared in 2012 [18] is a supervised method based on artificial neural networks to create these vector representations. The vectors are constructed so that the semantic similarity of words is measured in the vector space by the cosine of the angle between the vectors. To obtain these vectors for each word, a variety of techniques rely on training on

large volumes of textual data. Word2vec for example was designed for this task by a research team at Google. [19] The authors show that these vectors offer state-of-the-art performance for measuring syntactic and semantic similarities of words. Two architectures were presented. The continuous bag-of-words (CBOW) and the skip-gram model. The first one aims at predicting a word given its context (n words on the right and on the left) and the second one aims at predicting these n context words from the central word. This is an example of self-supervised learning. These approaches have been tested to vectorize sentences and paragraphs. The principle for Doc2VecC [20] is to represent each document from the average of its word embeddings and captures the semantic meanings of the document during learning.

C. Advanced methods

1. Transformers

The transfomer is a deep learning model introduced in 2017. Like RNNs, transformers are designed for translation and text summarization. The difference with RNNs is that the data does not need to be processed in order. This allows to parallelize the computations, to have a faster training and thus to be able to train on large corpora. These models have therefore taken the place of LSTMs based on recurrent neural networks and have led to the creation of models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-Training Transformer), which have been trained on large data volumes. These models are now widely used in the NLP domain to perform a wide variety of tasks, including information extraction and disambiguation. The BERT model, for example, can be used to create word embeddings. This then allows us to perform various tasks such as, classification, clustering, similarity evaluation.

2. Clustering & dimensionality reduction

Document clustering aims to identify groups of similar documents based on the internal characteristics of the texts, be it a literary genre, a set of topics or concepts reflected in the texts. This involves the use of descriptive text variables to maximize both the similarities between texts in the same group and to maximize the differences between representatives

of the groups of texts. Before running a clustering algorithm, a dimensionality reduction step may be necessary when working with high dimensional vectors.

However, the dimension reduction step may cause the data to lose signal. In this area, there are classical linear methods such as PCA and non-linear methods, which adapt to the local structure of the data. One of the latest techniques in this field is the t-SNE. It has been extended with the UMAP method (Uniform Manifold Approximation and Projection for Dimension Reduction) [21]. The UMAP algorithm competes with t-SNE for visualization quality and preserves more of the global structure with superior execution performance.

Among the clustering algorithms, the K-means algorithm is commonly used. DBSCAN [22] (density-based spatial clustering of applications with noise) is a data partitioning algorithm proposed in 1996. It is a density-based algorithm that relies on the estimated density of clusters to perform the partitioning [22] (see Figure 8). The algorithm is very simple and does not need to be told how many clusters to find. It can handle outliers by eliminating them from the partitioning process. The clusters do not have to be linearly separable (as for the k-means algorithm for example). However, it is not able to handle clusters of different densities. The HDBSCAN [23] algorithm searches for clusters of points similar to those of the DBSCAN method, except that it uses variable distances which allows to search for clusters of variable density depending on the probability of the aggregation. The HDBSAN method does not classify all points. Some can be left in the outlier category if no proximity is found with a cluster. Below are two examples of clustering performed by k-means and HDBSCAN.

	Flat	Hierarcical
Centroid / Parametric	k-means GMM	Ward Complete-linkage
Density/ Non-Parametric	DBSCAN Mean shift	HDBSCAN

Fig 9- Typology of clustering methods

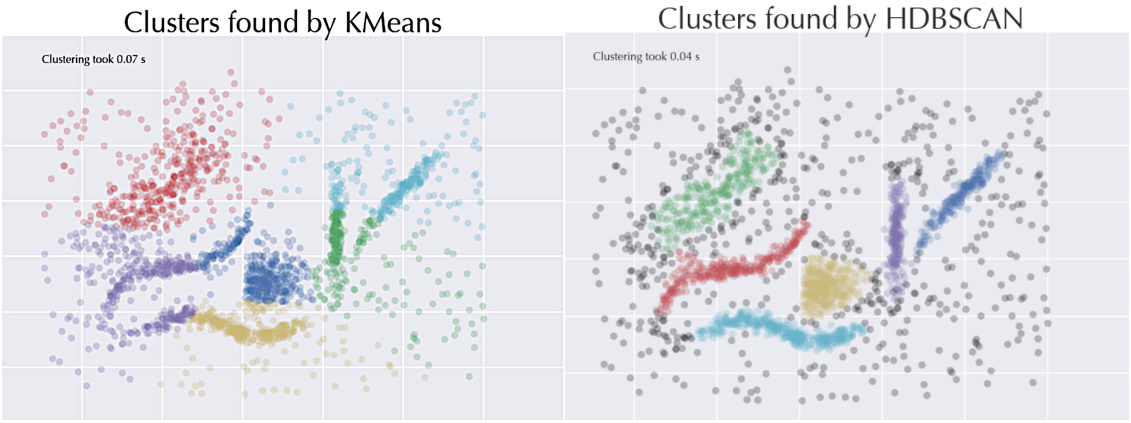


Fig 10- Comparison of clusters found with the k-means and HDBSCAN methods

VII. Latent Dirichlet Allocation

A. LDA model details

1. Algorithm

LDA is a generative probabilistic thematic model that aims at discovering latent thematic structures within a corpus. The latent thematic structure is expressed as themes and proportions of themes per document and is represented via hidden variables. The LDA generative method describes a random process based on probabilistic sampling rules applied to document elements [24]. In other words, the process aims at finding the underlying themes most likely to be at the origin of the documents. In the end, we obtain the posterior distribution that captures the hidden structure given the observed documents. Here is the procedure in detail:

1) For each topic $k \in \{1, ..., N\}$
    (a) draw a distribution over the vocabulary V, $\beta k \sim Dir(\eta)$
2) For each document d
    (a) draw a distribution over topics, $\theta d \sim Dir(\alpha)$
    (b) for each word w within document d
        i) draw a topic assignment, $zd_{,n} \sim Mult(\theta d)$, where $zd_{,n} \in \{1, ..., K\}$
        (ii) draw a word $wd_{,n} \sim Mult(\beta zd_{,n})$, where $wd_{,n} \in \{1, ..., V\}$

Each topic βk is a multinomial distribution over the vocabulary V and comes from a Dirichlet distribution $\beta k \sim Dir(\eta)$. Additionally, every document is represented as a distribution over K topics and come from a Dirichlet distribution $\theta d \sim Dir(\alpha)$. The Dirichlet parameter α represent the smoothing of topics within documents, and η denotes the smoothing of words within topics.

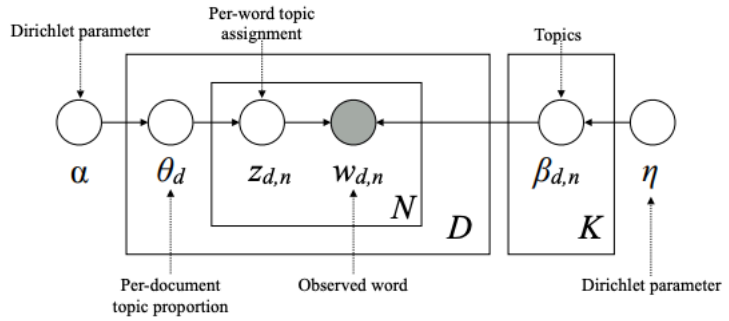


Fig 11- LDA graphical model. Nodes represent random variables, and the edges show their inter-dependencies. The grey node is the observed random variables. The other are hidden variables.

2. Topic coherence score

After approximating LDA's posterior distribution, the K topics are represented as multinomial distributions over V. In each topic distribution contains is assigned a different probability to each of the words. The words within topics with high probability are words that tend to co-occur more frequently [24]. LDA outputs a number K of topics. If K is low, it will result into topics too broad to capture the activities denoted in the documents. If K is too high, topics will be difficult to interpret. have been merged. Thus, it is important to consider a certain criterion to choose the right value of this parameter. A measure such as the predictive likelihood of held-out datasets has been proposed to evaluate topics' quality. However, it can negatively affect human interpretability and make topics less coherent. Topic coherence measures can be performed by analyzing the words that are related to the topic and its top N words. The goal is to develop a measure that can identify topics that are most likely to generate interest among humans. This measure should be able to predict the likelihood that a topic will generate interpretability [25]. The article proceeds to different topic coherence measure tests [26] and build a new one, Cv which aims at offering better interpretability, based on human topic ranking data. Cv is built as follow:

(i) building word pairs from the data,
(ii) estimation of pair likelihood,
(iii) quantifying how a word set entails another word set
(iv) aggregation into an overall topic coherence indicator

B. Clustering of business activities in Milton Keynes based on LDA

1. Text corpora

For the classic companies in Milton Keynes (not Startups), 2248 websites have been extracted and constitute the *CompHouseWeb* corpus. And for the companies listed on Crunchbase we have 320 sites that make up the *CrunchBaseWeb* corpus.

2. Pre-processing

Before proceeding with the analysis, we removed stopwords, which are common words in the English language (e.g., below, each, because, more, only, some, why, etc). These words do not bring any specificity to the texts of the corpus and therefore generate noise. Numerical data are removed. Then, the texts were tokenized, i.e. reduced to atomic elements.

Various pre-processing can then be applied to improve the results. For example, Porter's abbreviation algorithm removes the most common morphological endings of words in English [4] words and brings words of similar meaning under the same token.

Original words	Stemmed word
psychological, psychology, psychologic	psycholog
administration, administrative, administrate	administrators

Fig 12- Stemming examples with Porter's method

We also extracted the n-grams from the texts. That is, to put under a single token the most frequent co-occurrences of n words. We can have as bigrams, for example privacy_policy, because these two words usually appear together, or surrounding_area. This is used to reduce the text to as few features as possible before processing by LDA.

3. Topic modeling for CompHouseWeb corpus

a) Number of topics

We hope via this approach to obtain topics that are like industrial sectors. The standard industrial classification has 21 main classes and 88 subsections. As explained in the article [1] it is in this range of values that the best compromise in terms of number of subjects is probably found. We therefore undertake a test of the model for a number of topics varying between 5 and 50, with twenty computational runs at each iteration. We use the topic coherence measure based on the degree of semantic similarity between the high-scoring words of each topic. A high score will indicate that the topics tend to be interpretable and are not an artifact.

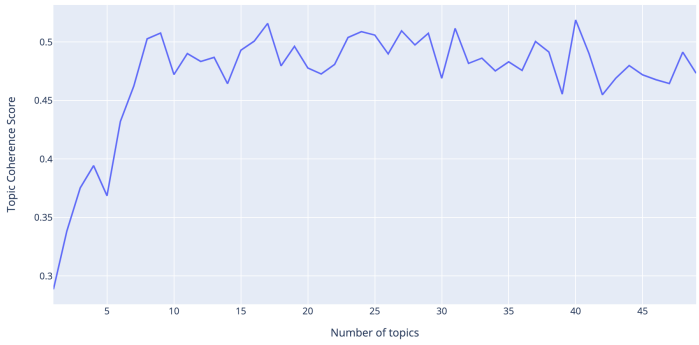


Fig 13- Topic coherence score - CompHouseWeb corpus

We observe that beyond 8 subjects, the coherence remains constantly above 0.45 whatever the number of selected subjects.

b) Resulting topics

Here below, the result obtained for a fixed number of 14 topics for which the first 5 words are displayed. We have also reported for each topic, the number of documents associated with the most representative word of the topic. If we look at the most present topic (service, cookie, solution, management, use), we notice that the terms are not specific enough to give us a precise idea of the activity. We must therefore enrich the list of stop words by adding the irrelevant words that emerge in the analysis. Several attempts are therefore necessary to

obtain a better result. We also notice that the terms appear with lexical similarities within the topics, which is a positive point. For example, "property, home, mortgage" represents companies associated with the real estate sector.

	Number of documents	Terms
0	149	system, product, equipment, solution, material
1	358	service, cookie, solution, management, use
2	43	door, window, trade, roof, art
3	145	training, course, class, learn, studio
4	51	coffee, glass, gift, watch, tea
5	165	property, service, home, management, mortgage
6	48	tea, health, home, patient, care
7	75	add, price, delivery, print, view
8	59	information, use, datum, personal, service
9	217	development, service, solution, take, work
10	321	home, care, service, child, clean
11	228	service, repair, vehicle, car, treatment
12	120	cookie, use, view, store, tax
13	269	home, book, store, order, table

Fig 14- Top 5 words of topics for CompHouseWeb corpus (14 topics)

	Number of documents	Terms	Supposed activity
0	197	recruit, now, appli, take, industri, retail, england, sustain, per, back	human resources
1	55	train, cours, learn, student, level, workshop, class, educ, skill, join	educational institution
2	41	autom, notari, process, public, technolog, document, industri, read, partner, global	unknown
3	53	deliveri, boiler, heat, ga, plumb, courier, logist, ship, quot, email	plumbing & heating
4	27	tune, abarth, car, fiat, vehicl, part, engin, key, merced, audi	vehicule dealer
5	10	speaker, sensor, power, research, cbd, system, rate, smart, sound, batteri	electronics
6	30	park, museum, safeti, visit, comput, centr, fit, tour, shirt, yoga	outdoor activities
7	111	properti, let, home, hous, mortgag, pcm, rent, agent, bedroom, buckinghamshir	real estate & services
8	24	system, use, applic, thermal, materi, control, comput, univers, led, scienc	engineering
9	15	travel, book, now, airlin, fli, holiday, flight, bum, pciaw, air	travel agency
10	138	open, read, peopl, health, covid, use, time, direct, counsel, post	library
11	13	truck, forklift, steel, conveyor, stainless, pallet, rack, section, warehous, denbigh	handling equipment
12	40	txm, frame, accessori, decor, furnitur, leather, watch, top, tabl, climb	decoration
13	82	recommend, thank, great, alway, remov, friendli, time, qualiti, profession, home	unknown
14	44	fibr, tattoo, treatment, skin, patch, page, cabl, facebook, plu, voir	skincare
15	148	electr, electrician, safeti, industri, instal, mainten, home, system, engin, commerci	electrical installations & maintenance
16	15	view, slide, bed, home, previou, next, bathroom, diamond, set, badminton	real estate
17	58	door, window, roof, price, glass, kitchen, trade, mug, doubl, christma	home design
18	47	skip, room, hotel, vat, home, wast, space, present, hire, view	hotels
19	232	account, consult, tax, financi, right, plan, make, home, email, use	accounting and consulting
20	40	hire, school, home, meet, wavendon, plant, centr, divis, facil, lift	recruitment
21	68	care, home, treatment, health, beauti, hair, live, donat, visit, patient	beauty & healthcare
22	39	church, job, commun, centr, regist, saab, healthcar, nurs, doctor, home	community
23	26	music, elit, structur, leo, nec, email, home, matti, pulvinar, use	unknown
24	87	children, dog, art, googl, ad, top, grass, onlin, blog, click	pet supplies & care
25	43	legal, advic, commerci, solicitor, law, employ, construct, partner, properti, privat	legal consulting
26	37	clean, floor, carpet, massag, cleaner, blind, commerci, belvoir, treatment, bodi	property maintenance, development, decoration
27	121	cooki, use, store, necessari, function, set, consent, inform, polici, privaci	web security
28	15	add, cart, view, oil, price, login, list, incens, pleas, wish	unknown
29	42	gift, wed, home, love, set, wax, care, beauti, walnut, dress	beauty
30	128	car, vehicl, book, repair, price, insur, make, mot, free, check	vehicule support
31	15	fit, bodi, train, andi, class, total, sport, trainer, membership, metabol	sports club
32	15	vacuum, coffe, pump, cup, spare, machin, valv, filter, special, print	specialised equipment
33	72	menu, food, order, tea, book, restaur, street, drink, park, pleas	restaurants
34	97	system, engin, web, secur, print, high, inspect, industri, brand, control	computer and data processing services
35	23	data, inform, person, use, may, parti, privaci, third, collect, process	web security

Fig 15- LDA applied to CompHouseWeb corpus, k = 36, with cluster labels

4. Topic modeling for CrunchBaseWeb corpus

a) Number of topics

We decide to apply the same procedure to *CrunchBaseWeb* which has a much smaller amount of text.

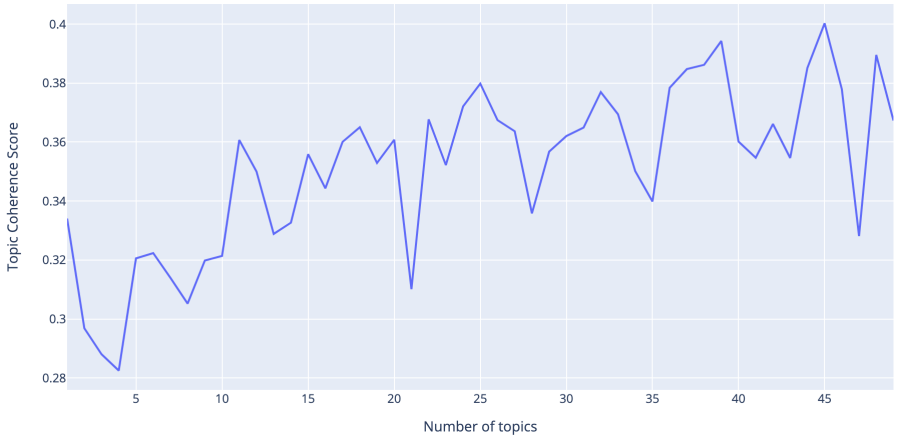


Fig 16- Topic coherence score - CrunchBaseWeb corpus

For this corpus of text which contains 320 documents, we notice that the consistency varies little according to the number of topics chosen. Moreover, it does not exceed 0.4. It is likely that this can be attributed to a lack of lexical richness or to a too great homogeneity of the corpus.

b) Resulting topics

Here are the topics that emerge, for k = 14. A larger number of topics is not relevant because this corpus contains only 320 documents.

	Number of documents	Terms
0	11	media, digit, social, seo, bed
1	28	cooki, use, necessari, consent, function
2	12	toggl, clamp, latch, engin, duti
3	12	read, technolog, light, envis, grow
4	17	properti, pcm, learn, let, add
5	44	intellig, leadership, skill, qualif, award
6	83	safeti, risk, fire, health, assess
7	13	energi, film, display, banner, print
8	14	bedroom, apart, amscan, knight, visibl
9	34	click, cast, time, celestra, view
10	22	therapi, blood, ozon, explor, recommend
11	11	aculab, data, technolog, secur, cloud
12	6	volunt, back, canal, speaker, make
13	12	sava, surveyor, kolbu, residenti, survey

Fig 17- Top 5 words per topics for CrunchBaseWeb corpus (14 topics)

We note that overall, the terms extracted denote themes that one would expect for a set of startups. Compared to the *CompHouseWeb* corpus, we find a greater number of terms related to digital, with words like cloud, data, media.

5. Review

This approach is based on the frequency of words that appear in the documents and not on the semantic links that connect the words. The LDA has made it possible to highlight the sectorial themes that appear in the companies' websites. This approach has the advantage of allowing mixtures of sectorial themes for each website and therefore for each company. As the algorithm returns the proportions of each topic for each document, a weighting of the topics according to their relevance to the implementation of 5G would allow us to build a relevance score for each company. For example, topics related to electronics, computers and

web technologies would have more weight than topics related to plumbing or restaurants. In the next chapter, we aim to improve this approach to obtain a more easily interpretable topic extraction.

C. Clustering of TechCrunch articles dealing with 5G based on LDA

1. Text corpora

We wish to use the Tech5G corpus, to test a knowledge extraction on these texts of a specific genre and dealing with 5G. The 557 articles of TechCrunch, available online at https://techcrunch.com/ constitute an interesting database to study to discover news related to 5G. These articles are aimed at investors and therefore offer information on new and existing opportunities and the most important fundraising events. For this reason, we are interested in the trends that emerge from these documents.

2. Number of topics

Below is the evolution of the coherence of the topics for k varying from 1 to 45. The graph is quite chaotic and gives a maximum coherence of 0.37 for a number of 18 topics. We could have chosen a lower number but preferred to obtain a larger variety of topics.

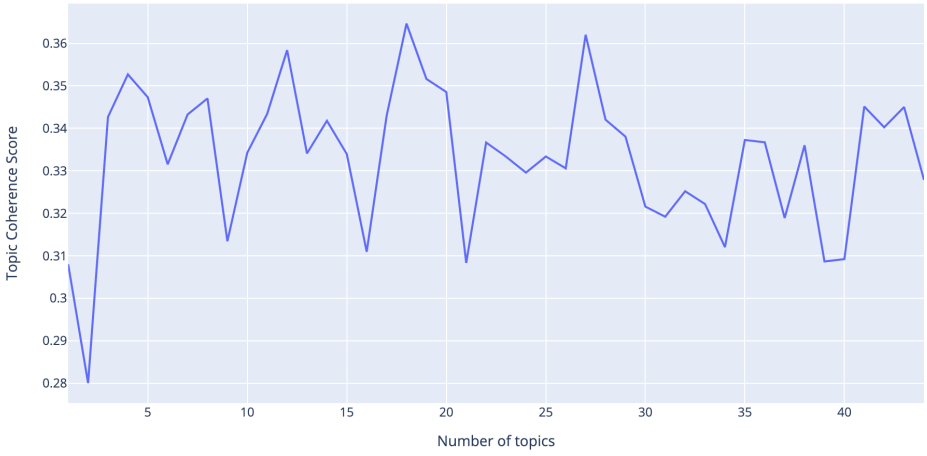


Fig 18- Topic coherence score - Tech5G corpus

3. Resulting topics

	Number of documents	Terms
0	85	company, technology, network, new, also, service, year, business, chip, market
1	105	phone, device, company, camera, new, also, year, well, big, time
2	10	qualcomm, company, shareholder, broadcom, acquisition, board, deal, chip, transaction, chipmaker
3	61	year, huawei, company, government, chinese, smartphone, market, network, sale, quarter
4	11	project, open_stack, company, kubernete, presentation, foundation, first, event, key, year
5	35	apple, company, new, device, pod, phone, year, time, pixel, event
6	4	network, carrier, system, mobile, government, security, card, payment, cell, also
7	8	broadband, rule, city, net_neutrality, service, wireless, company, internet, pai, deployment
8	3	device, network, connectivity, sim, starry, wireless, razer, user, chipset, first
9	26	app, game, apple, video, user, mobile, new, gaming, developer, platform
10	54	mobile, network, service, new, company, verizon, carrier, tv, also, consumer
11	28	investment, tech, startup, digital, also, new, industry, market, company, year
12	7	misinformation, company, claim, conspiracy, system, covid, vaccine, internet, also, virus
13	18	datum, security, also, technology, tech, chinese, risk, standard, issue, device
14	51	company, car, startup, vehicle, year, also, new, technology, system, platform
15	11	new, also, android, device, cloud, feature, enterprise, developer, app, container
16	38	company, year, new, also, investor, service, week, platform, firm, startup
17	2	accessibility, vision, blind, technology, sight, event, global, medium, session, tech

Fig 19- Top 10 words per topic for Tech5G corpus (18 topics)

The most quantitatively represented topics resulting from this modelling concern mobile telephony, with almost 1/5th of the texts dealing with this theme. Also, among the easily interpretable categories, we note the presence of documents about Huawei, Qualcomm, gaming, security, network operators and conspiracy theorists. Finally, this analysis of TechCrunch articles does not reveal any information that is useful for targeting industries. We will therefore consider another method of knowledge extraction and if it does not yield satisfactory results, we will qualitatively select the industry sectors of interest.

VIII. Transformer-based clustering

A. Introduction

In this chapter, we undertake a corpus analysis based on state-of-the-art NLP techniques: transformers. [27] This type of deep learning model is designed to handle sequential data and to perform a variety of language processing tasks such as translation. They have come to replace models derived from RNN (Recurrent Neural Networks) such as LSTM and exploit the attention mechanism. During the learning phase, the states associated with tokens (atomic elements of sentences) are updated, considering the states of previous tokens. The attention layer allows access to all previous states and weights them according to their relevance to the token being processed. The Bidirectional Encoder Representations from Transformers (BERT), designed by a team from Google [28] team, exploits both directions around each token. This tends to extract the contexts in which the words appear. This model available in opensource has been trained on huge amounts of data and allows free applications of this model. It has been adapted for topic extraction via the BERTopic library, which takes advantage of various state-of-the-art Machine Learning techniques. For these reasons, we have undertaken a topic extraction using this library

B. BERTopic architecture

The library proposed by BERTopic includes the use of various state-of-the-art algorithms, including the creation of embedding sentences with BERT, dimensionality reduction with UMAP [21]clustering with HDBSCAN [16]and then the use of more classical techniques to generate the words representing the classes via c-TF-IDF. Below is the architecture proposed in the BERTopic model, which has the advantage of being adaptable to the expected task.

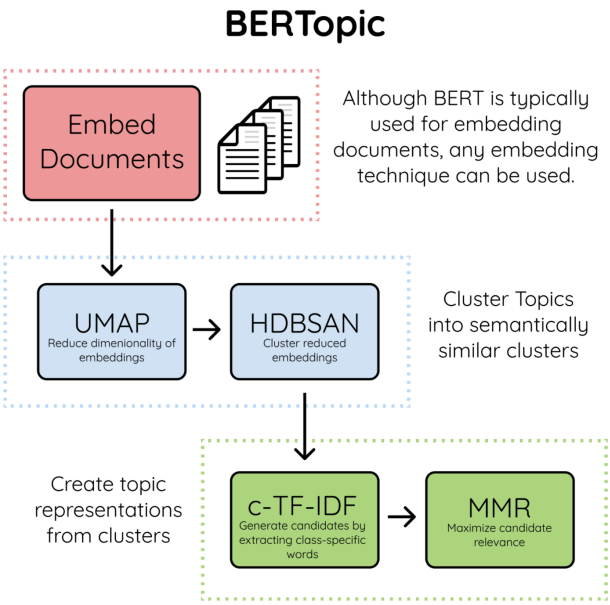


Fig 20- BERTopic architecture [29]

BERTopic for the document embedding part uses a Siamese network architecture. Two input sentences are passed to a layer of the BERT model which creates an embedding sentence for each sentence. Then, the embeddings of the pair of sentences are used as inputs to compute the cosine similarity. After a re-training phase, the algorithm aims to decide whether the input sentences are contradictory, implied by each other or simply independent. After this procedure, sentences embeddings are created for the never seen text, then aggregated to create a vector corresponding to the document. The proposed algorithm has the following 3 phases:

1 - Creating document embeddings

- Using BERT to extract embeddings from documents

2 - Clustering of documents

- Reduction of the dimensionality of embeddings with UMAP
- Grouping of documents into clusters to have semantically similar groups

3 - Representation of themes

- Extracting themes from classes with c-TF-IDF
- Improved consistency of terms with Maximal Marginal Relevance

1. Sentence-BERT

BERT is a generic model that does not require any task-specific architecture: it can be fine-tuned on a new dataset by simply adding an output layer. BERT is built with a Transformer architecture and pre-trained on two types of tasks: predicting hidden words in a sentence and predicting the next sentence in a text. The authors of BERT do not plan to extract sentence vectors from their model, but they demonstrate that simple transfer learning (extraction of word vectors without fine-tuning used as input to a new model) [30] can match the state of the art for a named entity detection task.

Sentence-BERT proposes not universal vectors, but a fine-tuning architecture of the BERT model specifically adapted to produce lexical sentence plunges suitable for certain types of tasks. This model modifies BERT into a Siamese network, with a final layer depending on the type of task on which the network is trained.

2. Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique which, like t-SNE, can be used to make visualisations, but also to make general non-linear reductions. The algorithm is based on two assumptions about the data:

-the data is uniformly distributed on a manifold
-the manifold is locally connected

Following these assumptions, it is possible to represent the variety using a fuzzy topological structure. The integration is determined by looking for a low-dimensional projection of the data whose topological structure most closely approximates the fuzzy topological structure. The solid mathematical foundations ensure that the algorithm is robust and interpretable. They are also applied to more complex problems in unsupervised learning.

3. Hierarchichal DBSCAN

HDBSCAN (Hierarchichal DBSCAN) is a clustering algorithm proposed by Campello et Al. in 2013. It assumes that:

-density-based clustering algorithms, such as DBSCAN, only cluster according to a global density threshold, which will prevent finding clusters of too variable density.
-Hierarchical clustering algorithms are also interesting but may have a hierarchy that is too complex and difficult to interpret.

Another problem encountered is the multiplication of parameters, influencing the result (for example, the number of classes must be specified for k-means). Note that a clustering algorithm is different from a partitioning algorithm, like k-means. The aim of the latter is to associate any element with one of the k groupings, by minimising the intra-clustering distance. In our definition of clustering, we allow ourselves to have points that do not belong to any clustering: they are considered as noise.

In short, HDBSCAN is a mix between a hierarchical clustering algorithm and DBSCAN. It will allow to consider clusters of different densities, requires only few parameters, gives particularly reliable results. Moreover, a powerful implementation integrated with sk-learn has been proposed.

4. From TF-IDF to c-TF-IDF

TF-IDF is a method for generating features from text documents [24] that is the result of multiplying two terms:

1- Frequency of terms (TF)
2- Inverse Frequency of Documents (IDF)

TF term frequency is simply the number of occurrences of a word in a document relative to the number of words in the document. Inverse document frequency extracts the

informativeness of certain words by computing the frequency of a word in a document relative to its frequency in all other documents.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

The term wij is high in a given document j if the term i is frequent there compared to the rest of the corpus. On the contrary, this indicator is low for words that do not bring specificity to a document. The TF-IDF considers documents without class distinction. We can modify the TF-IDF by grouping documents of the same class, as if they were a single document. In this way we can describe the classes with terms that constitute a topic for each class, with the advantage of having weights for each term.

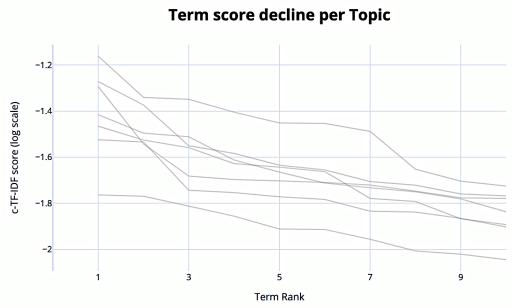
C. Clustering of business activities in Milton Keynes using BERTopic

1. Precautions

To obtain the result obtained below, we applied to the corpus as a preprocessing, the cleaning of special characters and the removal of documents in languages other than English. We removed all texts shorter than 80 words because they often correspond to errors or do not provide enough information. The corpus then counts 2055 websites. Similarly, words such as: 'mk', 'milton' , 'keynes' , 'world' , 'work' , 'uk' , 'service' , 'search' , 'contact' , 'use', etc., have been removed and only influence the representation part of the clusters based on c-TF-IDF and thus essentially favor the interpretability of the categories obtained. The number of clusters is set by the similarity of the terms we want within a category. The threshold is set at a similarity of 0.9. The terms that do not meet this criterion form the outlier category. To achieve the result, we played with the dimensionality reduction parameters of UMAP and the parameters of HDBSCAN. As before, we manually labeled the different categories, to make them correspond to activity domains.

2. Term score decline per topic

Each category predicted by HDBSCAN groups a certain set of documents that share similar semantic features. Since the documents have been clustered in a space that does not allow for interpretation, c-TF-IDF computation is performed on each of the clusters to make sense of them. For each group, we sort the terms by decreasing rank of c-TF-IDF score and represent on a vertical logarithmic scale the score related to the term of rank n. This allows us to see for each theme, the marginal contribution of the addition of the word. Beyond 10 words, we notice that the contribution of a word to the group is low.



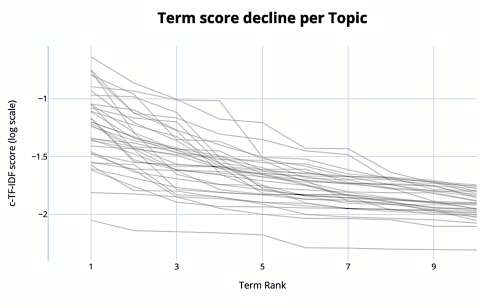


Fig 21CompHouseWeb corpora - term score decline Fig 22- CrunchBaseWeb corpora - term score decline

3. Resulting topics

Topic Id	Count	10 most representative words	Supposed activity
-1	771	business, services, management, information, privacy, customers, email, courses, 01908, clients	Outliers
0	95	car, vehicle, cars, mercedes, tyres, tyre, benz, vehicles, wheel, volkswagen	Vehicule dealer
1	73	food, menu, sauce, chicken, roast, adipisicing, dishes, restaurant, pizza, delicious	Restaurants
2	68	business, accounting, accountants, accountancy, accounts, financial, accountant, finance, corporate, audit	Accounting
3	63	sensor, cnc, sensors, machining, machines, aluminium, welding, photoelectric, inspection, metal	Industrial automation
4	61	licensee, licensed, construction, property, projects, residential, licensees, license, homes, landscaping	Property management
5	60	marketing, business, brand, services, clients, management, sales, customers, consulting, businesses	Management consulting
6	59	property, rent, valuation, estate, landlords, tenants, tenant, landlord, bedrooms, sell	Real estate
7	52	recruitment, candidates, jobs, job, manager, recruit, employee, clients, vacancies, salary	Recruiting
8	45	wedding, bridal, clothing, bridesmaids, wear, dress, womens, dresses, salon, suit	Event management
9	43	personal, vacuum, hotels, gift, information, privacy, home, furniture, categories, site	Unknown
10	41	cleaning, clean, window, removals, cleaners, cleaner, house, removal, cleans, cleaned	Cleaning
11	40	club, classes, training, gym, badminton, membership, camp, sessions, rooms, clubs	Sports
12	38	heating, plumbing, boiler, gas, engineers, boilers, repairs, plumber, radiators, plumbers	Heating & plumbing
13	36	care, support, autism, families, carers, residents, dementia, community, assistance, resident	Home healthcare
14	32	doors, blinds, windows, door, window, kitchen, hardwood, curtains, showroom, wood	Refurbishment
15	32	hours, message, 01908, place, app, testimonials, table, appointment, pm, menu	Unknown
16	29	cloud, services, microsoft, security, backup, silverbug, cyber, infrastructure, office, digital	Cloud infrastructure
17	26	logistics, freight, courier, transport, delivery, parcel, shipping, express, cargo, ccnp	Logistics
18	26	printing, print, frames, framing, printers, printer, banners, cartridges, flyers, google	Printing
19	26	office, centre, rooms, stables, offices, location, facilities, studios, home, hotel	Office rental
20	24	clinic, massage, acupuncture, therapy, physiotherapy, treatments, pilates, treatment, therapist, cupping	Health care
21	24	law, solicitors, legal, mediation, disputes, litigation, divorce, dispute, solicitor, issues	Legal
22	23	church, prayer, christian, ministry, worship, fellowship, muslim, faith, churches, muslims	Religion
23	23	incense, art, memorabilia, lighting, furniture, wooden, decorative, tables, ceramics, resins	Home decor
24	22	electrical, electrician, electricians, contractors, lighting, electrics, alarms, maintenance, installations, emergency	Electrical maintenance
25	21	construction, engineering, asphalt, concrete, geotechnical, roads, weighbridge, surveyors, ground, rhinophalt	Construction
26	21	skin, treatments, facial, brow, facials, acne, face, peel, surgical, peels	Beauty
27	21	mortgage, mortgages, financial, regulated, insurance, debt, finance, loans, lenders, adviser	Financial services
28	20	courses, students, training, cpfi, counselling, student, skills, college, tuition, seminar	Training
29	18	hair, salon, nail, hairdressing, gel, barbers, hairdressers, prices, nails, lacy	Hairdressing
30	16	taxi, airport, taxis, transfers, booking, vehicles, cab, cars, travel, airports	Transportation
31	14	dog, dogs, turf, breeding, pet, groomers, grooming, puppy, animal, pawty	Grooming
32	13	flooring, carpet, carpets, cfs, tiles, laminate, kitchen, showroom, polyflor, floor	Refurbishment
33	12	school, malvern, academy, students, brighton, learn, education, admissions, london, student	Higher education
34	12	fibre, energy, cabling, carbon, cable, crowdcube, deployable, cables, assemblies, mounting	Energy
35	12	reset, confirm, email, pizza, bookings, close, cancellation, beard, food, barbers	Food
36	11	children, support, charity, adoption, families, donate, family, childrens, parents, community	Charity
37	11	valves, systems, pumps, dryers, equipment, industrial, products, refrigeration, mixers, compressors	Industrial equipment
38	11	safety, logistics, protection, industrial, dangerous, security, transport, emergency, supply, rail	Industrial safety
39	10	security, alarm, alarms, safety, emergency, extinguishers, patrol, key, protection, parking	Property safety

Fig 23- Top 10 words per category for CompHouseWeb corpus - BERTopic

Topic Id	Count	10 most representative words	Supposed activity
-1	186	business, services, management, contact, privacy, systems, marketing, customers, technology, cookie	Outliers
0	17	design, website, marketing, business, services, clients, app, apps, privacy, agency	Digital marketing & SEO
1	16	security, services, cloud, microsoft, cyber, 365, secure, consultancy, ibm, ccl	Cloud data services
2	15	products, cards, balloons, packaging, accessories, balloon, gift, supplies, decorations, bags	E-commerce
3	15	business, accountants, accounting, taxi, companies, coupons, accountancy, businesses, payroll, accountant	Business intelligence
4	15	technology, vehicle, connectivity, amn, waam, mobility, connected, technologies, telematics, mobile	Internet of things
5	14	property, rent, seo, sale, leaflet, removals, houses, house, search, estate	Property management
6	10	envisics, marketing, events, automotive, holographic, business, robots, robotazia, technology, patent	Robotics & automation

Fig 24- Top 10 words per category for CrunchBaseWeb corpus - BERTopic

The result obtained with BERTopic shows a much better homogeneity of the terms within each category, compared to LDA. Almost all the topics could be interpreted very easily, except for two categories that do not make sense. We therefore expect to have a much better targeting of activities using this model even if the very high number of outliers penalizes the approach. The result is a qualitatively very good result for processing documents in a way that uses the semantics of the sentences and brings out coherent and tangible categories with respect to the real activity of the companies.

D. Clustering of TechCrunch articles dealing with 5G using BERTopic

1. Resulting topics

Below are the topics that stand out within the clustered articles. The results are qualitatively very satisfying because they are very easily interpretable. We also find dominant topics found with the LDA, concerning for example Huawei. A number of 5G related products are highlighted that were not as prominent with LDA. This includes connected devices made by Samsung, Apple and Google, cloud-related services, chip manufacturers such as Intel, the application development and gaming sector, autonomous vehicles, and telephone operators. It can also be noted, the presence of a category represented by terms such as "covid", "conspiracy", "information", "disaster", "virus" that has a priori nothing to do with 5G technology but echoes articles of the magazine addressing growing conspiracy theories around this field. Finally, this approach proved to be very robust in identifying keywords that are both highly relevant and consistent for each of the document categories. Furthermore, we note that the use of this source highlights business opportunities related to 5G that are at the top of the value chain of this technology. This naturally includes network operators and mobile and chip manufacturers. As 5G is rolled out, other sources of opportunity will emerge, but these are not really represented here. So, this analysis has not been conclusive in highlighting these new, more subtle use cases.

Topic Id	Count	10 most representative words
-1	118	company, companies, tech, technology, industry, mobile, startups, vr, broadband, startup
0	99	huawei, china, chinese, eu, european, technology, chinas, huaweis, risk, billion
1	61	phone, apple, pod, apples, phones, smartphone, devices, device, mini, pods
2	60	samsung, galaxy, device, fold, samsungs, devices, camera, foldable, smartphone, phone
3	51	company, cloud, fund, bank, funding, investment, startups, companies, investors, startup
4	38	wireless, network, cities, mobile, networking, bandwidth, verizon, connectivity, internet, networks
5	26	qualcomm, broadcom, qualcomms, shareholders, company, companies, chipmaker, takeover, nxp, stockholders
6	25	app, apps, apple, startup, facebook, startups, google, developer, developers, apples
7	17	intel, chip, intels, company, chips, computing, wireless, companies, technologies, technology
8	17	merger, deal, carriers, wireless, network, broadband, verizon, transaction, fcc, communications
9	16	covid, digital, coronavirus, internet, information, conspiracy, web, disaster, ai, virus
10	15	verizon, verizons, service, company, customers, business, aol, subscribers, revenue, services
11	14	pixel, google, googles, android, devices, hardware, device, phone, camera, battery

Fig 25- Top 10 words per category for Tech5G corpus - BERTopic

2. Visualise the hierarchical structure of the topics

A ward linkage function is used to perform the hierarchical clustering based on the cosine distance matrix between topic embeddings.

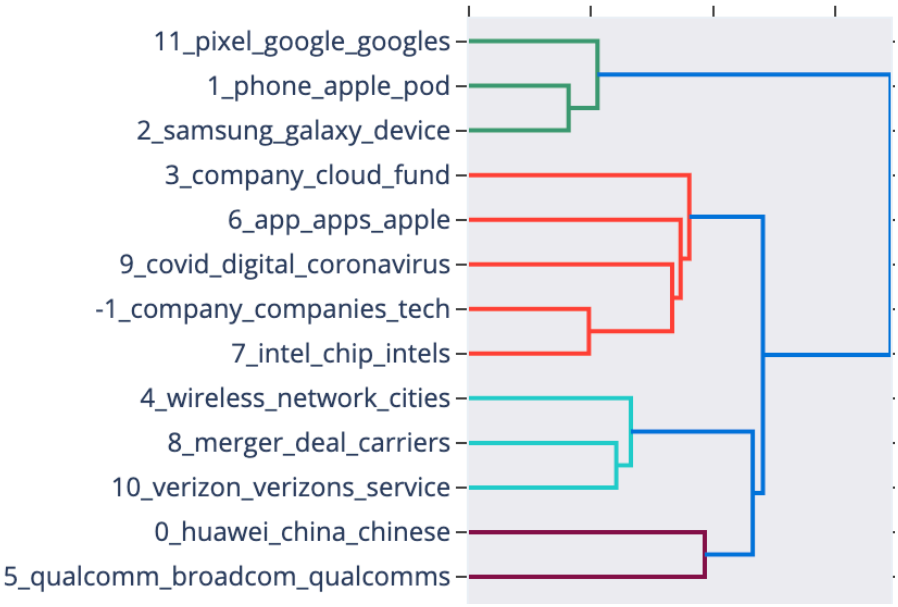


Fig 26- Hierachy of topics about 5G - BERTopic - Tech5G corpus

We can then visualize the proximities between the different subjects. From the point of view of the interpretation of this tree, the groupings make sense. We can therefore adjust the granularity of the clustering according to the diversity of topics we want.

IX. Cluster visualization

A. Preliminary remark

In this section, we undertake the geo-spatial visualization of the target area for 5G. We select from the business areas highlighted with BERTopic those most related to the deployment of 5G. This includes, among others, IoT, logistics or cloud services. The companies belonging to these semantic clusters are geo-mapped and then highlighted by a density estimation to highlight the target area in Milton Keynes.

B. Interest of the business cluster concept

A business cluster is a geographical area where companies from the same industrial sector are concentrated. These clusters allow, according to [14] the companies located there to benefit from a comparative advantage and to increase their productivity [31]. This would bring to the companies of the cluster, a competitive advantage based on knowledge, relations, and motivation, compared to distant competitors. Typical industrial clusters include Silicon Valley in the field of high-tech or the Champagne region in France for the wine business. In these regions, companies may be linked by common skills or technologies. This concentration effect tends precisely to accentuate itself, attracting in its field of gravity: talents, institutions, and investors.

This notion is interesting because, in our scenario, identifying target areas for 5G can be like identifying possible industrial clusters. An industrial cluster can indeed be favored by the existence of exchange and communication infrastructures: road network, rail network or communication means, such as fiber or cellular network. However, there is no standard methodology for identifying these clusters. They depend on each geographical and economic context. Moreover, the standardized industrial codes often used in industrial cluster studies do not necessarily reflect the activity of companies. Several limitations have been highlighted, in particular by [1] :

- The classification is old and not adapted to new technologies
- The referencing done by the companies themselves is not very precise: some declare themselves in a more general activity.
- Companies can have their activity spread over several different sectors
- The activity of a company can change over time

Range of SIC Codes	Division
0100–0999	Agriculture, Forestry and Fishing
1000–1499	Mining
1500–1799	Construction
1800–1999	not used
2000–3999	Manufacturing
4000–4999	Transportation, Communications, Electric, Gas and Sanitary service
5000–5199	Wholesale Trade
5200–5999	Retail Trade
6000–6799	Finance, Insurance and Real Estate
7000–8999	Services
9100–9729	Public Administration
9900–9999	Nonclassifiable

Fig 27- SIC main divisions [29]

C. Target activities grid

We use the activity clustering done with BERTopic on the CompHouseWeb and CrunchBaseWeb corpora to target the companies. We will not use in this approach the topic modeling done for Tech5G because it did not allow to highlight precise application cases. We therefore qualitatively identify the activities most related to 5G. The rest of the activities are thus discarded, to keep exclusively the clusters denoting applications for 5G. We have thus kept 3 clusters for each of the two corpora and are shown below with their corresponding words.

Activity	Terms
Cloud data services	security, services, cloud, microsoft, cyber, 365, secure, consultancy, ibm, ccl
IoT	technology, vehicle, connectivity, amn, waam, mobility, connected, technologies, telematics, mobile
Robotics & automation	envisics, marketing, events, automotive, holographic, business, robots, robotazia, technology, patent

Fig 28- Selected clusters from CrunchBaseWeb - BERTopic

Activity	Terms
Industrial automation	sensor, cnc, sensors, machining, machines, aluminium, welding, photoelecric, inspection, metal
Cloud infrastructures	cloud, services, microsoft, security, backup, silverbug, cyber, infrastructure, office, digital
Logistics	logistics, freight, courier, transport, delivery, parcel, shipping, express, cargo, ccnp

Fig 29- Selected clusters from CompHouseWeb - BERTopic

Each cluster refers to a set of websites. The companies are themselves mapped to these websites, with sometimes several companies associated with the same website, which happens for example when several establishments exist for a company. We have therefore filtered the sites of these clusters and then made a join with the table of correspondence between companies and website. We obtain a total of 320 companies associated with the CompHouseWeb topics and a total of 52 companies for those associated with the CrunchBaseWeb clusters. Both sets refer to 314 sites.

D. Map of the targeted businesses

Below is a map of the companies we have targeted and grouped by activity cluster.

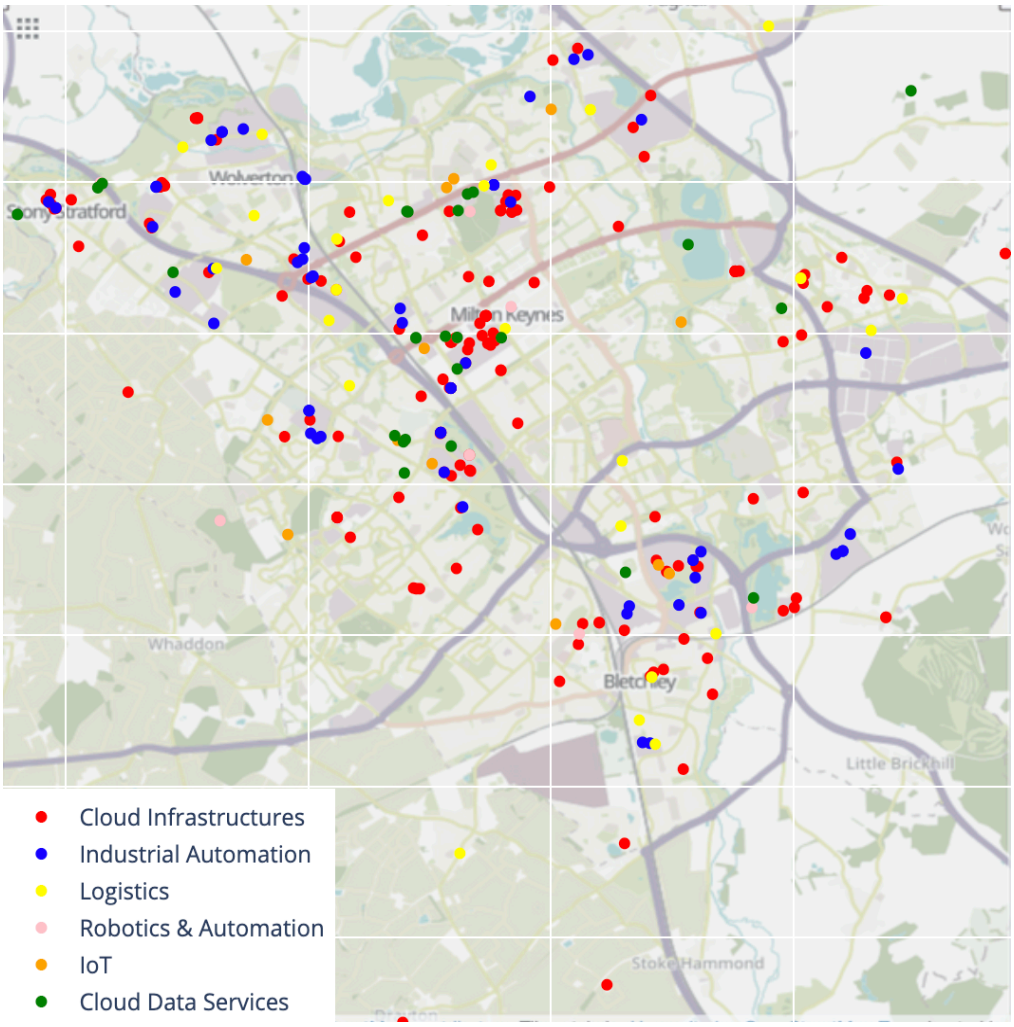


Fig 30- Map of targeted companies by activity in Milton Keynes - based on 6 clusters generated with BERTopic

E. Suggested target area for 5G

Here we use the Kernel Density Estimation method to estimate the concentration of points of interest and highlight the target area for 5G deployment. We consider a uniform weighting for all firms belonging to the 6 identified clusters. We chose a bandwidth of 0.1 which is a good compromise. We notice the highlighting of areas of the city that could not be distinguished with the approach based on the density of startups only.

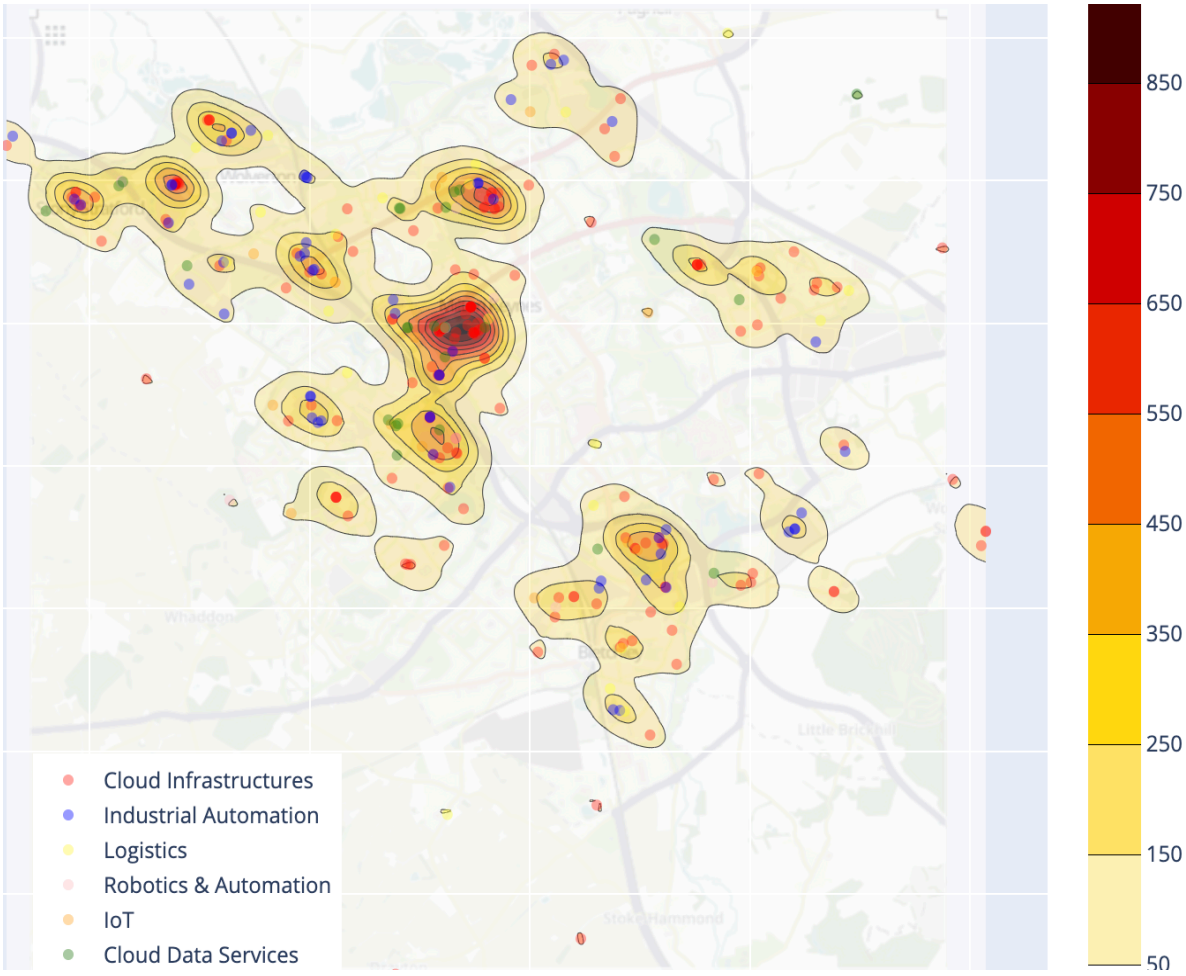


Fig 31- Heatmap of 5G related businesses in Milton Keynes - Kernel Density Estimation with a 0.1 bandwidth

X. Conclusion

A. Findings

The cluster concept has been attracting the interest of policy makers and the attention of academic researchers for years. The concept is so pervasive in the study of modern economies and global trade that it attracts the attention of many different disciplines. Through this thesis, we have proposed an analytical methodology through which the web is used as a source of dynamic information about business activities in the context of a defined geographical area. This work can be used in other contexts to extract the semantics of business activities and enable geo-spatial analysis that informs decision makers. We addressed the problem of identifying target areas for 5G, using natural language processing techniques that showed a very satisfactory knowledge extraction on the semantic level and succeeded in representing the target area through a heatmap that integrates our activity modeling.

B. Future work

This approach is a first step to succeed in detecting in text the activity of a company. We have shown that the use of natural language processing techniques, based on transformers, allows to improve the quality of the clusters obtained. However, this approach has only allowed to reflect the essential themes of the texts. We believe that a different approach would make it possible to highlight the semantics specific to the business domain, by highlighting, for example, the type of service or product, and the characteristics of the business, to better situate their activity in the economic landscape. This would require, for example, the creation of a specific framework for the description of the activity. The processing of a larger quantity of data and spread over all the industrial sectors would make it possible to situate the activity in relation to the industry.

C. Closing remarks

This master thesis has been a rewarding opportunity to explore different natural language processing methods and to learn how to implement internet data collection techniques. We hope that this work will provide a basis for a local analysis for 5G deployment in MK. Finally, this work is an approach of geo-spatial analysis from a data science perspective, and it is not intended to substitute an expert knowledge. The methodology is adaptable to other regions to geomap a specific business activity.

Bibliography

[1] S. Papagiannidis, «Identifying industrial clusters with a novel big-data methodology: Are SIC codes (not) fit for purpose in the Internet age? Computers and Operations Research,» *Elsevier,* 2017.

[2] Yogesh Gupta, «A Review on Important Aspects of Information Retrieval,» *World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering,* 2013.

[3] Z. Harris, «Distributional structure,» 1954.

[4] M. Porter, «An algorithm for suffix stripping,» 1980.

[5] M. Hanumanthappa1, «A survey on Information Retrieval System for Online Newspapers,» 2017.

[6] P. Gupta, «EVOLVEMENT OF MOBILE GENERATIONS : 1G To 5G,» *Electronics and Communications Department,* 2013.

[7] «Alcatel et le Coréen KT partenaires dans la 5G mobile,» 2015. [En ligne]. Available: https://www.lesechos.fr/2015/05/alcatel-et-le-coreen-kt-partenaires-dans-la-5g-mobile-264890.

[8] A. Gohar, «The Role of 5G Technologies in a Smart City: The Case for Intelligent Transportation System,» *Department of Electrical Engineering and Computer Science, University of Stavanger,* 2021.

[9] «5 Real Life Use Cases of 5G Ultra-Reliable Low-Latency Communication,» 2020. [En ligne]. Available: https://www.section.io/engineering-education/five-real-life-use-cases-of-5g-ultra-reliable-low-latency-communication-urllc/.

[10] «5G AND SMART CITIES: SMARTER SOLUTIONS FOR A HYPERCONNECTED FUTURE,» 2021. [En ligne]. Available: https://www.reply.com/en/industries/telco-and-media/5g-smart-cities.

[11] P. Orosz, «QoS Guarantees for Industrial IoT Applications over LTE - a Feasibility Study,» *IEEE International Conference on Industrial Cyber-Physical Systems,* 2019.

[12] «TECHNOLOGIES CLÉS Préparer l'industrie du futur,» *Ministère du numérique,* 2020.

[13] M. Thelwall, Webometrics, Annual Review of Information Science and Technology, 2006.

[14] M. Porter, «Clusters and the New Economics of Competition,» 1990. [En ligne]. Available: https://hbr.org/1998/11/clusters-and-the-new-economics-of-competition.

[15] T. Hofmann, «Probabilistic Latent Semantic Indexing,» *International Computer Science Institute, Berkeley, CA &,* 1999.

[16] T. Liu, «Hierarchical Latent Tree Analysis for Topic Detection,» *Department of Computer Science and Engineering The Hong Kong University of Science and Technology,* 2014.

[17] P. Celard, «LDA filter: A Latent Dirichlet Allocation preprocess method for Weka,» 2020.

[18] T. Mikolov, «Efficient Estimation of Word Representations in Vector Space,» 2013.

[19] T. Mikolov, «Efficient Estimation of Word Representations in Vector Space,» 2013.

[20] M. Chen, «EFFICIENT VECTOR REPRESENTATION FOR DOCUMENTS THROUGH CORRUPTION,» 2017.

[21] L. McInnes, «UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction,» 2020.

[22] M. Ester, «A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,» 1996.

[23] L. McInnes, «hdbscan: Hierarchical density based clustering,» 2017.

[24] S. Syed, «Examining Topic Coherence Scores Using Latent Dirichlet Allocation,» 2017.

[25] J. Chang, «How Humans Interpret Topic Models,» *Advances in Neural Information Processing Systems,* 2009.

[26] M. Roder, «Exploring the Space of Topic ¨ Coherence Measures,» *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining,* 2015.

[27] T. Wolf, «Transformers: State-of-the-Art Natural Language Processing,» *Association for Computational Linguistics,* 2020.

[28] J. Devlin, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,» *arXiv e-prints,* 2018.

[29] M. Gr, «BERTopic,» [En ligne]. Available: https://maartengr.github.io/BERTopic/tutorial/algorithm/algorithm.html.

[30] B. Mazoyer, «Représentations lexicales pour la détection non supervisée d'événements dans un flux de tweets : étude sur des corpus français et anglais,» 2019.

[31] A. Rodríguez-Clare, «Clusters and comparative advantage: Implications for industrial policy Author links open overlay panel,» 2007.

[32] «Wikipedia,» [En ligne]. Available: https://en.wikipedia.org/wiki/Standard_Industrial_Classification. [Accès le 2021].

[33] M. Du, «Natural language processing system for business intelligence,» 2017.