UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga
(Final project paper)

Sledenje inovacijam in tehnologijam v umetni inteligenci

(Tracking innovations and technologies in artificial intelligence)

Ime in priimek: M.Besher Massri
Študijski program: Računalništvo in informatika
Mentor: doc. dr. Branko Kavšek
Somentor: prof. Dunja Mladenić
Delovni mentor: Marko Grobelnik

Koper, avgust 2021

Ključna dokumentacijska informacija

Ime in PRIIMEK: M.Besher MASSRI

Naslov zaključne naloge: Sledenje inovacijam in tehnologijam v umetni inteligenci

Kraj: Koper

Leto: 2021

Število listov: 50 Število slik: 18

Število prilog: 1 Število strani prilog: 7 Število referenc: 16

Mentor: doc. dr. Branko Kavšek

Somentor: prof. Dunja Mladenić

Delovni mentor: Marko Grobelnik

Ključne besede: vizualizacija podatkov, rudarjenje besedila, taksonomija, inovacije

Izvleček:

Diplomsko delo predstavlja sistem za sledenje inovacijam in tehnologijam od njihovega nastanka v akademskem svetu do njihove možne uresničitve v poslovnem svetu. Sistem uporabljamo orodje iz umetne inteligence (UI) za prepoznavanje in analizo možnosti tehnologij. Pokažimo, kako lahko sedanje vroče teme v UI spremljamo do inovacij in napredka, ki so se zgodile pred leti. Naš sistem spremlja stanje tehnologij v več fazah, vključno z raziskavami, patenti, novicami in trgom dela. Trenutna literatura se večinoma osredotoča na sledenje inovacijam samo na eni stopnji razvoja, običajno se osredotoča na patente, raziskave ali trg dela ločeno. Vendar naš system, kolikor nam je znano, prvi javni sistem, ki sledi razvoju tehnologij po več različnih kanalih.

Key words documentation

Name and SURNAME: M.Besher MASSRI

Title of final project paper: Tracking Innovations and Technologies in Artificial Intelligence.

Place: Koper

Year: 2021

Number of pages: 50 Number of figures: 18
Number of appendices: 1 Number of appendix pages: 7 Number of references: 16

Mentor: assist. prof. Branko Kavšek, PhD

Co-Mentor: prof. Dunja Mladenić

Working mentor: Marko Grobelnik

Keywords: data visualization, text mining, taxonomy, innovation

Abstract:

This thesis presents a system for tracking innovations and technologies from their inception in academia to their possible realization in the business world. We apply the system to the Artificial Intelligence domain to identify and analyze the landscape of AI technologies. We illustrate how current hot topics in AI can be traced back to innovations and progress that happened years ago. Our system monitors the status of AI technologies in multiple stages including research, patents, news, and the job market. Current literature mostly focuses on tracking innovations on one stage of development only, usually focusing on patents, research, or job market separately. However, our system is the first - up to the knowledge of the author - public system that tracks the evolution of AI technologies across the different channels.

Acknowledgement

The project was supported by the Organisation for Economic Co-operation and Development (OECD), which enabled me to use data sources that otherwise wouldn't be available, as well as the Artificial Intelligence Department in Jozef Stefan Institute that enabled access to additional technologies and resources. I thank my mentors and colleagues at the institute for their help and advice.

Contents

List of Figures

Appendices

List of Abbreviations

i.e.	that is
e.g.	for example
AI	artificial intelligence
CLI	command line interface
API	application programming interface
TSV	tab separated values
CSV	comma separated values
DAG	directed acyclic graph

1 Introduction

The rapid development of technologies caused an expansion in innovation, which prompted a need for systematic and continuous monitoring of the development of innovations and technologies in any domain through a series of dimensions that define the environment or the ecosystem in which the domain is located and affected.

Given that innovations usually occur in the academic environment and can be observed through publications. Initial ideas often appear first at less important academic events and journals, and later progress to higher-ranked conferences and journals. Many innovations die out in academia due to insufficient potential, and some of them survive and continue to the next stages. In particular, one narrative of stages which we could understand as a "journey of an innovation" throughout the innovation life cycle:

- An innovation typically appears in the academic world;
- projects are started around the innovation;
- the innovation gets possibly patented;
- companies are established around innovation;
- companies get investments, possibly in several rounds;
- investments influence the job market;
- market reacts to the quality and possible impact of the innovation;
- public and expert perception gets formed;
- media starts publishing about the innovation and companies;
- educational institutions integrate innovation in their curricula,
- policymakers regulate the innovation; and
- to close the cycle, funding agencies start creating new funding opportunities to create space for follow-up innovations.

Each of the above stages has its stakeholders (from scientists to policymakers) who contribute to the "journey" in their specific ways. The "journey" or path of innovation can last from few years, up to decades. Many of the innovations get lost on the way due to several reasons, like lack of potential or sometimes politics.

The main hypothesis that this paper is trying to answer is whether innovations paths or cycles like the one mentioned above can be automatically tracked through time and whether there are other possible paths for innovations. For that purpose, the paper presents a system that enables users to follow the state of innovation in any topic over time. That includes extracting relevant concepts from literature, using text-mining to track the concepts across the different data sources, and monitor their evolution through phases and time.

Due to the breadth of the global innovation ecosystem, and the limitation of this work, the thesis's main focus will be on a narrower field of Artificial Intelligence, as currently one of the hottest topics spreading horizontally across many fields of research, technology and impacting society in various ways. In addition, it will be using selected resources only, namely research publications, patents, news coverage, open-source code, and the labor market. However, the approach will be generic and will have the potential to expand it to other fields and resources.

Looking at existing systems, the innovation ecosystem is typically observed and analyzed only locally and fragmented focusing on individual stages, without a holistic view of the complete life cycle. Most of the research focuses on the patents as the main driver of innovation therefore use them as a proxy for it, as seen in [9, 5]. Other work aims at tracking open-source software by connecting open source communities and user feedback, exhibited in QAs from Stack Overflow [3]. Meanwhile, solutions like Gartner hype cycle [7] provide the intended solution but without much explainability or the ability for user exploration, which limits its usefulness.

Having such systems (possibly with extension to other data sources) can be used in three levels:

- Organizational: informing policymakers about the state of technologies and areas of innovation, so that they can create policies that enable their perspective countries/organizations to be on the frontier of innovation

- Financial: helping VC investors to learn about the hot trends and technologies that might have a high potential so they can invest in them.

- Academic: helping researchers to identify gaps and opportunities where innovation can be made to be at the forefront of research.

2 Methodology

This chapter contains some overlap with my earlier work on the OECD AI observatory [16] as well as the DataBench EU project [2].

The system consists of several steps starting from crawling and processing the data, into building the visualization dashboard, as seen in figure 1. Below is a description of the process:

- Collect and process data sources: this include build crawling scripts, cleaning and formatting the data, etc.
- Identify the list of concepts, innovations, and technologies from research: using literature as a source of innovation and extract relevant concepts from it.
- build the taxonomy out of the corpus of vocabularies: using subsumption method to build a hierarchy out of the concepts themselves.
- Document annotation: annotate the documents of all data sources with taxonomy concepts and calculate the necessary results.
- Ingesting data into elastic search: for enabling efficient search through results, using Logstash.
- Analytical dashboard: using the results stored in the database to build visualizations that unlock the hidden insights from the data.

Each of the steps will be discussed in detail in the next sections. As for the analytical dashboard it will have its own chapter.

2.1 Crawling and processing data sources

The Data Sources used for the system cover research and development, news, and industry and include:

- Research publications and patents from Microsoft Academic Graph;
- News articles from Event Registry system;

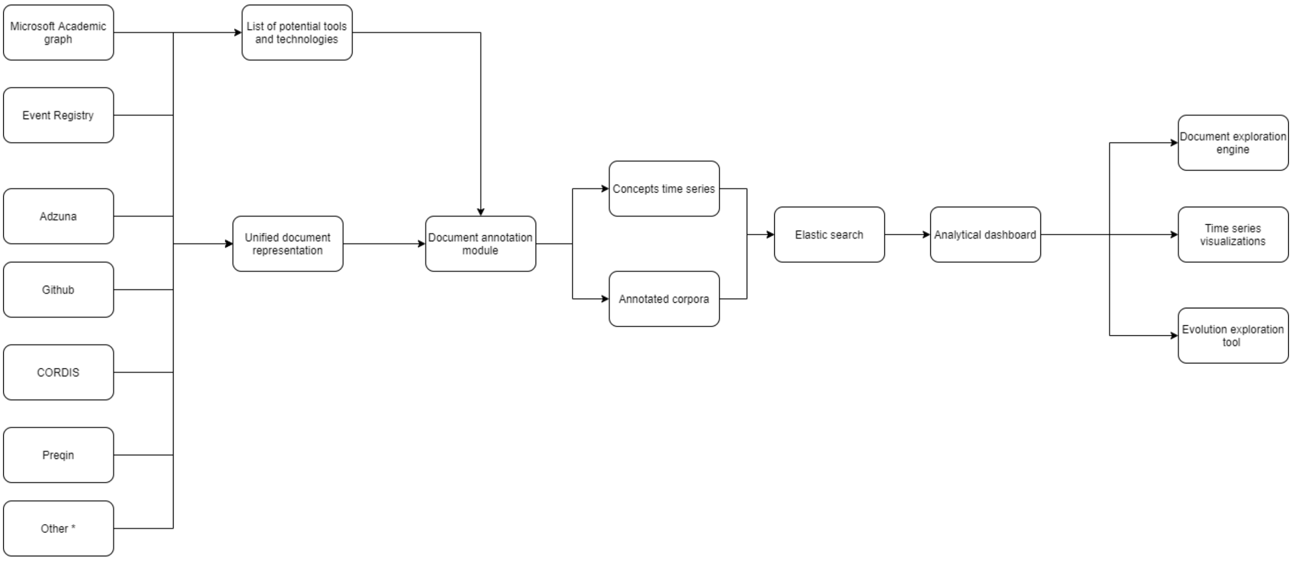

Microsoft Academic graph
Event Registry
Adzuna
Github
CORDIS
Preqin
Other *
List of potential tools and technologies
Unified document representation
Document annotation module
Concepts time series
Annotated corpora
Elastic search
Analytical dashboard
Document exploration engine
Time series visualizations
Evolution exploration tool

Figure 1: Solution architecture

- Job postings from Adzuna service;
- Open source projects from GitHub

Since the system ingests different forms of datasets, a custom crawler and processor have to be built for each one.

2.1.1 Research and Patents: Microsoft academic graph

The Microsoft Academic Graph (MAG) [15] is the largest publicly available database of scientific publications. It contains more than 265 million records (As of August 2021) that span from the early 1800s up to the present time and contains several paper types including journal articles, conference papers, books, book chapters, and others (mostly being from repositories like Arxiv). In addition, MAG provides large corpora of patents that are collected from several public repositories. MAG database is made as a heterogeneous graph containing information about the journals, conference, citation relationship, and fields of study.

The database is provided in the form of dumps in an azure container on a biweekly basis. Each dump contains the entire snapshot of the graph and not just the additions from the last version. Therefore a python script was made that would check once a day for a new version and if available would download it using Azure CLI. The data comes in the form of a SQL-like dump, i.e. a table for each entity type (e.g. papers, authors, institutions, etc.) in a TSV format. For the latest MAG schema, please refer to [12].

For this tool, we used papers that fall under the artificial intelligence or machine learning topics or one of their subtopics. The period of data was limited up until 1980

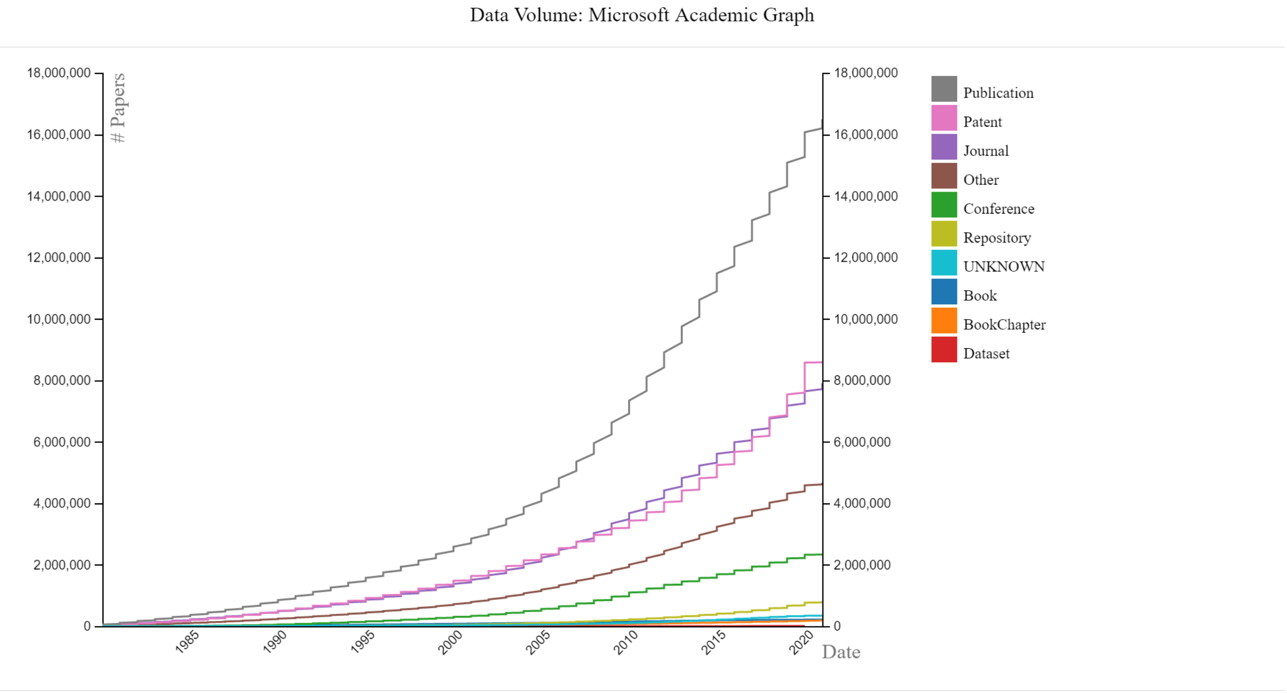


Figure 2: Number of AI research papers by paper type in time (monthly basis)

as most publications before were noise from the topic classification. As a result, 25m items have been examined. The time series of publication per paper type can be viewed in figure 2. Please note that ”Publication” type refers to the aggregate of all types except patent, highlighting the difference between patent and research publication, which will be used across the paper. An example of a research paper document can be seen in the appendix.

2.1.2 News articles: Event Registry

For collecting news articles about artificial intelligence, the Event Registry system has been used. Event Registry [10] is a global news monitoring system and news intelligence platform that collect, annotate, cluster and analyze news articles and world events from all over the world. Event Registry consumes and analyzes news data in multiple languages. It currently supports news from the following languages: English, German, Spanish, Portuguese, Italian, as well as 40 other languages. Its cross-lingual feature is powered by the Wikifier tool [1], which uses Wikipedia knowledge base to annotate text with relevant Wikipedia concepts, by using the multilingual nature of Wikipedia concepts(pages) across languages. The result of the annotation is a list of concepts and URLs to Wikipedia pages normalized to the English base if available. Using the concepts, Event Registry provides a cross-lingual search through articles by any

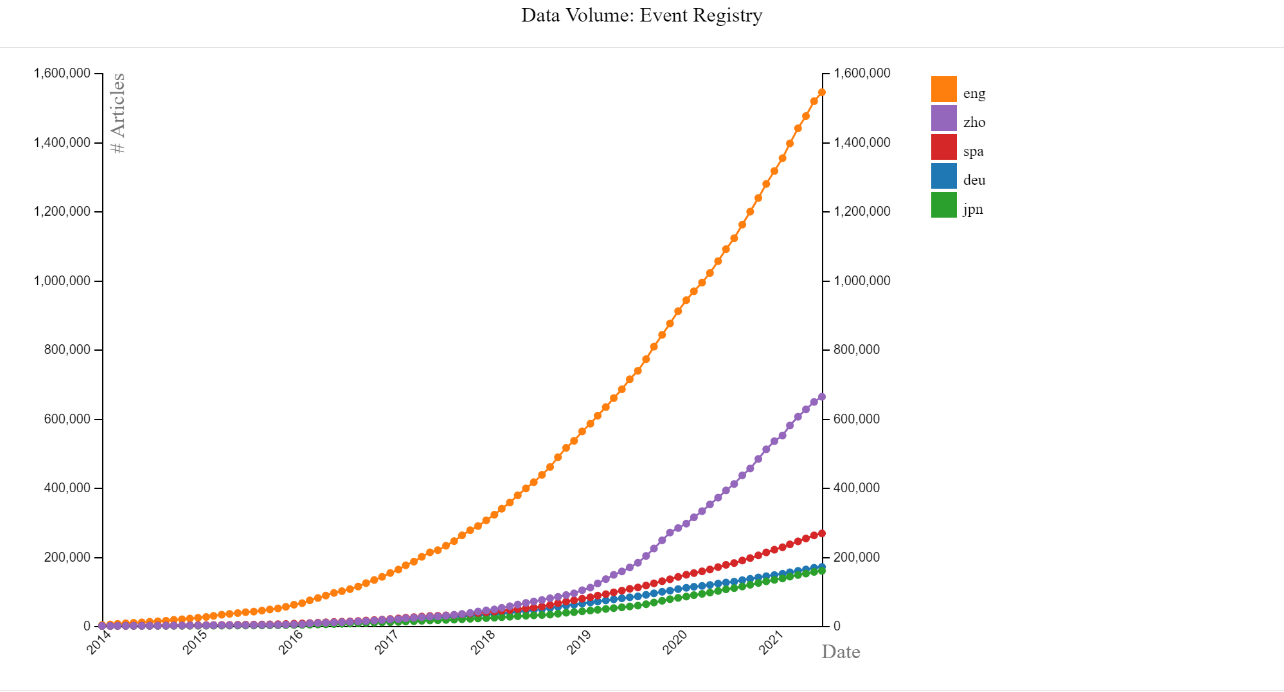


Figure 3: Number of AI news articles by language in time

term that has any corresponding Wikipedia concept, in addition to the usual keyword, language specific search. The data has been collected using the Event Registry python SDK library (`https://github.com/EventRegistry/event-registry-python`) using the ”get-articles” endpoint. The result is a list of documents in JSON format each representing a single article. To filter for the AI articles, ”Artificial intelligence” has been used to fetch all the relevant articles. The resulting corpus contains 3.4m articles in multiple languages that span from 2014 until the present time. In figure 3 we see the time series of AI news articles by language. An example of a news article can be seen in the appendix.

2.1.3 Job postings: Adzuna.com

The job posting data comes from the service Adzuna.com [8], a job search engine for job postings. Adzuna covers countries such as Austria, Australia, Brazil, Canada, France, Germany, India, Italy, Netherlands, New Zealand, Poland, Russia, Singapore, United Kingdom, USA, South Africa. The job ad data is multilingual and published in English, German, Portuguese, French, Hindi, Italian, Dutch, Polish and Russian. Each job posting is classified by Adzuna into one of the following categories: Consultancy, Charity Voluntary, Property, IT, Legal, Customer Services, Teaching, Other/General, Accounting Finance, Retail, Manufacturing, Hospitality Catering, Healthcare Nurs-

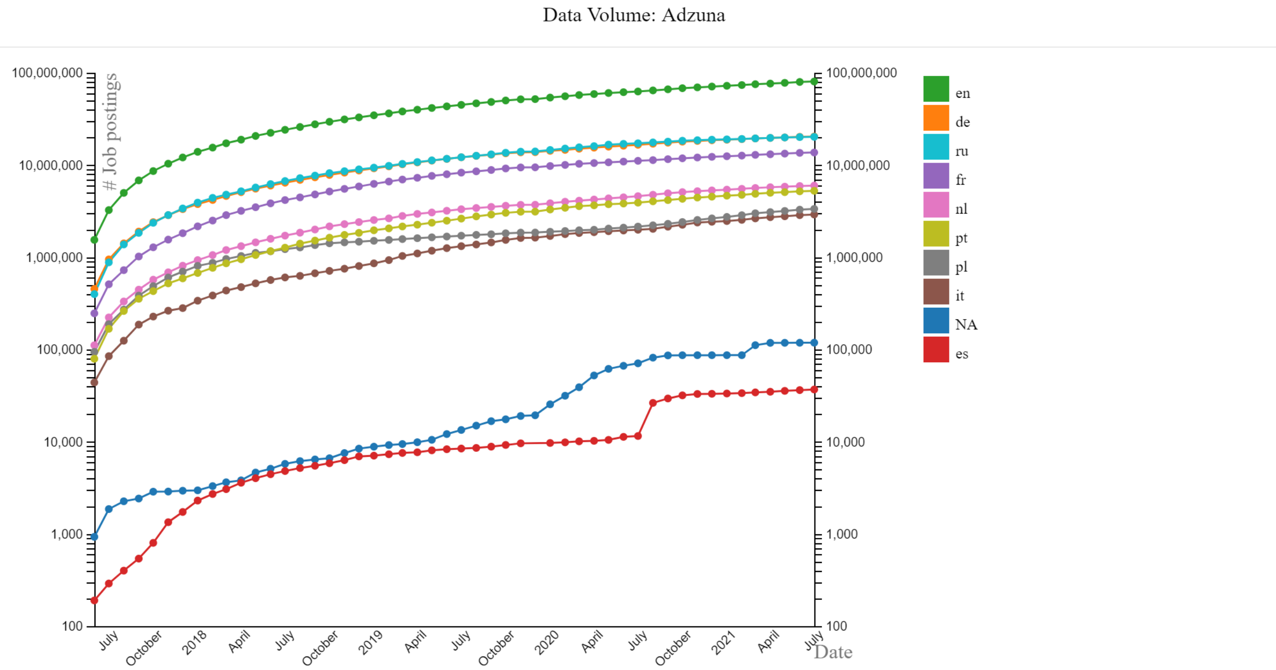


Figure 4: Number of IT job postings by language in time

ing, Trade Construction, Domestic help Cleaning, Creative Design, Logistics Warehouse, HR Recruitment, PR, Advertising Marketing, Social work, Travel, Energy, Oil Gas, Maintenance, Scientific QA, Graduate, Engineering, Part time, Unknown, Sales, Administration.

The data is provided in an amazon container by the Azuna team in an XML format. However, the data obtained were processed and wikified by my colleague Jakob Jelencic in the Jozef Stefan institute who is working on this dataset for his Ph.D. I received the data in a CSV format that contains all the relevant attributes as well as the top 10 Wikipedia concepts (extracted using the same wikifier tool used for news) for each article. The corpus was filtered to those with IT categories. The final corpus contained 152m job posting that spans from 2017 till now, as seen in figure 4. An example of a job posting can be seen in the appendix.

2.1.4 Open Source Code: GitHub

GitHub is a development platform that allows for hosting open source as well as business projects [6]. The users can host and review code, manage projects, and build software alongside 50 million developers. Each repository contains information like title, description, list of text files, # stars, # forks, list of topics, list of programming languages, etc. For data collection, a list of all the repositories that contain "artificial

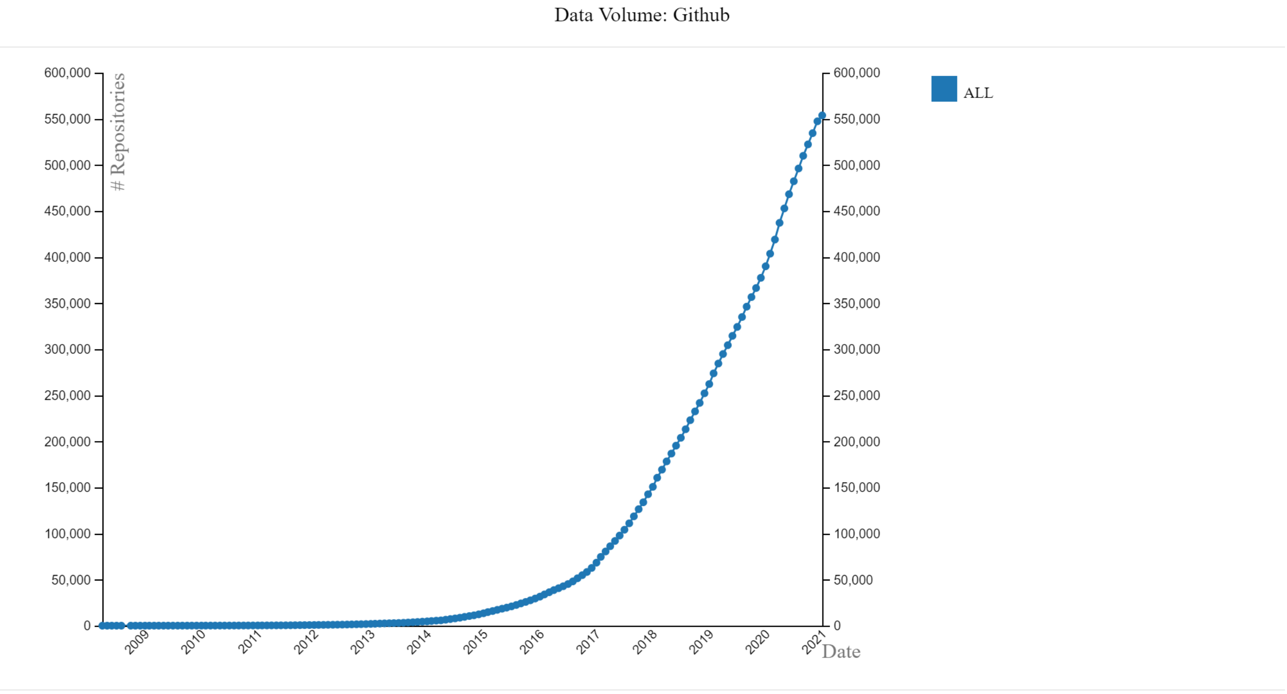


Figure 5: Number of AI open source repositories in time

intelligence", "deep learning", or "machine learning" in either the textual data or the topics and has at least 5 stars or forks has been obtained. Using this list, the Official GitHub GraphQL API has been used to collect information about each of the repositories. The range of data used is from 2008 till January of 2021, as seen in 5 with 550 thousands repositories collected in total. An example of a GitHub repository document can be seen in the appendix.

2.1.5 Common document representation

To simplify document representation, a common document representation has been introduced. This representation contains most of the shared attributes across the data source. The list of the attributes are as follows:

- id: the id of the document
- title: the title of the document
- text: the text/body/content of the document
- created date: the timestamp when this document was created
- locations: list of all geographical locations that this document mention or corresponds to. The locations are free text, or country codes, or Geo names ids

- topics: list of categories/tags/topics that this document is annotated with.
- wiki URLs: list of Wikipedia concept URLs that this document is annotated with.
- authors: list of authors of this document
- affiliations: list of affiliation(e.g. institutions, companies, agencies, etc.) that corresponds with this document
- other: any additional information that doesn't fit one of the above

Using this representation allows for analyzing textual content across time, location, and topic allowing for a wide variety of analyses. An additional advantage of using this representation is that it allows the system to ingest other textual resources without the need to modify the system. The only thing needed is a configuration file containing the new source information like the directory, format, and the mapping between the common documents attributes and their equivalent in the new data source.

2.2 Extraction of innovation concepts

Once the data sources are crawled and processed, the next step is to extract all the possible relevant terms from the literature that would ideally include all technologies, tools, and concepts that have the potential to drive innovations; building on the hypothesis that for most innovative tools, technologies, and concepts exists as a phrase in the literature, e.g. "recurrent neural network", "transformer", "TensorFlow", etc. With that in mind, a procedure has been developed that would process each document from Microsoft Academic Graph, generating all the possible n-grams up to a certain length, and then applying multiple steps of filtering.

At first, the textual description (i.e. title + content/abstract if available) of each MAG document has been tokenized and filtered from any punctuations, and other symbols. Then, lemmatization and part of speech tagging have been applied to each word to merge the different forms of the word like "network" and "networks", as well as to remove verbs since the target vocabulary is more of abstract terms and names. The WordNet library has been used for lemmatization and part of speech tagging, which is a lexical database of semantic relations between words and has corresponding libraries in multiple programming languages [4].

After the mentioned word filters have been applied, the list of n-grams with lengths in the range [1,3] has been generated. The resulting n-grams might include partial phrases that do not have a meaning on their own or don't correspond to a single concept.

To mitigate that effect, the ngrams were matched against the Wikipedia database of concepts and all the n-grams that don't correspond to a Wikipedia concept/page have been filtered out. Finally, as the interest is in tracking innovations (hence, there should be enough research interest first) and to cut the long tail of resulting terms, we filtered all the resulting vocabulary with document frequency less than a certain threshold [1].

2.3 Building taxonomy

Using the list of vocabulary, a hierarchical dependency is needed to form the taxonomy. For that purpose, a similar methodology has been used to that of building MAG topics hierarchy [14], which is based on the subsumption concept introduced in [13]. except that in this case, the matched concepts are unweighted. More formally, let **a** and **b** be two concepts from the generated vocabulary, and Let R be the following:

$$R = \frac{|A \cap B|}{|A|} - \frac{|A \cap B|}{|B|} \tag{2.1}$$

Where **A**, **B** is the set of documents tagged with concepts **a**,**b** respectively. Concept **a** is said to fall under concept **b** if **R** falls in a certain range[2]. Due to the nature of this approach, the same concept can fall under two different concepts, hence the resulting taxonomy is a directed acyclic graph (DAG), rather than a tree. A directed acyclic graph (DAG) is a directed graph that contains no cycles.

2.4 Document annotation

By this step, we have the documents of all data sources in the common representation as well as the taxonomy to tag the documents with. A document can be annotated with a concept in two ways:

- If the concept exists in the textual description of the document (title + text)
- If the concept exists as one of the topics of the documents

The concept-tagging via topics allows enables tagging of non-English documents, as in Adzuna and Event Registry. This is possible since the list of concepts are Wikipedia concepts, and each taxonomy concept is also a Wikipedia concept.

[1] The threshold has been empirically set to 100

[2] The range is set empirically to [0.1,0.5]

2.4.1 Annotated corpora

The first result out of the annotation is the annotated corpora itself. The resulting files contain a file per data source, each containing the list of documents annotated with at least one concept. Each document is represented by the id of the document, its title, text, and the list of concepts that this document has been tagged with.

2.4.2 Concept time series

Another important result for tracking concepts is the time-series statistics of the concepts. Each data source produces a metric or a set of metrics that are tracked in time. More specifically, for each concept, we were tracking the following metrics across time:

- Number of research papers
- Number of news articles
- Number of GitHub repositories
- Number of job posting

Those values were calculated and aggregated monthly. In addition, data volume metrics representing the overall volume of data against each metric (i.e. each document is considered) were calculated to help to put numbers in the context of the corpus used.

2.4.3 Tracking evolution

Tracking evolution across data sources means monitoring how each concept value increase or decrease in a stage/data source compared to other stages/data source. To achieve that, values have to be on the same denominator across data sources, i.e. there is a need for a normalization step. In addition, the values tend to increase over time. Hence, there are two possible steps for normalization each conveys a different purpose.

- normalizing w.r.t the total volume for each resource at each point in time: this helps track the number of documents as a ratio of the total corresponding value. This is essentially calculating the probability distribution of the concept.
- normalizing w.r.t the time-span: by having the largest value equal to 100 and the smallest equal to 1 (with zeroes remaining as zeroes). This achieves the intended normalization that allows for comparison between data sources

Hence, the evolution score can be defined as the probability distribution of a concept, normalized to a score from 0 to 100.

2.5 Ingesting into Elasticsearch

After results have been calculated, the data is ingested into a database to allow building efficient analysis and visualizations. To consider the choice of database, the list of operations to be supported by the dashboard has to be considered. Most notably, the following properties are required:

- Efficient search across text description
- Fast aggregation and filter queries
- Fast insertion and deletion is not a priority, as the data is ingested each period of time (biweekly to monthly) and the entire database is replaced.

For that purpose, Elasticsearch seemed to be the best solution, as most of the other solutions don't support fast search queries out of the box. Having two different types of databases could be an option, like elastic search for text queries and PostgreSQL for filter and aggregation, but that would be an unnecessary complication.

As for efficient ingestion of data into elastic search, the log stash tool has been used, which is a data processing pipeline that ingests data from a multitude of sources, apply transformations on it, and output it or store it in a database [11].

3 The AI concepts monitoring dashboard

An interactive visualization dashboard was built on top of the database to enable a visual exploration of the results, powered by libraries and tools like D3, Bootstrap, search point. The dashboard provides analysis on different levels and from different perspectives enabling in-depth exploration as well as big picture analysis. The main sections in the dashboard are data volume series, taxonomy explorer, concept evolution, resource time series, and documents explorer.

3.1 Data volumes

The data volume section provides a first insight into the data used by the system. It contains four different charts that correspond to the four data sources used in this paper (research publication and patents both provided by Microsoft Academic Graph). Each chart provides the time series of the number of documents, either as a total or by a certain category, as follows:

- Microsoft Academic Graph: total number of AI-related research per paper type, that includes conferences, journals, patents, among other
- Event Registry: total number of AI-related news articles per language
- Adzuna: total number of job postings per language.
- Github: total number of open source repositories

At first, the data volume charts represent the total volume of the corpus data in each data source. These charts give the big picture about the amount of data we're dealing with to gain perspective when examining the volume per concept. The charts were omitted as they are the same charts used in the data sources section in methodology.

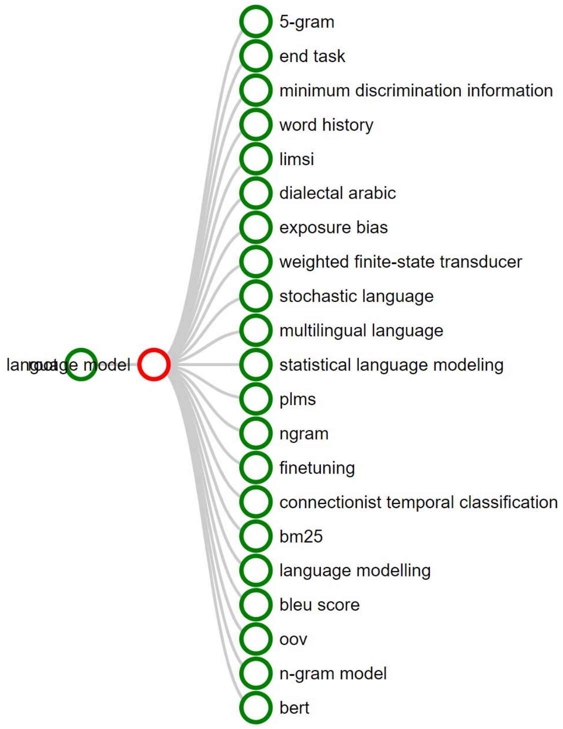


Figure 6: The list of concepts that falls under the concept ”language model”

3.2 Taxonomy explorer

The generated vocabulary consists of 176832 concepts that contain a wide variety of categories, from narrow and specific like the names of algorithms, like linear regression, support vector machine, to topics and fields like a neural network, deep learning, and expert systems, as well as tools and libraries, like TensorFlow, PyTorch, etc, etc. To explore the taxonomy, the taxonomy explorer tool can be used, by initially selecting a concept, the list of concepts was that fall under the selected concept will be displayed in a tree-like graph. In figure 6 we see the list of concepts that were determined to fall under the language model concept. While there are some errors, it catches several concepts that can be considered language models, like BERT and n-gram, and stochastic language among others. Some others are concepts that are usually associated with language models (and other stuff), like fine-tuning. The tool enables navigation through the concept tree by clicking on the concepts, or by using a search box.

3.3 Concept Evolution tool

The main tool in the dashboard is the concept evolution tool, Which enables exploring a single concept across the different stages of development. It’s based on the evolution time series explained in the methodology. By plotting the evolution time series of the

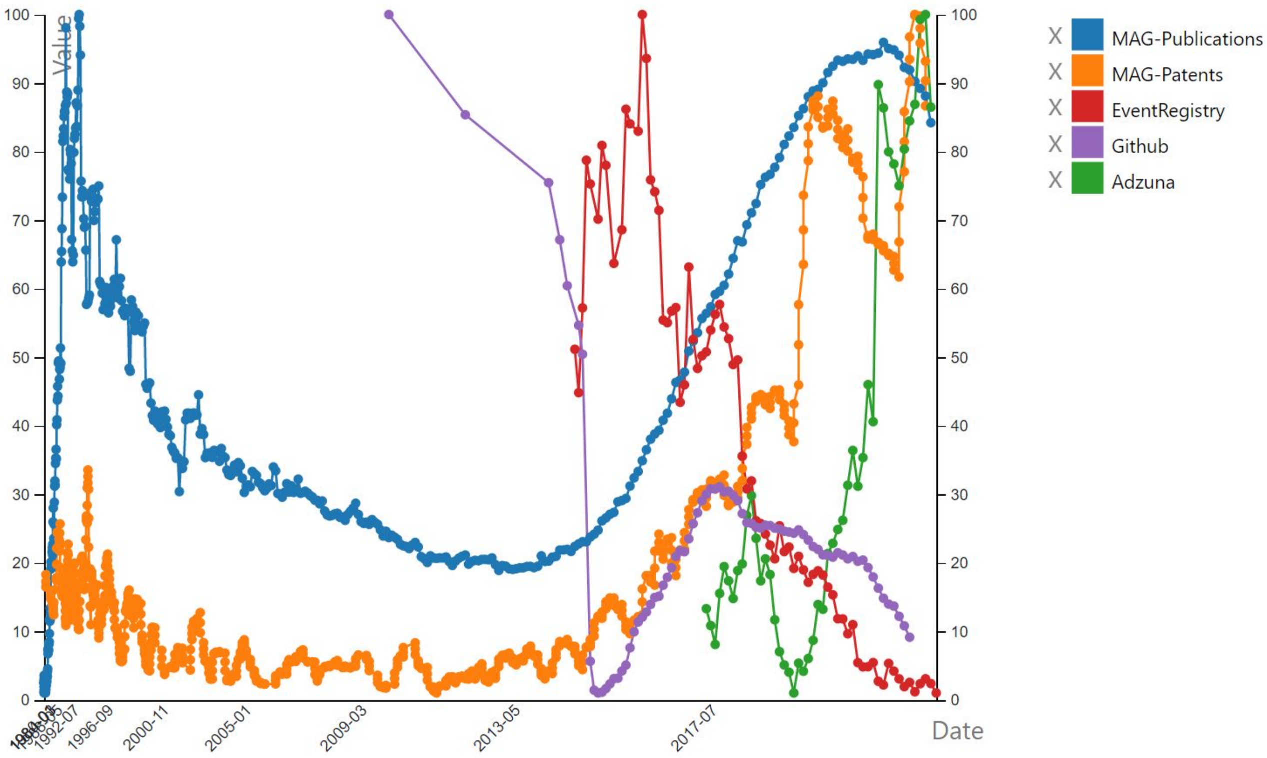


Figure 7: Evolution tracking of recurrent neural network

selected concept in each data source, we manage to observe the status of the concept in each stage at the same time, and based on that, identify the stage in which the concept lies. For example, in figure 7 we observe the evolution graph of the recurrent neural network concept. We can identify the concept inception in the late 1980s, where some spark in publications and patents as a result. However, as with other neural network-based models, the model required an amount of data and hardware requirements that did not exist at that time. Hence, the decline of actions. Things didn't start again until the last decade with the emerges of neural network architectures again, then the research continued with increasing open-source content and occasional spikes of news that correspond to important events. Finally, the industry demand starts increasing and peaking last year.

By inspecting the group, there are two points would need more investigation, the first being the spike of open source code represented by Github at the beginning of 2009, and the reason behind the spikes in news, i.e. what are the events that led to that. To help answer these types of questions, two things have been introduced. First, the annotated corpora have been used to explain spikes and tendencies in the chart. By providing a sample list of the documents annotated with the selected concept at the date that corresponds to the selected data point. In addition to the annotation, The tool provides different options for customization to enable exploring from different angles, amplifying peaks, and fixing anomalies. The full list of options are as follows:

- Date Range filter: enable selection of a specific time range, and calculating the evolution based on that. This is helpful to remove the points caused by noise and zoom in on the useful area to provide a better chart.

- Normalization technique: provide four options when it comes to normalization, based on the two levels of normalization provided in the methodology:

  - No normalization: this is used to show the absolute numbers. This makes the lines is hard to compare across data sources, but provide an idea about the difference in volumes of documents annotated with the concepts across data sources.
  - Normalize by time only: this enables simple comparison between the different stages, by putting them on the same denominator. This option is used when the question to be answered is in the form of ”Does this concept/technology/tool still used/required/talked about or not”
  - Normalize by data volume: this enables tracking the change in ”market share” of a certain concept w.r.t to the ”market”. It enables observing the change in trends of concepts across time.
  - Normalize by data volume and time: this is basically what is referred to as the evolution score where the interest is on both the emerging/decline or a certain concept w.r.t itself and to other concepts.

- Normalization power: when normalizing values across time. Instead of using a linear scale to scale numbers, use a power scale of exponent 2 or 3. This has the effect of amplifying high values and inhibiting low values, which helps in identifying peaks.

- Min cutoff percentage: to remove small values that are likely caused by noise, use the min cutoff option to set a percentage cutoff. a 10 percent cutoff threshold means any data point that has less than 10% of the maximum data point value in the same data source time series will be set to 0. This is particularly useful to eliminate noise in the early stages of the data source where there are not many items, hence the evolution score will be high but wrong.

- Smoothing: since the data points are calculated per month, some peaks might arise that are more reflective of the data source rather than the concept, like in research when peaks appear in the months with big conferences. A good default value for smoothing is set to be the average of the year it falls in, which helps stabilize results.

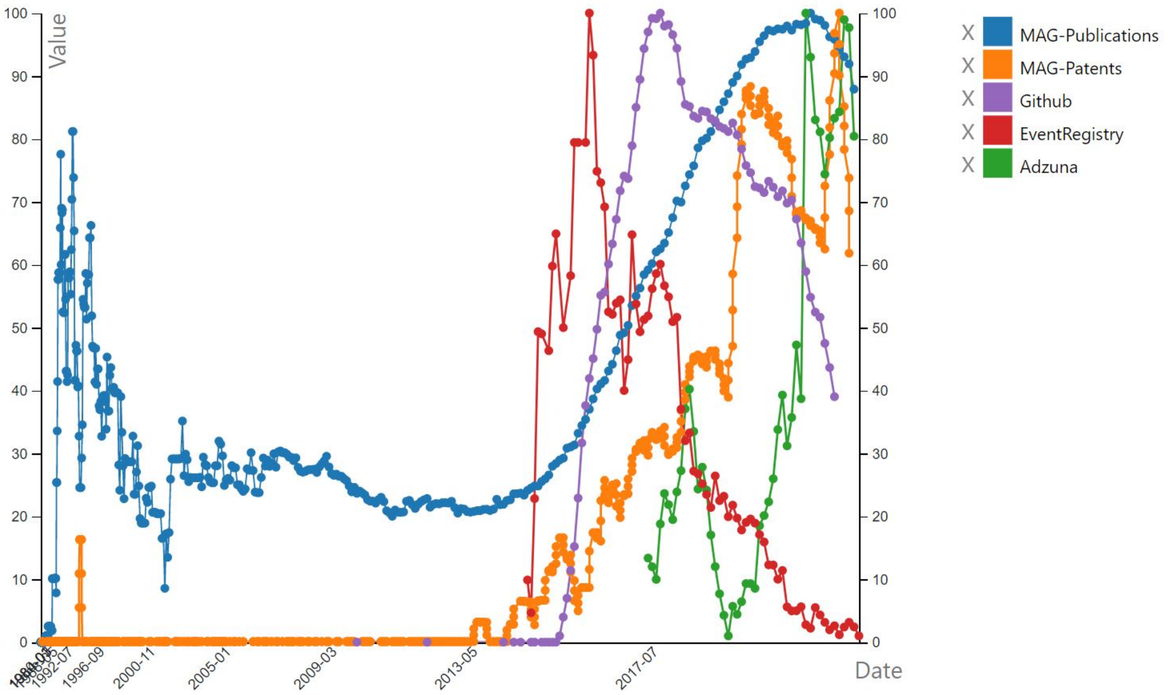


Figure 8: Evolution tracking of recurrent neural network, with min cutoff percentage set to 5%

When inspecting the peak in Github evolution scores at the start of Github corpus time span - by checking the list of repositories tagged with it and comparing it with the non-normalized version - we notice that repositories are outliers in which a repository was created at that time but then the READ ME file updated to include new information, which included the concept, and as the repository creation date was used, this is tagged as something that happened in the past. On the other side, at the beginning of the Github time span, AI repositories were relatively low compared to now, which made the ratio to data volume high enough to be the top in terms of evolution score. This can be fixed by applying the min cutoff threshold to remove the outliers, as seen in figure 8. However, it's important to note that this cut-off threshold is applied to all data sources at once hence the change in the shape of other data sources' lines.

3.4 Resource time series

While the concept evolution tool provides an evolution of a single concept across stages, this tool enables comparing different concepts against the same data source, thus enables a more in-depth comparison between the performance of similar concepts. In figure 9, we observe the difference in evolution score in research between expert systems, neural networks, and deep learning topics. The graph shows the wide use of

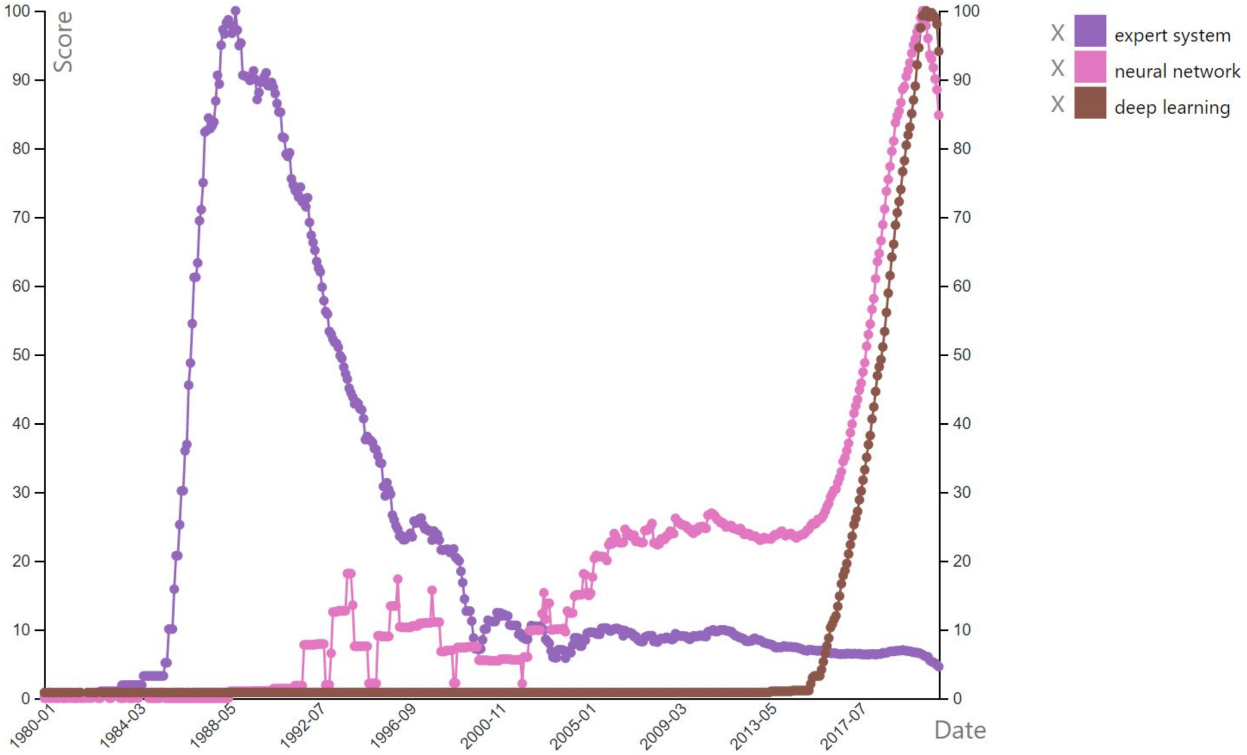


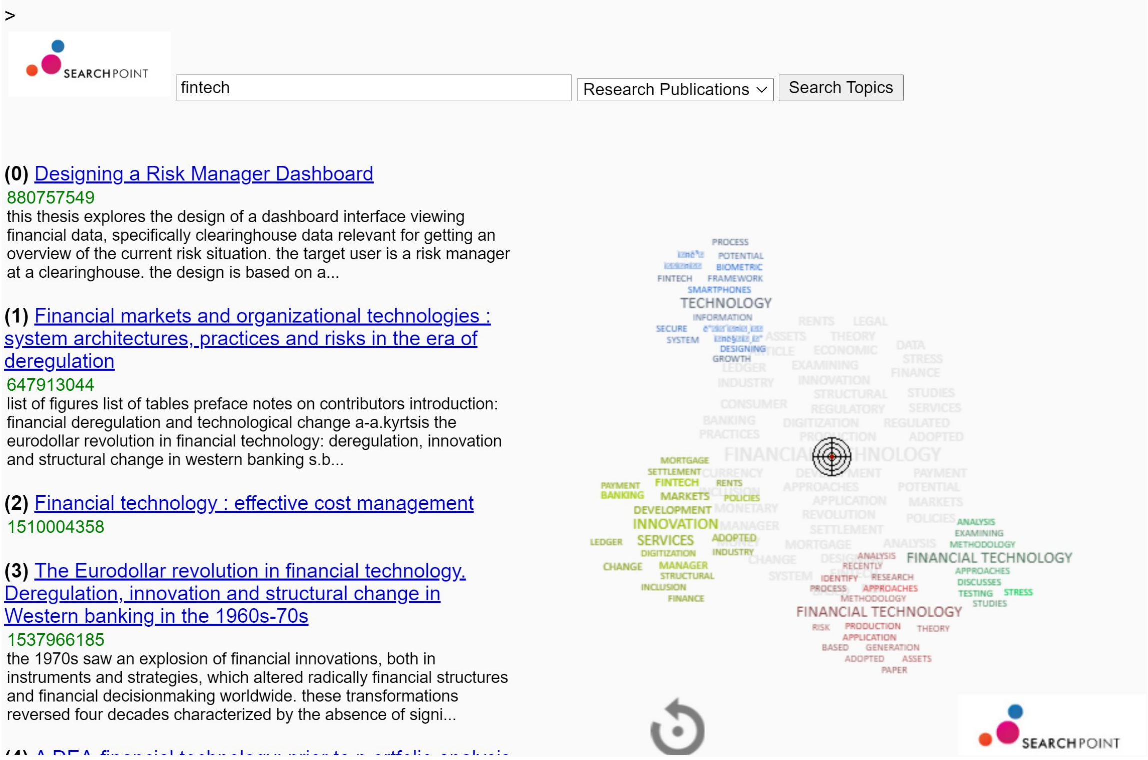
Figure 9: Comparison between the evolution of expert system, neural network, and deep learning in research

expert systems in the early days of Artificial intelligence and its decaying towards the modern-day, as opposed to the deep learning field which was coined in the last decade. As for neural networks as a whole, some signals can be seen in the early days, similar to the ones found in recurrent neural network graphs, but then only emerging later as well.

The tool provides similar features and options to the concept evolution tool, hence, they will not be explained again. More on the tools is provided in the discussion.

3.5 Document explorer

The document explorer tool utilizes the annotated corpora and provides a search engine to help to explore concepts on the document level. For this purpose, the search-point tool has been used, which is a tool developed by the Artificial Intelligence Department at Jozef Stefan Institute. The tool provides a search interface by searching through keywords and then using elastic search to get the results, it then categorizes the results in multiple clusters based on their content. For this system, the tool search was changed to a concept-based search. By selecting a concept and a data source, the search result will return the list of documents annotated with that concept. An example of this



(0) Designing a Risk Manager Dashboard
880757549
this thesis explores the design of a dashboard interface viewing financial data, specifically clearinghouse data relevant for getting an overview of the current risk situation. the target user is a risk manager at a clearinghouse. the design is based on a...

(1) Financial markets and organizational technologies : system architectures, practices and risks in the era of deregulation
647913044
list of figures list of tables preface notes on contributors introduction: financial deregulation and technological change a-a.kyrtsis the eurodollar revolution in financial technology: deregulation, innovation and structural change in western banking s.b...

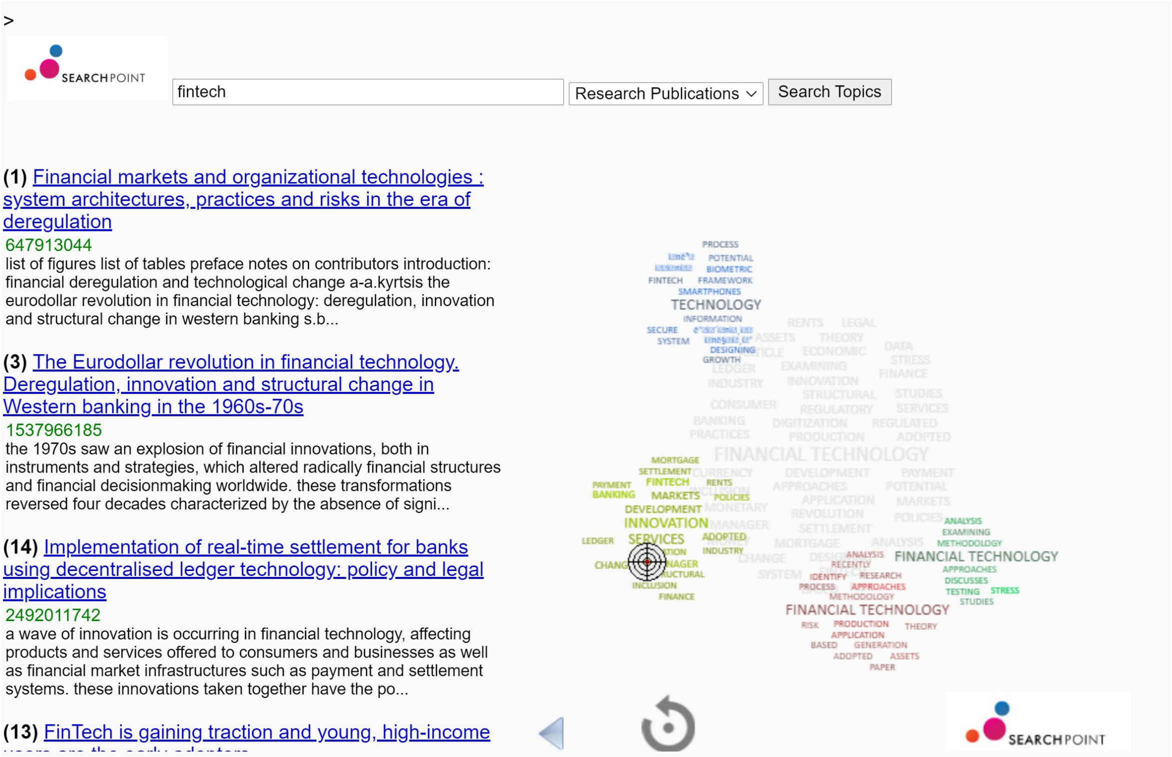
(2) Financial technology : effective cost management
1510004358

(3) The Eurodollar revolution in financial technology. Deregulation, innovation and structural change in Western banking in the 1960s-70s
1537966185
the 1970s saw an explosion of financial innovations, both in instruments and strategies, which altered radically financial structures and financial decisionmaking worldwide. these transformations reversed four decades characterized by the absence of signi...

(4) A DEA financial technology: prior to portfolio analysis

Figure 10: List of research publications related to FinTech using the search point tool, and how the results change towards the topics pointed to

can be seen in figure 10, where the list of research papers that are related to the FinTech concept was shown. On the right, we see that the content was clustered into four main clusters, highlighting innovation, technology, and financial technology as the main subcategories. By moving the cursor to any of the clusters, the results be will be reordered to focus on the contents that fall under that subcategory, as seen in figure 11



>

fintech | Research Publications | Search Topics

(1) Financial markets and organizational technologies : system architectures, practices and risks in the era of deregulation
647913044
list of figures list of tables preface notes on contributors introduction: financial deregulation and technological change a-a.kyrtsis the eurodollar revolution in financial technology: deregulation, innovation and structural change in western banking s.b...

(3) The Eurodollar revolution in financial technology. Deregulation, innovation and structural change in Western banking in the 1960s-70s
1537966185
the 1970s saw an explosion of financial innovations, both in instruments and strategies, which altered radically financial structures and financial decisionmaking worldwide. these transformations reversed four decades characterized by the absence of signi...

(14) Implementation of real-time settlement for banks using decentralised ledger technology: policy and legal implications
2492011742
a wave of innovation is occurring in financial technology, affecting products and services offered to consumers and businesses as well as financial market infrastructures such as payment and settlement systems. these innovations taken together have the po...

(13) FinTech is gaining traction and young, high-income

Figure 11: List of research publications related to FinTech, focused on the innovation subcategory

4 Discussion

In this chapter, we will explore the insights that can be provided by the system by investigating two different use cases. Then we will move onto comparing different paths of evaluation. Finally, a list of challenges and problems will be discussed and solutions and plans for future work will be provided.

4.1 Monitoring the status of a concept: Tensorflow as an example

The main usage of the system is to be able to track a particular concept, whether it is tracking its evolution across stages, comparing it with similar concepts in different stages or data sources, or investigate the documents that relate to it. In this section, an example use case of how that can be achieved is shown through the concept Tensorflow. At first, to get a big picture about the status of Tensorflow, the topic evolution tool will be used. We observe in figure 12 the evolution of the Tensorflow as a concept across stages. Its inspection is marked by a news peak, which corresponded to the event of Google releasing Tensorflow to the public, as it was being developed and used in-house until that date. The publishing of its repository sparked a spike in open source repositories as the framework was published at Github, people started forking it or using it for their purposes, citing Tensorflow in the process. The spark in open source development was soon followed by many research papers being published that use the tool for their experiments, and later patents of systems that use Tensorflow. Meanwhile, the industry started picking up on the trends by requiring job applicants to have it as a skill. The graph was generated using both normalization techniques explained in the methodology, with smoothing applied to the average per year. Based on the evolution score of each data source in the last year, Tensorflow can be considered as a mature framework with active use in industry and research, as indicated in the high value of industry and research.

To put things into perspective, there is a need to compare Tensorflow with other similar concepts, i.e. other ML framework. In figure 13, we compare the amount of research related between TensorFlow, PyTorch, and mxnet as a ratio to the data volume, where the volume of papers of Tensorflow surpasses that of pytorch and mxnet.

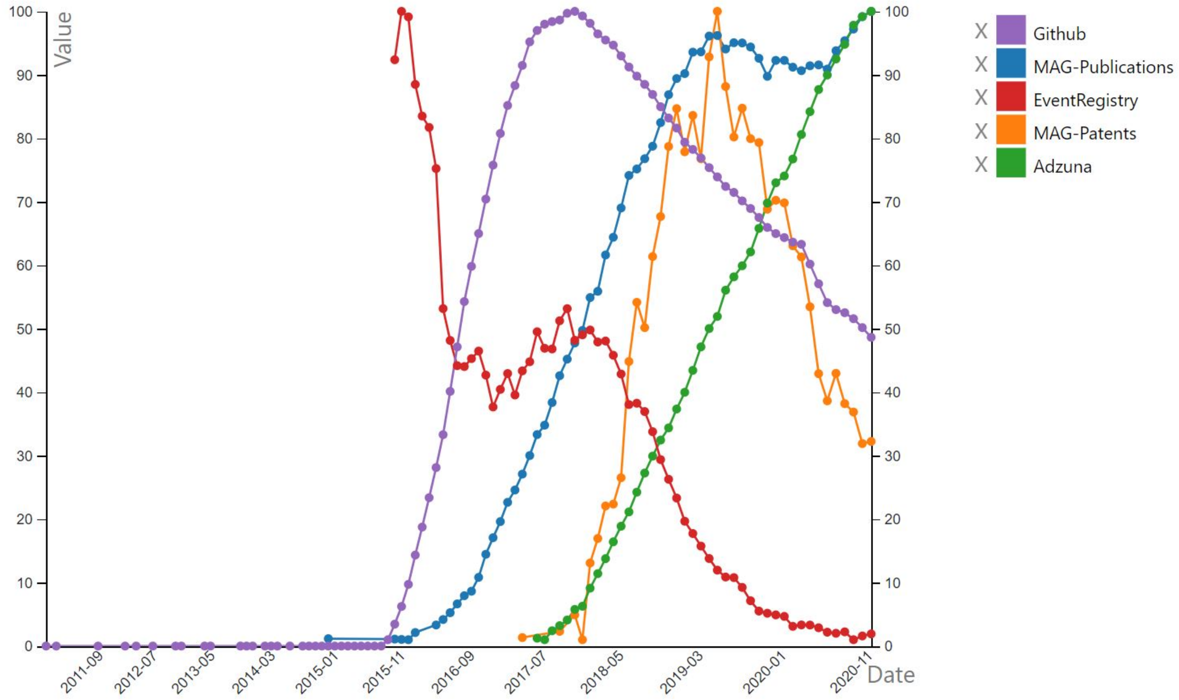


Figure 12: Evolution chart of Tensorflow concept across data sources

However, a different perspective emerge when we compare their trendiness or evolution score (i.e. normalization per data volume AND time) in both research and open source code reveals that Pytorch is becoming more trendy and its user base will likely surpass that of Tensorflow. Finally, to dig deeper and explore the documents that were annotated with Tensorflow, we use the document explorer tool. In figure 15 we observe the list of patents that used TensorFlow.

4.2 Comparing different algorithms and model architectures

Another area of interest is to compare the traditional machine learning algorithms and how they compare to neural net architectures. For this experiment, we chose to track the following algorithms/architectures: logistic regression, linear regression, support vector machine, random forest, decision tree, recurrent neural networks, convolutional neural network, and LSTM. We see in figure 16, that traditional algorithms used to be dominant until the last decade, where neural network architectures start gaining traction, mainly due to the rise of enough hardware and data to support them. A less obvious insight is that the rise of logistic regression surpassing all others. One might argue that this is due to the usage of logistic regression as a baseline. However, given the fact that the text used is for the most part titles and abstracts only, that reason

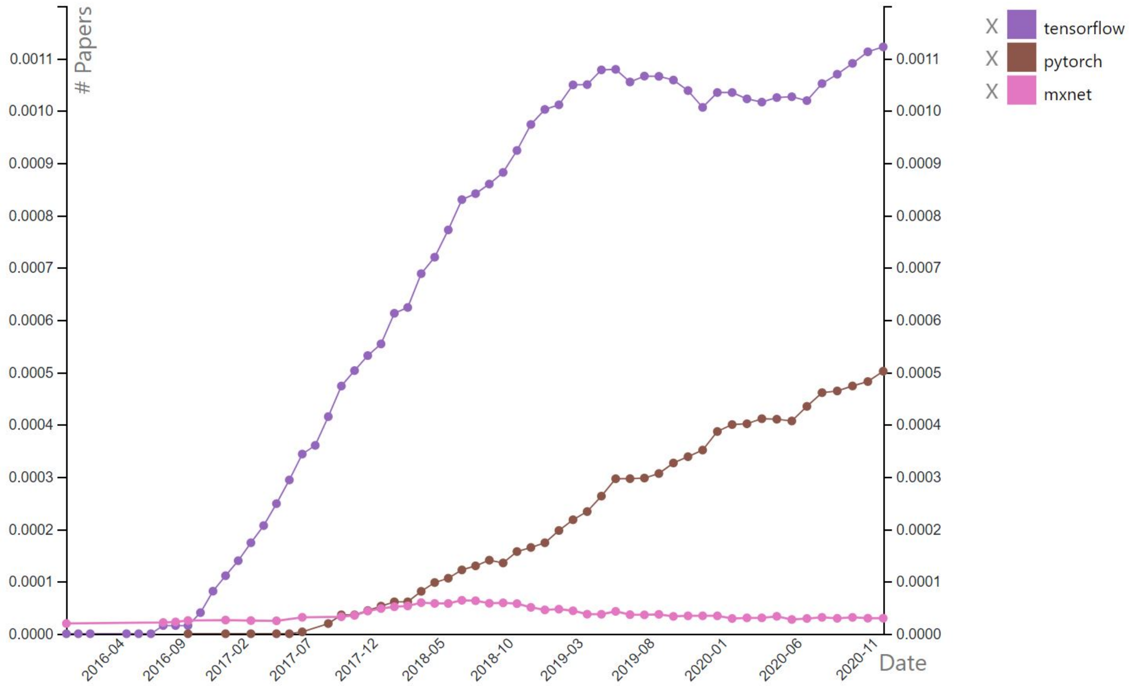


Figure 13: Comparison between Tensorflow, Pytorch, and MXNet, in research as a ratio to data volume

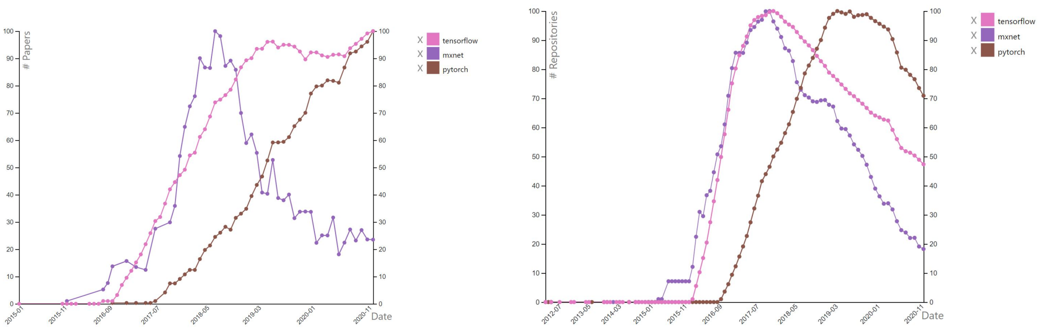
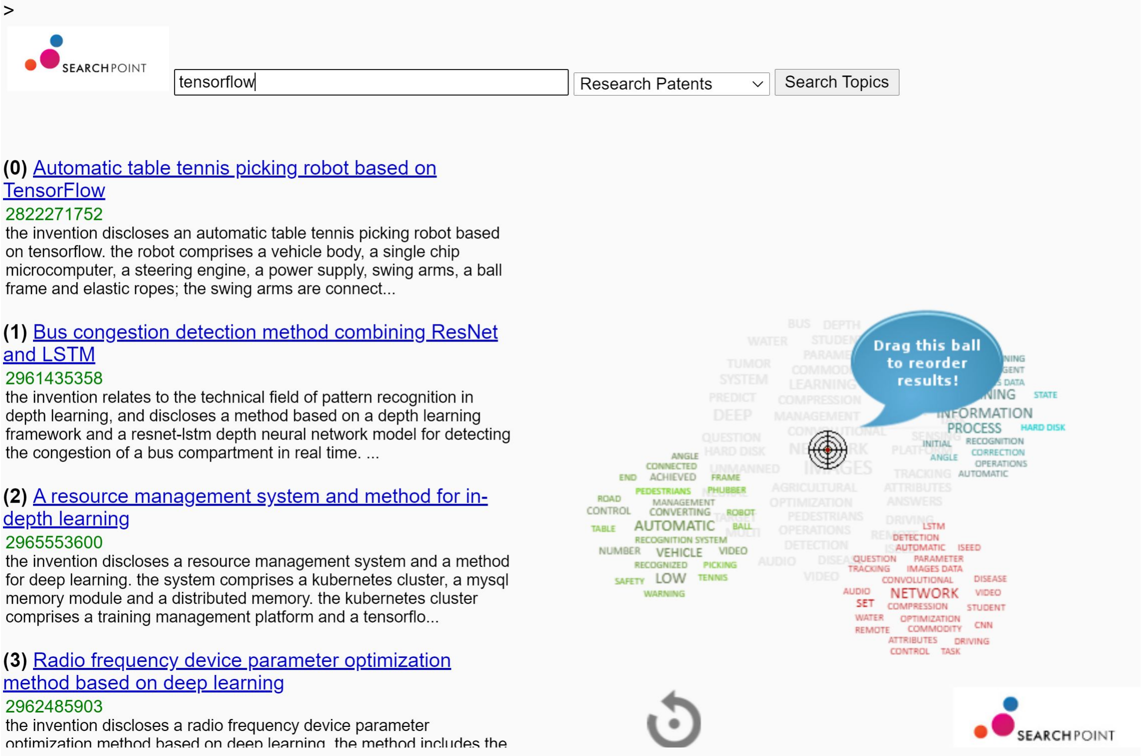


Figure 14: Comparison between evolution score of Tensorflow, Pytorch, and MXNet in research and open source code



(0) Automatic table tennis picking robot based on TensorFlow
2822271752
the invention discloses an automatic table tennis picking robot based on tensorflow. the robot comprises a vehicle body, a single chip microcomputer, a steering engine, a power supply, swing arms, a ball frame and elastic ropes; the swing arms are connect...

(1) Bus congestion detection method combining ResNet and LSTM
2961435358
the invention relates to the technical field of pattern recognition in depth learning, and discloses a method based on a depth learning framework and a resnet-lstm depth neural network model for detecting the congestion of a bus compartment in real time. ...

(2) A resource management system and method for in-depth learning
2965553600
the invention discloses a resource management system and a method for deep learning. the system comprises a kubernetes cluster, a mysql memory module and a distributed memory. the kubernetes cluster comprises a training management platform and a tensorflo...

(3) Radio frequency device parameter optimization method based on deep learning
2962485903
the invention discloses a radio frequency device parameter optimization method based on deep learning. the method includes the

Figure 15: Example patents that are tagged with TensorFlow concept

is of little effect (as this information is usually stated in methodology or results). By using the annotation feature, we can see that most of those papers come from social sciences, which makes sense as they are practitioners of AI that apply ML algorithms in their research domain rather than developing new algorithms or architectures. By exploring through news coverage, we see a different picture, we observe the dominance of convolutional and recurrent neural networks, as well as decision trees, as seen in 17. The dominance of CNNs is due to their usages in autonomous vehicles, social media apps (face filters), as well as voice assistants. All of these usages have direct effects on the public and therefore have a larger news coverage. As for the decision tree concept, it is mainly due to its usage in conversational bots. The fact that those algorithms are widely used today across different domains hints that these algorithms reach a level of maturity where they became the mainstream, as opposed to the hot new thing as in some neural network architectures.

Different stages sometimes use different notations

One particular problem that might arise is that the concept or technology is usually referred to it by its abbreviation. Several notable examples include RNN, for recurrent neural networks, and CNN for convolutional neural networks, LSTM, BERT, etc. While usually in research authors tends to use the full name (at least the first time it is

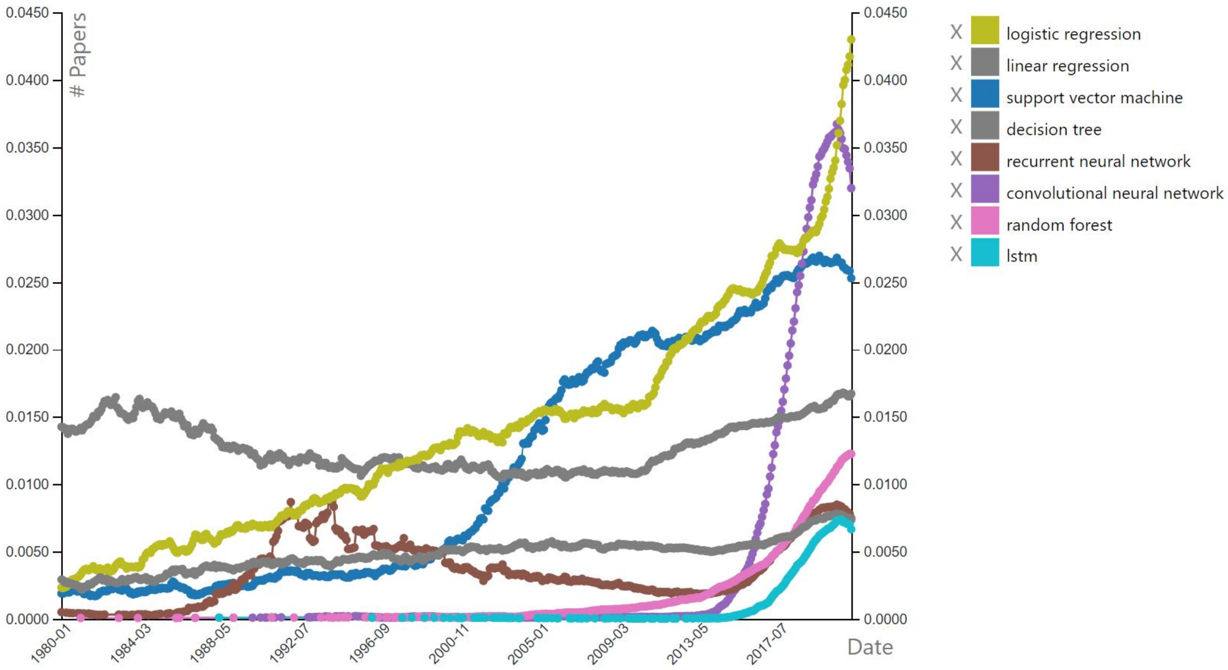


Figure 16: Comparison of different algorithms and model architecture in research, as a ratio to the total volume

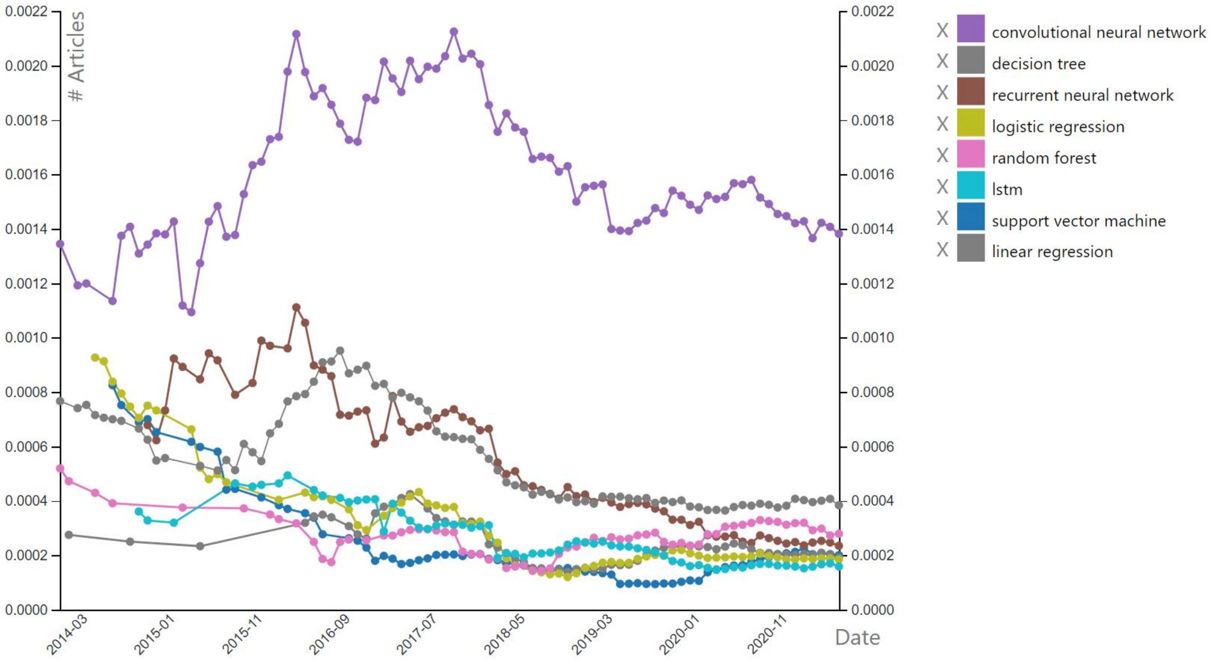


Figure 17: Comparison of different algorithms and model architecture in news, as a ratio to the total volume

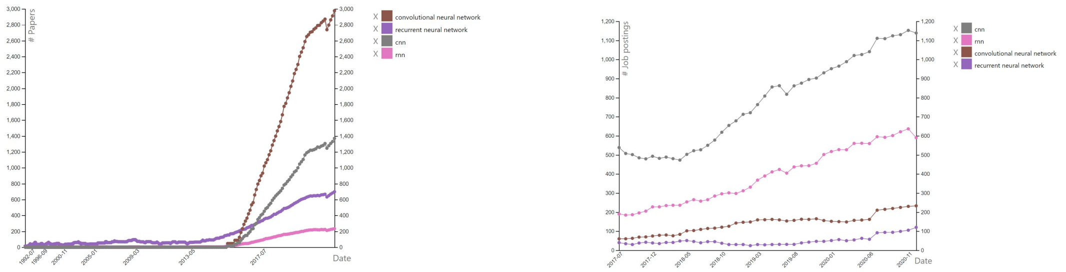


Figure 18: Comparison between Convolutional and recurrent neural network and their abbreviations in research vs job posting

mentioned) job postings might and open source code are usually less rigorous and might use the abbreviation only. An example of this is shown by comparing the research time series vs the job posting time series of convolutional neural networks and recurrent neural networks. One way of solving it requires finding heuristics to group the concepts with their abbreviations by checking their co-occurrences. However, grouping them might result in conflicts of concepts, as abbreviations have different meanings depending on the context. One notable example is CNN which refers to the model architecture as well as the news agency. Therefore, there is a need to eliminating text match methods (and depend on semantics tagging only, instead of the current hybrid method), before applying such grouping.

4.3 Evolution behaviours

Based on the previous two examples, different behaviors have been observed when it comes to the evolution graph. In the case with algorithms and model architecture, the concept inception was from research, which sparked patents, then the news and open source collaboration, and then finally a surge in labor market demand for that skill. This is the typical path that was hypothesized where things start from academia and end with industry. However, a different pattern has been observed in tools like Tensorflow and Pytorch. These frameworks were both developed in-house by a large company, then they were opened to the public with good PR support hence the spike in news coverage. The effect of this behavior made the evolution phases more or less in reverse, and the focus on the practical usage (by industry and practitioners) was as strong and sometimes stronger than research. Of course, after some time more research and patents will emerge but that might be on using the tool or framework more than developing it since the development phase was done in private in earlier stages. This type of scenario occurs mainly in tools developed by companies and then open-sourced to the public. Similar behaviors might occur when comparing other tools like Angular vs React vs Vue, albeit with the data focused on web development rather than artificial intelligence.

Given the fact that two different evolution scenarios have been identified, one might argue that more scenarios might exists, especially when adding other dimension through other data sources, and that the inception origin might be one of the main factor in determining the path of evolution.

4.4 Main challenges and future work

Regarding taxonomy concepts, several points need to be addressed. First, regarding lemmatization, if the word doesn't exists in the wordnet dictionary, which would happen with technical terms, the word will be left as it is, resulting in two different but matching terms, like blockchain and blockchains. One way this could be mitigated is if an additional merging step would be created so that if some words have the same prefix with a suffix that is not part of the word (as in blockchain vs blockchain), which could be done either using fixed rules (which is manageable considering that we only deal with English concepts) or a better way is to depend on the mapping of Wikipedia concept to their page ids, which is currently not used (the only check performed currently is for the existence of a page with that title), which is something that could be investigated in future work. Another point regarding the Wikipedia filter is the inherent use of it and the assumption that all possible relevant concepts exist, which would be the case for early technical concepts. However, as we're more interested in concepts that span across stages, such effect is minimal. Finally, even after all these steps, some concepts are just abstract that does not correspond to technology or innovation. Therefore the current system concepts should be taken with the knowledge that it focuses more on recall rather than precision and that the system is more of an exploratory tool where some investigative work has to be done by the user of the system.
Regarding taxonomy building and grouping concepts into a hierarchy, the current method has been inspired by the way it was done for MAG fields of study. However, a couple of differences in the approach made these results are not as good, with the main one being the concept discovery process that used an in-house knowledge base that uses Wikipedia entities [14], which is different from the one used here, as the goal was to identify innovation concepts from literature text. Another point is that they fixed the first two levels which are usually the most important, which is something that could be tested with the current approach. Another strategy for taxonomy building and identifying topics is to use the Wikipedia categories instead, since each concept corresponds to a Wikipedia page, and hence has a list of categories, which is another point for future work.

Despite the fact the current dashboard tracks concepts against simple metrics only, the system with its current state is equipped to answer many other interesting questions. For example, using the location dimension, a comparison among countries can be drawn, for example, we can observe the most prominent players in each technology, and how their impact change in time, what are the research communities in the area of natural language processing and which institutions are collaborating, etc, etc. Some of those questions are already answered in some of my previous work with the OECD AI observatory, however, this system enables us the ability to delve into specific areas, and technologies in AI and observe them ACROSS datasets rather than focusing on AI in general in each data source separately.

As it has been stated above, the current status of the system has more of an exploratory nature, which only explains the past and shows some possible signals for the future. However, one key element that the system is lacking is modeling, i.e. to build upon the data and train models that would help detect and predict the emerging tools on its own rather than depending on the user investigation. This would require having a clear definition of how emerging and innovative are calculated, then decide on different phases and finally build a model to classify against those phases. Furthermore, in the cycle of innovation, many sources were not used in this thesis, like public funding and private investments and policy regulations. Using such resources would improve the overall quality and usefulness of the system.

5 Conclusion

In summary, the paper present a system for identifying a taxonomy of technologies, tools, and concepts from literature as well as annotating and tracking the taxonomy concepts against multiple data sources. The system is equipped with an analytical dashboard with interactive visualization that enables tracking and exploring the results. The system has been built as a proof of concept of how technologies can be observed and tracked across stages of development using text data. However, the system can be extended with additional components out of the box using a module that is based on the common representation document.

From the results, two different behaviors of evolution have been identified, which are research-driven innovation and company-driven innovation, while each exhibited different behavior, those that fall under the same type showed a similar path. By studying the latency of innovation propagation through stages in different scenarios, one might be able to predict possible paths based on inception conditions. This might require considering scenarios based on areas of industry, as each industry exhibits different behaviors and dynamics.

6 Povzetek naloge v slovenskem jeziku

Inovacije in tehnologije so ključi, ki državam in civilizacijam omogočajo izboljšanje in napredek. V sodobnem svetu je mogoče te inovacijske koncepte izslediti in spremljati iz več virov, ki so običajno na voljo na internetu. Z nadziranjem teh tehnologij se lahko raziskovalci pravilno odločijo, katera področja se bodo lotili, odločevalci in investitorji pa lahko obvestijo v katere tehnologije in podjetja vlagati. V te diplomi predstavljamo sistem, ki lahko iz digitalne literature opredeli pomembne koncepte, orodja in tehnologije ter spremlja kako se spreminja njihov razvoj na različnih stopnjah. Uporabljamo metode iz področja umetne inteligence in sledimo konceptom v patentih, raziskavah, novicah, računalniški kodi in na trgu dela. Glavna hipoteza naloge je, da inovacije in tehnologije večinoma izhajajo iz raziskav (ali vsaj obstajajo v raziskavah) in se nato manifestirajo v drugih fazah. S sledenjem razvoju konceptov na vseh teh stopnjah dobimo pregled stanja vsakega tehnološkega in inovacijskega koncepta ter pomagamo napovedati, kaj bo v prihodnosti. Za vire podatkov smo uporabili Microsoft Academic Graph (MAG) za patente in raziskave, system EventRegistry za novice, Github za računalniško kodo in Adzuna za objave delovnih mest. Sistem lahko opišemo v treh glavnih korakih: izvlečemo koncepte iz literature in jih vgradimo v taksonomijo, s temi koncepti označimo korpus dokumentov v vseh virih in z obdelanimi podatki sestavimo analitično in raziskovalno nadzorno ploščo. Najprej za izločitev pojmov uporabimo naslov in besedilo iz raziskovalnih člankov kot začetni korpus besedila. Opredelimo vrsto filtrov za besedo, na primer filtriranje nelatinskih besed z uporabo oblike leme in odpravljamo glagole. Za ustvarjanje stavkov so bili izvlečeni n-grami od dolžine 1 do 3. Za filtriranje stavkov, ki ne ustrezajo smiselnemu konceptu, so bili koncepti preverjeni v bazi znanja Wikipedije in odpravljen je bil vsak koncept, ki ne ustreza obstoječi strani ali konceptu, seznam pa je bil dodatno filtriran do tistih konceptov, ki so bili omenjeni v najmanj 100 dokumentih. Seznam konceptov je bil nato uporabljen za ustvarjanje taksonomije s konceptom podvzetja, ki je postavil koncept A kot nadrejenega za koncept B, če je na seznamu dokumentov, ki vsebujejo koncept B, tudi koncept A. Pogoj je bil sproščen, da se omogoči nekaj negotovosti v zaznavanju koncepta. Po izdelavi taksonomije so bili dokumenti označeni s taksonomijo. Dokument je bil označen s konceptom, če koncept obstaja neposredno v besedilnem atributu ali v eni od tem in konceptov, ki jih ima dokument, označevanje po temah omogoča odkrivanje konceptov iz večjezičnega korpusa, saj so bile teme normalizirane v angleščini. Na podlagi anotacije so bili izračunani trije različni rezultati, od katerih je vsak odvisen od prejšnjega. Prvi rezultat so označeni korpusi v vsakem viru podatkov. Označeni korpusi vsebujejo seznam konceptov za vsak dokument, pa tudi nekaj metapodatkov o dokumentu, kot so ID, naslov, datum objave in povzetek. Iz označenih korpusov so bile izračunane časovne serije, ki omogoča primerjavo obsega podatkov med različnimi koncepti v istem viru podatkov. Iz surovih časovnih vrst lahko izračunamo razvoj ali trendnost teme po virih podatkov, kar je tretji rezultat. Časovne vrste evolucije so bile izračunane z uporabo dveh nivojev normalizacij. Prva je da se sčasoma normalizira z izračunom razmerja med številom dokumentov, označenih s konceptom na določen datum, in virom podatkov glede na celotno količino tega vira podatkov v istem obdobju. Drugi nivo je normalizacija vira podatkov z normalizacijo vsake časovne vrste koncepta v vsakem viru podatkov na določenem območju (npr. [0,100]), kar omogoča primerjava virov. Da bi omogočili učinkovito iskanje po rezultatih, so bili podatki o rezultatih vneseni v bazo podatkov Elasticsearch z orodjem Logstash ETL. Izbira med drugimi možnostmi Elasticsearch je posledica učinkovitega iskanja po besedilu, kar je velik delež rezultatov. Na bazi podatkov je bila zgrajena interaktivna nadzorna plošča za vizualizacijo, ki omogoča vizualno raziskovanje rezultatov, na podlagi knjižnici kot sta D3 in Bootstrap. Nadzorna plošča ponuja analize na različnih ravneh in iz različnih vidikov, kar omogoča poglobljeno raziskovanje in globalno analizo. Glavni deli nadzorne plošče so bili količinski podatki, raziskovalec taksonomije, razvoj tem, časovne vrste virov in raziskovalec korpusov. Sprva grafikoni predstavljajo skupno količino korpusnih podatkov v vsakem viru podatkov. Ti grafikoni pokažejo celotno sliko o količini podatkov, s katerimi se ukvarjamo, da bi pridobili perspektivo pri preučevanju obsega na koncept. Drugo orodje je za raziskovanje taksonomije, ki omogoča grafično krmarjenje po taksonomiji. Uporabnik lahko poišče koncept in si ogleda seznam konceptov, ki spadajo vanj, ter skupno število dokumentov, označenih z njim. Tretji del vsebuje orodje Topic Evolution. Kar omogoča raziskovanje enotnega koncepta na različnih stopnjah razvoja ter uporablja zgoraj navedeno časovno vrsto evolucije. Orodje ponuja različne prilagoditve, kot je filtriranje za dana časovna obdobja, uporabo različne ravne normalizacije, glajenje in uveljavljanje mejnih vrednosti. V drugem delu predstavimo razvoj enega samega koncepta po stopnjah, orodje omogoča primerjavo različnih konceptov z istim virom podatkov, to pa omogoča primerjavo med uspešnostjo podobnih konceptov - kot sta TensorFlow in Pytorch - v vsaki stopnji razvoja. To orodje ponuja podobne možnosti prilagajanja kot orodje za razvoj tem. Za podrobnejše raziskave o označeni vsebini je bilo uporabljeno orodje iskalne točke, ki omogoča iskanje, katerega za raziskovanje besedilnega korpusa uporablja Elasticsearch. Orodje je bilo uporabljeno za iskanje označene vsebine tem za raziskovanje rezultatov na ravni dokumenta. Med preučevanjem taksonomije in njenih konceptov je bilo na primer ugotovljenih več vprašanj, čeprav je bilo ujemanje lematizacije uporabljeno za združevanje pojmov z različnimi oblikami (omrežje, omrežja) npr. blockchain in blockchains. Poleg tega, čeprav je metodologija taksonomije ujela nekaj uporabnih odvisnosti (npr. bert je primer jezikovni model, TensorFlow je primer okvir poglobljenega učenja), so bile nekatere odvisnosti obratne, na primer strojno učenje je primer nevronskega omrežja. Drug način za oblikovanje taksonomije je mogoče uporabiti z uporabo Wikipedia ali wiki kategorij, saj so vsi ustvarjeni koncepti koncepti Wikipedije. Sedanji sistem je raziskovalni in deloma pojasnjevalni, vendar prikazuje le trenutno stanje in je odvisen od posameznega uporabnika. Ne podaja natančne vrednosti za meritve zrelosti ali samodejno primerjavo med podobnimi koncepti. Zato ne ponuja neposrednih vpogledov, niti napovedi o prihodnjem stanju. Poleg izboljšanja taksonomske strukture so te točke izboljšave, ki se jih bomo lotili v prihodnem delu.

7 Bibliography

[1] Janez Brank, Gregor Leban, and Marko Grobelnik. 2017. Annotating documents with relevant wikipedia concepts. `https://ailab.ijs.si/dunja/SiKDD2017/Papers/Brank_Wikifier.pdf`. (2017).

[2] [n. d.] Databench: big data benchmarking. [Online; accessed August-2021]. (). `https://www.databench.eu/`.

[3] Qiang Fan, Huaimin Wang, Gang Yin, and T. Wang. 2015. Ranking open source software based on crowd wisdom. *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 966–972.

[4] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

[5] Ashkan Fredström, Joakim Wincent, David Sjödin, Pejvak Oghazi, and Vinit Parida. 2021. Tracking innovation diffusion : ai analysis of large-scale patent data towards an agenda for further research. *Technological forecasting and social change*, 165, 120524. Validerad;2021;Nivå 2;2021-01-11 (alebob);Finansiär: Evald and Hilda Nissi Foundation. DOI: `10.1016/j.techfore.2020.120524`.

[6] [n. d.] Github: where the world builds software. [Online; accessed August-2021]. (). `https://github.com`.

[7] Gartner Inc. 2012. Hype cycles. (2012). `http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp`.

[8] [n. d.] Job search - find every job, everywhere with adzuna. [Online; accessed August-2021]. (). `https://www.adzuna.com/`.

[9] Pantelis Koutroumpis, Aija Leiponen, and Llewellyn Thomas. 2021. Digital instruments as invention machines. *Communications of the ACM*, 64, (January 2021). DOI: `10.1145/3377476`.

[10] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web* (WWW '14 Companion). Association for Computing Machinery, Seoul, Korea, 107–110. ISBN: 9781450327459. DOI: `10.1145/2567948.2577024`. `https://doi.org/10.1145/2567948.2577024`.

[11] [n. d.] Logstash: collect, parse, transform logs. [Online; accessed August-2021]. (). https://elastic.co/logstash.

[12] Microsoft Research. [n. d.] Microsoft Academic Graph data schema. https://docs.microsoft.com/en-us/academic-services/graph/reference-data-schema. [Online; accessed August-2021]. ().

[13] Mark Sanderson and William Croft. 1999. Deriving concept hierarchies from text. In (January 1999).

[14] Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A web-scale system for scientific knowledge exploration. In *Proceedings of ACL 2018, System Demonstrations*. Association for Computational Linguistics, Melbourne, Australia, (July 2018), 87–92. DOI: 10.18653/v1/P18-4015. https://aclanthology.org/P18-4015.

[15] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, B. Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. *Proceedings of the 24th International Conference on World Wide Web*.

[16] [n. d.] The oecd aritficial intelligence ai observatory - oecd.ai. [Online; accessed August-2021]. (). https://oecd.ai.

Appendices

A Example Documents of Data Sources

A.1 Example of a MAG Research Document

```
{
    "ID":2963403868,
    "docType":"conference",
    "title":"Attention is All you Need",
    "abstract":"The dominant sequence transduction models
       are based on complex recurrent orconvolutional
       neural networks in an encoder and decoder
       configuration. The best performing such models also
       connect the encoder and decoder through an
       attentionm echanisms. We propose a novel, simple
       network architecture based solely onan attention
       mechanism, dispensing with recurrence and
       convolutions entirely.Experiments on two machine
       translation tasks show these models to be superiorin
        quality while being more parallelizable and
       requiring significantly less timeto train. Our
       single model with 165 million parameters, achieves
       27.5 BLEU onEnglish-to-German translation, improving
        over the existing best ensemble result by over 1
       BLEU. On English-to-French translation, we
       outperform the previoussingle state-of-the-art with
       model by 0.7 BLEU, achieving a BLEU score of 41.1.",
    "year":"2017",
    "citation":23,433
}
```

A.2 Example of an Event Registry News Document

```
{
    "uri": "6628233443",
    "lang": "eng",
    "dateTime": "2021-06-30T23:52:00Z",
    "dataType": "news",
    "url": "https://www.wallstreet-online.de/nachricht/14104234-brainchip-takes-a-look-at-what-ml-and-ai-achieve-with-arm-fellow-jem-davies",
    "title": "BrainChip Takes a Look at what ML and AI Can Achieve With Arm Fellow Jem Davies",
    "body": "An experienced technologist and architect, Davies has run the ML group within Arm for the past 4 years, driving the technology inflection across the company and guiding the direction of the company as it supports evolving workloads. In the eighth episode of the series, Davies and Telson discuss the process that technology undergoes before implementation at the consumer level, how AI on devices is helping to evolve the direction of the technology, and how these transformations can ultimately benefit society....", "source": {
        "uri": "wallstreet-online.de",
        "dataType": "news",
        "title": "wallstreet:online", "description":
        "location": {
            "type": "place",
            "label": {"eng": "Berlin"}
        },
    },
    "concepts": [
        {
            "uri":"http://en.wikipedia.org/wiki/Machine_learning",
            "type": "wiki",
            "score": 5,
            "label": {"eng": "Machine learning"}
        },
        {
            "uri": "http://en.wikipedia.org/wiki/Artificial_intelligence",
            "type": "wiki",
            "score": 5,
            "label": {"eng": "Artificial intelligence"},
        }
    ],
    "eventUri": "eng-6899585",
    "sentiment": 0.5921568627450979,
    "relevance": 100
}
```

A.3 Example of an Adzuna Job Posting Document

```
{
    "id": 1590424001,
    "cat_id": 2,
    "loc_id": 191269,
    "date": "2020-07-01 19:28:03",
    "text": "Web UI Software Developer. WEB UI SOFTWARE DEVELOPER RESPONSIBILITIES  Create web interfaces, using standard HTML/CSS practices, incorporating data from various Back End databases and distributed services.  Create well-designed, tested code using best practices for website  development, including mobile.  Interact with stakeholders to work quickly and effectively to complete  small edits requested by users, develop plans for larger projects and suggest new solutions to improve existing websites.  Develop and maintain Back Office services including: batch processing,  clearing, allocations; interacting with various 3rd party services. QUALIFICATIONS  Bachelor's degree in a technical field or equivalent work experience  Experience with web UI libraries and charting libraries(Bootstrap,
        D3, etc.)  Working knowledge of web Servers like
        Apache.  Working knowledge of Dockers and Containers
        .  Strong Python programming skills.  Comfortable
        with UNIX environment/basic Unix commands.  Basic
        database knowledge (Sybase, MySQL, InfluxDB).  Solid
         experience with PHP, HTML, CSS, Javascript.
        Ability to work under pressure within a dynamic
        trading environment.  Attention to detail for
        problem solving and code robustness. If this is an
        opportunity that you're interested in please email
        your resume to: (see below)",
    "salary": "USD 80000.00 to 125000.00 per annum",
    "curr": "USD",
    "company": "Request Technology   Kyle Honn",
    "country": "US",
    "Language": "en",
    "wikifierConcepts": [
        {"concept": "Unix","pageRank": 0.1378616407144},
        {"concept": "MySQL","pageRank": 0.12597253668128},
        {"concept": "HTML","pageRank":
           0.12535407381015098},
        {"concept": "InfluxDB","pageRank":
           0.107068096694765},
        {"concept": "PHP","pageRank": 0.10671901410967},
        {"concept": "Python (programming language)","
           pageRank": 0.0933367462579915},
        {"concept": "Batch processing","pageRank":
           0.0871312901153647},
        {"concept": "Bachelor's degree","pageRank":
           0.0866039527894413},
        {"concept": "Sybase","pageRank":
           0.0682187342471607},
        {"concept": "Cascading Style Sheets","pageRank":
           0.0617339145729283
        }
    ]
};
```

A.4 Example of a GitHub repository Document

```
{
    "repository": "tensorflow/tensorflow",
    "contents":
    [
    {"name": "README.md",
    "path": "README.md",
    "sha": "63d85ce2df4a9ae0a7303a627f28561410d0ddf1",
    "size": 19859,
     "url": "https://api.github.com/repos/tensorflow/tensorflow/contents/README.md?ref=master", "download_url": "https://raw.githubusercontent.com/tensorflow/tensorflow/master/README.md"}},
     {"name": "RELEASE.md",
     "path": "RELEASE.md",
    "sha": "d73bf4ca0462e91839169074cdf0cde0485e333c",
    "size": 316026,
     "url": "https://api.github.com/repos/tensorflow/tensorflow/contents/RELEASE.md?ref=master",
    "download_url": "https://raw.githubusercontent.com/tensorflow/tensorflow/master/RELEASE.md"}}],
    "file_contents":
    {"README.md":
    "[TensorFlow](https://www.tensorflow.org/) is an end-to-end open source platform\nfor machine learning. It has a comprehensive, flexible ecosystem of\n[tools](https://www.tensorflow.org/resources/tools),\n[libraries](https://www.tensorflow.org/resources/libraries-extensions), and\n[community](https://www.tensorflow.org/community) resources that lets\nresearchers push the state-of-the-art in ML and developers easily build and\ndeploy ML-powered applications.\n\nTensorFlow was originally developed by researchers and engineers working on the\nGoogle Brain team within Google's Machine Intelligence Research organization to\nconduct machine learning and deep neural networks research. The system is\
          ngeneral enough to be applicable in a wide variety
          of other domains, as well.\n\nTensorFlow provides
          stable [Python](https://www.tensorflow.org/api_docs/
          python)\nand [C++](https://www.tensorflow.org/
          api_docs/cc) APIs,...",...},
    "created at": "2015-11-07T01:19:20Z"
}
```