3. Building a bottom-up industrial taxonomy

In this section we summarise the results of an initial exploration of an approach to build a bottom-up industrial taxonomy using business descriptions from websites. It involves the following steps:

1. Extracting keywords and keyphrases (KW/KP) from business descriptions
2. Creating KW/KP networks based on their co-occurrence in business descriptions
3. Decomposing these networks into communities of frequently co-occurring KW/KPs that may refer to industries

We have tested this approach with samples of companies from two 4-digit SIC codes. They are:

- 6201: Computer Programming Activities
- 7490: Other Professional, Scientific And Technical Activities Not Elsewhere Classified

We have selected these two SIC codes because they illustrate two use cases where a bottom-up taxonomy may be particularly valuable: identifying sub-sectors that are developing or adopting emerging technologies and operating in new markets (in `6201`) and decomposing uninformative "Not elsewhere classified" sectors into more meaningful subsectors.

We outline the methodology in subsection 1 and emerging finding in subsection 2. Subsection 3 discusses limitations and next steps.

1. Methodology

We (pragmatically) define a sector as a set of companies that produce similar goods and services. One way to identify such sectors in a 'bottom-up', empirical way, is to look for co-occurrences of terms referring to goods and services in the same company description. Doing this requires building a vocabulary of relevant terms.

We have explored two avenues for this:

a. Top-down approach: United Nations Standard Product and Service Classification (UNSPSC)

The UNSPSC is a regularly updated taxonomy of products and services maintained by the UN. It contains 4 levels of aggregation (*segment*, *family*, *class* and *commodity*), with just under 148,000 commodities at the highest level of resolution.

This taxonomy could offer a principled and interpretable approach to identify mentions to products and services in business descriptions that we can then use to build our co-occurrence network. Unfortunately, we find that the terms used in UNSPSC have low overlap with the vocabulary in business descriptions. The main reason for this is that UNSPSC terms are often long, detailed and specific (see table 1 for some random examples).

Segment	Commodity
Live Plant and Animal Material and Accessories and Supplies	Dried cut white snapdragon
Food Beverage and Tobacco Products	Canned or jarred kapia peppers
Power Generation and Distribution Machinery and Accessories	Aluminum triplex service drop cable
Drugs and Pharmaceutical Products	Tizanidine
Healthcare Services	Drainage of right spermatic cord with drainage device, percutaneous approach
Healthcare Services	Bypass right kidney pelvis to ileocutaneous with autologous tissue substitute, open approach
Food Beverage and Tobacco Products	Canned or jarred summer blush nectarines
Healthcare Services	Dilation of right pulmonary vein with intraluminal device, open approach
Healthcare Services	Reposition right lower femur with internal fixation device, percutaneous endoscopic approach
Healthcare Services	Drainage of right ethmoid bone with drainage device, percutaneous endoscopic approach

Table 2: UNSPSC category examples

It is unlikely that firms selling these goods and services would mention them using the same language as UNSPSC in the descriptions that we have access to. The table illustrates that the taxonomy is highly unbalanced, with one of the families (*Healthcare Services*) comprising 73,000 commodities. We also find that the taxonomy does not include terms related to emerging technologies such as *Artificial Intelligence* or *Machine Learning*, an aspect of company activities that we are specially interested in capturing in the taxonomy.

b. Bottom-up approach: Automated keyword and keyphrase extraction methods

An alternative strategy is to exploit information about word positions in company descriptions in order to build sets of keywords and keyphrases (sets of frequently occurring words such as *web developer* or *building surveyor*) that we can then use to build our co-occurrence network. As part of this test, we have explored three algorithms to extract these keywords / keyphrases (KW/KPs).

- `RAKE` (Rapid Automatic Keyword Extractor) is an algorithm that identifies frequently occurring sets of terms in documents that are not uninformative stopwords or separated by punctuation (Rose et al, 2010).
- `YAKE` (Yet Another Keyword Extractor) exploits further patterns in documents such as the position of a candidate KW/KP in a document (e.g. upweighting those that appear at the beginning) or the range of sentences where a candidate KW/KP occurs (Campos et al 2018).
- `KeyBERT` uses a pre-trained language model to identify KW/KPs in the document that are closest to it in vector space (more semantically similar / better able to summarie the document) (Sharma and Li, 2019).

We note that all these algorithms operate at the document rather than corpus level. One potential avenue for further exploration is to adopt an ngram extraction strategy that identifies combinations of words that happen across the corpus at a higher rate than would be expected if they were statistically independent.

We use open source implementations of these three algorithms in order to test their suitability. Next section outlines our pipeline for this.

3. Results

Figure 1 present the steps in our pipeline.

a. Creation of industry vocabulary

We begin with company descriptions extracted from business websites for our two selected sectors (this comprises 5,499 `6201` companies and 10,000 randomly sampled `7490` companies). We apply our three KW/KP algorithms to these descriptions and extract their frequencies in each of the corpora focusing on KW/KPs that occur at least twice in the corpus (terms that are present in a single company description are irrelevant for producing a co-occurrence network). We also flag those KW/KP that are present in a concatenation of the UNSPSC taxonomy (and which we would expect to be specially related to products and services).

This yields a list of KW/KPs for each 4-digit SIC and extraction methods with descriptive statistics presented in table 2.

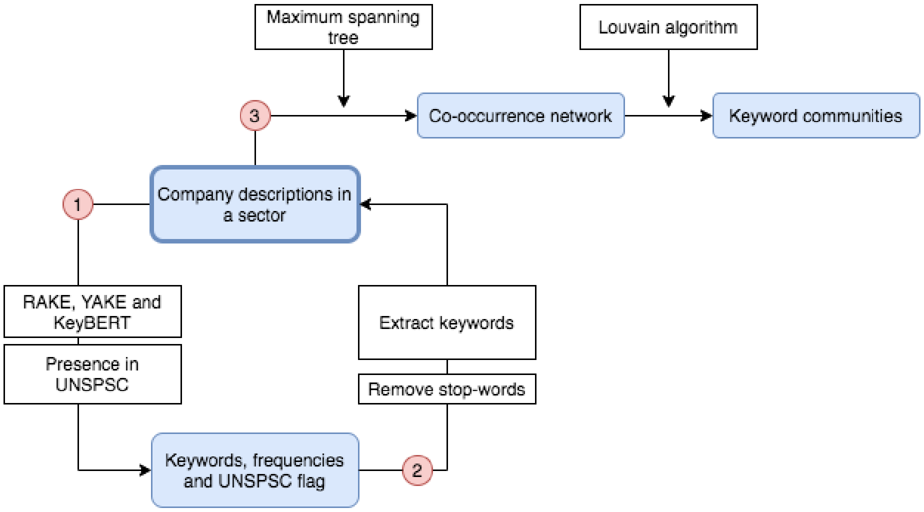

Figure 1: Figure: Data processing and analyis pipeline

sic4	method	count	mean	std	min	25%	50%	75%	max
6201	keybert	423	4.89835	4.12846	3	3	4	5	50
6201	rake	2890	4.02595	6.64891	2	2	2	4	134
6201	yake	3655	4.45253	8.81292	2	2	2	4	289
7490	keybert	816	5.06005	5.14883	3	3	3	5	75
7490	rake	5506	4.2223	7.80572	2	2	2	4	289
7490	yake	6991	4.48705	8.33889	2	2	2	4	221

Table 2: Descriptive statistics for keywords

Some observations: * 7490 yields more KW/KPs than 6201. This may be partly explained by the fact we use a larger corpus of descriptions, and by the greater degree of heterogeneity in the company descriptions of that catch-all category. * KeyBERT yields fewer KW/KPs than alternative approaches. * The frequency distribution is highly skewed. Most keywords appear in very few documents (the top frequency quartile ranges between 2 and 4 words depending on the method). * In terms of performance, RAKE and YAKE are significantly faster than KeyBERT. RAKE takes 20 seconds to run over 10,000 7490 company descriptions, YAKE takes 13 minutes and KeyBERT four hours (see figure 2, noting the logarithmic scale in the horizontal axis).

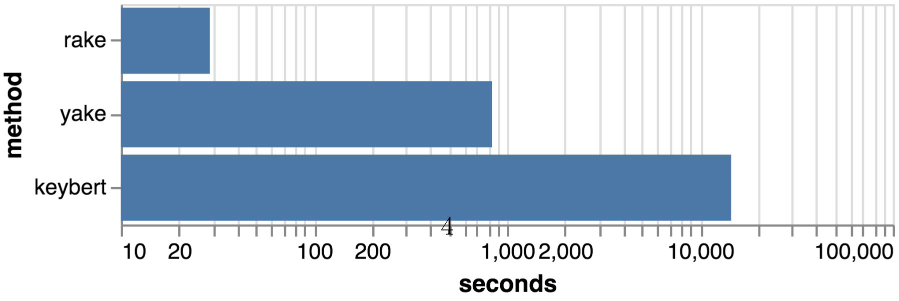

Figure 2: Figure 2: Extraction algorithm performance

b. Filtering of terms

```
effective   efficient   proven   track record   best   solution
offer   company   client   partner   solutions   leader   offer
dedicated   prices   business   businesses   success
```

We also remove a long tail of low frequency KW/KPs with less than 5 occurrences in the corpus.

In Figure 3 we show the most frequent KW/KP in our corpus by SIC code and keyword extraction method. It shows that our extraction method seems to be capturing terms related to various industrial activities (*software*, *software development*, *training*, *health and safety*, *design*, *building surveying*, *management consultancy* and *social media* among others). We also note some leftover short and uninformative terms and promotional terms such as *wide range* (we have decided not to implement a rule to remove all short terms for now because this would also take out short acronyms often used in the ICT industry such as CRM - Customer Relationship Management - or ERP - Enterprise Resource Planning - that we have noticed in the data).

It also appear that, in general, KeyBERT is better at capturing longer and more detailed KPs describing what a business does.

In Figure 4 we present the same information but only including terms that are not featured in UNSPSC according to our naive filter. It shows that while implementing the filter would help us to remove some irrelevant terms such as the location of a business in the UK, they also would lead us to lose many KW/KPs that are informative about what a business does and therefore relevant for the construction of our taxonomy. For this reason, we have decided not to apply the UNSPSC filter and keep all KW/KPs.

c. Network building and community extraction

We use our prototype industry vocabulary to build, for each of the SIC-4 codes we are exploring, a KW/KP co-occurrence network where every node is a KW/KP in our filtered set, and the weight of the edges represents the number of co-occurrences between connected nodes in the same industry description.

Our plan is to decompose this network into tightly connected communities of KW/KPs that might represent sub-sectors within our SIC codes. In order to simplify this process, aand following the strategy used by Hausmann and Hidalgo (2009), we build a maximum spanning tree of this network that only includes the highest weight edges that keep the network fully connected. We then apply the louvain community detection algorithm to the resulting network (this algorithm looks for partitions of a network that optimise its modularity i.e. extent to which the network can be decomposed into sub-groups including nodes that are tightly connected with each other and weakly connected with those in other subgroups).

Table 1 presents the resulting communities and related keywords for the SIC code `6201` and Table 2 focuses on `7490`.

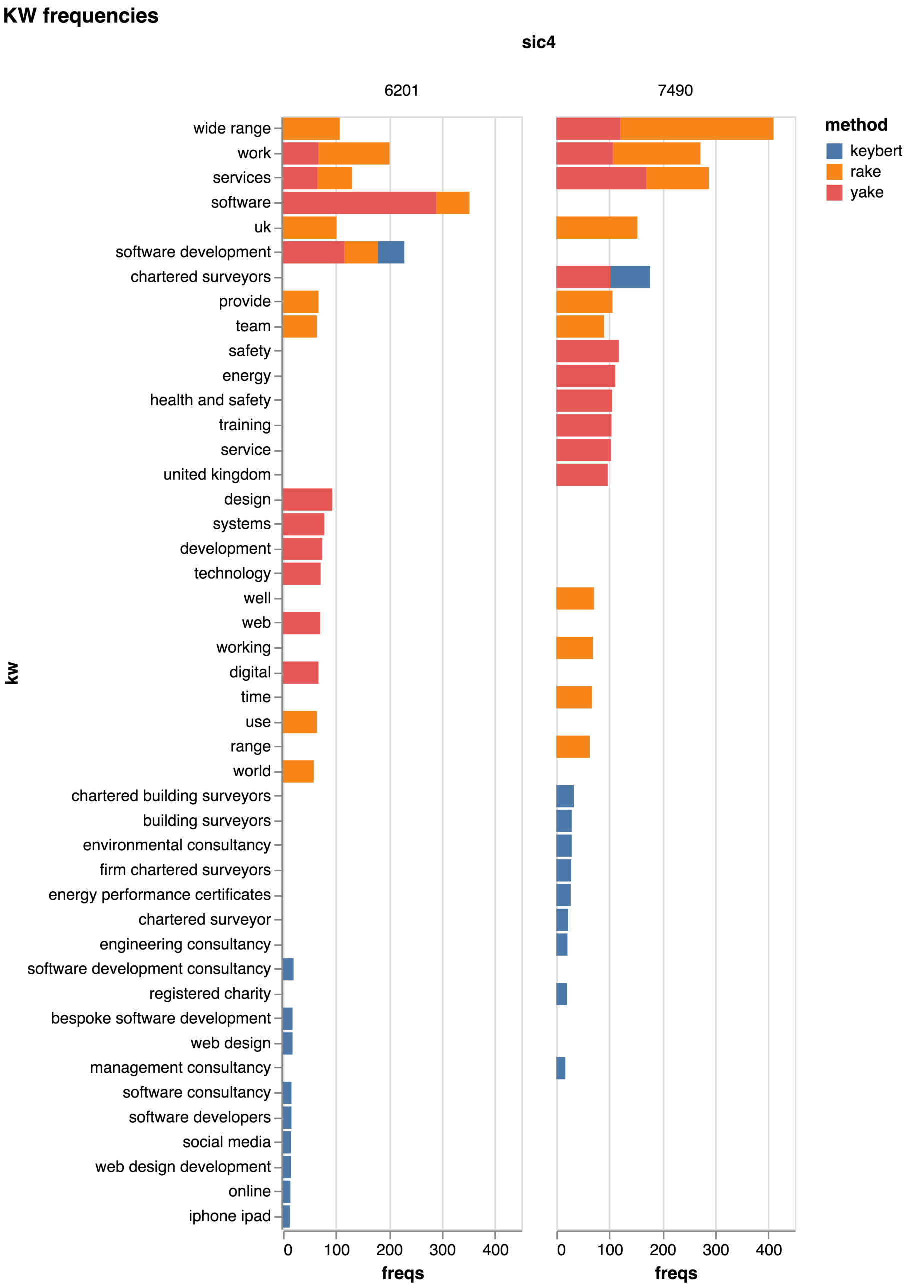

Figure 3: Figure 3: KW/KP frequencies

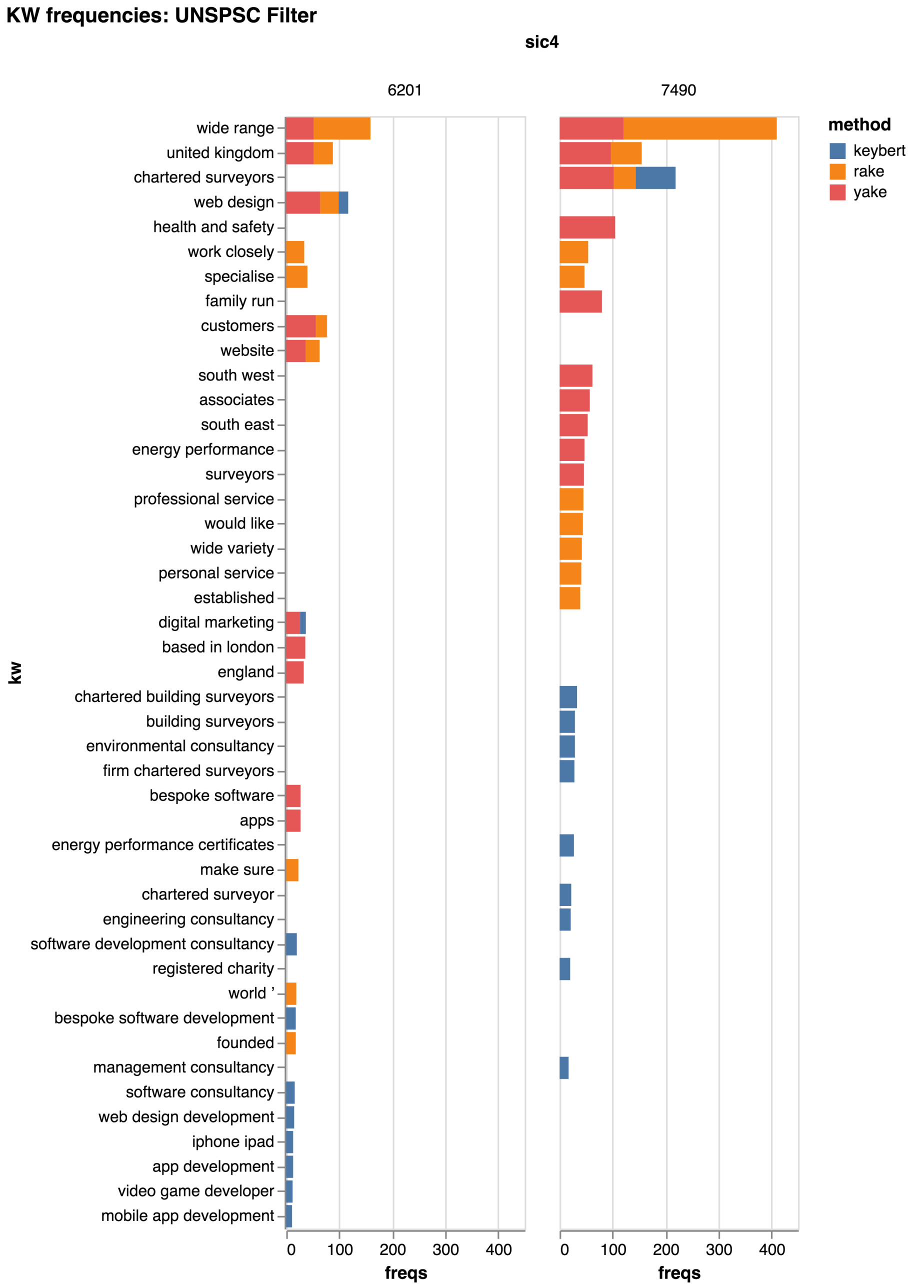

Figure 4: Figure 4: KW/KP frequencies (not featured in UNSPSC)

These preliminary results suggest that the approach that we have piloted here has potential, but also some of the challenges that we need to tackle in order to develop a robust and intepretable bottom-up industrial taxonomy.

1. We note that the community detection extracts 18 communities in `6201` and 24 communities in `7490`. This suggests that there is significant potential to increase the sectoral resolution of 4-digit SIC codes.
2. When we look at the content of different communities we detect some evidence of sectoral coherence. Some examples of this include:
    - `6201`: Communities 2 (Artificial Intelligence, Machine Learning and Data), 3 (app development), 4 (fintech), 7 (content management systems), 8 (digital marketing), 10 (enterprise software), 12 (web development), 13 (video games and virtual reality)
    - `7490`: Communities 0 (oil and energy), 2 (solicitors), 7 (engineering and surveying), 8 (knowledge intensive consultancy including pharmaceutical and environmental), 15 (flood and safety assessments), 20 (chartered surveying, 24 (energy assessment)
3. We detect several irrelevant communities capturing uninformative terms such as locations, promotional and marketing terms and meaningless business language. Some examples include:
    - `6201`: Communities 9, 14, 16
    - `7490`: Communities 3,4,6,10,17,18,22

Index	keywords
0	2016 profession built environment clear collect consultancy services bring full range technical support consultancy firm idea today development house end software delivery robotic process automation experience working wide range software consultancy found allows us hard work software development hou…
1	access founder host limit machine learning managing director right time test control data science transform network usa computer systems director latest software sap version extensive experience artificial intelligence code ireland stock control systems ltd. logic power data systems practice managem…
2	every project always looking rest expect full service within budget software products friend would love strive big june target audience got competitive advantage money work together passion may plan turn ongoing support email consult small team surrounding areas one stop shop good tech budget give u…
3	android app app development app developers app store app stores apple app store store mac ios and android mobile apps retail user experience design sign web and mobile ipad ipod touch small software innovative products mobile app mobile app development mobile devices mobile applications user experie…

Index	keywords
4	privately owned software specialist specialist knowledge financial services staff april two decades suit team members united kingdom developing software proud wealth product development experience developing software enables us financial services industry group board field united states effect custo…
5	asset health and safety citi leading provider life cycle take advantage care real vision real world real time family owned take care providing software health asset management asset management software software engineering life work hard make life easier commit mission reduce costs healthcare sector…
6	digital signage form start back strongly believe development team talk perform thing journey cutting edge technology cutting edge top tool page point computer science red research and development think see fast year would like tailor individual needs paul whole range touch made job daily basis sale …
7	based in glasgow blog web services multi edinburgh publish content content management system content management systems web site establish web site design management system management systems new website site web sites content management data management…
8	latest technology every stage based in london brand media media services social media digital marketing enabled us seo main focus headquartered in london london unique blend digital agency london based digital marketing agency offices in london online presence marketing agency team of designers info…
9	sole traders specialise in providing excellent customer bespoke software take pride outstanding customer project managers pride limited was formed consulting firm technical consultancy take great pride technical expertise customer satisfaction customer service software limited work closely software …
10	sql server consultancy based cloud cloud based microsoft certified cloud computing excel software house based software cloud based software crm line latest technologies microsoft technologies software development experience erp consulting services microsoft microsoft windows dynamics nav erp softwar…
11	sme sell pleased to announce week always happy open source software service provider free ruby on rails full stack price order pay happi open open source day speak…
12	internet marketing design and development milton keynes web developer web developers web development website design web design development services graphic design internet web designers web technologies wordpress search engine optimisation web presence agency based ibm creative web design agency des…

Index	keywords
13	games industry newcastle upon tyne fun game developer based studio based augmented reality game development studio virtual video game developer game developer design studio interactive entertainment video games written bristol game development games studio video games industry pc play based in brist...
14	tree state local long long term technologies limited first class service run event wide variety worked together high standard local authorities facilities management august home art young people computer repair deliver software family run local government computer services friendly service long esta...
15	independent software vendor independent software financial sector software developer software developers software developers based small independent...
16	england south west sussex north west east of england north south north america north east west birmingham east north wales middle east north of england kent west london south wales west yorkshire unique needs west midlands south east based in north west sussex...
17	human resources private sector public sector public sector organisations net sector public and private...
18	bespoke web solve problems custom software development application development custom software web applications web application development web application web based bespoke web applications custom...

Table 3: Output KW/KP communities for SIC code 6201

Index	keywords
0	aberdeen equipment manufacturers passion core development projects united arab emirates portable appliance testing event fully qualified gas industry group key conservation volunteers ltd. member network oil oil and gas locksmith services radiation protection providing support rail reach share speci...
1	customer care age party wall etc young people social social enterprise social enterprises continuous professional development social media pet message across dog friendly service take pride domestic and commercial home counties young traditional values agency specialising taken care interior interio...
2	also provide wide selection personal injury solicitor door specialist knowledge sustainability consultancy employment law support services law legal estate planning legal advice profession legal profession provide training expert witness expert witness services many different industries lasting powe...

Index	keywords
3	based in west west yorkshire practice specialising service provided north east england edinburgh north america ireland england based in edinburgh england and wales sussex north london county durham based in cheshire birmingham based in manchester established in january norfolk consultancy practice b...
4	highly skilled team europ planning process build import work across development services independent advice relationship industry experience satisfied customers employ recruitment consultancy short private individuals long long established aspect sustain systems limited complete service trust us new...
5	career planning consultancy energy sector planning services private and public private sector private sectors local authorities property developers house team provide professional public and private public sector bodies public sector organisations public sectors senior management public sector reduc...
6	life sciences person us today write matter david committed to providing detail project inception range of sectors new start let us know specific requirements step effective way internet advice and guidance wider community next day service work closely problem know avail also work closely every stage...
7	engineering consultants project management services comprehensive range project manager quantity surveyor contract professional quantity surveying provide a range construction consultancy engineering design tailored service surveying consultancy cost consultancy quantity surveying service structural...
8	idea cutting edge data consultancy based pharma consulting pharmaceutical companies test unit pharmaceutical industry planning applications ecological consultancy european policy makers intellectual property present united states ecological surveys research institutes question hub aim to provide eco...
9	contact personalised service rang due please contact us professional indemnity insurance please note professional ecologists set suit touch would expect respond quickly would sales and marketing game individual requirements like make every effort skill plan professional team abl start would love ass...
10	unique blend impact impact assessment york small team number one priority environmental performance contractor sound strong reputation special occasion innovative approach natural environment proud to announce account action enable us top priority fully trained professional approach knowledge and ex...

Index	keywords
11	impartial advice give us customer satisfaction mobile phone phone phone call away specialist skills design service design services please feel free security services price fully insured john lead st albans mind peace of mind excellent service landscape design give every one public liability insuranc…
12	lake district please see seen working closely delighted to announce vast range plumb star southampton forward local community april look look forward launch back take advantage services page get back flexible approach perform pleased to announce see page…
13	paul small family small family run family run stop shop husband and wife safety training courses experience within third parties family owned run training provider training courses owned and run electron consultants was formed training consultancy one stop shop shop…
14	pest pest control renewable energy consultancy supply chain management power farm print product range limited was founded worked hard professional manner exhibit iso reduce costs supplier system add value safety management systems control custom iso 9001 size primary objective solar renewable energy…
15	flood risk assessments fully committed insurance approved practical approach appartment blocks carryout fire carryout fire risk assessments fire fire protection financial conduct authority expertise evaluating many kinds private estates approved standards insurance approved standards shopping areas …
16	high wycombe truly believe head office london property maintenance middle east milton keynes south south east london specialise in providing residential properties whole richard work ethic bespoke service sale friendly approach platform private limited south wales south west south coast south east s…
17	world class privately held first class real student right place live beyond positive difference excel past midland first class service real estate class usa new zealand class service grow heart world…
18	great range customer service perfect deal dna soon great customer great customer service place last two decades save money job work place seo decades experience peak private landlords range of companies outstanding customer service take place began trading multi excellent customer britain tech award…
19	friendly and professional personal and professional personal service qualified chartered surveyor professional advice professional expertise professional property services professional service professional services chartered accountant property industry property management property owners provide a …

Index	keywords
20	party wall matters wall surveyors quantity surveying services surveying services building surveying practice chartered building surveying independent building surveying chartered surveying building surveying building surveying services building surveyors based professional building surveying propert…
21	extensive knowledge wealth management wealth of experience knowledge and experience charter experience and expertise wealth experience and knowledge…
22	independent consultancy providing independent consultancy independent consultancy specialising consultancy providing consultancy specialising…
23	consulting limited economic development diverse range environmental consultancy services specialist consultancy industry sectors design consultancy leading provider leading providers provide consultancy provide consultancy services provide support engineering consultancy occupational hygiene trainin…
24	energy assessment energy assessors energy performance certificate performance certificate performance certificates commercial energy commercial energy performance energy performance certificates epc estate agents energy performance independent estate…

Table 4: Output KW/KP communities for SIC code `7490`

3. Next steps

a. Improve KW/KP extraction

Our first step is to improve KW/KP extraction. In particular, we need to remove terms that do not refer to a business' productive activities. These include promotional and marketing expression, business bona fides and locations. Some potential avenues to do this include:

- Expanding our stopword vocabulary with a larger set of terms perhaps including names of places in the metadata of ONS geographical datasets or named entity recognition for specific locations. This expansion could be done manually or using keyword expansion methods such as `word2vec` that identify terms that are semantically similar to those in a seed list.
- Using parts-of-speech tagging to identify KW/KPs that are likely to be making promotional statements. As an example, adjective phrases such as *professional manner* and adverbial phrases such as *friendly service* are likely to fall in that category.
- Removing short KWs/KPs

- Using a document-wide ngram extraction strategy to expand our list of KWs and KPs.
- Removing high degree nodes from the co-occurrence network once this has been created

b. Scale up KW/KP extraction

Our improved KW/KP extraction protocol will need to be scaled over a corpus of more than half a million business descriptions matched with Companies House. This will require parallelisation, particularly to speed up the deployment of KeyBERT, which yields longer and more informative KW/KPs but taking orders of magnitude more time than alternative approaches such as RAKE and YAKE.

We also need to scale up our strategy for network creation and community building. In particular, we need to decide whether to build a network of *all* company descriptions and decompose it into communities (perhaps using a hierarchical community detection strategy), or take some elements of the SIC structure as given and look for subsectors inside it (along the lines of what we have done here). The first approach is more in line with the notion of building a wholly bottom up industrial taxonomy with a completeyly different structure from SIC. The second approach is computationally cheaper because it applies community detection to smaller networks eg defined around KW/KPs extracted at the 4-digit SIC.

c. Develop a protocol to assign companies to sectors

Having generated a set of communities with KW/KPs related to sectors, we will need to find a way to label these concisely. One option we are piloting is to query the KW/KPs in a sector community against wikipedia in order to obtain the closest article.

We also need to develop a strategy to classify companies into these bottom-up communities. One potential strategy to do this is to generate a labelled dataset of companies where a high share of KW/KP belong to a single community and train a supervised machine learning model using KW/KPs as features to predict community labels along the lines of what we have done in the first part of the project. This model can then be applied to the whole corpus. This approach would generate, for each company, a vector of probabilities for *all* bottom-up sectors in our labelled dataset, which we can use to identify companies operating in multiple sectors, and to meaasure similarities and differences between sectors independently from their position in the SIC taxonomy.