

Trends Tool, a tool for abstract textual data analysis used for additive manufacturing trend

Bachelors: Industrial engineering and management, graduation paper.

University of Twente, Enschede, The Netherlands

Author: C.C. Derks

Supervisors: PhD N. Knofius and DR A.B.J.M. Wijnhoven.

Abstract

This paper contributes towards addressing the gap between supply and demand in data science. The proposed solution to high demand and low supply is to make a tool that: separates technical proficiency from domain knowledge in the required skillset of the analyst, enlarging the amount of people that qualify for the job, increasing supply. The tool is also meant to streamline the data science process as a whole, resulting in a faster process, decreasing demand. The tool is designed to address high-level abstraction analysis goals of textual data such as tweets, emails, articles and publications. The tool was designed to analyze trends but can be used for other purposes. The end from using the tool can be described as a very extensive segmentation. A better understanding of the researched concept can be gained. The tool is designed using an example scenario in analyzing Additive Manufacturing as a whole. In conclusion, the tool has some modularly defined functionalities that are understandable and easy to use for someone with domain knowledge. Making it possible for a person without technical proficiency to conduct complex analysis.

Keywords

Google Alerts, Additive Manufacturing, KDD, Data mining, Web scraping, Text mining, Trends Tool, Topic Modelling, Stanford NLP, Gensim, Python.

Introduction

Knowledge discovery from databases (KDD or data/text) mining) is in high demand. As this article about data science jobs in Europe suggests (Dataconomy D. Dutta, 2019) there is a lack of supply in data scientists. Also, demand is suggested to increase (SearchBusinessAnalytics B. Holak, 2019).

KDD consists of many tedious and unspecialized tasks. To analyze data, the analyst often needs to first go through some kind of data retrieval, cleaning and preprocessing. Then for each research goal, different modelling algorithms are considered and handpicked. Running algorithms often requires picking the right parameters for which several trial runs are required which can take up unexpected amounts of time. After discovering something doesn't work as desired or new knowledge comes to light, the process might start over again. It is difficult to optimize or even plan this process. Streamlining this process will lower the demand for data scientists because data scientists can operate more efficiently. (YANG and WU 2006).

Besides that, a data scientist needs to have enough understanding of two very different proficiencies. Domain knowledge is needed to make decisions (what is interesting information) and a technical proficiency for algorithms, statistics is needed (how to get this information). This makes for a required skill set composed of both technical and domain knowledge which is harder to come by than those proficiencies separate in one person's skillset. This makes it hard to come by potential data scientists which hurts supply in data analysis. Reducing the required technical proficiency makes it easier to find employees to do KDD and might also result in more insightful analysis, since it's also easier to find more specialized people in a certain domain. (DeBortoli, Müller, and vom Brocke 2014), (Jiang, Klein, and Means 1999). This paper tries to contribute towards a solution.

Problem

The perceived problem is that businesses cannot satisfy KDD needs because of high demand and low supply in data scientists.

Goal

The goal of this paper is to both lower demand and increase supply by designing a tool that enables employees with low technical proficiency to do data analysis and streamline the process as a whole. This increases supply because the analyst does not need to have a lot of technical proficiency; and decreases demand as the process goes faster -- no intermitted programming is needed, just configuring and processing.

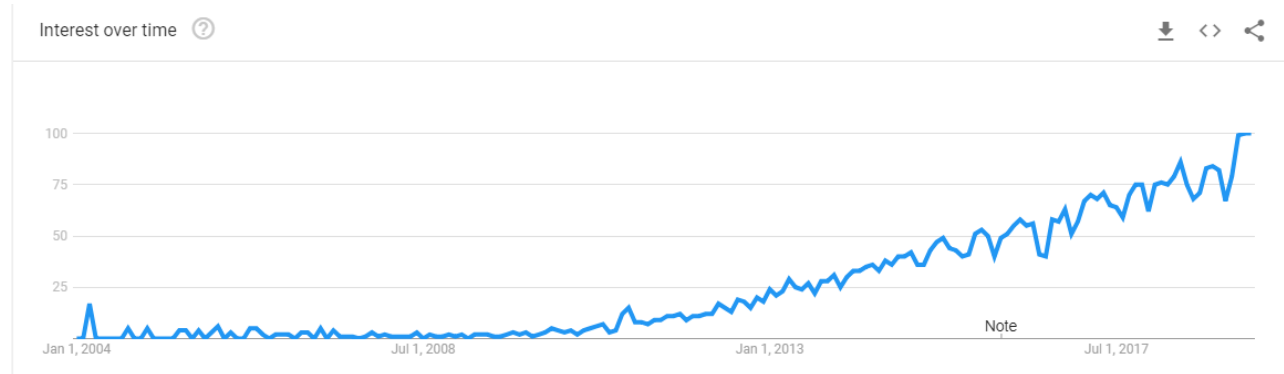
The entire scope of data analysis is too big to make a tool in one paper. We will make a tool to analyze textual data, with a high level abstraction analysis goal (non-specific). This way the tool will cater many research goals, since there is a lot of textual data and doing non-specific data analysis can at the very least be a good start of the KDD process.

To design the tool we will go through the KDD process using Additive Manufacturing (AM) as example. However, the tool generalizes to other topics and research goals without the need to program the steps again.

In section 1, we shortly motivate our choice to focus on AM and establish the related research goal. Then in section 2, we go into the details of the data source and discuss our options on how to approach the KDD process. In section 3, we formulate an approach to reach our research goal and apply it once to AM as an example. In section 4 we will discuss what the tool can do in its current implementation and what potential it has with more implementation. Section 5 concludes this paper and can be read as a management summary.

1. Additive manufacturing

Additive Manufacturing (AM) as a manufacturing technique has been around for forty years (Bradshaw, Bowyer, and Haufe 2010). And, AM has been growing in google trends since around October 2009 (“Additive Manufacturing - Explore - Google Trends” n.d.). Below the graph.



Graph 1: Google Trends: “additive manufacturing”, global interest, since 2004

Over time, AM seems to only become more interesting as it offers many opportunities. For instance, better unmanned aerial vehicles or lighter planes (for less fuel consumption) (Ferro et al. 2016). These opportunities come forth from AM enabling more complex, lightweight designs and also infrastructural advantages like reducing transport distances and stock on hand (Rüßmann et al. 2015). Right now, the AM industry is valued at 7.3 billion dollars (T. T. Wohlers et al., 2018). Additive Manufacturing went through many breakthroughs getting where it is now (T. Wohlers and Gornet 2016), there are many different techniques and the manufacturing industry has slowly but surely been using AM in more and more ways. Wohlers yearly report describes this into much detail each year. In the literature, this development has been analyzed based on surveys, sales figures, literature reviews and expert opinions, e.g. (See Wholers report series) and (Thompson et al. 2016). Key results from such research are insightful pieces of knowledge that go into detail. How industry changes and why. However, what this kind of research might overlook are things that are happening on a larger scale. For that purpose, we suggest applying our KDD tool that allows potentially to uncover large-scale trends. Furthermore, data mining techniques have not been used yet to take a look at AM as a whole. Accordingly, we define our research goal for AM as:

Take a look at additive manufacturing as a whole using data mining techniques

2. Background of the tool

As we figure out how to best analyze AM, we will describe our tool. This way the tool is demonstrated based on a real scenario of KDD and its functionality becomes clear.

In this section we will assess our options: the data set and KDD-techniques are discussed.

Data mining approach should be based on two main factors. The main goal of the problem to be solved and the data set (Gibert et al. 2010). The problem is the narrow perspectives used to study AM and the goal is to identify quantitative trends using data mining.

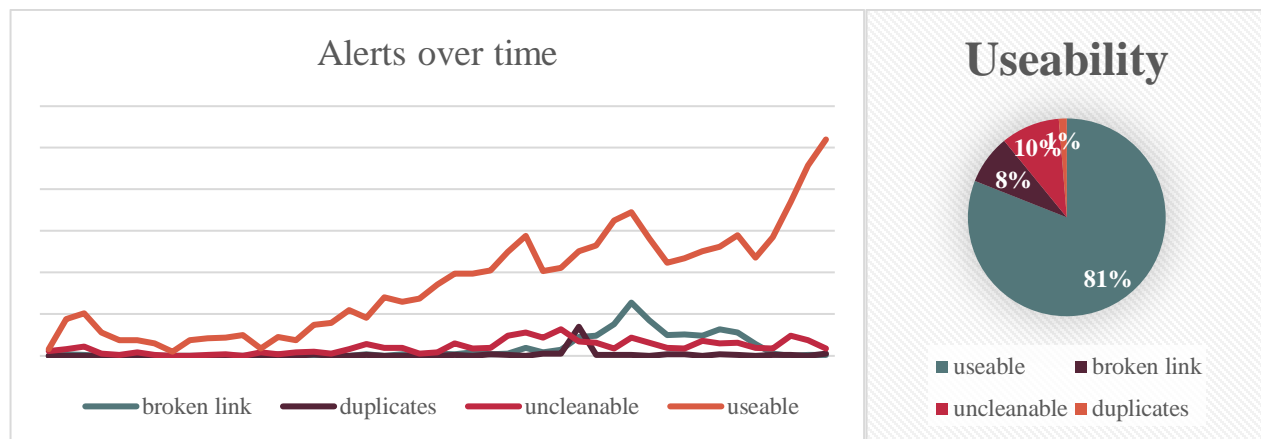
2.1 The data set

In this study we use several .mbox files with emails from Google Alerts. Google Alerts is a service offered by Google where you register your email address and one or more search strings (“Google Alerts - Meldingen van Interessante Nieuwe Content Op Internet” n.d.). Google will continuously send new search results for the searched string in emails (let’s call them alerts). After collecting years’ worth of these alerts, it can be considered to apply data mining with them. These alerts come with one or more results and each result has, among other things, a link to the actual result somewhere in the internet (like a regular google search result). Besides the information inside the alert, we also use the webpage of this link. Before the webpage is useable as a source of information, the irrelevant information needs to be removed (this activity is commonly called web scraping). This results in pieces of data as shown in Figure 2 in which each part of the informational structure is highlighted in **bold**.

Alert#: 3 **Link#:** 4 **Date received:** Fri, 28 Aug 2015 07:02:08 +0000
Title: 3D Printing and Additive Manufacturing Market Size, Shares, Analysis, T...
Source: Business Wire **Additional Tag(s):** (press release)
Example Text: ALBANY, N.Y.--(BUSINESS WIRE)--ResearchMoz has added a report titled \xe2\x80\x9cGlobal and China 3D Printing and Additive Manufacturing Market ...
Link: <http://www.businesswire.com/news/home/20150827005468/en/3D-Prin...>
Cleaned webpage:
ALBANY, N.Y.--(BUSINESS WIRE)--ResearchMoz has added a report titled “Global and China 3D Printing and Additive Manufacturing Market 2015-2025: Industry Analysis, Shares, Size, Trends, Growth, Technologies, Key Players and Forecast” to its research report database. 3D printing has received much attention in the press over recent years. Hyped as the technology to bring about a 3rd industrial revolution, 3D printing technologies were in fact invented in the early 80s. They remained a niche technology until the expiration of a key patent in 2009 allowed many startups to emerge offering cheap consumer-level 3D printers. A media frenzy in 2012 thrust 3D printing into the limelight and major players are reporting dramatic growth in everything

Figure 1: Data piece example

Some of the links don’t work, some of them are duplicates and some webpages can’t be cleaned, Figure 3 shows an overview of the received data pieces over time as well as the portion of data that is useable:



Graph 2: Alerts received over time and data usability (Aug-2015 until Apr-2019)

Webpages are cleaned by dividing the HTML into parts at devisor characters *TAB, *NEWLINE and “|”. For each part is determined whether it’s a good, neutral or problematic part. A good part has words matching with words from the description in the Alert email. A problematic part has a percentage of programming symbols that is too high. A neutral part has neither. Then the HTML is cleaned by picking out the good parts and all adjacent neutral parts ended with problematic parts.

Cleaning arbitrary HTML was a challenge. This process was formulated after a lot of trial and error. This was the third and best approach to cleaning arbitrary HTML pages. Trial runs with different programming symbols percentages were ran and many different parameters to classify parts as good neutral or bad was experimented with.

2.2 Google Alerts as a data source for trend analysis

The methods used by Google to provide the Google Alerts service are unknown. We don’t know how the webpages are identified and we also don’t know when an eligible webpage is accepted as a webpage to be sent as an Alert. Google might even change these methods and not let anyone know. Therefore, we need to treat the alerts as approximation for the underlying trend. As for relevancy of the data pieces, we can only rely on the reputation of Google. According to a Forbes article, the reputation of google alerts might not be so good (Hill 2013). However, we are not using google alerts to monitor the reputation of a name, we are using it to monitor an industry. We don’t need to see every mention of AM on the internet, we want to see all the relevant articles about AM. Furthermore, it has to be underlined that our tool can also be used in combination with other data sources. The only requirement is textual data.

2.3 Data mining techniques

The first stage(s) of data mining process is about understanding the thing to be researched (here: AM), preparing the data (web scraping) and pre-processing data (Natural Language Processing). In this case, understanding about AM is reached through reading a number of data pieces data first after cleaning it (in the end, also going through the data mining process this added up to more than one hundred pieces of data). This seems the most convenient way, because this gives both insight into the data as well as AM. The cleaning of the data pieces is further explained in the appendix.

After reading through some of the data, we get a general understanding of the commonly mentioned concepts. Commonly occurring phrases are related to the bigger concepts as summarized in Figure 3.

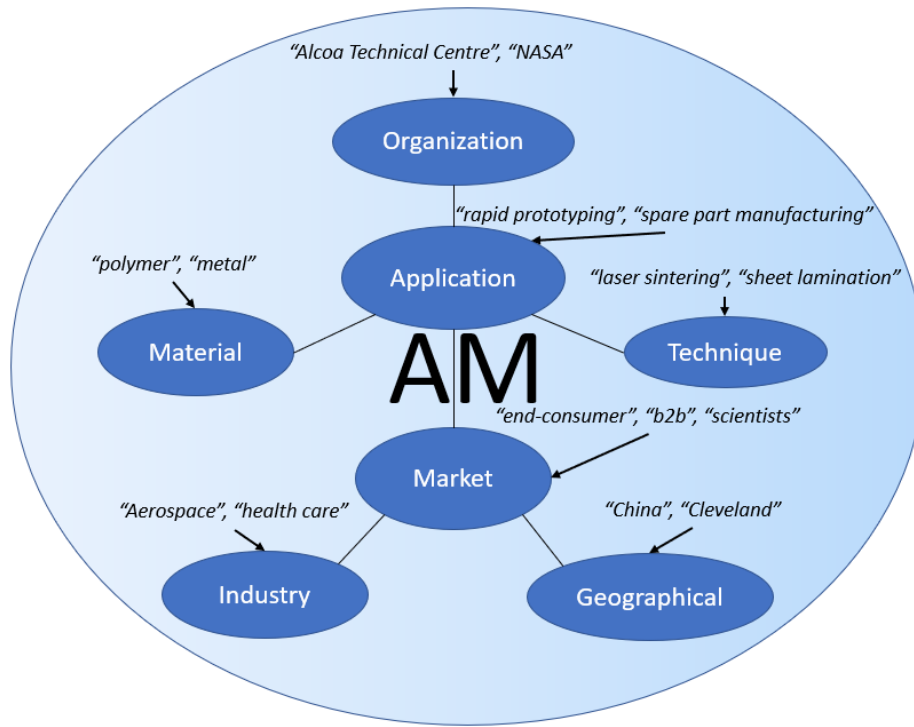


Figure 2: Commonly mentioned concepts

The data pieces commonly speak about AM in a sense of how it progresses in terms of AM being adopted in a new way or how a technique might make something possible to do better with AM. Sometimes an organization is mentioned that is using this new application of AM. Also often referred to are places and industries where AM is applied. Hence, in this figure, an application can be that an organization is using a certain kind of AM approach (material and technique) in a certain market to supposedly, more efficiently reach certain goals. These goals are not considered because it seems too complex to relate different goals to each key concept: Material, Organization, Technique, Market, Industry, and Geographical. Furthermore, these goals are researched a lot better in a non-KDD approach, with for instance expert opinions and surveys. For example, an article about Airbus adopting AM because AM allows parts to be lighter with the same performance. It is possible to extract information like Material: Aluminum, Organization: Airbus, Industry: Aerospace. But it is too complex to for each data piece extract information like: lighter parts. Because this dimension would be composed by a very complex and wide variety of arguments to adopt AM, other arguments can be related to: supply chain, material cost, lead time and more. When compared to dimensions chosen this dimension is too complex.

We need to find a way to interpret the data with the computer. First, we need to structure the data by segmenting it. This way we can discover trends on more specific parts of AM. To segment the data a couple of approaches were considered: a trained named entity recognition (NER) classifier, topic modelling using latent Dirichlet allocation modelling (LDA) and finally data piece tagging. We discuss each approach subsequently.

Named entity recognition (NER)

Named entity recognition is the concept of an algorithm pointing out the names of certain entities and knowing what kind of entity the name is for. For instance “Bill Gates” is a name of a person. “NASA” is a name of an organization and “China” is a name of a location. This concept can be implemented in many ways. The most prominent one is the

implementation of Stanford (Manning et al. 2014). It is possible to train your own classifier with this implementation, recognizing your own entities like “steel” as material (“The Stanford Natural Language Processing Group” 2006).

A trained NER classifier (an algorithm that marks words as a certain entity, for instance ‘iron’ as metal) was not used for initial segmentation because there was too much uncertainty in the results it would yield. However, this technique is used later with a default classifier, this part is written about a self-trained classifier. The manually trained classifier would doubtfully classify the data pieces correctly on its own. In order to find out its accuracy of this method a lot of time would have to be spend on training a classifier first. This would cost by approximation 10-30 minutes per data piece with at least 100 data pieces plus the time it takes to make all the programming work. Optimistically that equals approximately $(100 \times 20) / 60 + 8 = 41$ hours. Realistically speaking, during this process things tend to not work in an expected way and problems need to be solved, maybe some part of the tagging needs to be done again because of mistakes which adds a huge uncertain potential extra amount of time to that estimation (8 up to another 41 hours by estimate). Pessimistically speaking that uncertain potential amount of extra time can be considered double.

Topic modelling with latent Dirichlet allocation (LDA)

Topic modelling is the concept of extracting underlying topics in a body of data. This can also be done in many ways. LDA is accredited to work well for a (relatively) small data set, which is the case for us (“Gensim Tutorial - A Complete Beginners Guide – Machine Learning Plus” n.d.). And there is a specific implementation of LDA that has shown to work well in practice -- the Mallet implementation (“Gensim Topic Modeling - A Guide to Building Best LDA Models” n.d.). After feeding the data to the LDA algorithm you get a number of topics out of it. The analyst needs to declare beforehand with how many topics the data needs to be represented. The best amount of topics can be estimated using coherence scoring, the more coherent the topics are, the more reasonable that amount of topics is. Coherence scoring is only indicative of higher quality, there is no way to tell in an absolute way. A topic is represented by words with weights. The weight of the words determines how distinctive that word is to that topic. The meaning of the topic needs to be inferred from the words that are important to that topic. If “acquisition”, “Google”, “laser” and “sintering” are the words with the highest weight in a topic, the topic is very likely about an acquisition of google that has something to do with laser sintering. To verify what a topic is about it is good to look at data pieces that represent that topic with a big weight. Each data piece resembles each topic with a certain weight as well. Reading the data pieces that have a big weight in representing a topic probably describe the topic.

LDA is used later in the process. But, LDA was not used for segmentation of AM because of two reasons. First, after testing, the results did not seem of enough quality, a lot of fine tuning would be necessary for better quality (filtering words that are used by the algorithm, meaningless words were used too often to represent distinct groups). Initial results show that very general words (meaningless to a topic) are given big weights (words like “printing”, “additive”, “manufacturing” while all the data is about these topics). This can only be solved by filtering those words out. Then, there is the nature of data being spread over time. One topic can be visible over the entire time span. However, more realistically there are probably more local trends. Hence, it seems like a bad approach to try to identify all topics at once from the entire time line. Furthermore, there is the problem that earlier trends probably have less data pieces dedicated to them and later trends more data pieces, simply because of AM gathering more and more attention over time.

Secondly, the structure this would result in is unclear. It would not have clear dimensions. If we would use LDA as a segmentation method, it would result in a couple of chaotic topics. Not all of the topics make sense and some of them do. The ones that do can be wildly different. It is hard to compare these topics. There is no real structure within these topics besides that they probably have something in common.

Bigrams

Bigrams are two words that appear together commonly. Bigrams are two words that together mean something completely different than when they are mentioned separate. An example is for instance “long_term”. These are considered as standalone words in LDA. This is important later in the paper as it is not only used as part of LDA but also for another purpose.

Tagging

We would formulate one or more conditions for the groups and if a data piece qualifies, the data piece is tagged as part of the associated group. One data piece can be tagged with any number of groups. The groups are defined so that data pieces are distinguished with clear and understandable dimensions.

Tagging was chosen as the best method to make an initial segmentation of the data. This was chosen because it would result in very clear dimensions. The dimensions are derived from figure 3. It was estimated that it was very practically doable to tag data pieces as part of the metal or polymer material groups. Also tagging locations and organizations seems very plausible using the NER algorithm. In the end, the previously discussed methods were not used for initial segmenting but they are used in formulating the right conditions for data pieces to be tagged. So for a good condition to tag data pieces with the China location tag, the NER algorithm needs to have identified a location with name “China” in that data piece. The LDA topic modelling is also used but in a later stage, which is explained later.³

Analyzing complex trend: additive manufacturing

In this section we go into depth as to how AM was analyzed also discussing how the tool plays a role in the process.

The first step in this segmentation approach (tagging) is determining the tags and conditions per tag. The concepts from Figure 3 are used as a basis to start from. Then tags are defined as more specific instances of those concepts. For instance, big organizations get their own tag, material is divided in metal and polymer, market gets scientific, business to business and end-consumer tags. See below: Figure 4.

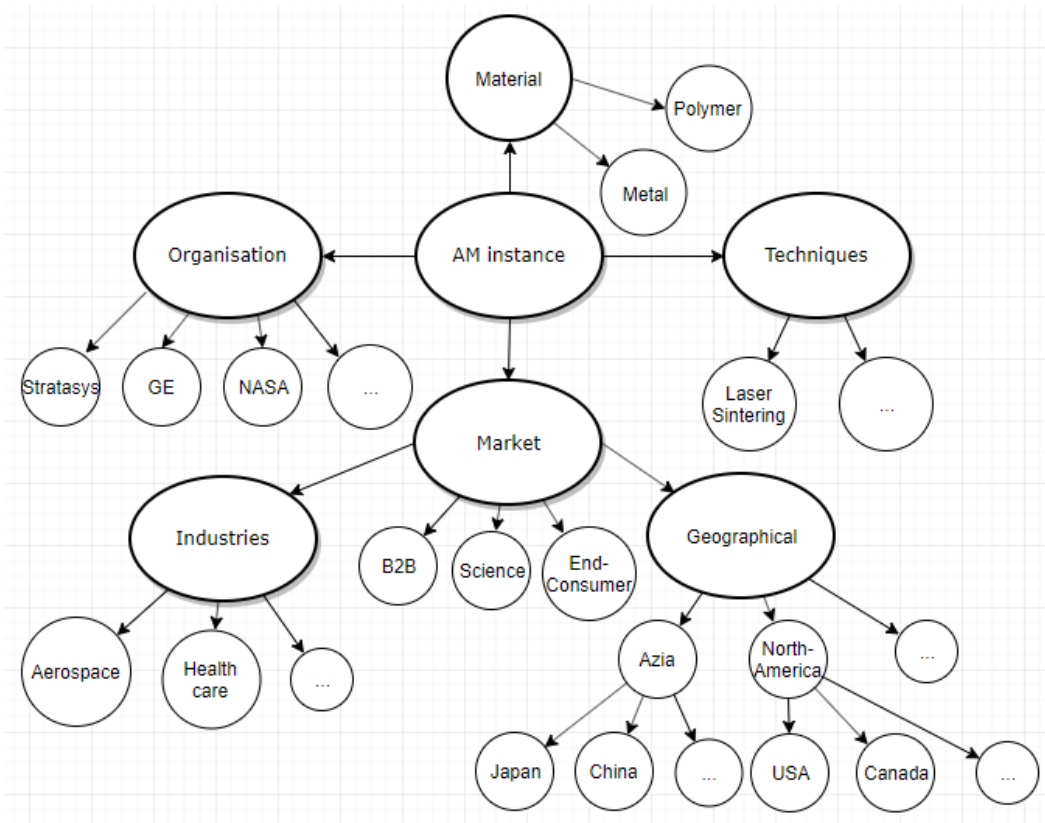


Figure 3: Concepts with their tags.

3.1 Segmentation

In Figure 4, “AM instance” is the center concept. Each data piece is an instance of AM, in other words, it contains some information about AM. The other concepts are circumstances of an AM instance. Each AM instance can partially be described by these concepts.

All the choices in constructing the concept network are based on reading through the data set by the analyst. In these choices, a balance between measurability and going into detail as much as possible is considered. It is possible to expand the concept network with concepts that would tag data pieces that address certain AM pros or cons, like slow production time or high customizability. However, the way in which the pros and cons of AM manifest into instances of AM seem too unpredictable to put in concepts to be tagged. The way authors write about progress as a result of AM has too many different words attached to it. Also the relations between the words matter too much for this model to take note of. It is important to know what ‘improved’ or ‘low cost’ is about in an article. Relations are not implemented in our tool.

3.1.1 Strategy

The strategy of tagging the data pieces is based on the nature of the data, we don’t know a lot of where it came from and we are certain we don’t have all the data. We can only take indications out of the data. That is why we don’t want to tag as much as possible, we want to tag accurately. A condition can only belong to one tag, because none of these

tags mean another tag should apply as well. After the conditions are defined for each tag, we let the computer tag the data pieces.

3.1.2 Assigning conditions to tags

The conditions for each tag were assigned using the tool. The tool essentially consists of four boxes as shown in Figure 5. The first box contains the defined concepts with their parent annotated (“USA: North America”, “Norht America: Geographical”) Secondly, there is a list of algorithms. Each “algorithm” is a way of looking at the data and finding conditions in the data, the backgrounds of these are explained in Section 2.3 Data mining techniques. Thirdly, there is a list of conditions (with the amount of data pieces that qualify the condition annotated). Note that the conditions change depending on the algorithm. And finally, there is a box that shows an example data piece that qualifies the chosen condition, so that the analyst can read the context in which the condition is met. Of course, there is also a box to put arguments in and several buttons to run functions. For example, in Figure 5, the ‘Bigrams’ algorithm is selected which results in a list of recognized bigrams in the data with the amount of data pieces that contain the selected bigram annotated, then “long_term: 284” is selected and an example data piece in which the bigram “long_term” was found is displayed.

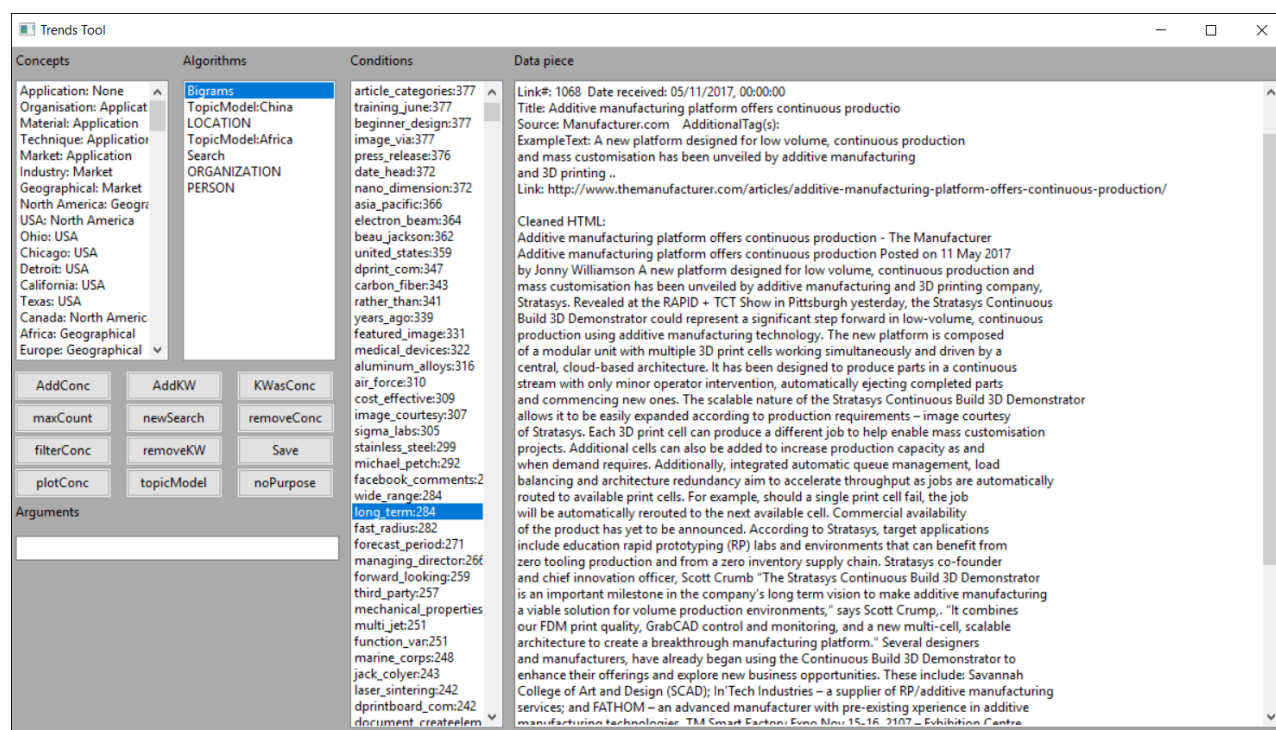


Figure 4: Trends Tool prototype

3.1.3 Algorithms

The algorithms are methods to go through the data and recognize potential conditions to tag data with. Three methods are used to construct potential conditions. First a Stanford NER default classifier to identify Organizations, Locations and Persons. Then bigrams that were formulated in the LDA process. Finally, the analyst can search through the data

themselves, the amount of data pieces the word appears in is counted and example pieces can be watched. NER and bigrams were discussed earlier in the paper under data mining techniques.

NER

The NER default classifier makes it possible to tag the data with geographical tags like Europe, Asia or Germany. Also often occurring organizations were tagged like NASA. Some organizational tags from the NER algorithm also helped with different concepts. NER would recognize “Automotive” as an organization while that was a major condition for segmenting the industries.

Bigrams

Bigrams was useful for many purposes, for instance “consumer_goods” was very helpful to tag this market. In figure 5 “laser_sintering”, “medical_device” and “aluminum_alloys” are visible. These bigrams help segment the laser sintering under techniques, healthcare under industries and metal under materials tags respectively.

Search

The search method gives the option to define your own keyword, to examine the data better as well as segment the data in general.

Topic Model

The topic model lines under algorithms is explained in the next section. Topic Modelling can be used as a supplement for the search method but this is explained later.

3.2 Trend discovery

All the tags visible in Figure four were constructed and tagged using the tool. This was done in a interactive manner. When a concept is not well defined enough it might have too little or too many data pieces that are tagged with the concept. When the analyst notices this, conditions can be altered.

Now that we made distinctions between pieces of data by tagging them, we can discover trends within desired parts of the data. In the literature, identifying trends is commonly done by topic modelling using the bag of words method and an algorithm (Glance, Hurst, and Tomokiyo n.d.), (Zhang and Li 2011). Topic modelling can be done in various ways, all of them seek to reveal hidden topics in data.

In our data set, new things about AM are described in each data piece, thus a group of data pieces that are describing a similar new thing can be interpreted as a trend. A big amount of similar changes is a trend. This can in our case be quantified with an amount of data pieces that talk about a trend plotted over time.

To identify these similar changes we will use the earlier explained LDA algorithm, then it was considered as a segmentation method, now we consider it as a trend discovery method. LDA is again explained from this perspective\

The algorithm reads the data you give it and makes several topics based on that. One topic is represented by several words with weights. Most of the time there is a clear relationship between those words, sometimes it's less clear and sometimes there is no relationship. Depending on the number of groups you tell the algorithm to make these topics represent actual topics. The coherence score is also computed, the closer to 1 this score is the more likely it is the best number of topics to choose.

The analyst can choose tags and a time frame to do the topic modelling with. If the analyst is only interested in metal additive manufacturing or thinks there might be an interesting trend within the data pieces that are tagged with metal, the analyst chooses to do topic modeling only using data tagged with the “Metal” tag.

The analyst can add or remove concept specific stop words, words to not use in the topic modelling. Words that are describing the concept to not be used in the topic modelling. If you are topic modelling within the confines of the location “USA” you don’t want to segment your topics also using the phrase “USA”.

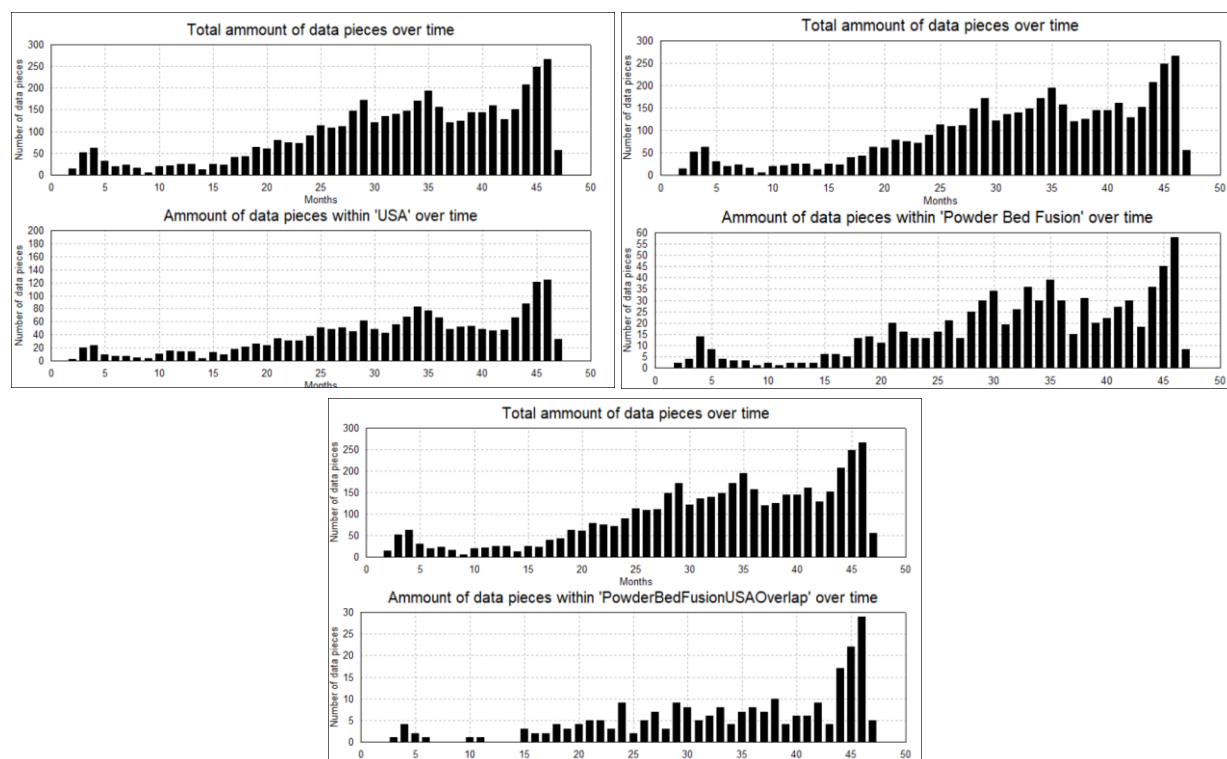
The results of the topic modelling are displayed in the tool for the analyst to look at if the analyst sees some seemingly interesting topics, the analyst can choose to look to data pieces that resemble that topic. This way the analyst can discover a trend. If many data pieces resemble a topic and this topic clearly is a trend, a trend is discovered.

The topic modelling can also be used to refine the concepts and key words for tagging. When a certain word is show to be very distinctive for a topic within a concept, it might be a good idea to make a sub concept within which trends can be discovered.

3.3 Trends in AM

To see what the tool can do in its current implementation we will take a look at what trends we can discover using the tool. We will use the tool once, as an example.

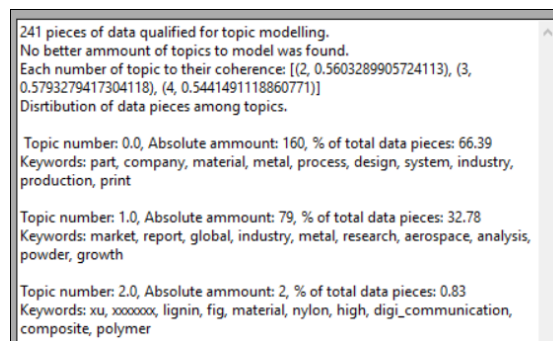
Say we want to know more about Powder Bed Fusion techniques used within the USA. We use the constructed concepts to see what data pieces are tagged both with USA and laser sintering. See below the data pieces over time for both tags and them together, take note of the vertical axis numbers that change:



Graph 3: USA and Powder Bed Fusion Data over time

The conditions of the USA tag are all coming from the NER default classifier to recognize locations: ['USA', 'States', 'US', 'U.S.', 'U.S.']. The conditions of Powder Bed Fusion came from the Search and Bigram algorithms: ['dmls', 'ebm', 'sls', 'slm', 'selective_laser', 'powder_metallurgy']. The first four came from reading an article about AM techniques and they are abbreviations of four sub techniques under Powder Bed Fusion. The last two conditions are bigrams of phrases related to Powder Bed Fusion. From figure 6 we can already see that 20% of the total in the last spike is accounted for with Powder Bed Fusion and that USA accounts for about half the total data pieces. It is not surprising that the overlap of the two is about half the amounts of Powder Bed Fusion alone, at least in that last spike.

From here we can use the topic modelling on the constructed concept which is the overlap of two previous concepts: USA and Powder Bed Fusion. We start with three as a topics parameter since we don't have many data pieces and there are probably not many different trends to identify in such a niche. See below the result text:



```

241 pieces of data qualified for topic modelling.
No better amount of topics to model was found.
Each number of topic to their coherence: [(2, 0.5603289905724113), (3,
0.5793279417304118), (4, 0.5441491118860771)]
Distrtubition of data pieces among topics.

Topic number: 0.0, Absolute ammount: 160, % of total data pieces: 66.39
Keywords: part, company, material, metal, process, design, system, industry,
production, print

Topic number: 1.0, Absolute ammount: 79, % of total data pieces: 32.78
Keywords: market, report, global, industry, metal, research, aerospace, analysis,
powder, growth

Topic number: 2.0, Absolute ammount: 2, % of total data pieces: 0.83
Keywords: xu, xxxxxx, lignin, fig, material, nylon, high, digi_communication,
composite, polymer

```

Figure 5: Topic model result text 1

In total 241 data pieces qualify for the topic modelling. The topic model with 3 topics scores the highest coherence score so this one is chosen as a result. From the topics we can see that the last one is sort of a fluke. The most important keywords are: “xu” and “xxxxxx” which is probably duo to bad data cleaning. Luckily only two data pieces fall under this topic (0.83%). So the Powder Bed Fusion technique in the geographical Market of USA is basically divided into two topics by the tool. Both of them are related to metal and industry. One is more about analysis, research, and aerospace and the other is more about industry production, process and production. From this point we can see that the software has divided the data pieces into a more production process business section and a more aerospace related scientific side which talks about global market growth. To confirm our own interpretation we are going to look into the most representative data pieces for both of these topics. The most representative data piece of the presumed business like section is an article about a new launch of NASA. The market growth section is exactly about that, market growth, in a stock market sense.

We are not satisfied with just these two general sections so we are going to topic model again but then with a substantially larger amount of topics. Let's pick 20. See below:

```

Topic number: 0.0, Absolute ammount: 13.0, % of total data pieces: 5.39
Keywords: gkn, research, company, system, aircraft, center, innovation, standard, national, corporation

Topic number: 1.0, Absolute ammount: 40.0, % of total data pieces: 16.6
Keywords: market, global, report, research, industry, analysis, growth, forecast, share, size

Topic number: 2.0, Absolute ammount: 35.0, % of total data pieces: 14.52
Keywords: market, metal, report, powder, analysis, global, industry, segment, type, chapter

Topic number: 3.0, Absolute ammount: 7.0, % of total data pieces: 2.9
Keywords: design, printer, company, digital, engineering, metal, model, system, rapid, center

Topic number: 4.0, Absolute ammount: 7.0, % of total data pieces: 2.9
Keywords: report, market, digi_communication, patent, occur, europe, company, transaction, regulate, lead

Topic number: 6.0, Absolute ammount: nan, % of total data pieces: nan
Keywords: beam, fabrication, system, structure, size, intensity, time, approach, energy, light

Topic number: 9.0, Absolute ammount: 2.0, % of total data pieces: 0.83
Keywords: machine, hybrid, metal, laser, tool, part, component, process, design, high

Topic number: 10.0, Absolute ammount: nan, % of total data pieces: nan
Keywords: company, industry, base, production, large, material, system, printer, xu, market

Topic number: 11.0, Absolute ammount: nan, % of total data pieces: nan
Keywords: automotive, system, space, engine, nasa, electronic, rocket, vehicle, sensor, launch

Topic number: 12.0, Absolute ammount: 14.0, % of total data pieces: 5.81
Keywords: lignin, nylon, composite, high, material, viscosity, melt, sample, temperature, filament

Topic number: 13.0, Absolute ammount: 13.0, % of total data pieces: 5.39
Keywords: part, process, production, design, make, industry, support, work, build, component

```

Figure 6: Topic model result text 2

The first 14 topics are displayed (because of a bug topic 5 and 7 were not numbered correctly and are displayed at the bottom, these topics can still be reviewed like the rest). Some of these topics are stereotypical and speak for themselves. We can see again a NASA topic and a topic about global markets but also many other topics. We can see topic 4 about digi_communication with something about a patent and topic 12 with something about Nylon ligament. Let's read into those two. Topic 4 seems like a dud because Digi Communications refers to a financial holding and not something with digital communication that has to do with AM. The data piece that is most representative for topic 12 is a text that goes into depth about material properties used in AM.

This was an example on how the tool can be used. In the next section we will discuss the tool more.

4. Outlook

In this section we will discuss what the tool can do in its current implementation and what potential it has with more implementation.

Email and HTML parsing

First, we parsed emails. We went through the emails in a mailbox (.mbox) file that only contained google alerts emails. The emails are structured data, there were clear titles, sources, example texts and links. Then we went on the internet to retrieve the HTML corresponding to links and cleaned the HTML. The scale of the data allowed us to save it in a file and load the entirety of the data with our RAM. To support larger bodies of data, another way of saving and loading data needs to be implemented.

This first part resulted in a structure for data pieces to be saved in a regular file and methods to parse through google alerts and HTML. The HTML was cleaned using the example text from the email it came in, to identify relevant parts. This means that for data from links without this beforehand knowledge we got from the emails, another cleaning function needs to be written. The Google Alerts cleaning function can be improved.

There are no functions for importing other data from other textual data sources although these are supported with the tool. These can be requested to be written per case as these are not hard to program. Its also possible to make a standard importer from .csv files or .json files with default names for columns for certain attributes, like timestamps.

Concept network

Then, we defined a network of concepts to structure the data. An analyst does not have to start by making a network of concepts. The analyst might choose to do topic modelling on all the data and define the concepts from there according to the emerging topics. The concepts function is quite modular and needs no additions besides a better graphical display. Refining the concepts until sufficiently specific trends can be discovered.

The concept network function can have many uses. It can be used to exclude or include data pieces to the topic modelling or potential other newly implemented algorithms. The concept network is a way to segment the data as a goal or a history of reaching the specific trends an analyst wants to reach.

Period based

It should be made possible to put a period constraint on a concept so that only data pieces from a certain period qualify.

Sub-concept

What isn't implemented yet are sub concepts, concepts that are composed from a subset of their parents concepts. This would be a good addition.

Concept combining

In the current implementation potential conditions are removed from the list when they are used for a concept, for practical use. If I want to formulate a concept that has constraints from two already consisting concepts, I need to take the constraints from both. This is not supported yet by the GUI. This was done with a function for the example use in section 3.3.

Concept visualization

A visualization of different concepts and how they overlap would give great insight. For our example, AM it would be visible in what markets what techniques are used, what materials are used for what markets. A visualizer that shows how two defined dimensions overlap would be of great value.

Tagging conditions

Data pieces were tagged with a concept if they met one of the conditions attached to that concept.

There are many kinds of conditions possible for data pieces to be part of a concept. For instance the occurrence of more than one word together within certain proximity, say within three sentences. It is even possible to annotate part of speech and extract relations within data pieces which can then be used in the constraints for tagging. For simplicity, none of this was implemented in the initial implementation. This is the first thing that should be improved.

Topic Modelling

After making a segmentation with tagging, we used LDA topic modelling to look at what trends we can find in the data. Parameters like stop words and number of topics to model needed to be determined.

Topic Model optimization

Filtering words out of topic modelling with the GUI is not implemented yet, these should be saved per concept specifically, since its best to make a list of not used words per concept. These are words that are not distinctive within the concept, they are distinctive for the concept. For example, we want to make a concept that is about financial markets. Within this segment words like growth, market, stock, etc are meaningless. These words will probably be used to tag data pieces with the financial markets tag. All the data pieces within this concept should talk about these

words. However, within other segments we don't want to segment the data taking account of these words since they are considered in another concept. For short, Topic Modelling should be customizable what words it doesn't use. By default it shouldn't use words that are distinctive to other concepts if it is not topic modelling for those concepts.

Topic representation matrix saving

Right now just the most representative data piece is saved per topic. It would be useful if its an option to save representation of all data pieces for all the topics. This would make it possible to look at a dimension like Techniques or Geographical and asses the adherence to a topic model. If a well defined topic model has one topic for each industry using the topic modelling optimization the adherence of locations can be assessed to industries. This would be a less deterministic way of comparing versus the Concept Visualization and potentially yields different insights.

Result exporting

After one of the comparison functions (topic modelling to concept and concept to concept) have been implemented it would be nice to export results into a CSV file.

5. Conclusion

In conclusion, Trends Tool became a tool that enables an analyst without technical proficiency to discover trends within the analysts domain knowledge as long as there is data of sufficient quantity and quality.

The combination of the very modularly programmed functionalities: Concepts, tagging conditions, topic modelling and plotting allows for a very wide and unspecialized array of potential uses. Trends Tool can be used for any length of text data. Publications, tweets, articles, blogs, YouTube comments and subtitle logs of YouTube videos can all be analyzed using this tool.

These very basic implementations of tagging, a concept network and topic modelling led to a couple of insights. The proposed additional functionalities to be implemented in section 4 can greatly increase the effectiveness of the tool. When implemented correctly, the tool can potentially be very effective for two commercial applications. Tracking industry trends like our example AM and analyzing consumers, since their preferences are often expressed in textual messages. There are also scientific applications this tool can be used for. Given there is a data source, a researcher can gain a better understanding of what is researched.

The tool uses very widely applicable techniques and thus is not able to do the entire job for some purposes. However, this tool does make analysis very interactive. It is possible to learn a lot about your data very quickly. The Tool is simple enough to use to make more in depth analysis an option with constrained time.

5.1 A streamlined process

Trends Tool in its current implementation supports importing data with a google alerts .mbox file. The user only needs to specify the path to the file and the rest of the process is standardized in the Tool. The process it structures didn't change, it goes a lot faster using this tool though, which makes it a more predictable and manageable process.

5.2 Separating technical and business proficiency tasks

There are countless possible improvements or slight adaptations that require pure technical proficiency to implement. Many of which are mentioned in section 4. After implementation of an improvement or adaptation a someone with domain knowledge can use the tool to get the desired results. The tool is programmed in such a way that it is easy for a programmer to add extra algorithms to propose constraints, make more ways to implement more kinds of data, and to run more algorithms on the data for certain desired results. The functions just need to be called in the routine to construct the data objects and the data objects need to be added to the current structure in a correct way.

6. References

- “Additive Manufacturing - Explore - Google Trends.” n.d. Accessed April 18, 2019.
[https://trends.google.nl/trends/explore?date=all&q=additive manufacturing](https://trends.google.nl/trends/explore?date=all&q=additive+manufacturing).
- “Beautiful Soup Documentation — Beautiful Soup 4.4.0 Documentation.” n.d. Accessed May 18, 2019.
<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>.
- Bradshaw, S, Adrian Bowyer, and P Haufe. 2010. “The Intellectual Property Implications of Low-Cost 3D Printing.” *ScriptEd* 7 (1): 5–31. <https://doi.org/10.2966/SCRIP.070110.5>.
- Debortoli, Stefan, Oliver Müller, and Jan vom Brocke. 2014. “Comparing Business Intelligence and Big Data Skills.” *Business & Information Systems Engineering* 6 (5): 289–300.
<https://doi.org/10.1007/s12599-014-0344-2>.
- Ferro, Carlo, Roberto Grassi, Carlo Seclì, and Paolo Maggiore. 2016. “Additive Manufacturing Offers New Opportunities in UAV Research.” *Procedia CIRP* 41 (January): 1004–10.
<https://doi.org/10.1016/J.PROCIR.2015.12.104>.
- “Gensim Topic Modeling - A Guide to Building Best LDA Models.” n.d. Accessed July 1, 2019.

<https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/>.

“Gensim Tutorial - A Complete Beginners Guide – Machine Learning Plus.” n.d. Accessed May 24, 2019.
<https://www.machinelearningplus.com/nlp/gensim-tutorial/>.

Gibert, Karina, DA Swayne, W Yang, AA Voinov, A Rizzoli, and T Filatova. 2010. “Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation.” *Iemss.Org*, no. Kdnuggets 2006: 1–9.
[http://www.iemss.org/iemss2010/papers/W11/W.11.02.Classification of Data Mining techniques and intelligent assistant for choice - KARINA GILBERT.pdf](http://www.iemss.org/iemss2010/papers/W11/W.11.02.Classification%20of%20Data%20Mining%20techniques%20and%20intelligent%20assistant%20for%20choice%20-%20KARINA%20GILBERT.pdf).

Glance, Natalie S, Matthew Hurst, and Takashi Tomokiyo. n.d. “BlogPulse: Automated Trend Discovery for Weblogs.” Accessed May 18, 2019. <http://portal.eatonweb.com>.

“Google Alerts - Meldingen van Interessante Nieuwe Content Op Internet.” n.d. Accessed April 23, 2019.
<https://www.google.nl/alerts>.

Hill, Kashmir. 2013. “‘Google Alerts’ Are Broken.” *Forbes*, July 2013.
<https://www.forbes.com/sites/kashmirhill/2013/07/30/google-alerts-are-broken/#77802bd42ff7>.

Jiang, James J., Gary Klein, and Tom Means. 1999. “The Missing Link between Systems Analysts’ Actions and Skills.” *Information Systems Journal* 9 (1): 21–33.
<https://doi.org/10.1046/j.1365-2575.1999.00050.x>.

Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny

Finkel, Steven J Bethard, and David Mcclosky. n.d. “The Stanford CoreNLP Natural Language Processing Toolkit.” Accessed May 1, 2019.
<https://www.aclweb.org/anthology/P14-5010>.

Rüßmann, Michael, Markus Lorenz, Philipp Gerbert, Manuela Waldner, Jan Justus, Pascal Engel, and Michael Harnisch. 2015. “Industry 4.0: The Future of Productivity and Growth in Manufacturing Industries.”
http://www.inovasyon.org/pdf/bcg.perspectives_Industry.4.0_2015.pdf.

“The Stanford Natural Language Processing Group.” n.d. Accessed April 23, 2019.
<https://nlp.stanford.edu/software/CRF-NER.shtml>.

Thompson, Mary Kathryn, Giovanni Moroni, Tom Vaneker, Georges Fadel, R. Ian Campbell, Ian Gibson, Alain Bernard, et al. 2016. “Design for Additive Manufacturing: Trends, Opportunities, Considerations, and Constraints.” *CIRP Annals - Manufacturing Technology* 65 (2): 737–60.
<https://doi.org/10.1016/j.cirp.2016.05.004>.

Wohlers, Terry, and Tim Gornet. 2016. “History of Additive Manufacturing.”
<http://www.wohlersassociates.com/history2016.pdf>.

Wohlers, Terry T., Ian (Specialist in three dimensional printing) Campbell, Olaf Diegel, and Joseph Kowen. n.d. *Wohlers Report 2018 : 3D Printing and Additive Manufacturing State of the Industry : Annual Worldwide Progress Report*.

YANG, QIANG, and XINDONG WU. 2006. “10 CHALLENGING PROBLEMS IN DATA MINING

RESEARCH.” *International Journal of Information Technology & Decision Making* 05 (04): 597–604.
<https://doi.org/10.1142/S0219622006002258>.

Zhang, Zhongfeng, and Qiudan Li. 2011. “QuestionHolic: Hot Topic Discovery and Trend Analysis in Community Question Answering Systems.” *Expert Systems with Applications* 38 (6): 6848–55. <https://doi.org/10.1016/J.ESWA.2010.12.052>.

10. Appendix

All additional and technical information can be found at the git hub page of this project:

<https://github.com/sisko444/Trends-Tool>