UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Tõnis Hendrik Hlebnikov

Towards a Knowledge Graph of Internet Memes

Bachelor's Thesis (9 ECTS)

Supervisor(s): Riccardo Tommasini, PhD

Tartu 2021

Towards a Knowledge Graph of Internet Memes

Abstract:
In this thesis, the notion of considering internet memes as rich units of information is presented. Memes and internet memes are defined and discussed. Their research merits are indicated and the research questions are framed using the Macro-Meso-Micro framework. The micro level leads to the definition of requirements to enable such research. This results in the proposal of creating a pipeline to facilitate the construction of a meme knowledge graph. The foundation for such a pipeline is devised, and its initial components are created. The description of the pipeline provides an assessment of requirement compliance. Ultimately, the created components facilitate the collection of a rich internet meme dataset.

Keywords:
Internet memes, memes, knowledge graph, web scraping

CERCS: P175 - Informatics, systems theory

Internetimeemide teadmiste graafi suunas

Lühikokkuvõte:
Selles lõputöös esitletakse ettekujutust käsitleda internetimeeme kui rikkalikku informatsiooni ühikuid. Meemid ja internetimeemid defineeritakse, ning on aluseks arutelule. Täpsustatakse nende väärtus teaduse jaoks ning uurimisküsimused püstitatakse Makro-Meso-Mikro raamistikuga. Mikro tase võimaldab püstitada nõuded sellise teaduse võimaldamise jaoks. Sellest tulenevalt pakutakse välja internetimeemide teadmiste graafi konstrueerimist võimaldava konveieri loomine. Konveieri alustalad töödatakse välja, ning esmased komponendid luuakse. Konveieri kirjeldamise käigus hinnatakse nõuetele vastavust. Lõpetuseks võimaldavad loodud komponendid kokku koguda rikkalik internetimeemide andmestik.

Võtmesõnad:
Internetimeemid, meemid, teadmiste graaf, veebi kraapimine

CERCS: P175 - Informaatika, süsteemiteooria

Contents

1 Introduction

From neural networks to cognitive robotics, computer sciences demonstrates a long history of observing and learning from natural sciences, nurturing simple intuition into fruitful applications. This is a symbiotic relationship, as natural sciences too have greatly benefitted, with a prominent example being genomic computing. Gene sequencing, a massive feat of both understanding and engineering, brought ample breakthroughs to biology. But what if a similar symbiosis was achievable with social sciences?

With the largely qualitative nature of social sciences, the opportunities for computer science to *measure something* have been few and far between. And this would likely remain so, discounting the sudden discovery of a method to quantify thoughts, concepts, and emotions. But still, *what if*?

With perhaps a similar question in mind, the concept of *memes* was derived. They would be the cultural equivalent of genes, both the singular units and the building blocks of larger things. The dawn of the digital age brought with it a new culture, one that spans the entire globe - *internet culture*. This more expansive culture would require equally expansive memes, something that would spread faster and further, something that persists within the ephemeral internet. This something would be *internet memes*, abstract concepts given (slightly more) tangible form.

The proposal, then, is to quantify these internet memes and utilize them as research vessels, hoping to follow the gene-meme analogy to similar breakthroughs in social sciences. A more substantial discussion of this notion is provided within section 3. This discussion culminates with indicating the requirements for such research and noting their potential fulfillment by a novel technology known as knowledge graphs. Following the discussion, the first steps towards systematic meme analysis are taken in section 4.

2 Background

This section contains the necessary knowledge and definitions to understand the content of the thesis, alongside brief additional discussion for context.

2.1 Memes

Memes, as originally defined by Richard Dawkins in 1976, are a "unit of cultural transmission, or a unit of imitation" [11]. In different terms, memes are concepts and customs, patterns and behaviors, attitudes and habits - the building blocks of culture, science, religion, and even society itself. It is everything that is transmitted between individuals by means of imitation, in the loosest definition. Dawkins elaborates upon this further by drawing an analogy to genes - both being, in a sense, self-replicating entities that propagate themselves through people and, thus, through time.

Understandably, this definition has been a source of much controversy. Susan Blackmoore, in her work "The Meme Machine" [4], identifies (and argues against) three problems with memes, of which two are immediately relevant:

1. The unit of a meme cannot be specified
2. The mechanism for copying and storing memes is unknown.

This provides clear limitations for quantitative study and is the basis of much criticism.

2.1.1 Internet memes

Patrick Davison, in his 2012 essay *The Language of Internet Memes*, proposes the following definition:

Definition 2.1 (Internet meme)**.** a piece of culture, typically a joke, which gains influence through online transmission [10].

It is important to reinforce that humor is not a requirement, as internet memes have also been observed as the vehicle for political propaganda [24], hate speech [18] and traumatic confessions [29]. The significant part of this definition is *online transmission*, which requires the meme to be encoded into an internet viable medium. Whether this medium is an image, text, or video, it is readable data.

This additional property now allows us to discern discrete units and observe the mechanisms of replication and storage in action, at least within the online environment - opportunity for quantitative analysis.

2.1.2 Image macros

A final level of granularity is required to reason about internet memes in a sufficiently precise manner. Image macros are perhaps the most representative subgenre of internet memes and are often regarded as the epitomical internet meme. The singular required property is for the meme to be encoded into an image (file). Typically, the resulting image consists of:

1. A background image that is chosen as such that it is immediately recognizable by the intended audience and provides them context
2. Superimposed text as a caption, containing the message and sometimes additional contextual information.

A singular format does not exist; however, most share the property of being multimodal constructions of text and image [9, 31].

2.2 Knowledge Graph

Semantic Web, envisioned by Tim Berners-Lee in 2001 [3], imagines a web of data where resources are interconnected in a meaningful and machine reasonable way. Major milestones of this endeavour are massive datasets of semantically linked information such as DBpedia[1], Wikidata[2] and YAGO[3]. With the advent of the Google Knowledge Graph [27], these datasets have now been rebranded in fashion - as knowledge graphs. A proposed definition for knowledge graphs follows:

Definition 2.2 (Knowledge graph)**.** A graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities. [14]

This definition alludes to the capabilities of knowledge graphs. Through the establishment of ontologies and rules, knowledge graphs can provide insight through logical inference and deduction, creating meaningful information where none was explicitly provided [14]. A final important feature is that knowledge graphs, when following a universal entity naming schema, are interoperable, enabling the extension of one through another [14].

[1] https://www.dbpedia.org/
[2] https://www.wikidata.org/
[3] https://yago-knowledge.org/

3 Problem statement

The purpose of this section is to distinguish memes as units rich in information and indicate the precedence and challenges of their systematic analysis. Considering the breadth of this topic, the Macro-Meso-Micro framework [19] is appropriated and utilized to assist reasoning. The framework separates research concerns into three levels of magnitude and reveals the steps required to achieve the broader goal.

Meme analytics would be a significant undertaking, requiring solutions to multiple sub-problems. Utilizing the Macro-Meso-Micro framework, the problems can be divided as follows:

Macro: How to enable *meme analytics*?

Meso: How to enable *meme sequencing*?

Micro: How to construct a *knowledge graph of internet memes*?

The following subsections provide further elaboration for each level, with the final one outlining the requirements for the first step.

3.1 Meme analytics

Considering the broad definition and abstract nature of memes, it is difficult to undertake a broad, all-encompassing approach. Therefore, a better-defined approach would be through image macros, with the intention of setting a foundation that could eventually include other types of memes. Image macros, even as the least nebulous category of memes, prove demanding for systematic analysis.

Image macros are a form of human communication comparable to natural language in their unstructured nature, rich in information that is effortlessly comprehensible for humans and notoriously unintelligible for computers. This is further complicated by multimodality [31] and commonly even incompleteness, often requiring external context for the synthesis of meaning. Previous work has shown some success in extracting semantic entities from image macros [12], and this is the foundation to the extraction of meaning.

Beyond the dissection of singular memes lies the understanding of the meme pool - the diffusion of memes and the notion of virality. Memes spread through the medium of humans. On the internet, this is accomplished by sharing and re-posting memes on social networks. It is, as of yet, unknown why memes are circulated and what are the required factors in order for a meme to become viral. Work has been done in these areas, with some factors being identified [16] and used to measure [7] or even successfully predict the virality [20, 1] of memes within the meme pools of some social networks. However, a site-agnostic approach would be required in order to validate the universality of these findings.

The result of circulating a message is exposure. This is the basis of advertisement and propaganda, both active fields of research. Success of advertisement is often measured in sales, a clear and observable metric that reflects the purpose of the circulation. Thus, considering the result orientation of advertising and that internet memes have been noted as an emerging mode of advertisement [8], the potential influence of internet memes is undeniable. Even with recognition and parallels drawn to leaflet propaganda [24], the research activity has not extended to impact measurement for internet memes as a whole.

In order to enable such research, an environment must be developed. This environment would facilitate relevant queries, providing contextually meaningful information about image macros and, ultimately, memes. However, in order to discuss such an environment, it is necessary to first discuss its population.

3.2 Meme sequencing

The sequencing of the complete human genome was first publicly proposed in 1984 with the primary goal of providing insight into the mechanism of cancer [13]. The proposal was met with both praise and controversy, with one concern being technical feasibility [15]. With advancements in computational biology and genetics in the following years [17], the Human Genome Project was deemed feasible and launched in 1990 [6, 2]. The project concluded in 2003 with the successful sequencing of major parts of human DNA, which enabled significant breakthroughs within biology and medicine [15]. The computational challenges of the project can even be deemed as the rise of bioinformatics [21, 28].

Building upon the gene-meme analogy and recognizing the impact of computer science in biology, what if memes could pave the way for a similar breakthrough? Memes have already been recognized as a potential vehicle of research within the domains of digital culture research [26], politics [30], cognitive neuroscience [22], and even information warfare theory [25], all in addition to the discipline of memetics itself. Having additionally been described as "the building blocks of your mind, the programming of your mental 'computer'" [5], the potential research benefit for social sciences is tremendous.

The first step, much like with gene sequencing, would be to identify the components of a meme. With image macros, a potential avenue for this reveals itself through the Semantic Web initiative. Tools have been developed, to enable entity extraction from both text [23] and images[4], and even, to an extent, image macros themselves [12]. However, the complete deconstruction of any meme would, at the very least, require context, understanding of its application, and the composite meaning of the entities contained.

Definition 3.1 (Meme sequencing)**.** The process of determining the entities, their relations, the application domain, and the meaning in the context of a given meme instance.

[4] `https://cloud.google.com/vision`

While complete meme sequencing is currently out of reach, the technologies available provide enough to explicate an environment that will eventually facilitate this information and its processing.

3.3 Requirements analysis

The creation of such an environment will ultimately unfold a perpetual endeavour, with the environment evolving alongside progressing research and accumulating requirements. With both this and the abstract nature of (internet) memes in mind, attempting to predict the concrete requirements of the final environment in hopes of constructing it outright would likely prove unsuccessful. Therefore, a more practical approach would be to instead focus on enabling it, to define strictly the first step and its requirements. This brings us to the preliminary requirements:

ER1: The environment must enable frequent content updates and perform them on an adjustable interval.

ER2: The environment must enable the inclusion and processing of exotic data.

ER3: The environment must enable the addition of novel enrichment and augmentation tools.

ER4: The environment must be scalable, to meet increasing demand in storage and processing.

The objective is to establish the adaptability and malleability of the environment and its contents. The environment should be inherently modular, allowing the integration of new data sources and their appropriate tools regardless of languages or frameworks. This proposes additional requirements for the data within the environment:

DR1: The data model should be flexible and welcome modification.

DR2: The data model should be schema-agnostic, allowing insertion of any data.

DR3: The engine must enable semantic querying, to benefit from inference.

These requirements exclude relational databases and hint at something which would simultaneously be less structured and yet allow for expansive and meaningful querying. The solution to this is the amalgam of graph databases and semantic information: knowledge graphs.

Finally, an initial data source must be chosen. The following requirements are particularly critical for the initial source, yet deem consideration for any future sources as well. The source requirements are as follows:

SR1: The source must provide (extractable) relevant semantic information.

SR2: The data must be plentiful.

SR3: There must be a distinction between instances and templates, with the latter being clearly differentiable.

SR4: The data must be accessible, either through an API or be provided in a format that is homogeneous enough to enable scraping.

SR1 and ***SR3*** provide the platform for disassembling memes into their component entities and relations. ***SR2*** is instrumental to this, as the likelihood of finding shared component entities (relationships) increases with quantity. ***SR4*** assists with a quicker development cycle. The initial source holds additional significance, as it will determine the first challenges for the environment, and therefore its first capabilities.

With the establishment and subsequent consideration of the initial requirements, the proposed solution is the creation of a modular pipeline that would ultimately facilitate the construction of an internet meme knowledge graph. The components of the pipeline should stay loosely coupled and open to individual modification without compromising the pipeline. This would be enabled by containerization and external orchestration.

4 Contribution

This section details the construction of the internet meme knowledge graph pipeline, describes the challenges faced, and outlines its initial architecture. Multiple technologies were utilized and more considered. Brief introductions will be provided to these technologies, along with the reasoning behind their choice. The section concludes with a summary of the collected data.

4.1 Data discovery

With the increasing prevalence of internet memes, multiple sources were to be considered and evaluated regarding the requirements. The available image macro sources come in three distinguishable categories, with each providing distinct advantages and disadvantages:

Social media: Large social content sharing platforms, such as Reddit[5], 4chan[6], Instagram[7], Twitter[8], and Facebook[9].

Generators: Internet meme specific social media platforms that additionally provide the tools to create image macros, such as Imgflip[10].

Encyclopedias: Websites dedicated to providing more substantial information about memes, such as KnowYourMeme[11], The Meme Wikia[12] and Memeing Wiki[13].

The categories and their respective advantages and disadvantages are discussed further in the following subsections. Ideally, the final pipeline would facilitate ingestion of all possible sources, yet for the purpose of its construction, a single source is sufficient.

4.1.1 Social media

Social media platforms are the natural habitat for internet memes and thus provide the largest sample. However, they are seldom restricted to memes, additionally providing a platform for other content. Reddit's primary meme-focused subreddits host a cumulative average of around 10,000 submissions per day[14], and 4chan well above that with 200,000

[5] https://www.reddit.com/
[6] https://4chan.org/
[7] https://www.instagram.com/
[8] https://twitter.com/
[9] https://www.facebook.com/
[10] https://imgflip.com/
[11] https://knowyourmeme.com/
[12] https://meme.fandom.com/
[13] https://en.meming.world/
[14] https://subredditstats.com/

per day[15]. Of the larger and less meme-focused platforms, Twitter averages 500,000,000 submissions per day[16], and recent post count metrics for Facebook and Instagram are unavailable.

It is difficult to obtain metrics for exactly what fraction of the circulated content is made up of image macros, given the large data variety. Furthermore, these platforms seldom provide meaningful information alongside image macro instances, beyond titles or comments. Accessibility requires consideration as well, as for example while Reddit[17] and 4chan[18] provide APIs, the entire history of the platforms' content is unavailable. This is likely due to the data volume and velocity of such platforms. These issues cumulatively render social media platforms unfit for the initial steps, yet significant for future inclusion and analysis.

4.1.2 Generators

Internet meme generation sites are the source of most image macros, as the tools provided do not require particular proficiency with image editing. This is accomplished by providing common templates for users to caption. Necessitating the differentiation between templates and instances, lists of instances per template are often available. Additionally, these sites typically feature a timeline of generated images and sometimes even commenting and voting functionality.

The most notable of these platforms is Imgflip. Instances are available with accompanying tags, instance-template links, and plain-text captions. There are over 1,000,000 captionable templates available, according to Imgflip. Unfortunately, large sets of these are nearly identical without a provided method of discovering such similarities. Additionally, the entire history of the site is unavailable, with the API being restricted to only the latest instances.

4.1.3 Encyclopedias

Efforts have been made to collect and catalogue internet memes, resulting in what are essentially internet meme encyclopedias. These websites strive to explain internet memes, providing background, origins, applications, and, in the case of Meming Wiki, even interpretations of their meaning. Usually a collaborative effort, the information is provided by volunteers as unstructured text. These encyclopedias do not provide APIs, with the exception of the platform's auto-generated one in the case of The Meme Wikia.

The most prominent internet meme encyclopedia is KnowYourMeme. It has both the largest catalogue and typically the most information per meme. Additionally,

[15] `https://4stats.io/`

[16] `https://www.internetlivestats.com/`

[17] Reddit API: `https://www.reddit.com/dev/api/`

[18] 4chan API: `https://github.com/4chan/4chan-API`

KnowYourMeme has created a taxonomy of internet memes, which is represented by parent, child, and sibling relations between the memes. KnowYourMeme therefore meets the requirements ***SR1***-***SR3***, and ***SR4*** sufficiently, with its content presentation being homogeneous enough to enable scraping.

4.2 Pipeline

The pipeline is designed to facilitate the entire process of knowledge graph construction, from data access and ingestion to its processing, storage, and eventually application. It is composed of multiple loosely coupled modules, following the necessity for adaptability outlined in ***ER2*** and ***ER3***. An architectural diagram of the modules currently populating the pipeline is provided in figure 1. These modules, excluding database engines, are primarily written in **Python**[19], a popular high-level programming language, chosen due to familiarity and its rich ecosystem of data engineering related tools.

Data is moved between modules by way of intermediary storage. **MongoDB**[20] was chosen for this, as it is a document-oriented database, which does not enforce a schema. This is essential due to the unpredictable nature of the data, outlined in ***DR2*** and ***ER2***. MongoDB additionally satisfies ***ER4*** via sharding, which distributes the data over multiple instances, while enabling queries through a single endpoint. For the actual knowledge graph, and the satisfaction of ***DR1*** and ***DR3***, a suitable graph database must also be included. This would be **Neo4j**[21], a graph database that enables the storage of semantic information through a flexible and optional schema consisting of constraints and indexes. This allows for the encoding of ontologies and semantic rules, and logical inference thereby, through the use of built-in tools. With the addition of sharding, similarly to MongoDB, Neo4j provides a more than suitable platform for knowledge graph implementation.

The aforementioned modularity and the requirement for scalability, outlined in ***ER4***, are further addressed by the containerization of the disparate modules. Containerization is achieved with the utilization of **Docker**[22], which is a tool that enables the separation of applications from their environments, by introducing an intermediary virtual environment. This virtual environment serves as a *container* for the application, which allows the bundling of dependencies and results in agnosticism towards the host system, as long as it is running Docker. The additional benefit of this is that clones may be deployed, allowing for horizontal scaling.

Coordination, scheduling, and management of these modules, or the *orchestration*, is performed using **Apache Airflow**[23], a workflow management platform. Airflow requires

[19] `https://www.python.org/`
[20] `https://www.mongodb.com/`
[21] `https://neo4j.com/`
[22] `https://www.docker.com/`
[23] `https://airflow.apache.org/`

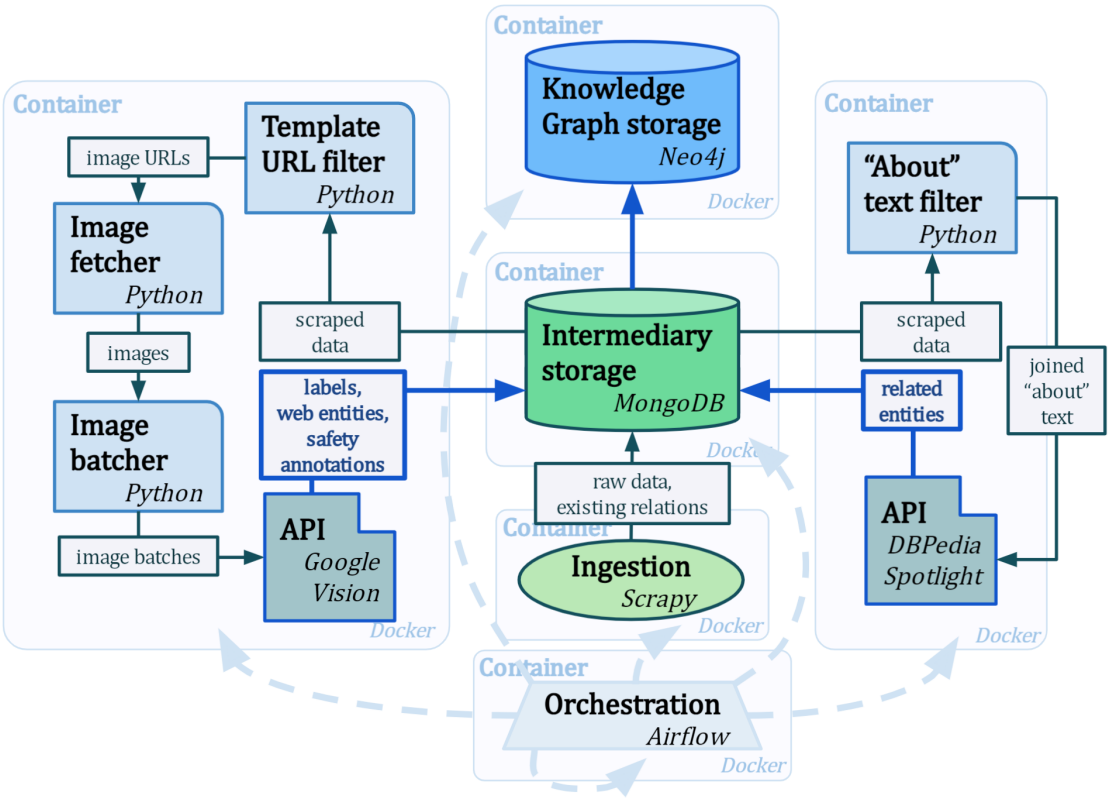

Figure 1. Pipeline diagram.

the definition of *tasks*, which are then, alongside their interdependencies, modelled into a directed acyclic graph. The nodes of this graph are then executed in their defined order and frequency by the *scheduler*, enabling constant updates as per ***ER1***. The execution is monitored, with implementable error handling in the event of task failure. Finally, in-built support for Docker containers is of significant benefit to the pipeline. This enables flexibility among the modules and provides scalability, satisfying ***ER3*** and ***ER4***.

With these tools provided, the pipeline itself is indifferent towards the specifics of its inhabitants, especially so should they reside in Docker containers. The first modules of the pipeline, and thus its application, are further described in the following subsections.

4.2.1 Access and ingestion

KnowYourMeme provides the largest and most structured catalogue of internet memes and their semantics, and yet lacks an API or any method of programmatically querying this data. The solution to this would be *web scraping*, the automated extraction of data from websites. This requires the establishment of what exactly is to be scraped, which follows:

1. Title
2. Category - content is separated into categories: people, events, and more, in addition to memes.
3. Addition timestamp
4. Latest update timestamp
5. Template image URL
6. Status - representing quality control, for example as confirmed, in research or *deadpool*.
7. Origin - meme origin, either as websites or media they are determined to originate from.
8. Year
9. Type - KnowYourMeme categorizes the memes into distinct types.
10. Tags
11. Additional references
12. Parent - presented as "Part of a series on", representing KnowYourMeme's internal meme taxonomy.
13. Siblings - presented as "Related Entries"
14. Children - presented as "Related Subentries"
15. Search terms - provided in an iFrame to Google Trends.

Additionally to the previous, of importance is also the text body of the article itself. This would require extra logic to separate the content into sections and subsections, as on the site, with separate extraction of all text, images and URLs. Excluding the latter and the relations, most is relatively simple to extract, requiring little extra logic, demonstrated by figure 2.

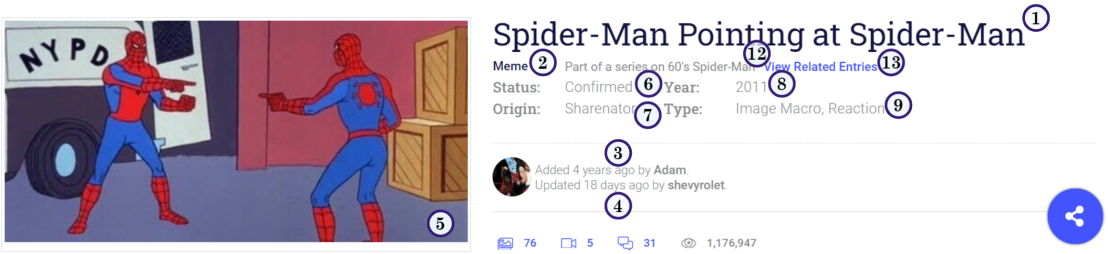

Figure 2. Header on KnowYourMeme page with most of the relevant information, annotated.

The first attempts of this were made using **Requests**[24], an HTTP request library for Python, and **Beautiful Soup**[25], a Python library to parse, and navigate HTTP documents. Colloquially, the pair of these are the usual suspects for most small-scale web scraping. With these, a pair of functions were written. The first would search for a given term using the site's built-in search engine, returning a set of URLs from the search results. The second would take an URL of a content page and parse it, according to hard-coded HTML element selectors, and return a Python dictionary object containing most of the relevant

[24] `https://docs.python-requests.org/en/master/`
[25] `https://www.crummy.com/software/BeautifulSoup/`

information, outlined above. This solution was successful in performing searches and extracting the resulting information. With this initial success, the scope broadened and the new objective would be to the extract information for all of the memes. For this, a more powerful framework was necessary.

Scrapy[26] is a powerful and feature-rich web scraping framework for Python. It is inherently asynchronous and contains appropriate tooling for any likely issue within the domain of web scraping. Most importantly, Scrapy abides by proper web scraping etiquette, otherwise known as *robots.txt* compliance. Scrapy works in a way of providing an environment for *spiders*, which are essentially Python scripts detailing the parsing and navigation directives of relevant websites. The previous parsing logic was ported over, necessitating translation to Scrapy compliant XPath selectors, and other small adjustments. A larger change was the inclusion of navigation directives, meaning that the spider would navigate the search results in their entirety, through pagination, and initiate parsing only on content pages. This was further complicated by the format that entity relationships are provided on KnowYourMeme. The siblings and children of a meme are provided on separate pages, with further pagination. With the extraction completed, Scrapy includes a pipeline to process and store the extracted items.

Outside of the parsing and navigation, additional problems were solved. The information on KnowYourMeme is contributed by volunteers, and this has resulted in many inconsistencies within the content. Examples of this include errors within HTML tags, missing and misidentified sections, and non-UTF-8 characters, all of which required additional logic to accommodate. With the growing data velocity, **ScraperAPI**[27], a rotating proxy service, was also included. Finally, with no inherent MongoDB support, a *plug-in* had to be written. With this, logic was also added to restrict complete parsing to only the pages that either lacked a database entry or showed a difference between the most recent update and the last recorded update. The scraped data was ultimately added to the MongoDB temporary storage with its KnowYourMeme URL as the index.

4.2.2 Enrichment

The ingested data, while rich in information, retains most of it in its unstructured part: the template image and article body. This includes the semantic information and entities necessary for knowledge graph construction. Extraction of this critical information would require a distinct approach for both mediums.

Image entity extraction Google provides image entity extraction as a service through its **Vision AI**[28]. It is accessible through an API and there are also *wrapper* libraries

[26] `https://scrapy.org/`

[27] https://www.scraperapi.com/

[28] `https://cloud.google.com/vision`

provided for multiple programming languages, including Python. The basic usage is straightforward: upload an image or provide its URL, specify the relevant features, and the API returns what was requested, along with confidence ratings where applicable. The features that were deemed relevant are label detection, explicit content detection, and web entity detection. The latter of these is perhaps of most significance, as the web entities returned are provided with *IRIs*, or internationalized resource identifiers. Identifiers such as these will eventually provide *alignment* between the meme knowledge graph and other knowledge graphs. Additionally, the selection of this feature essentially performs a reverse image search, providing URLs to matching and partially matching images from other sources - the instances of our templates.

The URL approach was the first attempted, proving unsuccessful as Google was unable to access the images. To overcome this, the images must first be downloaded. Thus, a module was created, which would download the image files from the template URLs stored in the intermediary database. In anticipation of large transmission, the downloader was written utilizing ScraperAPI and **AIOHTTP**[29], a Python framework that provides an asynchronous HTTP client. However, Google enforces strict rate limits on performed requests, which render the sequential image-by-image approach unpractical. A batch processing endpoint is provided as the alternative for scenarios of more voluminous data. This endpoint requires the images to be encoded into *Base64*. Accordingly, a batch-maker was added to the module, which would additionally handle the encoding. This was further supplemented with the **Python Client for Google Cloud Vision**[30], which enables reasoning about both the request and response of the API at runtime. The results of these queries are stored within the intermediary database, alongside the raw data, using the original KnowYourMeme URL as the index.

Text entity extraction The article body is comprised of multiple sections, providing information about the origin and spread of a given meme, and attempting to describe the meme itself. The latter often residing within the typically laconic *About* section, which presumably contains the information considered most relevant, such as featured characters, origination events or portrayed emotions. Following this presumption, it is then important to extract these entities. This is achieved with **DBPedia Spotligh**[31], a tool that recognizes DBPedia resources in text and returns the associated entities, along with confidence ratings, through an API.

To utilize this tool, a module was created. The module requests the *About* section data from the intermediary storage and joins all relevant text into a single string. This string is then sent to the API for annotation, with the result being stored within the intermediary database, using the original KnowYourMeme URL as the index. As this process is

[29] https://docs.aiohttp.org/

[30] `https://googleapis.dev/python/vision/latest/index.html`

[31] `https://www.dbpedia-spotlight.org/`

applied for all of the memes within the database, efficiency required consideration and as such, AIOHTTP is again utilized for its asynchronicity. A simple rate limiter is also included, to avoid efficiency becoming misuse.

4.3 Data

A large volume of data has been amassed, both through the initial scraping of KnowYourMeme and the subsequent enrichment. This section provides an overview of the data, along with some noteworthy observations.

KnowYourMeme has, throughout the years of its existence, developed a sort of schema, which is not enforced in any technical manner, but instead as convention. This schema dictates the existence of most of the relevant features determined in section 4.2.1. Without any sort of enforcement, the articles do not provide strict compliance. Table 1 provides an overview, presenting the percentage of articles which include the feature, out of a total of 28,799. Similarly of note is the internal taxonomy of KnowYourMeme, revealed through the relevant features (12, 13, 14). The most linked to memes are, in order, internet slang, image macros, and lolspeak. This makes sense, as these can be considered categories, even though KnowYourMeme does not make such a distinction.

Table 1. KnowYourMeme conventional schema compliance.

feature	1	2	3	4	5	6	7	8
frequency	100%	100%	99%	100%	100%	100%	100%	89%

feature	9	10	11	12	13	14	15	About
frequency	50%	51%	23%	42%	42%	49%	68%	36%

DBPedia Spotlight detected a total of 38,801 entities, with 10,262 of them distinct, from the 10,318 about sections provided. The most common, in order of occurrences, were image macros, japanese language, internet memes, YouTube[32], and catchphrases. It makes sense for these entities to be popular within about sections, as they often detail the origin and nature of memes.

Google Vision AI was more successful, determining 419,002 labels, of which 4,722 were distinct, and 197,582 entities, of which 42,746 were distinct. The most common labels were font, art, happy, gesture, and event. The most common entities were image, internet meme, Know Your Meme, meme, and humour. This aligns well with the general notion, that internet memes are frequently a vessel for humor.

[32]https://www.youtube.com/

5 Conclusion

This thesis presented the value of internet memes as a research vessel and detailed the construction of a pipeline to enable such research.

Section 3 introduced the concept of regarding internet memes as rich data awaiting analysis. This was assisted by the Macro-Meso-Micro framework, which enabled the deconstruction of the very broad question of how to enable systematic analysis of internet memes. This entailed the definition of meme sequencing, which would enable a novel quantitative approach to social sciences. However, meme sequencing itself, being out of reach for the current technology, would require the construction of a pipeline to facilitate it. The initial requirements for such a pipeline were presented in section 3.3.

Following the problem statement, section 4 describes the construction of this pipeline, starting with a discussion of what data to include as the foundation. The internet meme encyclopedia KnowYourMeme was chosen, as it provides the largest and best-detailed catalogue of internet memes, satisfying the requirements set for it in section 3.3. A description of the pipeline's architecture follows, with an outline of the used technologies and their brief introduction. The requirements are addressed, providing the reasoning for the choice of each component. With the encompassing pipeline defined, the next step would data ingestion. This was detailed in section 4.2.1, including solutions to multiple problems resulting from inconsistent data. To enrich this data, two additional modules were introduced, which queried external APIs for semantic information about the ingested data. This information would contain entities found in other knowledge graphs, providing the basis for alignment between these knowledge graphs and the eventual internet meme knowledge graph. These modules would provide utilization for the pipeline, demonstrating its application. Finally, the section was concluded in section 4.3, providing a summary of the data alongside notable observations.

Ultimately, this work has resulted in the creation of a flexible pipeline that welcomes future development, and one of the richest meme datasets, awaiting analysis. The primary future prospect for the pipeline is its completion regarding the knowledge graph. For the dataset, the prospect is analysis, for it is unknown what breakthroughs await within the memes. This is also the promise of systematic meme analytics in general, to provide a window into the mind, with no predictions of what there may await.

References

[1] Kate Barnes, Tiernon Riesenmy, Minh Duc Trinh, Eli Lleshi, Nóra Balogh, and Roland Molontay. Dank or not? analyzing and predicting the popularity of memes on reddit. *Applied Network Science*, 6(1), Mar 2021.

[2] Amy Bennett. Computers helped drive breakthrough in human genome sequencing, Dec 2000.

[3] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.

[4] Susan J. Blackmore. *The Meme Machine*. Oxford University Press, 1999.

[5] Richard Brodie. *Virus of the mind: The new science of the meme*. Integral Press, 2004.

[6] Heidi Chial. Dna sequencing technologies key to the human genome project. *Nature Education*, 1(1):219, 2008.

[7] M. Coscia. Competition and success in the meme pool: A case study on quickmeme.com. *ArXiv*, abs/1304.1712, 2013.

[8] Tamás Csordás, Dóra Horváth, Ariel Mitev, and Éva Markos-Kujbus. *4.3 User-Generated Internet Memes as Advertising Vehicles: Visual Narratives as Special Consumer Information Sources and Consumer Tribe Integrators*, pages 247–266. De Gruyter Saur, 2017.

[9] B. Dancygier and L. Vandelanotte. Internet memes as multimodal constructions. *Cognitive Linguistics*, 28(3):1515 –2017–0074, 2017.

[10] Patrick Davison. *The Language of Internet Memes*, page 120–134. New York University Press, 2012.

[11] Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976.

[12] Abhimanyu Dubey, Esteban Moro, Manuel Cebrian, and Iyad Rahwan. Memesequencer: Sparse matching for embedding image macros, 2018.

[13] R. Dulbecco. A turning point in cancer research: sequencing the human genome. *Science (New York, N.Y.)*, 231(4742):1055–1056, 1986.

[14] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid,

Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs, 2021.

[15] Leroy Hood and Lee Rowen. The human genome project: big science transforms biology and medicine. *Genome medicine*, 5(9):79, 2013.

[16] Michael Johann and Lars Bülow. One does not simply create a meme: Conditions for the diffusion of internet memes. *International Journal of Communication*, 13(0), 2019.

[17] M. J. Kelly. Computers: the best friends a human genome ever had. *Genome*, 31(2):1027–1033, 1989.

[18] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes, 2021.

[19] Jeffrey R. Lacasse and Eileen Gambrill. *Making Assessment Decisions: Macro, Mezzo, and Micro Perspectives*, pages 69–84. Springer International Publishing, Cham, 2015.

[20] Chen Ling, Ihab AbuHilal, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Dissecting the meme magic: Understanding indicators of virality in image memes, 2021.

[21] N. M. Luscombe, D. Greenbaum, and M. Gerstein. What is bioinformatics? a proposed definition and overview of the field. *Methods of information in medicine*, 40(04):346–358, 2001.

[22] Adam McNamara. Can we measure memes? *Frontiers in evolutionary neuroscience*, 3:1, 2011.

[23] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: shedding light on the web of documents. In Chiara Ghidini, Axel-Cyrille Ngonga Ngomo, Stefanie N. Lindstaedt, and Tassilo Pellegrini, editors, *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011*, ACM International Conference Proceeding Series, pages 1–8. ACM, 2011.

[24] Joshua Troy Nieubuurt. Internet memes: Leaflet propaganda of the digital age. *Frontiers in Communication*, 5:116, 2021.

[25] Greg Rowett. The strategic need to understand online memes and modern information warfare theory. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018.

[26] Limor Shifman. Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of computer-mediated communication: JCMC*, 18(3):362–377, 2013.

[27] Amit Singhal. Introducing the knowledge graph: things, not strings, May 2012.

[28] Sabu M. Thampi. Introduction to bioinformatics, 2009.

[29] Jacqueline Ryan Vickery. The curious case of confession bear: the reappropriation of online macro-image memes. *Information, Communication & Society*, 17(3):301–325, 2014.

[30] Dominic D. Wells. You all made dank memes: Using internet memes to promote critical thinking. *Journal of Political Science Education*, 14(2):240–248, 2018.

[31] Eline Zenner and Dirk Geeraerts. *One does not simply process memes: Image macros as multimodal constructions*, page 167–194. De Gruyter, 2018.

Appendix

I. Accompanying files

- The pipeline's **source code** and usage instructions are available on github[33].
- The **datasets**, in JSON format, are available on Nextcloud[34]

[33] https://github.com/Scytheface/memelord
[34] https://owncloud.ut.ee/owncloud/index.php/s/2LosgCo4bTjGM8n

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Tõnis Hendrik Hlebnikov**,
*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Towards a Knowledge Graph of Internet Memes,
   *(*title of thesis)

   supervised by Riccardo Tommasini.
   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Tõnis Hendrik Hlebnikov
03/08/2021