University of Zurich^UZH

Detecting and Mitigating Social Biases in Knowledge Bases

Thesis January 30, 2018

Silvio Fankhauser
of Langwies GR, Switzerland

Student-ID: 10-719-649
silviofankhauser@gmail.com

Advisor: **Bibek Paudel**

Prof. Abraham Bernstein, PhD
Institut für Informatik
Universität Zürich
http://www.ifi.uzh.ch/ddis

Acknowledgements

Danken möchte ich in erster Linie meinem Betreuer Bibek Paudel, der mich mit Rat und Tat unterstützt hat. Seine wertvollen Inputs und Ideen hatten einen wesentlichen Einfluss auf den Verlauf und das Ergebnis dieser Arbeit. Ausserdem möchte ich mich bei Prof. Abraham Bernstein bedanken, dass ich die Arbeit am Lehrstuhl für Dynamic and Distributed Information Systems an der Universität Zürich schreiben durfte. Ein besonderer Dank geht natürlich auch an meine Eltern, ohne deren Unterstützung ich mein Studium an der Universität Zürich niemals hätte verwirklichen können.

First of all, I want to thank my advisor Bibek Paudel for his guidance and advice during the last six month. His valuable inputs and ideas had a substantial influence on this thesis. Furthermore, I want to thank Prof. Abraham Bernstein for having the opportunity to write my thesis at the Dynamic and Distributed Information Systems Group at the University of Zurich. Last but not least, I would like to thank my parents, for always supporting and encouraging me during my studies.

Zusammenfassung

Die vorliegende Masterarbeit untersucht soziale Verzerrungen in Wissensdatenbanken. Dabei werden unter anderem verschiedene Berufsgattungen und deren Zusammenhang mit dem Geschlecht, der ethnischen Zugehörigkeit einer Person oder regionalen Unterschieden untersucht. Es werden drei verschiedene Varianten aufgezeigt, solche Verzerrungen aufzudecken. Die Unterschiede einzelner Regionen oder der Geschlechterverteilung einzelner Berufe sind teils gravierend. Wir demonstrieren anhand von Experimenten mit zwei verschiedenen Wissensdatenbanken verschiedene Arten von Verzerrungen. Diese Arbeit soll dabei das Bewusstsein steigern, dass solche Verzerrungen für die Nutzer dieser Datenbanken einen grossen Einfluss haben können. Solch unerwünschte Effekte auszugleichen ist keine triviale Aufgabe.

Abstract

This master thesis is about investigation of social biases in the knowledge bases. We examine, among other things, different professions and their association with a person's gender, race or regional differences. We present three methods to detect such biases. The differences between single regions or the varying distribution regarding the genders and professions are significant. We demonstrate with experiments on two large and widely used knowledge bases the different kinds of biases they can contain. The purpose of this work is to raise awareness, that this social biases can have an impact on the usage of those databases, given that mitigating is not a trivial task.

Table of Contents

1

Introduction

Knowledge Bases (KBs) constructed from natural language corpora have proven to be useful in several Artificial Intelligence (AI) tasks, like relation inference, link prediction, entity disambiguation and resolution. With the availability of modern hardware and smart learning algorithms in recent years, many KBs are being created and expanded using unsupervised or minimally supervised learning methods. Some big and famous ones are Freebase, DBpedia, NELL and YAGO. They contain large collections of facts in the form of relations between entities and hence provide a structured or semi structured form of knowledge, in contrast to the unstructured form of natural language text. This makes them suitable for use in several other problems ranging from Information Retrieval and Recommender Systems to Expert Systems and the completion/expansion of KBs themselves.

As the learning algorithms involved in the construction of KBs depend on natural language text usually written by humans, they could potentially be affected by the human biases present in such text. Biases in themselves may not necessarily be a harmful property, since they might reflect the nature of human knowledge and language about certain things. As an example: the fact that a woman would be associated with relationships like grandmother or sister rather than grandfather and brother is the natural outcome of implicit knowledge about gender specific words. However, other biases that reflect social prejudices could be more problematic. For example, associating men with dominant roles and women to those considered to be of soft nature might help in perpetuating harmful prejudices present in human-written text.

In this work, we investigate whether and to what extent such social biases are present in machine-learned KBs. Further, we would also like to develop a way to mitigate these biases. The KBs differ from each other regarding the structure, the way they are built, the source of the data and many other properties. Due to such differences, we introduce three different methods to detect potential biases. We want to evaluate, which method is the best suitable in terms of performance, quality of results or flexibility, depending of a given KB.

A common research topic in Social Sciences are biases regarding the gender, profession or race of a person. We investigate, if we can find significant differences in our KBs related to those attributes, e.g. if the number of females with a certain profession differs from the number of males in the same profession. In addition, we inspect differences regarding the geographical origin of person associated with the profession and gender.

An example of such a bias is, that the number of female mathematicians in West Europe is significantly higher than in Africa. In all cases, the number of males was much higher in our datasets, what can be interpreted as a first bias. In order to make a finer exploration of biases, we did not only look at the absolute number of males or females, but we also examined the percentage of women or men practising a profession, compared to the total number of males or females in the dataset.

Two of the methods are based on analyses of a graph, built from the data of YAGO. The first method consists of different analyses of paths from one node to another. The second one is producing multidimensional vectors called embeddings and comparing those. The third method is using a query language developed to retrieve information from data in RDF format. DBpedia provides an endpoint for this query language called SPARQL. All three methods presented in this thesis revealed severe biases in the datasets. The impact of such biases on other systems using this KBs is very difficult to estimate, especially since 'de-biasing' a knowledge base is not a trivial task. We tried an approach using embeddings to achieve that. Unfortunately, this approach did not lead to the desired result.

One could argue that such biases are only reflecting the state of reality, where the ratio between males and females can vary as well, depending on the profession, but we discovered several biases, which are clearly present in the data of the KBs only. For instance, besides the over representation of males, the region West Europe is very dominant in both datasets.

In Chapter 2 of this thesis, we will have a look at related work, including some studies regarding biases in Social Sciences, an introduction of some famous knowledge bases and the concept of word embeddings. In Chapter 3, we will present tree different methods to detect social biases in knowledge bases. We do this for varying groups like persons of a certain gender, profession, race or geographical origin. Furthermore, we present one approach for mitigating such biases using word embeddings. Chapter 4 includes the results of the previously introduced methods and an evaluation of those methods. Chapter 5 finally covers the conclusions, limitations and propositions for future work.

2

Related Work

In this chapter, we want to explore, how biases in Social Sciences are examined and present some findings of related studies. In this thesis, we want to explore, if we can find similar biases like the ones appearing in the mentioned studies. Furthermore, we introduce the purpose, usage and origin of knowledge bases and present three different big and famous KBs.

2.1 Biases in Social Science

The term "bias" can be described as an inclination towards something. It does not have a negative connotation necessarily, but it can be used in the context of discrimination or prejudice. Finding biases is a frequently examined topic in social sciences. They examine for example biases associated with gender or ethnicity. The proportion of females entering in Bachelor studies in different faculties at Swiss universities is presented in Figure 2.1. The difference between the faculties is clearly visible. It seems, there is a preference for certain fields, depending on the gender. This can be considered as an example of a gender-bias present in society that reflects in educational and occupational preferences.

A possible way to assess implications towards certain object is the "Implicit Association Test" (IAT). It measures differential association of two target concepts with an attribute. The concepts appear in a two-choice task (e.g. flowers and insects as concept, pleasant and unpleasant as attribute). When the concept is highly associated with an attribute, like flower and pleasant, the response time of the participants attending this test was lower than for less associated categories. This time difference implicitly measures differential association of the concepts with the attributes [Greenwald et al., 1998].

With the IAT, it is possible to examine associations of the gender with different professions, sciences or race-related stereotypes. It was analysed if the membership in a group like males or females influence the individual preferences and performance. We saw in our sample in Figure 2.1 that the representation of women in certain fields is rather low. Nosek et al. verified in their study that the association of females with arts is significantly higher than with maths. Data from this study are correlational but the causality is not clearly proven. The casual priority of attitudes, identity or stereotypes is not evaluated [Nosek et al., 2002].

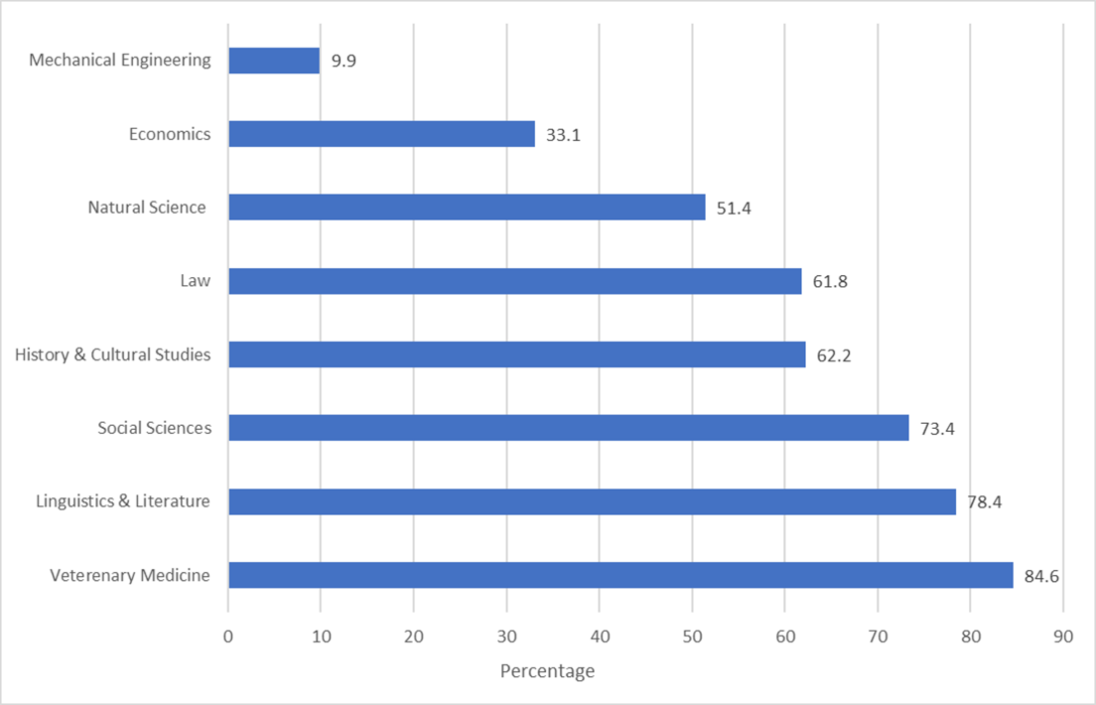


Figure 2.1: Percentage of women entered for Bachelor Studies per faculty. Switzerland 2010, Source: Bundesamt für Statistik

Almost 300'000 IATs from participants of several countries compared to the results of TIMSS (Trends in International Mathematics and Science Study) have shown a correlation between the IATs and the difference in the TIMSS performance. The stronger the association between male and science in a country, the larger the male advantage in science performance. The researchers could not determine whether the weaker performance of girls in science created the implicit gender-science stereotype or whether the stronger gender stereotype led to poorer female performance [Hill et al., 2010, pg. 77/78].

Another study examined the labor market discrimination depending on the name of an applicant. Job applicants with African-American names get far fewer callbacks for each resume they send out [Lavergne and Mullainathan, 2004].

2.2 Knowledge Bases

The term Knowledge Base (KB) was initially used for a specific part of expert systems. In 1965, NASA developed a program to determine the molecular structure of the Martian soil. Based on the concept of this program, the expert system era had begun. Those systems were built to solve complex problems in a specific domain, using different set of rules. The module containing those rules is called Knowledge Base [Leondes, 2002].

KBs play an important role in the Semantic Web, Recommender Systems or Artificial

Intelligence Systems. To access the desired information, computers need a structured collection of data [Berners-Lee et al., 2001]. Big KBs like YAGO, NELL or DBpedia are based on data retrieved from different sources like web pages, Wikipedia or natural language texts. They use machine learning algorithms to build an ontology, containing classes, entities and relations.

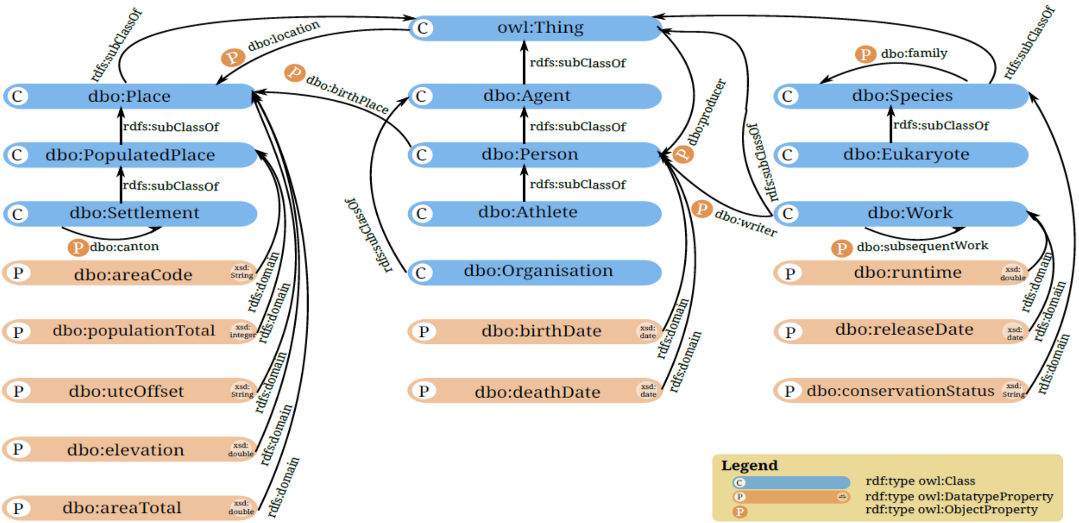


Figure 2.2: Snapshot of a part of the DBpedia ontology. [Lehmann et al., 2015]

With the technology nowadays, it is possible to gather huge amounts of data. The source of it is often written by humans and unstructured. The creators of the KBs use different approaches, sources and algorithms to extract and structure the data.

The DBpedia ontology consists of 320 classes which form a hierarchical taxonomy and are described by 1,650 different properties, extracted from Wikipedia [Lehmann et al., 2015]. The type of content that is most valuable for the DBpedia extraction are Wikipedia infoboxes. They are frequently used to list an article's most relevant facts, but other parts such as the text itself. The Wikipedia page is parsed into an Abstract Syntax Tree which serves as input for the extractors. DBpedia offers extractors for many different purposes (e.g. extract labels, abstracts or geographical coordinates [Lehmann et al., 2015].

DBpedia still has a substantial number of problems in quality. The error rate in a manual and semi-automatic assessment was 11.93%. The most flaws are based on incorrectly extracted object values, irrelevant extraction or broken links [Zaveri et al., 2013].

YAGO (Yet Another Great Ontology), another big KB, has knowledge of more than 10 million entities (like persons, organizations, cities, etc.) and contains more than 120 million facts about these entities [Mahdisoltani et al., 2014]. Its taxonomy is derived from Wikipedia, Wordnet and GeoNames and it assigns entities to more than 350'000 classes. Wordnet is a semantic lexicon for the English language. Yago uses the dictionary of nouns, which is structured in super- and sub-concepts. GeoNames is a geographical database, provided under a creative commons license by official public sources.

Besides the taxonomy, YAGO differs from DBpedia in the processing of new facts. YAGO's technique introduces new facts in addition to new entities, whereas DBpedia's main goal is to assign novel entities to DBpedia classes automatically [Mahdisoltani et al., 2014]. The correctness was evaluated manually in a sample of approximately 4'500 retrieved facts. 95.03% were judged to be correct [Yago, 2017]. Compared to DBpedia, the error rate is considerable lower.

The KB NELL (Never-Ending Language Learner) follows a different approach. It reads Webpages (initially 500 million pages) and extracts beliefs 24h/day, each day. To build this beliefs, several learning tasks are performed. Category classification or the distribution of noun phrases are part of this tasks. Each belief has a confidence percentage, of it being true. The learning algorithm aims to increase the confidence score over the time. Occasionally, humans interact with the KB over its web page (http://rtw.ml.cmu.edu/rtw/) and adjust the score manually [Mitchell et al., 2015]. The quality of NELL is barely comparable with the other KBs, because it's a running process, aiming to maximize the confidence scores over the time. When new beliefs are constructed, their confidence score is initially rather small and the belief itself possibly wrong.

2.3 Word Embeddings

Multidimensional vector representations of words, called word embeddings, were developed to improve natural language processing (NLP) applications. Many algorithms for NLP or machine translation are relying on rules or statistics. There are well known problems e.g. ambiguous words, and the word order in a sentence or contexts spanning over several sentences. The usage of Neural Networks combined with word embeddings in this field promises a significant improvement over the previously used approaches [Koehn, 2009]. There are many use cases for Neural Networks such as machine translation, recognition and detection in vision or speech, artificial intelligence or classification.

Word2Vec uses Neural Networks to build such word embeddings. Semantically similar words are mapped close to each other in a multidimensional vector space. With this mapping, it is possible to predict semantic relationships like: "Bern is to Switzerland as London is to ?". Usually, the Neural Network of Word2Vec is trained with data from natural language texts to build the vectors. [Mikolov et al., 2013a] The outcome depends heavily on the used corpora. Texts of different topics and length can lead to different vectors and as a result varying clusters of words in the vector space. A text about laws for example, will lead to different embeddings than a text about a love story.

Instead of natural language texts, is's also possible to use other sources as input. A knowledge base is basically a huge connected graph with interlinked nodes. A sequence of connected nodes from such a graph can serve as an input of a Neural Network to create embeddings [Ristoski and Paulheim, 2016] [Perozzi et al., 2014]

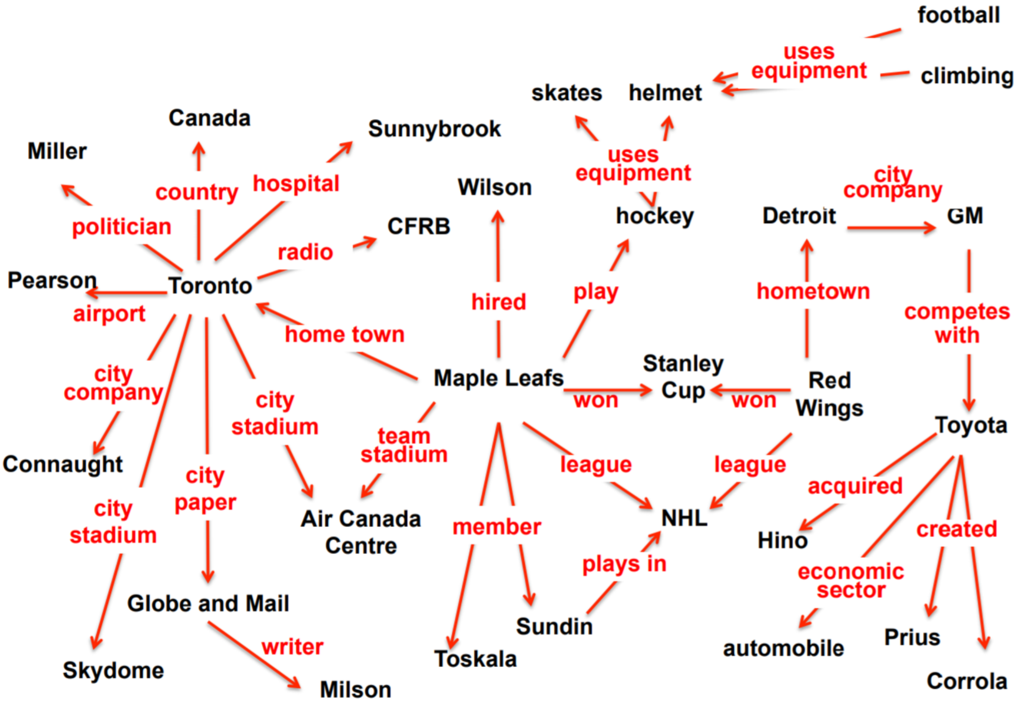


Figure 2.3: Fragment of the 80 million beliefs NELL has read from the web. [Mitchell et al., 2015]

3

Methodology

In this work, we focus on the Knowledge Bases DBpedia and YAGO. We will show three different methods to detect biases. For DBpedia, we will retrieve the desired information with the Query Language SPARQL, whereas on the other hand, use enumerations on the paths of a graph, built from the data of YAGO. The third possibility are vector representation of each entity in the KB, called embeddings. We will build them performing random walks on the graph and use those as an input for Word2Vec.

The basic idea is to retrieve data for certain professions, genders and countries and analyse the results. We will compare the number of entities found, grouped by gender, profession and region to see, if there are some significant differences in between. A significant difference will suggest the presence of biases.

Before we start with the retrieval of biases using different methods, we have to define several groups in order to know which properties we want to explore. Those groups are described in the next section.

3.1 Categories: Gender, Profession and Region

The two KBs we want to explore use different ontologies. YAGO's number of classes is more then 1'000 times higher compared to DBpedia. In this thesis, we focus on some examples, because an examination of all possible combinations would go beyond the scope. We chose the professions from different fields like sciences, arts and sports, based on their background (e.g. mathematical, social, artistic) to provide a broad overview.

Our first approach to build regions was to group the countries according to the continents they belong to. On a closer look, the cultural differences between the countries of a continent were too big. This could blur cultural biases in our results. So we divided the continent's regions based on similar cultural background. The assignment of a country to a region is not unambiguous and was also driven by the number of entities we found. Africa and South America for example had just a few results compared to Europe or North America, that's why this countries are assigned to one big region, despite potential cultural differences and the large number of actual inhabitants.

DBpedia and YAGO are using Wikipedia as input source. Besides possible biases from the authors of a Wikipedia article, the accessibility of the web page from certain countries can have an influence as well. Some countries like China, Vietnam, Iran, Russia, Arab countries and even more, have a severe censorship of the internet. But also other countries like some Latin American or African countries limit free access to certain webpages to some extend.[Warf, 2011] Another possible source of biases is the uneven distribution of internet access for the inhabitants (see figure 3.1). In some regions, the access is limited for big parts of the population.

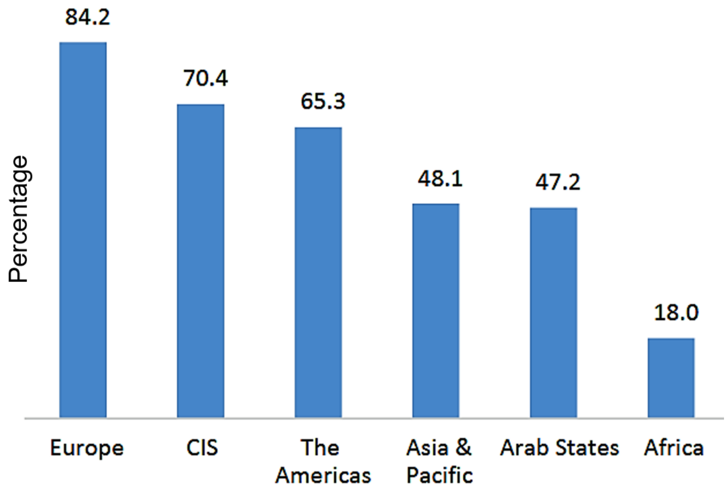


Figure 3.1: Proportion of households with Internet access, International Telecommunication Union [Union, 2017].

In this study, we used the following regions containing countries:

- *West Europe:* Austria, Belgium, Britain, Denmark, Holland, Netherlands, England, United Kingdom, Estonia, Finland, France, Germany, Iceland, Ireland, Latvia, Lithuania, Luxembourg, Norway, Scotland, Sweden, Switzerland, Wales, Monaco, Andorra

- *East Europe:* Belarus, Bulgaria, Czech, Hungary, Poland, Romania, Russia, Slovakia, Slovenia, Ukraine, Albania, Croatia, Serbia, Bosnia, Montenegro, Turkey, Georgia, Moldova, Yugoslavia, Macedonia, Moldavia, Armenia, Soviet Union

- *South Europe:* Italy, Portugal, Spain, Greece, Malta, Sicily, Cyprus, Vatican, San Marino

- *North America:* United States, Canada

- *Latin America:* Mexico, Dominican Republic, Cuba, Jamaica, Costa Rica, Salvador, Guatemala, Honduras, Nicaragua, Panama, Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Paraguay, Peru, Uruguay, Venezuela, Puerto Rico, Bahamas, Haiti, Dominica, Curacao, Barbados, Anguilla, Guyana, Belize, Guadeloupe, Martinique, Cayman Islands, British Virgin Islands, Suriname, Montserrat

- *East Asia:* Brunei, Burma, Cambodia, China, Indonesia, Japan, Korea, Laos, Malaysia, Mongolia, North Korea, Singapore, South Korea, Taiwan, Thailand, Tibet, Vietnam, Philippines
- *South Asia:* Afghanistan, Iran, Pakistan, India, Sri Lanka, Nepal, Bhutan, Bangladesh, Maldives
- *West Asia:* Oman, Saudi Arabia, Syria, Iraq, Emirat, Bahrain, Lebanon, Kuwait, Tajikistan, Kyrgyzstan, Kazakhstan, Azerbaijan, Turkmenistan, Uzbekistan, Qatar, Jordan, Yemen
- *Africa:* Algeria, Angola, Cameroon, Chad, Egypt, Ethiopia, Ghana, Ivori Coast, Kenya, Libya, Malaga, Mali, Mauritius, Morocco, Namibia, Niger, Nigeria, Congo, Senegal, Somalia, South Africa, Sudan, Tunisia, Zambia, Zimbabwe, Uganda, Gambia, Tanzania, Sierra Leone, Togo, Rwanda, Burundi, Burkina Faso, Equatorial Guinea, Malawi, Swaziland, Mozambique, Gabon, Djibouti, Rhodesia, Tangier, Madagascar, Botswana, Ivory Coast, Liberia, Guinea, Benin, Eritrea
- *Oceania:* Australia, Fiji, New Zealand, Solomon Islands, Guam, Samoa, Tonga, Vanuatu, Tasmania, New Guinea, Marshall Islands, Micronesia, Kiribati, Palau, Easter Island, Solomon Islands, Nauru, Tuvalu, French Polynesia, New Caledonia

Not all countries in the world are listed here. We assigned the countries present in the retrieved path from chapter 3.2.3. It's possible, that for certain nations, no relevant paths were found, in particular smaller or less developed nations. Israel belongs to West Asia, but its culture differs strongly from the other nations in this region. For that reason and the relative low number of retrieved paths, we decided to exclude path related to Israel.

With the following methods, we want to analyse those groups regarding the gender, profession, race and geographical origin in YAGO

3.2 Yago: Enumeration

A first method to detect biases is enumeration. We will build a graph using a dump of YAGO's dataset and investigate connections between certain nodes. The goal is to find all path from a start node to a target node and compare them with other paths with varying start and target nodes. We want to examine differences regarding genders, professions, race and regions as described previously.

3.2.1 Building a graph

YAGO provides the whole ontology as a download on their web page[1]. The uncompressed ontology has a size of 168GB in tsv-format. The results of this thesis are based on data

[1] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/

of the following dumps:

- Wikipedia (en, de, nl, fr, it, es, pl, fa, ar, ro): 2017-05-20
- Geonames: 2017-06-18
- Wikidata: 2017-05-22

Not all files are useful for our research. The English version of Wikipedia is by far the most comprehensive compared to the others. It has the highest number of articles and edits. The fact that it is an internationally recognised language, was another reason to focus our analysis on the English version of the data. To investigate the data and its relations, a graph from the two files ’yagoFacts.tsv’ and ’yagoTaxonomy.tsv’ was built. The structure of the files contained ids, entities and relations in between (see Table 3.1). The entities are represented by the nodes of the graph, whereas the relations are modelled as directed unweighted edges. The different types of relations are added as attributes to the edges. The graph can be defined as $G = (N, E, L)$, where N is a set of nodes, E is a set of directed edges and L are the labels of the edges.

A lot of nodes consisted of external websites with an edge of the type *’<hasWebsite>’*. Those nodes contain no useful information for this thesis, as they are leaf nodes in the graph and the links to external web sites are not evaluated. To speed up our further work, those nodes were ignored.

Table 3.1: Small part of the yagoFacts.tsv file.

<id_156a81d_z7a_1lfplq>	<Onchiam>	<isLocatedIn>	<Kerala>
<id_u0gksp_ab2_zz7558>	<Gregory_S._Martin>	<hasWonPrize>	<Order_of_the_Sword_(United_States)>
<id_9taxo5_1ul_1oe1uea>	<Wouter_Vrancken>	<playsFor>	<K.V._Kortrijk>
<id_gfqocs_dhj_27wsag>	<Anthony_Gilbert_(author)>	<diedIn>	<London>
<id_z13ez0_13l_1v6yd3x>	<Johan_Jacobsen>	<directed>	<The_Invisible_Army>
<id_4hyd17_1gi_d6762e>	<Ezekiel_Honig>	<created>	<Scattered_Practices>
<id_m65qnc_ice_1zg341>	<Wolfgang_Tillmans>	<hasWebsite>	<http://www.tillmans.co.uk>
<id_lp8jxt_ice_b1exj3>	<Absecon,_New_Jersey>	<hasWebsite>	<http://www.absecon-newjersey.org>
<id_qnzz48_1s2_1qsse35>	<Thomas_Brdari>	<isAffiliatedTo>	<VfL_Kirchheim/Teck>

The graph itself is composed of 7’276’455 nodes with 59’546’682 edges. The degree of the nodes are following the power law. There are few nodes with a very high degree and many with a low degree. We don’t distinguish between in or out degree, because of added inverted edges, what is described later in this chapter. To compare path probabilities, the adjacency matrix of the graph can be a useful tool. By computing and normalizing the scalar product of the matrix with itself, it is even possible to discover the probability of reaching a node from another in a selectable number of steps. Unfortunately, this procedure is only applicable for rather small graphs. The multiplication of a adjacency matrix in this dimension, by far exceeds the available memory of a common computer. So, we had to switch to another approach.

We want to find all paths on this graph between a start node to a target node with a defined maximal length. For the graph $G = (N, E, L)$, a start node $n_0 \in N$ and a target node $n_d \in N$, we search all paths P of depth d. In each step, we explore the outgoing edges $e(n_i)$ where $n_i \in N$ is the actual node we are exploring at the moment, starting with an index $i = 0$. The path will follow this pattern, building a Markov chain:

$$P = n_0 \rightarrow n_{i+1} \rightarrow ... \rightarrow n_d.$$

One possibility to achieve this are random walks on the graph. The walk starts at the defined starting node or from a set of starting nodes and selects randomly one of the edges to 'walk' to the next node, until it reaches the maximal path length.

First, we want to investigate paths from a gender, male or female, to a set of professions. The original structure of the graph contains directed labelled edges. Certain nodes had only a few outgoing, but thousands of incoming edges, like the node '<*male*>'. If we chose a start node like 'male' we can not reach all neighbours in one step. The same problem existed on many other nodes, therefore, we duplicated every edge in the graph, changed its direction and added it to the graph. To recognise such edges later on, the labels were marked with the postfix 'inverted'. Another solution would have been, to make the edges undirected, but this would cause an information loss, because the original direction would disappear.

$$John\ Doe \xrightarrow{\text{hasGender}} male$$

$$male \xrightarrow{\text{hasGender:inverted}} John\ Doe$$

3.2.2 Paths: Gender to Profession

With this enhancement, its possible to reach any node from any starting point in the graph. With start node '<*male*>', a big part of the graph is reachable within 3 steps (see Table 3.2). With random walks, it takes a lot of tries to hit the desired target nodes and it's not sure, that all relevant paths are found.

Table 3.2: number of nodes, reachable from start node '<male>' in x steps.

	number of nodes
1 step	934600
2 steps	1'367'817
3 steps	4'826'963

So, finally, we decided to use a depth first search algorithm to explore the graph. A single path can be found in $\mathcal{O}(V + E)$ (V=nodes, E=edges) time but the number of paths in a graph can be very large, e.g. $\mathcal{O}(n!)$ in the complete graph of order n. [Sedgewick, 2001]. Fortunately, our graph is not complete, despite having some nodes with a very high degree. The algorithm has a long runtime, but finds every relevant paths. As start nodes, we use the nodes '<*male*>' and '<*female*>'. The target nodes are the following:

- <wordnet_mathematician_110301261>
- <wordnet_engineer_109615807>
- <wordnet_dancer_109989502>
- <wordnet_computer_scientist_109951070>
- <wordnet_chemist_109913824>
- <wordnet_biologist_109855630>
- <wordnet_artist_109812338>
- <wordnet_writer_110794014>
- <wordnet_politician_110451263>
- <wordnet_poet_110444194>
- <wordnet_psychologist_110488865>
- <wordnet_physicist_110428004>
- <wordnet_volleyball_player_110759047>
- <wordnet_swimmer_110683349>

In the best case, we can reach a target node within two steps but often, a third step is necessary:

$$\begin{gathered} male \longrightarrow John\ Doe \longrightarrow mathematcian \\ male \longrightarrow John\ Doe \longrightarrow American\ mathematician \longrightarrow mathematcian \end{gathered}$$

After the first investigation, a lot of misleading paths are retrieved with this procedure. The node of a gender is connected with thousands of entities of unique persons. Those nodes are connected with the superordinate node '<*person*>'. Most of our target nodes are connected with '<*person*>' as well, what results in distorting paths. The same problem, but in a lower dimensions, exists for the node '<*scientist*>' and '<*athlete* >'.

The node <*person*>' is connected with several nodes representing different professions. The same also applies for '<*scientist*>' and '<*athlete* >'. Any path crossing one of this nodes can reach those professions. To omit such paths, this nodes are deleted from the graph (see Figure 3.2, highlighted red) resulting in a much lower number of paths. Without this modification, every existing person in the dataset is connected with every possible profession, like Donald Trump with Mathematician in our example. After deleting the mentioned nodes, only paths for persons related to the correct profession are retrieved. It's possible, that several valid paths exist passing the node of a unique person (highlighted green).

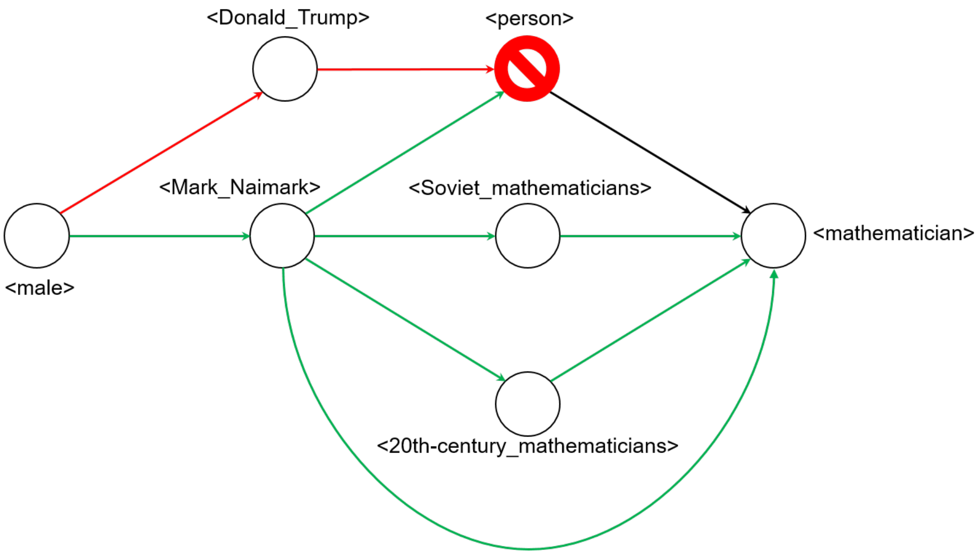


Figure 3.2: Example paths from YAGO, retrieved with enumeration in 3 steps from node '$<male>$' to '$<mathematician>$'.

3.2.3 Paths: Gender to Region

Within the 3-step-paths from a gender to a profession, we can find nodes with additional information. Some of them contain facts about the origin of the respective person, the time period when he was alive or more detailed information about the profession. The place of origin helps us to assign the persons to one of our previously defined regions. Several nodes contained geographical information, but we only considered nodes with the following structure in the name: <wikicat_Lithuanian_chemists>, of course with varying country name and profession. This pattern was chosen because other patterns contained more specific information (e.g. <wikicat_American_physical_chemists>). The node representing a specific individual is usually connected with several of such nodes (see Figure 3.2). Sticking to the described pattern reduces the number of paths per person, what results in a better representation of actual individuals in the dataset. Usually, a person is connected to only one node, matching the described pattern.

There is a big difference between the number of inhabitants of the single regions. The same is also true for the persons in YAGO. To compare the number of paths we found per region, the data must be normalized. First, we want to find the total number of persons per region, independent of their profession. To achieve this, all paths are retrieved with a 3-step-enumeration, starting from the genders (male and female) to the nodes representing a continent.

$$female \xrightarrow{\text{hasGender:inverted}} Asha\_Haji\_Elmi \xrightarrow{\text{isCitizenOf}} Somalia \xrightarrow{\text{isLocatedIn}} Africa$$

$$male \xrightarrow{\text{hasGender:inverted}} Ryland\_Blackinton \xrightarrow{\text{isCitizenOf}} USA \xrightarrow{\text{dealsWith}} Australia$$

$$female \xrightarrow{\text{hasGender:inverted}} Flora\_Nwapa \xrightarrow{\text{influences}} Amina\_Mama \xrightarrow{\text{isInterestedIn}} Africa$$

Again, the retrieved paths are not entirely matching our needs. Certain types of edges like '*<dealsWith>*', '*<isInterestedIn>*' or '*<influences>*' are present in the paths, but they are building misleading paths again. Therefore, paths with edges of this types are ignored in our results. To assign a person to a certain country, respectively region, edges of the following types must be present:

- <diedIn>
- <isCitizenOf>
- <isLeaderOf>
- <isPoliticianOf>
- <livesIn>
- <wasBornIn>

Unfortunately, this relations don't always connect nodes of the same types. Sometimes, the target node is a country, in other cases a city or a region within a country. In order to assign each retrieved path to a region, we have to find the associated country in a first step and allocate it to the previously defined regions.

To find the corresponding country, we used the Python module 'Geopy' [Geopy, 2015]. Via the Google Maps API, it is possible to query the desired information with a restriction of 2500 requests per day. The response was aligned with the countries from our regions. The retrieved paths from genders to continents revealed some paths afflicted with wrong connections. Two examples are the nodes of the countries '<United_States>' and '<France>' which are connected over an edge of the type '<isLocatedIn>' with the node '<Oceania>'.

$$United\_States \xrightarrow{\text{isLocatedIn}} Oceania \qquad \text{✗ false}$$

$$France \xrightarrow{\text{isLocatedIn}} Oceania \qquad \text{✗ false}$$

Those connections are obviously wrong. Such false entries lead to a big number of incorrect paths, especially when the country has a big number of inhabitants like the United States. We consider only the paths, where the retrieved country actually belongs to the target continent, ignoring all others.

The number of remaining correct paths serve as input to normalize the amount of paths found in chapter 3.2.2. We divide the number of path per profession and region by the total number of path per region.

3.2.4 Paths: US citizens of foreign descent to Profession

Besides possible biases based on the gender, we want to find out if the are biases depending the ethnic origin of a person. For this reason, we examined the origin of US citizens in combination with their profession. The US was picked for the purpose of the large number of persons related to this country. Furthermore, it's known for its history of immigration. This factor is represented in a list of nodes, describing the historical origin of the connected nodes representing an individual. Those nodes have the pattern: '<wikicat_American_people_of_Mexican_descent>', of course with different country names instead of 'Mexican'. Again, we used the depth-first-search algorithm to look up paths from those nodes to the professions. Like in the other cases, we grouped the countries in the described regions. An example of such a path looks like this:

$$American\_people\_of\_Chinese\_descent \xrightarrow{\text{type:inverted}} Man - Chung\_Tang \xrightarrow{\text{type}} engineer$$

3.2.5 Detecting noise

The correctness of the data in the knowledge base YAGO is not 100%, as the source data is created and manually maintained by humans. In our previous steps, we stumbled over some flaws. To distinguish between a bias or the natural outcome for a certain term, we have to elaborate the error rate based on our requirements, although its not the purpose of this thesis to evaluate the quality of the dataset. Some terms are clearly referring to a gender like mother, actress, uncle or king. In a perfect dataset, all of this terms should be associated to the correct gender. We are looking for wrong connections between the gender and the appropriate term. One possibility is to retrieve all 2-step-paths from a gender to King/Queen. Longer paths result in a futile search for noise, because they can contain relatives of the wanted person.

$$male \longrightarrow John\ Doe \longrightarrow King \qquad \checkmark$$

$$female \longrightarrow Jane\ Doe \longrightarrow King \qquad ✗ \text{ noise}$$

$$female \longrightarrow Jane\ Doe \xrightarrow{\text{hasCild}} John\ Doe \longrightarrow King \qquad ✗ \text{ not useful}$$

The precise terms of family relationships are often bound to a gender (father, brother, aunt, etc.). In YAGO, this relations are modelled in a gender neutral way, so it's pointless to search there for noise.

$$male \xrightarrow{\text{hasGender:inverted}} John\ Doe \xrightarrow{\text{hasChild}} John\ Doe\ junior$$

$$female \xrightarrow{\text{hasGender:inverted}} Jane\ Doe \xrightarrow{\text{hasChild}} John\ Doe\ junior$$

In the previously retrieved path from a gender to professions in Section 3.2.2, some intermediate nodes are referring to a certain gender. There are nodes distinguishing male and female dancers or artists. Some of the paths are traversing this nodes, when they shouldn't. If you are starting from the node 'male' and pass a node like 'Female_comics_artists', the path is very likely noisy. This kind of noise was discovered while investigating the path files. Once such nodes were identified, we counted their occurrences in the retrieved paths.

$$male \xrightarrow{\text{hasGender:inverted}} Laerte \xrightarrow{\text{type}} Female\_comics\_artists \xrightarrow{\text{subClassOf}} artist$$

By comparing the count of the noisy entries to the total number of paths found, we can compute the average error rate (see Table 3.3). Compared to the manual evaluation of approximately 4'500 facts what revealed a error rate of 4.97% [Yago, 2017], the error rate with 0.8% is approximately 5 times smaller. The sample size was 5'667 paths. Our evaluations only considered wrong entries regarding the gender, other flaws were ignored.

Table 3.3: Error rate for gender and profession.

Node	Total Count	Wrong Entries	Error Rate
Female Dancers	1392	9	0.006465517
Male Dancers	1324	12	0.009063444
Female Artists	2419	5	0.00206697
Kings	199	7	0.035175879
Queens	333	14	0.042042042
Total	5667	47	0.00829363

3.3 NELL: Enumeration

In the previous Section, we evaluated paths from genders to defined target nodes representing professions. In this Section, we will use a different KB and other start respectively target nodes. Instead of counting paths like before, we want to explore the distance of the target nodes from the start node.

In a study of Nosek et al, different combinations of classes and attributes are evaluated with the IAT.[Nosek et al., 2002] Caliscan et al. have shown that such biases are present in semantics derived from natural language corpora.[Caliskan et al., 2017]

We want to explore if similar associations are present in a KB. Instead of investigating a bias regarding the gender, ethnicity or profession, the choice fell on a universally accepted association: the comparison of plants and insects with pleasant and unpleasant adjectives. YAGO and DBpedia are mainly built from data of Wikipedia, whereby the info boxes are the most important source. Compared to natural language texts, the nodes of this two KBs usually don't contain any nodes representing adjectives. NELL was chosen because its sources are natural language texts of many different websites.

The data of NELL are provided on the web page[2] in different formats. For our analysis, the tab separated value file with every belief in the KB is used. We built the graph the same way we did it with the data from YAGO described in Section 3.2.1: The nodes are representing classes and entities of the KB, the relations are represented by the edges of the graph. Again, we added inverted edges for each existing edge. Some types of edges like 'concept:haswikipediaurl','concept:proxyfor' or 'concept:mutualproxyfor' are omitted, as they are related to web pages which are not providing any benefit for this analysis. The graph consists of 1'607'032 nodes and 4'414'532 edges. Like the graph for YAGO, the distribution of the node degrees is following the power law.

In a first step, a list of categories is defined. The choice fell on the same categories, which are used in the study of [Caliskan et al., 2017], including the associated terms.

- *positive words:* caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation

- *negative words:* abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, bomb, divorce, jail, poverty, ugly, cancer, evil, kill, rotten, vomit, agony, prison

- *flowers:* aster, azalea, bluebells, buttercup, carnations, clover, crocus, daffodils, daisy plant, gladiolas, hyacinth, iris flower, lilac, lily, magnolia flower, marigolds, orchid plant, peony, petunias, poppy plant, rose plants, tulip, violets, zinnia

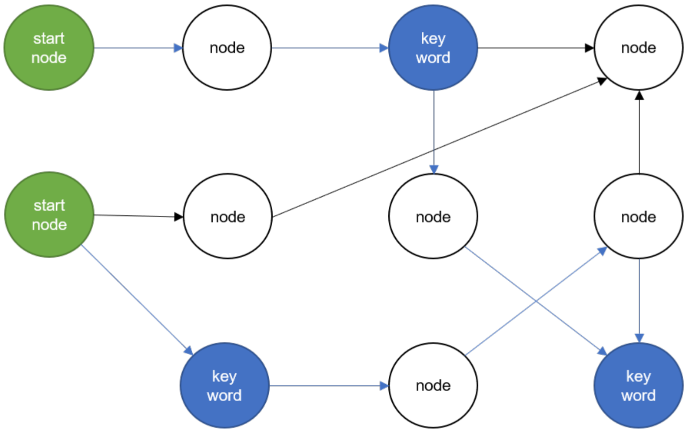
- *insects*: spider, tarantula, bedbug, cockroaches, fly, gnats, hornets, moth fly, roaches, ant, bee, beetle, caterpillar, dragonfly, fleas, locust tree borer, maggot, mosquito, termite, wasp, weevil, centipede

We want to investigate, how many steps it takes to reach a node containing a defined keyword, starting from a set of nodes, where all nodes belong to the same category (see Figure 3.3). The start nodes are sets of nodes corresponding to the entities in the list (flowers and insects). The target nodes aren't predefined. We lookup each node we are encountering during enumeration, if the name of the node contains a substring, matching any keyword in the positive or negative word list. If this is the case, we retrieve the distance between the start and actual node. Finally, we build a table containing the counts of this distances (see Figure 3.3).

A lot of paths are including dates. If a node is connected to the node date, all other dates and the connected events are easily reachable. For that reason, all paths including dates are omitted.

$$aster \xrightarrow{\text{generalizations}} plant \xrightarrow{\text{generalizations:inverted}} gold \xrightarrow{\text{atdate}} n2006 \xrightarrow{\text{atdate:inverted}} health$$

[2] http://rtw.ml.cmu.edu/rtw/resources



Distance	Nr of paths
1	1
2	1
3	0
4	2

Figure 3.3: Count of path with its length from a set of start nodes to nodes containing a keyword.

The results of this approach weren't satisfying. One reason is, that the target nodes of the paths weren't matching our expectations. Some nodes appear very often, like 'health' or 'dead', whereas adjectives like 'honest' are non-existent.

Another reason is the context within the retrieved paths. Some of them have a clear and comprehensible logical relation, considering only the start and target nodes. But mostly, the paths looked like this:

$$caterpillar \xrightarrow{\text{synonymfor}} musicartist : cat \xrightarrow{\text{agentcompeteswithagent}} fish : cat$$
$$\xrightarrow{\text{agentparticipatedinevent}} result \xrightarrow{\text{agentparticipatedinevent}} friend$$

We couldn't see any value in the content of such paths. The connections often passed nodes representing an umbrella term, like the node *result*, which have mostly a very high degree. As a result, the path can pass a high variety of nodes, blurring the logical relation between start and target node. The number of paths with a short length was very low. The longer the paths, the lower the logical context between the start and target node. That's the reason, why did not pursue the matter further.

3.4 DBpedia: SPARQL

SPARQL is a query language for RDF (Resource Description Framework). The name of this language is a recursive acronym for 'SPARQL Protocol And RDF Query Language'. RDF and SPARQL is recommended by the W3C. RDF is a standard model for data interchange on the Web. It extends the linking structure of the Web to use URIs and name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). [W3C, 2014]

DBpedia provides a public SPARQL endpoint. Like that, it is possible to access the data directly with a Web browser. Furthermore the dataset is downloadable in different file formats. The structure of the files is more complex compared to YAGO, because the different classes are represented in single files containing the associated entities. Each DBpedia entity is described by various properties. Those properties, called resources, are defined in the classes of the ontology. The class 'Person' for example contains a broad spectrum of 248 different resources. This includes for example the birth place, the eye color or the profession. Naturally, not all properties are set for each entity. With SPARQL, it is possible to query persons with desired resources.

3.4.1 Queries: Gender, professions and regions

One big obstacle querying data is the inconsistent usage of the different resources. One example are the resources 'profession' and 'occupation'. A profession like 'Lawyer' can be linked to a person via both resources. This is solvable with a union clause in the query. A more serious problem are missing links. A lot of occupations are linked to only one person. This unique resources are containing the information in text form, manually created by a human and they are not linked to the corresponding resources. Due to the large number of this unique strings, we decided to lookup this text to see, if they contain a substring matching the requested profession. An extract of the results can be found in Table 3.4.

Table 3.4: Extract from the result of querying the professions for the region Oceania

country	**gender**	**person**	**profession**	**occupation**
:Australia	male@en	:Leo_Baker	-	Writer, Director@en
:Tasmania	female@en	:Sally_Poncet	-	Antarctic biologist and ornithologist@en
:New_Zealand	male@en	:David_Garrett	:Lawyer	-
:New_Zealand	female@en	:Elsie_Locke	:Writer	-
:New_Zealand	male@en	:David_Lautour	:Musician	-
:Australia	male@en	:Alan_Jefferies	:Poet	-

Corresponding to the approach we followed in Section 3.2.3 to assign the correct region, we want to investigate several resources in DBpedia:

- birthPlace
- deathPlace
- livingPlace
- nationality
- citizenship

The countries are assigned to regions (see Section: 3.1). We used the same segmentation as for YAGO to make the results comparable. Because of the different ontology of the two KBs, it was not possible to query exactly the same resources. Some of the investigated professions in YAGO are not present in DBpedia or they are assigned to another class than 'Person'. For that reason, the queried professions differ.

- Artist
- Athlete
- Biologist
- Lawyer
- Mathematician
- Musician
- Poet
- Psychologist
- Writer

The attributes regarding the countries are combined with a 'UNION' statement. The same procedure has been implemented for the professions, whereas the classification into regions happens in the 'FILTER' statement at the end of the query. At this point, the desired profession is filtered as well (see Listing 3.1).

Listing 3.1: SPARQL query to retrieve persons with a profession modeled as recource

```
SELECT DISTINCT ?gender COUNT(?person) as ?count
        WHERE {
         ?person a dbo:Person.
         ?person <http://xmlns.com/foaf/0.1/gender> ?gender.

         {?person dbo:birthPlace ?country.}
        UNION {?person dbo:birthPlace ?place.
         ?place dbo:country ?country.}
        UNION {?person dbo:deathPlace ?country.}
        UNION {?person dbo:deathPlace ?place.
         ?place dbo:country ?country.}
        UNION {?person dbo:livingPlace ?place.
         ?place dbo:country ?country.}
        UNION {?person dbo:nationality ?country.}
        UNION {?person dbo:citizenship ?country.}

         { ?person <http://dbpedia.org/ontology/profession>
         ?profession.}
        UNION
         {?person <http://dbpedia.org/ontology/occupation>
         ?profession.}
        UNION
         {?person <http://dbpedia.org/ontology/
         dbo:otherOccupation> ?profession.}
        UNION
         {?person <http://dbpedia.org/ontology/occupation> ?occ.
         ?occ <http://dbpedia.org/ontology/title> ?value.}
        UNION
         {?person <http://dbpedia.org/ontology/
         dbo:otherOccupation> ?occ.
        ?occ <http://dbpedia.org/ontology/title> ?title.}

        FILTER(
                (
                ?country= <http://dbpedia.org/resource/
                United_States> ||
                ?country= <http://dbpedia.org/resource/
                Canada>
                )
                &&
                (( ?profession= <http://dbpedia.org/resource/
                Lawyer> ) || (CONTAINS(LCASE(?title), 'lawyer'))
                )
        )
}
```

The taxonomy of DBpedia structures the class person in a hierarchical manner. Some professions we are looking for are present within this structure. It's possible to query those subtypes instead of the class Person.

The advantage of this version is that there is no inconsistent usage of resources as we mentioned before. There is only one way to set a superclass and therefore, the potential number of issues in the manually created data is lower. Compared to the other query (Listing 3.1), the number of results is substantially higher. Not all professions are mapped in the taxonomy, so we focus on the following classes, which are hierarchically structured:

- Person
  - Athlete
    - Swimmer
    - Volleyball player
  - Artist
    - Actor
  - Politician
  - Writer
  - Criminal
  - Engineer
  - Scientist

Some of the classes are subclasses of another one, like 'Athlete' and 'Swimmer'. It follows that the results queried for the subclass are included in the results of the superclass (i.e. results for 'Athlete' contains also the results for 'Swimmer'). Compared to the first query, there is no need to combine and filter attributes to find the profession, because it is already defined by the class we are querying (in this case 'Athlete'). The rest stays basically the same (see Listing 3.2).

Listing 3.2: SPARQL query to retrieve persons of the subclass 'Athlete'.

```
SELECT DISTINCT ?gender COUNT(?person) as ?count
WHERE {
        ?person a dbo:Athlete.
        ?person <http://xmlns.com/foaf/0.1/gender> ?gender.

        {?person dbo:birthPlace ?country.}
        UNION {?person dbo:birthPlace ?place.
                ?place dbo:country ?country.}
        UNION {?person dbo:deathPlace ?country.}
        UNION {?person dbo:deathPlace ?place.
        ?place dbo:country ?country.}
        UNION {?person dbo:livingPlace ?place.
                ?place dbo:country ?country.}
        UNION {?person dbo:nationality ?country.}
        UNION {?person dbo:citizenship ?country.}

        FILTER(
        ?country= <http://dbpedia.org/resource/United_States> ||
        ?country= <http://dbpedia.org/resource/Canada>
        )
}
```

3.4.2 Queries: US citizens of foreign descent

In Section 3.2.4, we analysed the origin of US citizens and their professions for the KB YAGO. Data on this topic is present in DBpedia as well. The professions are staying the same as in the previous query. Instead of nationality, we retrieve the number of people living in the US whose origin were in different countries. Again, the results are grouped according to the regions (see Listing 3.3).

Listing 3.3: SPARQL query to retrieve American Writers from Austrian descent.

```
SELECT DISTINCT ?gender COUNT(?person) as ?count
WHERE
{{ ?person a dbo:Writer. ?person dct:subject
        <http://dbpedia.org/resource/Category
                :American_people_of_Austrian_descent>}
  UNION
  { ?person a dbo:Writer. ?person dct:subject
```

```
        <http://dbpedia.org/resource/Category
                :American_people_of_Belgian_descent>}
}
```

Of course, the substring 'Austrian' and 'Belgian' have to be replaced with the according country names. 'UNION' clauses are used to group the regions as defined in section 3.1

3.4.3 Queries: Cancer diseases per region

All previously retrieved informations concerning professions are dependent on cultural influences. Diseases like different kinds of cancer should not be influenced by culture. In order to compare cultural dependent with independent biases, we want to retrieve all persons suffering from cancer. The class 'Person' has different attributes relating to the death cause of a Person. To retrieve the data as comprehensive as possible, we combine those attributes with 'UNION' clauses.

Listing 3.4: SPARQL query snippet to retrieve people suffering from cancer.

```
{?person dbo:deathCause <http://dbpedia.org/resource/Cancer>.}
UNION
{?person dbp:causeOfDeath <http://dbpedia.org/resource/Cancer>.}
UNION
{?person dbp:deathCause <http://dbpedia.org/resource/Cancer>.}
UNION
{?person dct:subject dbc:Cancer_survivors.}
```

The chances of survival is strongly dependent on the country where the diseased live and its medical options. [Allemani et al., 2015]. Therefore, the survivors of a cancerous disease are added to the query as well. We selected several cancer types, based on the number of occurrences in DBpedia:

- Pancreatic cancer
- Kidney cancer
- Testicular cancer
- Breast cancer
- Colorectal cancer
- Lung cancer
- Skin cancer
- Cervical cancer
- Liver cancer
- Ovarian cancer

3.5 Yago: Detecting Biases using Embeddings

The third method we want to present, is based on multidimensional vector representations called embeddings. Deepwalk [Perozzi et al., 2014] is designed to derive such embeddings from a graph using Word2Vec. It creates lists of nodes using random walks. A random walk starts from a node, chooses randomly a neighbour, 'walks' to that node and selects the next one, until the defined length of the path is reached. Those lists are used as input for Word2Vec, which computes the resulting n-dimensional vectors called embeddings, using a neural network. In this thesis, we used the default setting producing a 64-dimensional vector. The walks are starting from each single node of the graph. The number of walks per node and the walk length can be defined and remains the same for each run.

Furthermore, Deepwalk provides the possibility to restart walks with a definable probability. This means the walk can jump back to the start node at any point of the walk with given probability and continue its walk from there, until the walk length is reached. [Perozzi et al., 2014]

$$\begin{gathered}\text{with restart: } Artist \rightarrow John\ Doe \xrightarrow{\text{(restart)}} Artist \rightarrow Jane\ Doe \rightarrow female \\ \text{without restart: } Artist \rightarrow John\ Doe \rightarrow Person \rightarrow Athlete \rightarrow Swimmer\end{gathered}$$

We perform random walks on the same graph, we built in section 3.2.1. The resulting list of nodes is used as an input for Word2Vec, which is based on a neural network. Word2Vec's architecture consists of nodes arranged in multiple layers (see Figure 3.4). The input layer consumes our word list, followed by two hidden layers with nodes performing linear activation functions. Finally, there is the output layer where a softmax function is applied. Each node of a single layer is fully connected through weighted edges with the nodes of the following layer, building a bipartite graph . The addressed functions are used to adjust the weights of the edges between the nodes of the single layers. The goal of this procedure is to predict the neighbours of the input token. [Mikolov et al., 2013b] This so called Skipgram model takes a sentence or, in our case, a sequence of nodes as an input. With a window size of 5, the next two tokens to each side of the input token are considered (see Figure 3.5).
Example:
Given the following sentence and a window size of five: "Once upon a time there was a farmer and his wife.". Word2Vec will try to predict the two tokens to the left and two tokens to the right of the input word. Lets assume, the input word is 'a'. The expected result is: 'Once', 'upon', 'time', 'there'. Results for further inputs are:

Input: $time \rightarrow upon, a, there, was$
Input: $there \rightarrow a, time, was, a$
Input: $was \rightarrow time, there, a, farmer$
etc.

The neural network is learning by back-propagating errors and adjusting the weights of its edges [Rumelhart et al., 1988].

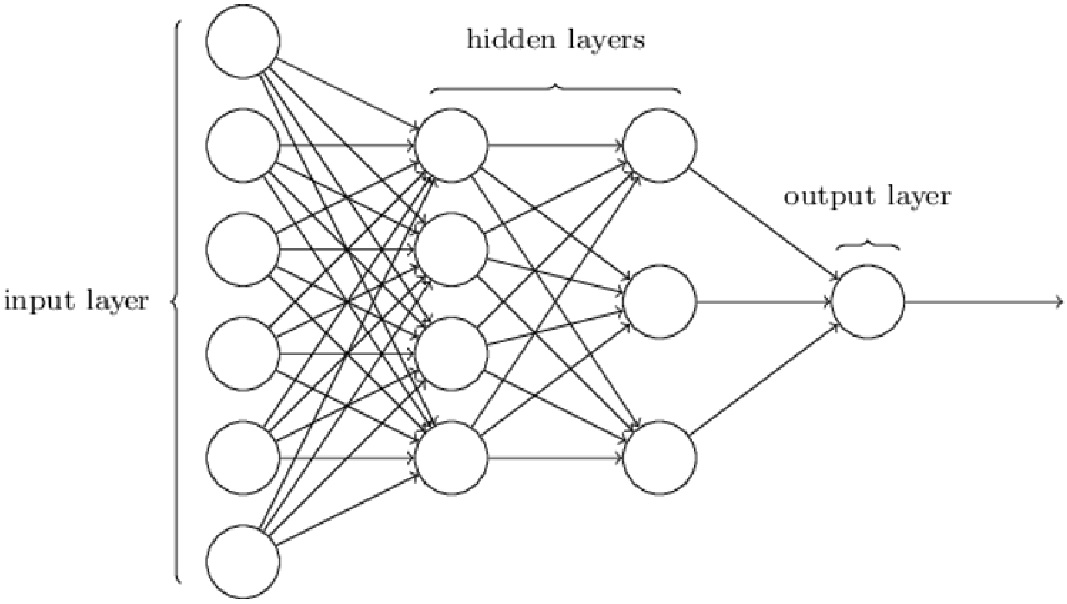


Figure 3.4: Architecture of a neural network, Source: http://www.shivambansal.com.

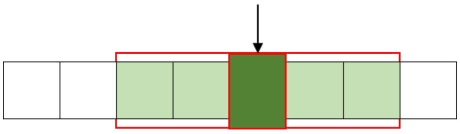


Figure 3.5: Skipgram model with window size 5.

The output file of this procedure contains a n-dimensional vector for each node in the graph. To compare those vectors, we compute the cosine similarity (Formula 3.1) between selected combinations. We want to see if a vector representing a certain profession is closer to the vector representing the genders 'male' and 'female'.

$$\text{similarity} = \cos\theta = \frac{A \cdot B}{||A||_2 ||B||_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3.1}$$

3.6 Yago: Mitigating biases using Embeddings

Ideally the cosine similarity of a profession and 'male' and the similarity of the same profession and 'female' are close to each other. This can be considered as unbiased. A big difference regarding the similarities indicates a bias. Our intention is to align the vectors, so that the cosine similarity is on a similar value. We don't want to change the functionality of Word2Vec, but we can adjust the input, i.e. the random walks.

The goal is to build a similar amount of walks from a profession to both genders. For this reason, the restart algorithm of the random walks is modified to reject and restart a walk under certain circumstances. We want to do that, if its start and endpoint are representing a known bias, e.g. *engineer* → *male*. With a probability α, we want to reject the path and restart it again (see Figure 3.6). Lets assume, that in a first try, 400 paths are reaching node A, 600 paths are reaching node B. 20% of the paths are rejected and start again from the same node. Because of the uneven distribution, more path from node B are restarted. The number of accepted paths for both nodes are more balanced after some iterations. The value of α affects the level and speed of the approximation massively.

As a refinement, we can assign different α values to the target nodes. Lets look at the same example with 1000 random walks and the same distribution as before after a first iteration. For node A, which is reached by a lower number of walks, we assign a lower α (10%), whereas B rejects 50% of all paths in each iteration (see Figure 3.7). After 3 iterations, node A already overhauled node B regarding the accepted number of walks. By choosing a big difference of α, it's easy to overstep the goal to approximate the number of walks for both nodes. Therefore, it is crucial to choose an appropriate probability to reject and restart the walks. The presented example is a simplified representation of the real graph. The start and intermediate nodes are connected to multiple other nodes as well. The restarted walk don't have to reach one of the target nodes again. It's likely, that the restarted paths won't take the same path again and stops at a 'foreign' end node. In that case, the walk will be accepted and not restarted again.

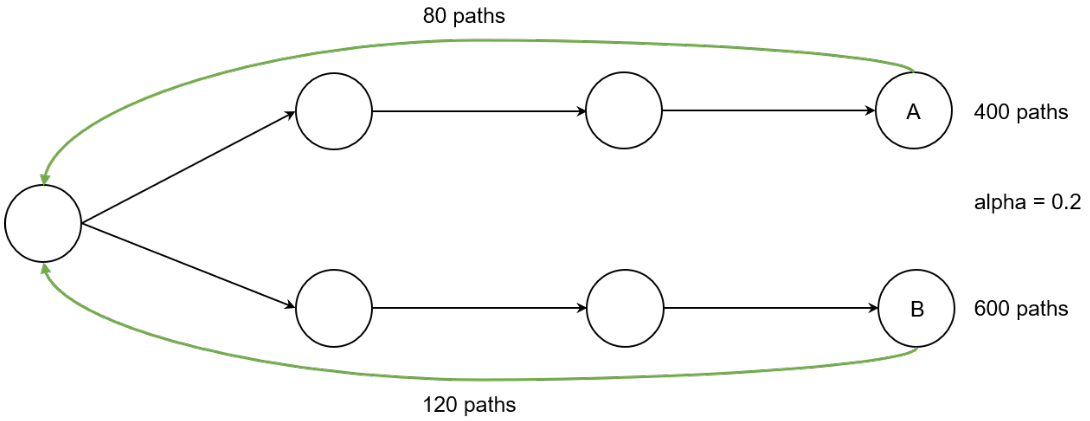
The target nodes in our graph are representing the genders. The node 'male' has the highest degree in the whole graph, but also 'female' has a very high value. The probability, that a random walk ends at one of those two nodes is comparably high. To find a suitable α, we want to include the degree of the target node. We want a higher α for nodes with a high node degree, in this case 'male', than for a node with a lower degree. The value of α in our algorithm is calculated with the following formula, whereas x can be adjusted to find the best performing setting.

$$\alpha = \frac{1}{degree^x}$$

Listing 3.5: modified random walk method.

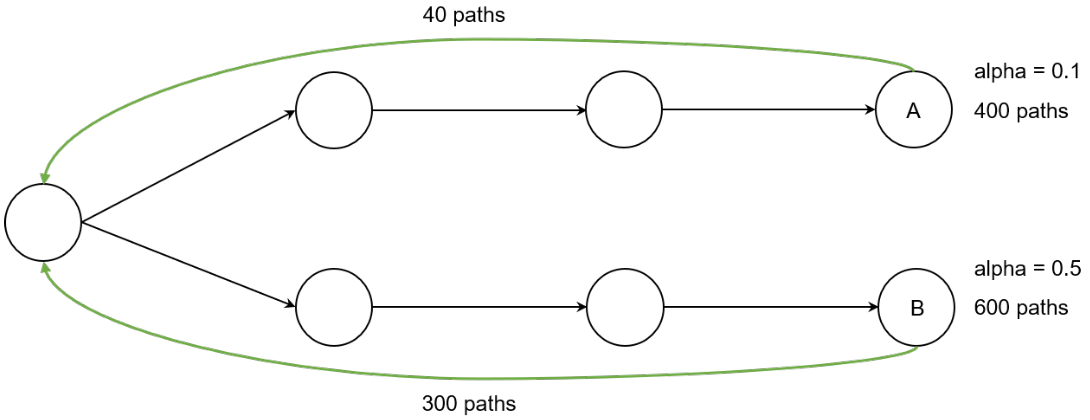
```
def random_walk(G, path_length, restart=0, rand=random.Random(), start=None):

        l = [
                (2121700, 19),  #engineer male
                (2160095, 21),  #dancer female
                (2434282, 19),  #chemist male
                (2434282, 21)]  #chemist female

        if start:
                path = [start]
        else:
                # Sampling is uniform w.r.t V, and not w.r.t E
                path = [rand.choice(G.keys())]

        while len(path) < path_length:
                cur = path[-1]

                if len(G[cur]) > 0:
                        if rand.random() >= restart:
                                path.append(rand.choice(G[cur]))
                        else:
                                path = [path[0]]
                else:
                        break

                if path_length == len(path):
                        nodes = (path[0], path[-1])
                        bias_nodes = ('', '')
                        if nodes in l:
                                bias_nodes = (nodes)
                                if (path[0] is bias_nodes[0] and path[-1] is bias_nodes[1]):
                                        alpha = 1/(G.degree(cur)**0.15)
                                        if rand.random() < alpha:
                                                path = [path[0]]
        return [str(node) for node in path]
}
```

For our tests, we will set our alpha only for the biased nodes, i.e a node representing a profession, which has a higher cosine similarity compared to both genders. In our case, we assign a value to alpha for *engineer* → *male*, *dancer* → *female* and *chemist* → *female/male*. Chemist is used as a control group to see the influence, if we apply this method to both nodes. All other walks are untouched and will be treated as usual. Of course, it is possible, that the random walk algorithm generates walk starting from a gender to one of those professions, but the chances are very low. The nodes representing a gender have very high degrees and we are starting only a rather small number of walks from each node. The chosen professions (engineer, dancer and chemist) have only a few entries compared to the total number persons in the dataset.



Iteration	Node A	Node B
1	320	480
2	64	96
3	12.8	19.2
Total	396.8	595.2

Figure 3.6: Random walk with restart: 1000 walks and the distribution after 3 iterations with a restart probability of 0.2.



Iteration	Node A	Node B
1	360	300
2	122.4	102
3	42.336	35.28
Total	524.736	437.28

Figure 3.7: Random walk with restart: 1000 walks and the distribution after 3 iterations with different restart probabilites for each node.

4

Results

In this chapter, we want to present the results of the previously introduced methods of Chapter 3. The order of the subsections is following the structure of the previous chapter. Additionally, we added a section comparing the different methods and results with each other.

4.1 Yago: Enumeration

In this section, we present the results for the knowlede base YAGO. We will analyse biases regarding different professions, the gender, possible regional differences or the race.

4.1.1 Gender to Profession

We have already seen the structure of the retrieved paths in section 3.2.2 and also the adjustments performed to extract meaningful results. The number of paths from 'male' to a professions is mostly higher than for 'female'. Overall, the number of retrieved path for 'male is ~2.5 times higher than for 'female'. This is no surprise as the degree of the two nodes is very different.

To make the results comparable, the number of retrieved paths is normalized by calculating the ratio.

$$Ratio = \frac{number\ of\ retrieved\ paths}{\frac{node\ degree\ of\ gender}{2}}$$

The node degree is divided by 2, because we doubled the degree while building the graph by adding the inverted edges. The Table 4.1 and Figure 4.1 show, that certain professions are present more often in our paths than others. Paths for writers and artists, followed by athletes were found most often. This is no surprise, because professions in this fields have usually a higher public attention than professions related to science or other fields. Because the source of the data of YAGO, Wikipedia, is containing mainly information about people of public interest. The chance that an actor, singer or author has an own article in Wikipedia is much higher than for a computer scientist or an engineer.

Table 4.1: Number of paths found per gender and the Ratio (number of paths per gender and profession divided by the node degree of the gender).

Profession	**male**	**Ratio Male**	**female**	**Ratio female**
Politicians	12394	0.013261288	304	0.00174448
Engineer	10662	0.011408089	349	0.002002709
Computer Scientist	3645	0.003900064	549	0.003150393
Chemist	6545	0.007002996	593	0.003402883
Physicist	10767	0.011520437	776	0.004453014
Biologist	3597	0.003848705	779	0.004470229
Psychologists	3242	0.003468864	1017	0.005835973
Mathematician	14949	0.015995078	1211	0.006949226
Dancer	2621	0.002804408	3996	0.022930726
Volleyball Player	25210	0.026974107	4313	0.024749805
Swimmer	17981	0.019239247	14473	0.083052151
Poet	42177	0.045128397	15591	0.089467704
Artist	112610	0.120490049	32375	0.185781343
Writer	168199	0.179968971	68911	0.395440252

The categories politician, chemist, engineer, physicist and mathematicians show a higher ratio for males. Others like Volleyball Player, Biologist, or Computer Scientist have a similar ratio for both genders whereas writer, artist, poet, swimmer, dancer and psychologist have a higher ratio for females.

In the category dancer, the total number of paths is higher for females, all others are dominated by males. The ratio regarding swimmers shows a unexpectedly big difference between males and females. A first assumption was, that synchronized swimming is included in this category, what is not very popular among males. Although that may be the case, the difference is just too big to be explained that way, because only 566 paths are crossing nodes related to synchronized swimming. That's only ∼4% of the discovered paths. When we look more closely at the at the data, we can see, that a female swimmer is connected to various other nodes, like for example <wordnet_swimmer_110683349>, <wikicat_Danish_swimmers>, or <wikicat_Olympic_swimmers_of_Denmark>. This results in a high number of paths. Compared to the male swimmers, the number of paths per person is higher for females.

The nodes between our start and end node are containing additional information. Some of them are describing the century, in which the person lived. The most complete set of such nodes are present for the categories poet and mathematician. In both professions, the total number of men is higher. Over a long time period, not a single female mathematician is present in the dataset. The first entries show up in the 18th century. Female poets show up earlier than mathematicians, but only a few. For both categories, the number of females are starting to rise from the 18th-century, but also the number of males is increasing (see Figures 4.2 and 4.3). There is a clearly visible male dominance in the historical data.

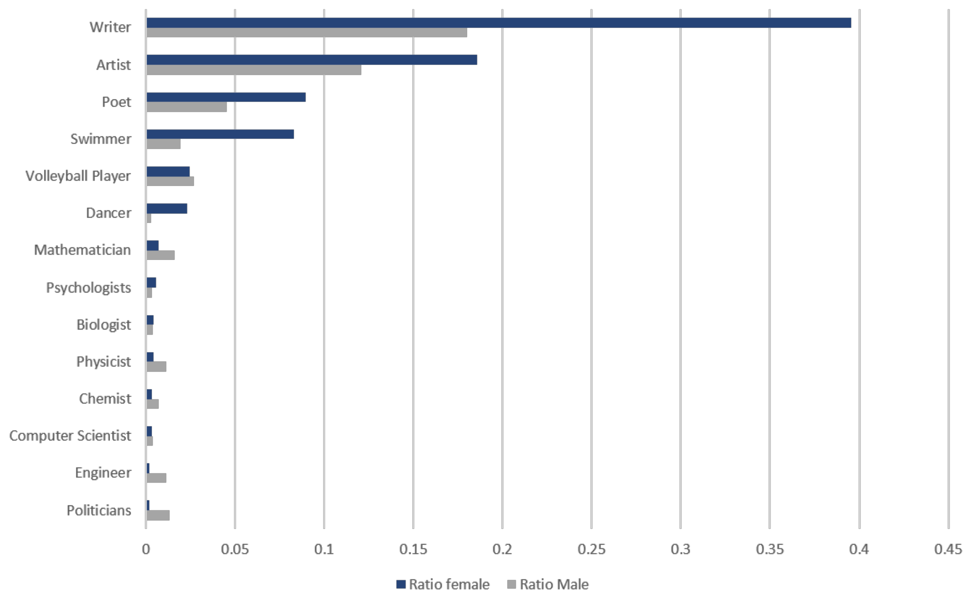


Figure 4.1: Ratios of the number of retrieved path per gender divided by the node degree of the gender.

4.1.2 Gender to Region

In order to make the comparison between the regions, we have to find first all paths for males and females and of course also all persons with the desired professions within this region. People from West Europe are strongly represented in YAGO, followed by East Europeans and North Americans (see Figure 4.4). The dominance of West Europe is rather surprising regarding the actual population of each region. Asia and Africa have a high population and also North America with a similar cultural background has significantly less entries in the KB than West Europe. The totals women's share is 30.5%, but in many regions, it is much lower. West and East Europe have the highest proportion of females with roughly 40%, whereas West Asia and Africa have the lowest with only $\sim$ 10%.

To normalize the number of retrieved path per region, we use the according total number for males respectively females. We divide the number of retrieved path per gender and profession by the number of total paths we retrieved per region. This ratio helps us to compare the values between the single regions. By watching the results for males, it sticks out, that some professions have a similar pattern while comparing the different regions. West Europe and North America have a rather high value, in comparison with the other regions for the professions artists, biologists, chemists, computer scientists, engineers and physicists. For females, it is similar, but the values are mostly lower. The difference between the regions, especially in science related fields is smaller. The Asian regions have a low number of paths for most professions. Writers and Poets have the highest appearance, but less than most other regions. It's visible, that the countries of

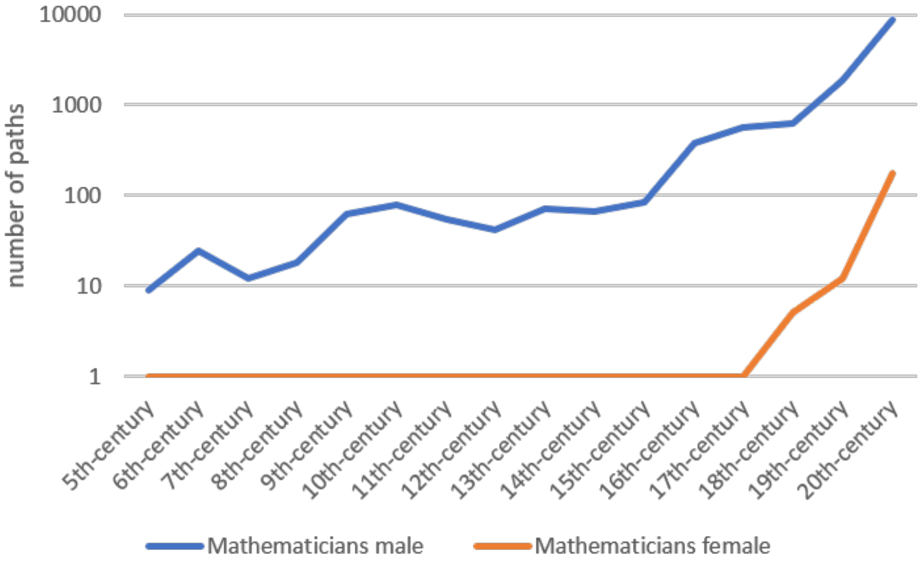


Figure 4.2: Number of paths per century for mathematicians.

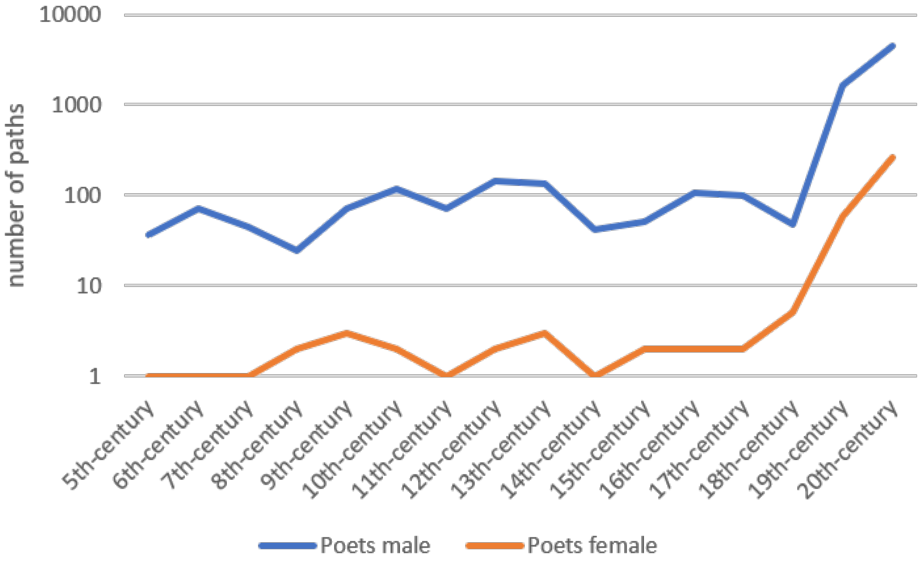


Figure 4.3: Number of paths per century for poets.

Asia, Africa and Oceania are under represented in this dataset.

In order to test, if the number of paths per gender and profession are independent, a Chi-squared test was conducted. Some regions have low counts for certain professions (lower than 5). In that case we performed the Fisher's exact test instead, which is more conservative than the Chi-squared test, but more suitable for low numbers. The outcome of this tests is the same for each single region: The p-values are very low (between 1.8508E-27 and 0). Therefore, we can reject the null hypothesis, that the number of retrieved paths for males and females are independent of the gender within the single regions.

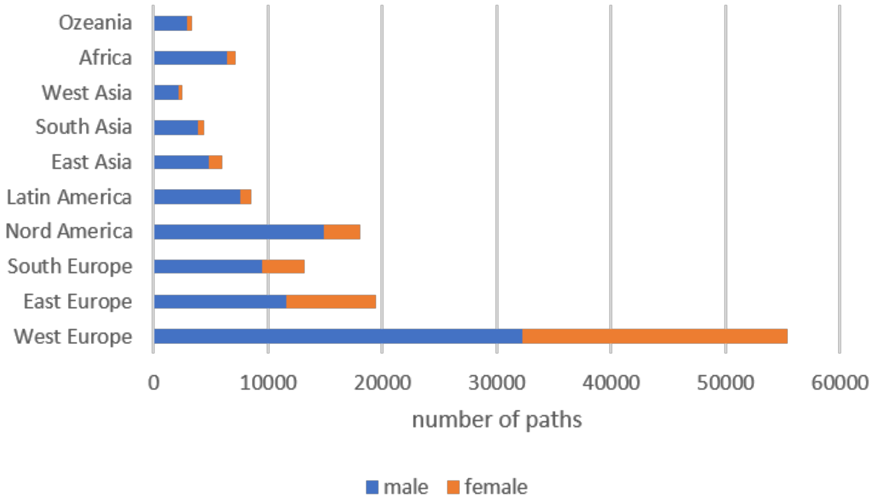


Figure 4.4: YAGO: Total number of paths found per region.

4.1.3 US citizens of foreign descent to Profession

We investigate the professions of Americans of foreign descent. Again, the number of people with West European origin is higher than for other regions (see Figure 4.6). Artists make up the biggest part of the paths, followed by writers. The other professions only have a very small proportion of all found paths, or are even non-existent in the results. Noticeable is the small number of Americans with an origin of a country from the region Oceania. The paths for African-American people have a surprisingly low number. In demographic statistics, the population of Latinos and African Americans was roughly 13% each in the year 2000, whereas the Asian diaspora was only about 3.6% [Grieco and Cassidy, 2001]. The amount of paths in our dataset for US citizens of Asian descent is relatively high. At the normalized number of paths, the people of south Asian descent attracts attention with a very high value of 0.9. This can be interpreted, that our professions, mainly artists and writers, are covering 90% of the total paths for this region.

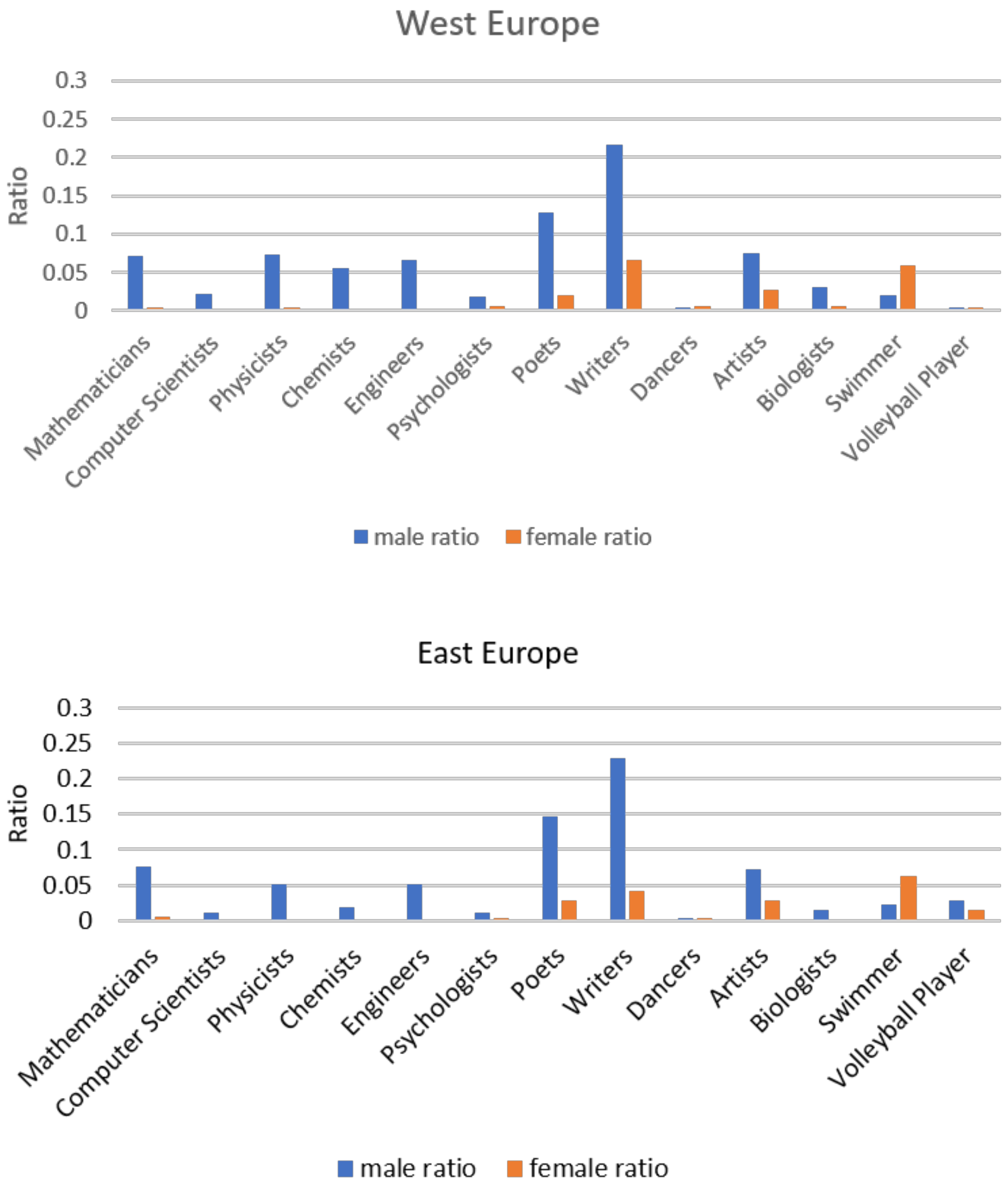
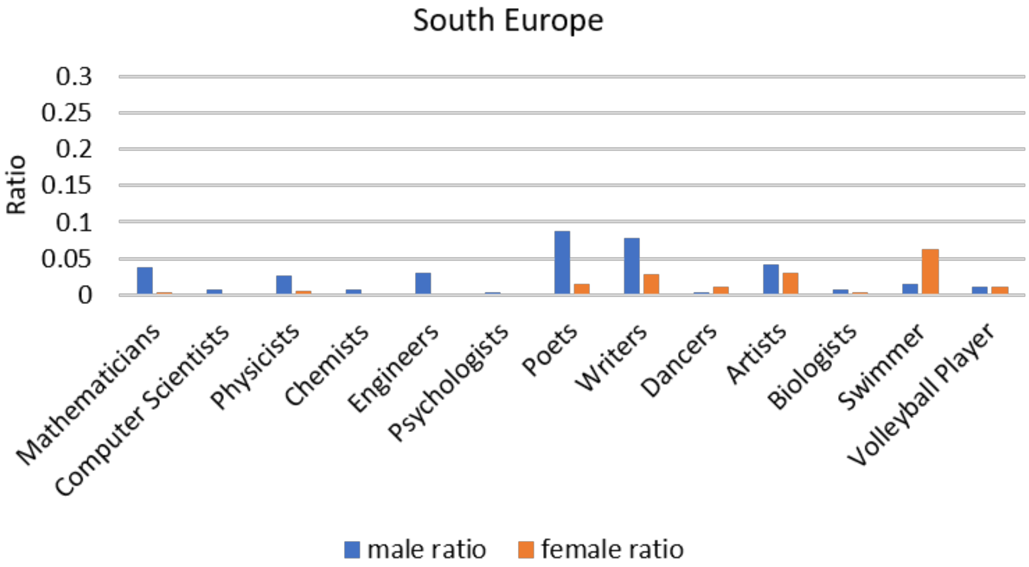
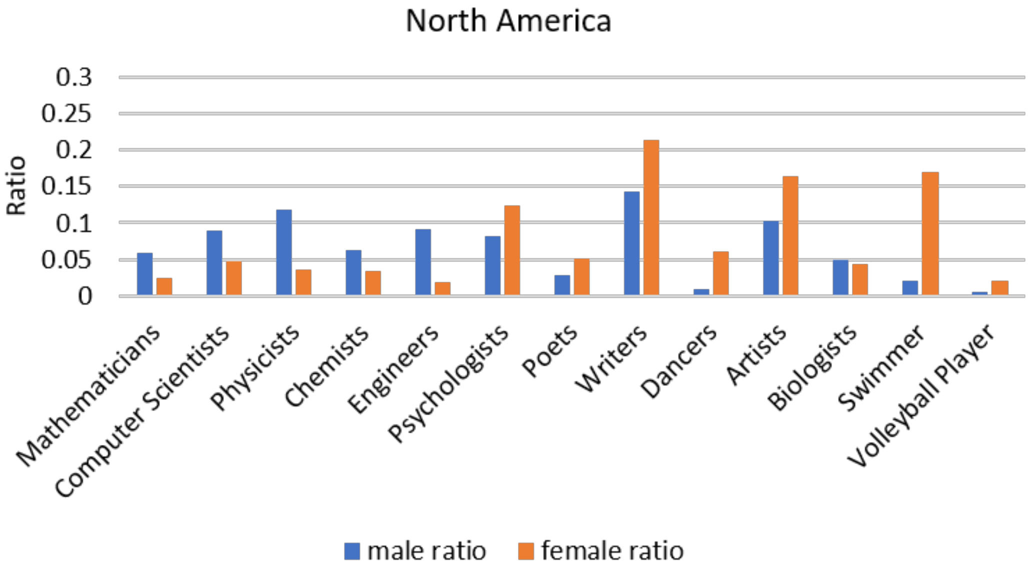
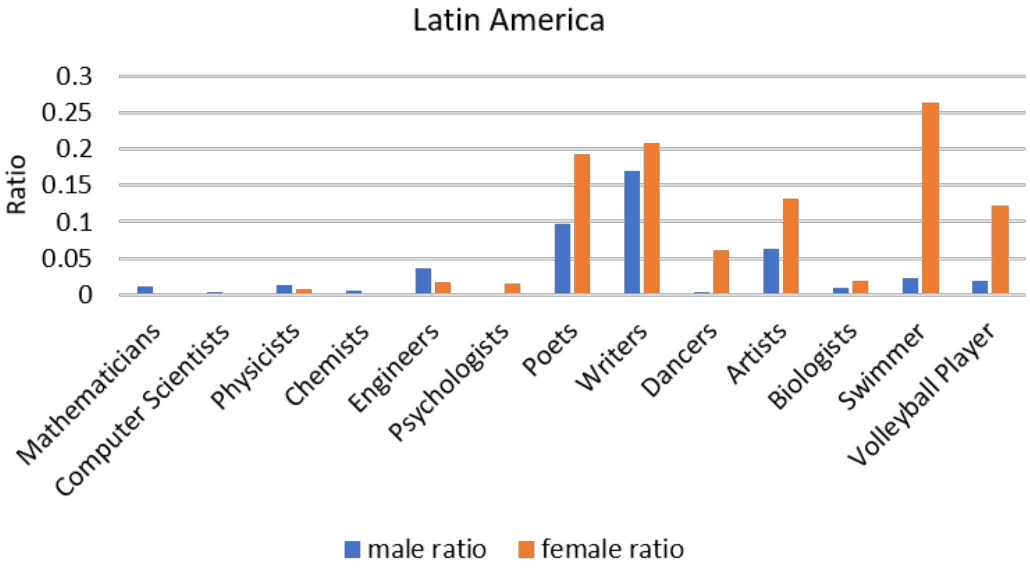
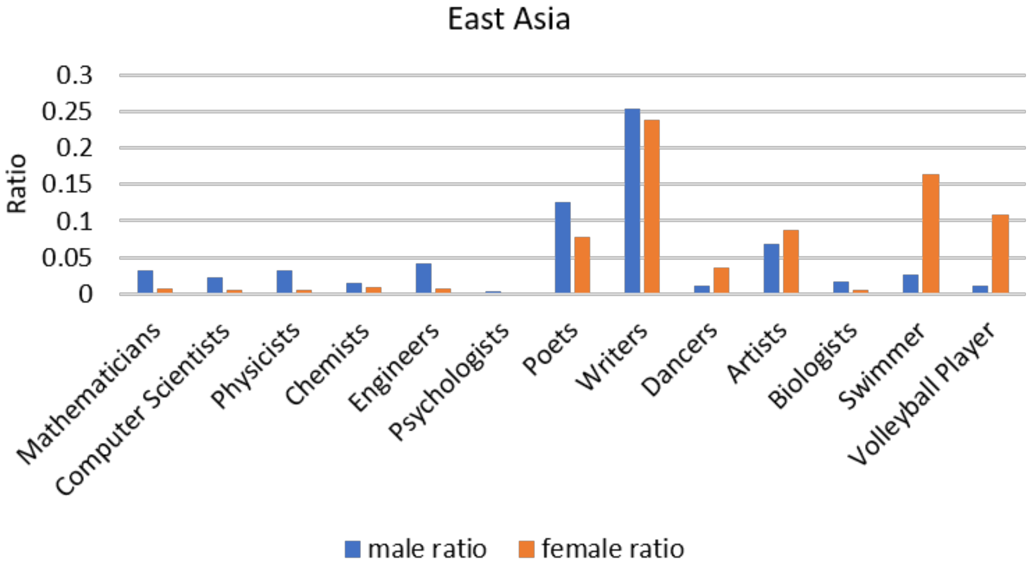


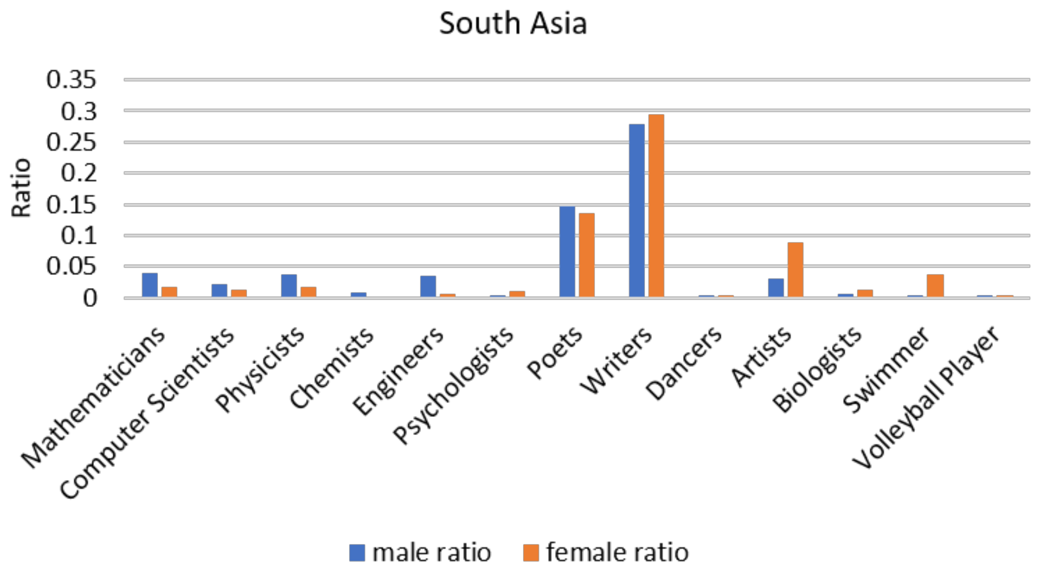
Figure 4.5: Yago: Ratios by region.

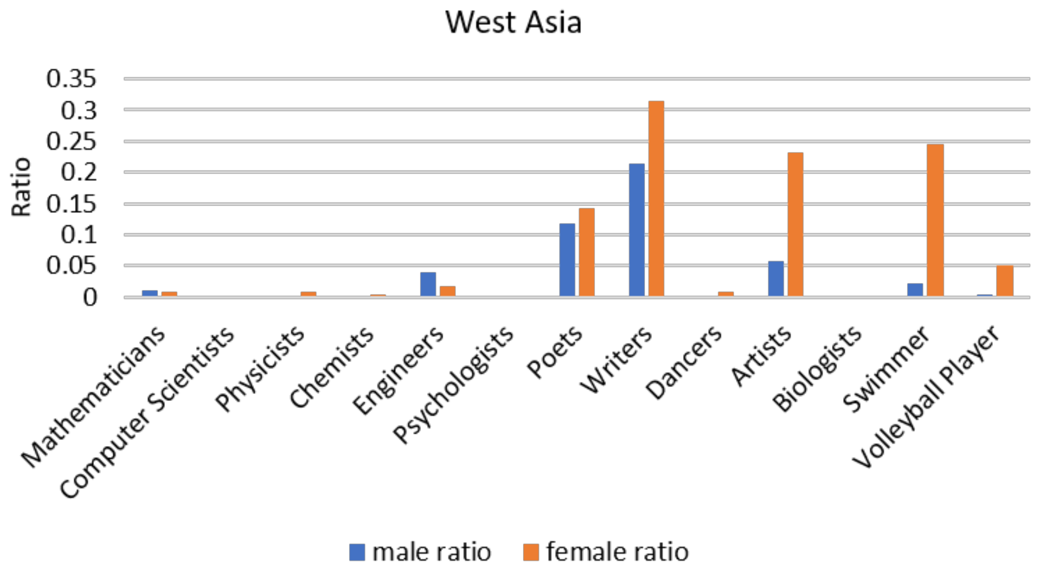

South Europe
Ratio
0.3
0.25
0.2
0.15
0.1
0.05
0
Mathematicians
Computer Scientists
Physicists
Chemists
Engineers
Psychologists
Poets
Writers
Dancers
Artists
Biologists
Swimmer
Volleyball Player
■ male ratio
■ female ratio

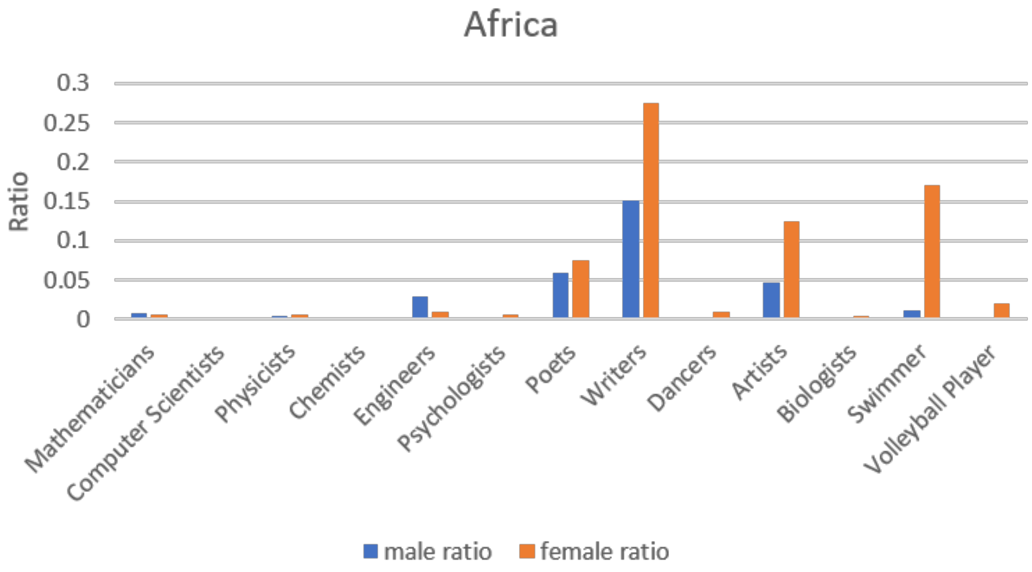

North America
Ratio
0.3
0.25
0.2
0.15
0.1
0.05
0
Mathematicians
Computer Scientists
Physicists
Chemists
Engineers
Psychologists
Poets
Writers
Dancers
Artists
Biologists
Swimmer
Volleyball Player
■ male ratio
■ female ratio

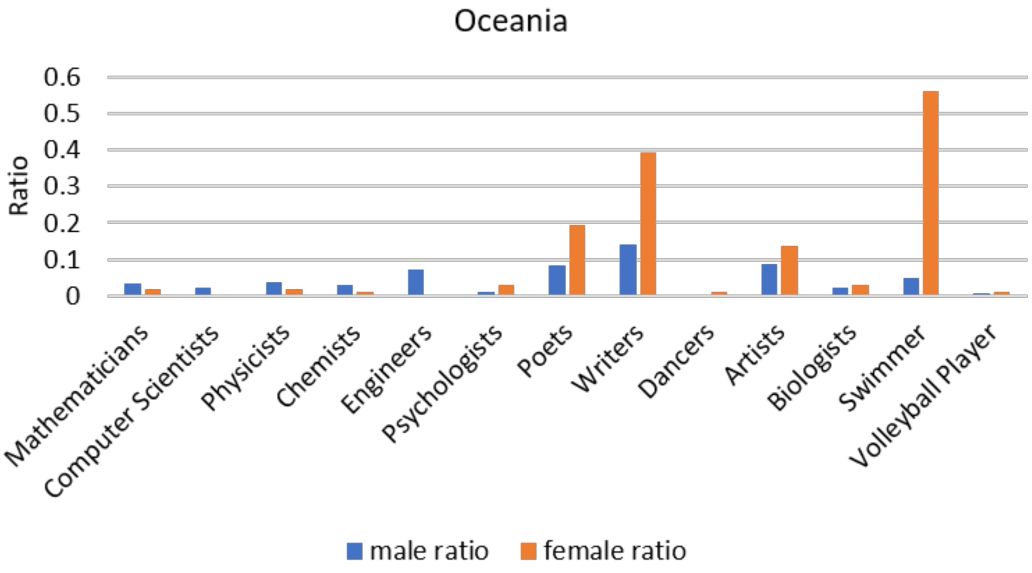

Latin America
Ratio
0.3
0.25
0.2
0.15
0.1
0.05
0
Mathematicians
Computer Scientists
Physicists
Chemists
Engineers
Psychologists
Poets
Writers
Dancers
Artists
Biologists
Swimmer
Volleyball Player
■ male ratio
■ female ratio


East Asia
Ratio
0.3
0.25
0.2
0.15
0.1
0.05
0
Mathematicians
Computer Scientists
Physicists
Chemists
Engineers
Psychologists
Poets
Writers
Dancers
Artists
Biologists
Swimmer
Volleyball Player
■ male ratio
■ female ratio


South Asia
Ratio
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Mathematicians
Computer Scientists
Physicists
Chemists
Engineers
Psychologists
Poets
Writers
Dancers
Artists
Biologists
Swimmer
Volleyball Player
male ratio
female ratio


West Asia
Ratio
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Mathematicians
Computer Scientists
Physicists
Chemists
Engineers
Psychologists
Poets
Writers
Dancers
Artists
Biologists
Swimmer
Volleyball Player
male ratio
female ratio


Africa
Ratio
0.3
0.25
0.2
0.15
0.1
0.05
0
Mathematicians
Computer Scientists
Physicists
Chemists
Engineers
Psychologists
Poets
Writers
Dancers
Artists
Biologists
Swimmer
Volleyball Player
■ male ratio
■ female ratio


Oceania
Ratio
0.6
0.5
0.4
0.3
0.2
0.1
0
Mathematicians
Computer Scientists
Physicists
Chemists
Engineers
Psychologists
Poets
Writers
Dancers
Artists
Biologists
Swimmer
Volleyball Player
■ male ratio
■ female ratio

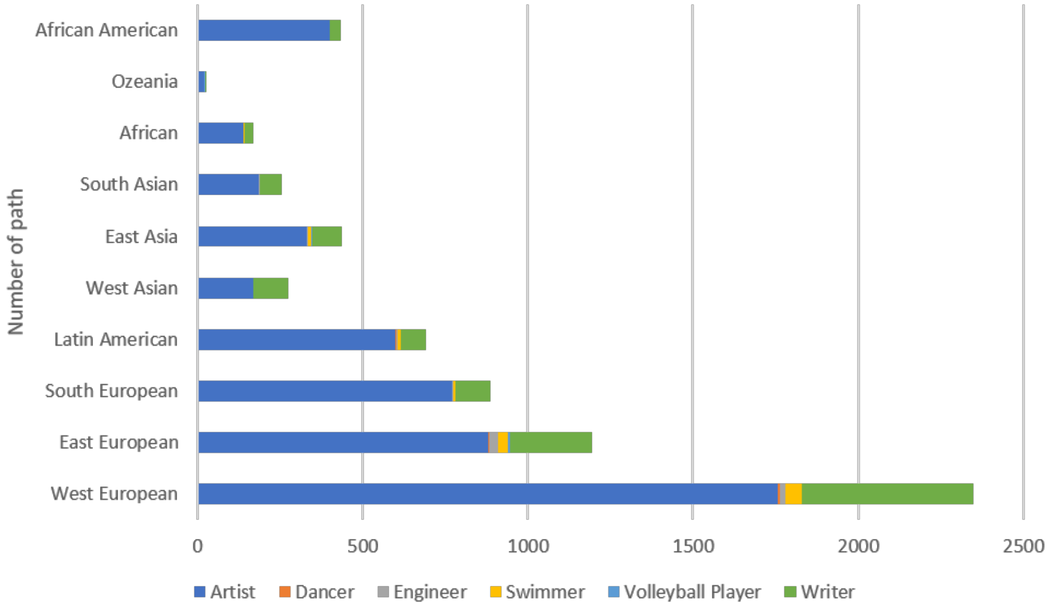


Figure 4.6: Yago: Number of paths for americans of foreign descent grouped by region.

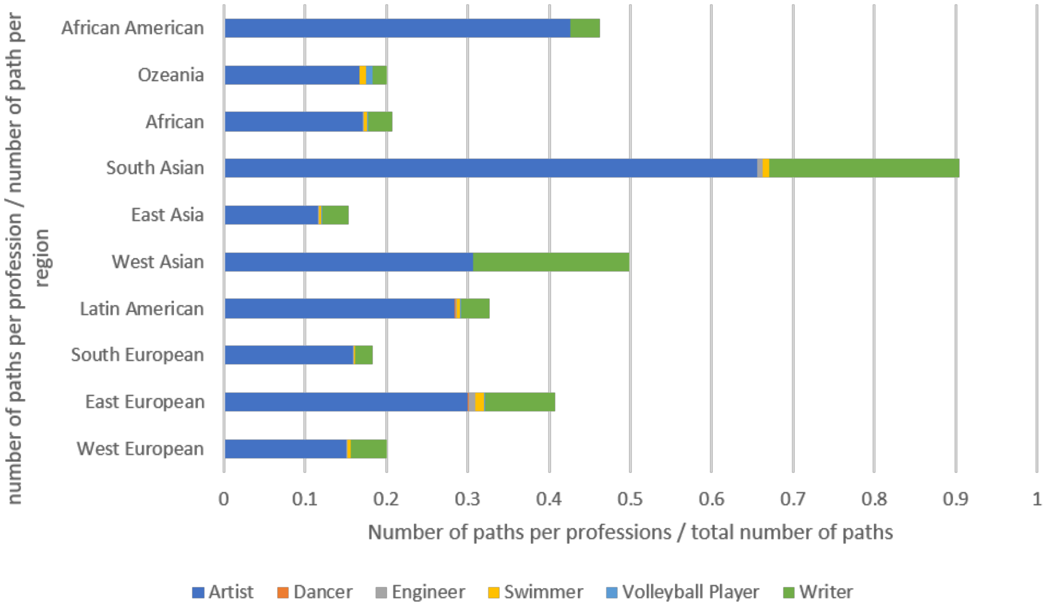


Figure 4.7: Yago: Ratio for americans of foreign descent grouped by region.

4.2 DBpedia: SPARQL

There are two different ways to query the professions of a person. In the first case, we want to find a profession modelled as an attribute of a person, in the second one, we lookup subclasses of the class 'Person' defined in the taxonomy of DBpedia. Overall, we can say, that we can find more results with the second approach. The attributes of a person are often not properly set or not set at all. Frequently, we can find equivalent relations to attributes like 'dbo:deathCause', 'dbp:causeOfDeath', 'dbp:deathCause', all of them linking an entity of a person to classes representing a death cause like a disease. This inconsistent usage of attributes makes it difficult to retrieve all relevant data. We solved this issue by UNION clauses in our queries. The concept of subclasses is less prone to errors regarding the manually created data from Wikipedia. The following results are the outcome of the queries containing the subclasses of 'Person'.

4.2.1 Gender to Profession

West Europe and North America have the highest total number of persons again. The most present profession of each region is Athlete, followed by Artist. Criminal and Engineer have the lowest number of entries. The ratio is computed the same way as in Section 4.1.2. We divide the number of persons per gender and profession by the total number of persons per gender (1'147'258 males, 217'962 females). The number of males is ~ 5.2 times higher compared to females and even in every retrieved profession, the males outnumber the females. The picture changes, when we look at the calculated ratios. Females have higher ratios for the professions Artist, Writer, Swimmer, Volleyball Player and Actor. Males have a higher ratio for Athletes, Engineers, Criminals and Scientists.

Table 4.2: Number of persons found per gender and the Ratio (number of persons per gender and profession divided by the total number of persons per gender).

male	**male Ratio**	**female**	**female Ratio**	
Athlete	393'472	0.342967319	46'131	0.211646984
Artist	68'139	0.059392918	22'992	0.105486277
Politician	49'153	0.042843894	4'267	0.019576807
Writer	32'888	0.028666612	10'831	0.049692148
Swimmer	5'067	0.004416618	4'239	0.019448344
Volleyball Player	4'756	0.004145537	4'678	0.021462457
Actor	4'443	0.003872712	4'334	0.019884200
Criminal	1'972	0.001718881	238	0.001091933
Engineer	1'489	0.001297877	13	0.000059643
Scientist	38'872	0.033882527	4'011	0.018402290

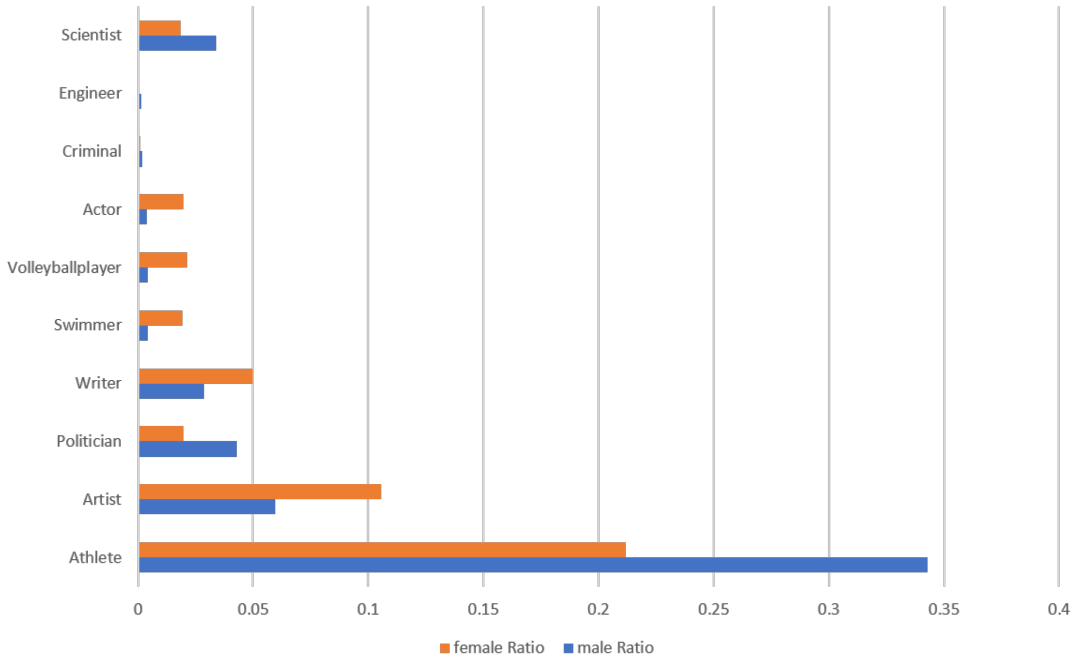


Figure 4.8: Ratios of the number of persons per gender divided by the total number of persons per gender.

4.2.2 Gender to Region

Every region has a much higher number of males, which is between 3.3 (East Asia) and 6.7 (South Europe) times higher than the number of females. This indicates again a strong bias towards males, independent of the region. The highest number of persons per region is West Europe with 459’922, followed by North America with 336’514. The lowest number of persons has the region West Asia with 15’543.

Athletes make up the biggest part in each region by far, with a higher male ratio (see Figure 4.9). One exception is East Europe, where the females ratio is slightly higher. The dominance of this profession can be explained with the popularity of sports all over the world and the public interest. Artists and Writers have also a high ratio in most regions, but the difference from the female to the male ratio is varying from one region to another. Mostly, the females have a higher ratio.

Criminals and Engineers have the lowest number of entries, but they have all a higher male value for both, the total number of entries and the ratio. The amount of Scientists is rather small, compared to professions like Athlete, Writer or Artist. Whereas males have a higher ratio in European, American and Asian regions, females have slightly higher ratios in Oceania and Africa for Scientists.

Like for YAGO, we performed a Chi-squared test to evaluate, if the retrieved number of persons is independent of gender or not. The results are basically the same as for YAGO: all p-values are extremely low (0 - 8.6143E-134), so we can reject the null hypothesis, that the number of persons per region is independent of the gender.

4.2.3 US citizens of foreign descent

Again, US citizens of foreign descent and their professions are evaluated. Like in YAGO, people with West European origin make up the largest part. Remarkable is the fact, that this region has the lowest total ratio of all regions (see Figure 4.10). In Section 4.1.3, we discovered, that two professions are dominating in every ethnic group in YAGO: Artists and Writers. In DBpedia, we have a more balanced distribution. Although Athletes and Artists have the highest numbers and ratios in every region, other professions have a remarkable share. This wasn't the case in YAGO. The ratios of the single professions are closer to each other compared to YAGO, what can be interpreted as less biased. The biggest differences between the regions can be found by Athletes. Oceania has a low number of persons, but the percentage of Athletes is very high. Compared to South Asia, which has also a low number of persons, the percentage is roughly 5 times higher. This means, a US citizen of Oceanian descent represented in DBpedia is 5 times more likley to be an Athlete than one with South Asian descent.

4.2.4 Cancer diseases per region

In order to investigate a potentially unbiased example, persons suffering from cancer were queried. This information is modelled as an attribute of a person. The results are not only including people died from cancer, but also cancer survivors, because otherwise, the medical possibilities of a country or the wealth of a diseased person could have an big influence. We wanted to exclude such social and regional biases as good as we can. The two most frequent types of cancers with roughly 500 entries each, are lung and breast cancer (see Figure 4.12). North America and West Europe have the highest number of diseased in each type. When we have a closer look at the ratios, the picture changes: The regional differences are much smaller. It is noticeable, that the share of West Europeans is much smaller, whereas East Asians have a proportional high share (see Figure 4.13).

Some types of cancer are strongly associated to a gender, which is not surprising. Entries for testicular cancer are only found for males, whereas ovarian cancer only for females. Breast cancer is mainly diagnosed for females, but also males can rarely affected. This is also reflected in our results (see Table 4.3).

Overall, we can say, that those entries are less biased regarding different regions. A possible cause for social biases could be privacy issues or, as mentioned before, the medical possibilities and wealth of a person. Looking at the genders, there are biases, but they are mostly the natural outcome of physical differences between man and women, as certain types of can merely occur for a particular gender.

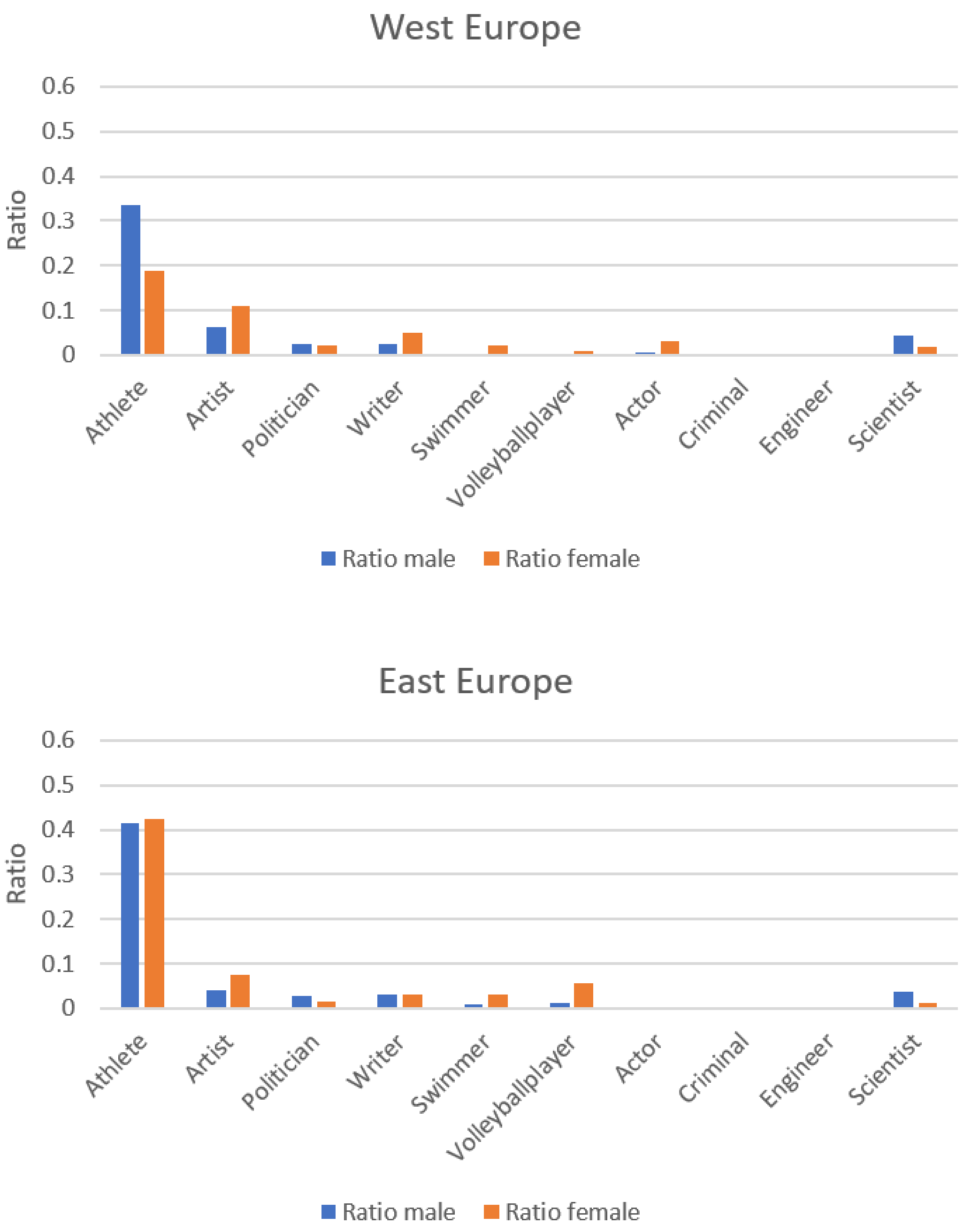
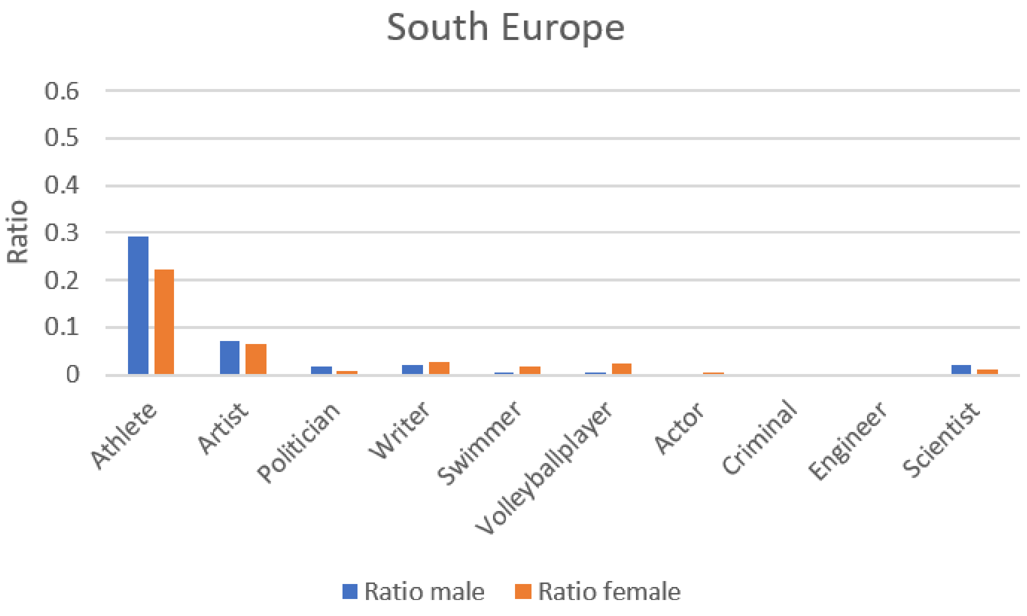
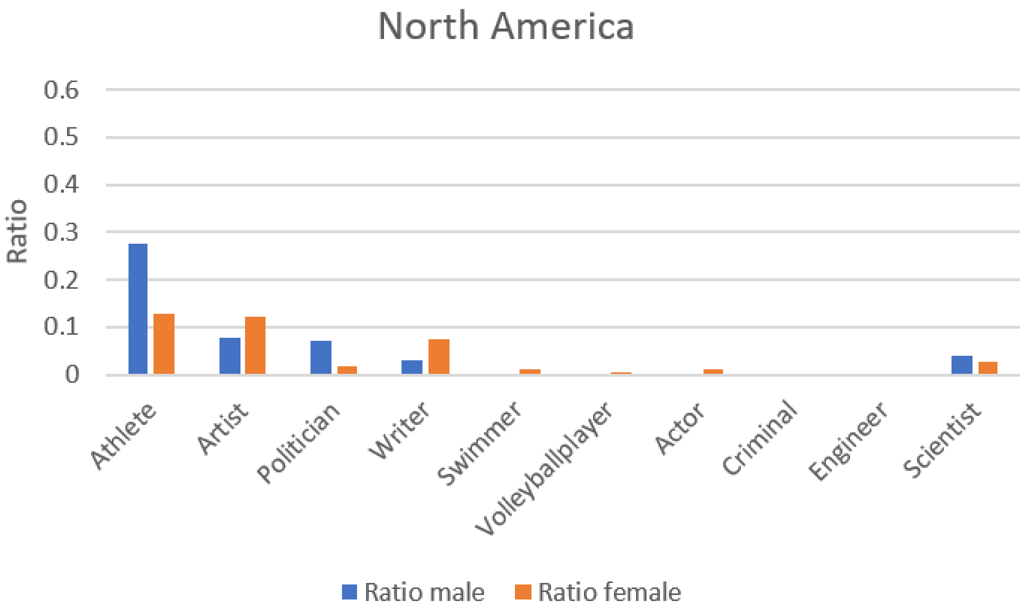
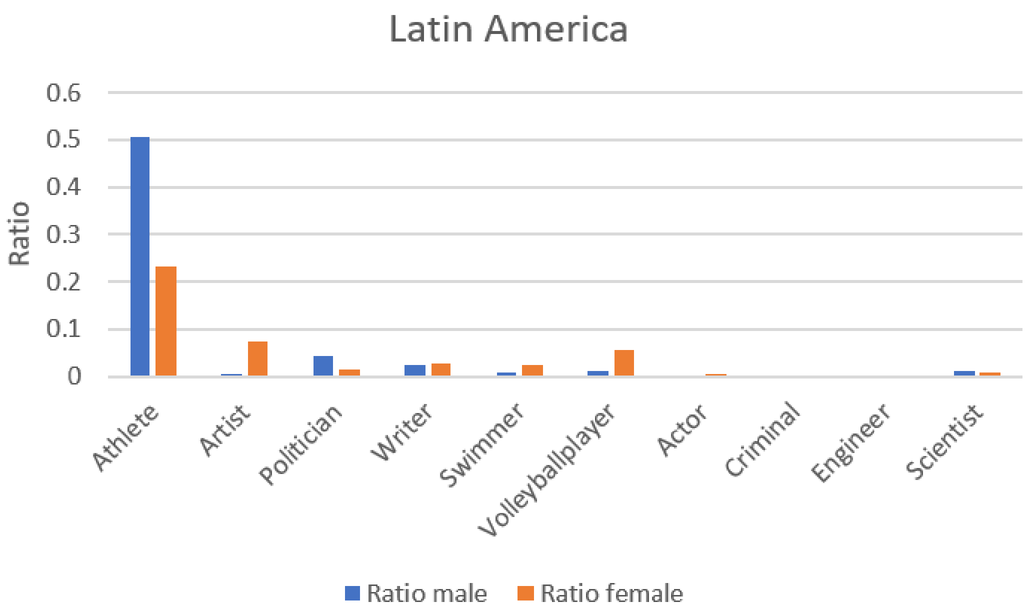
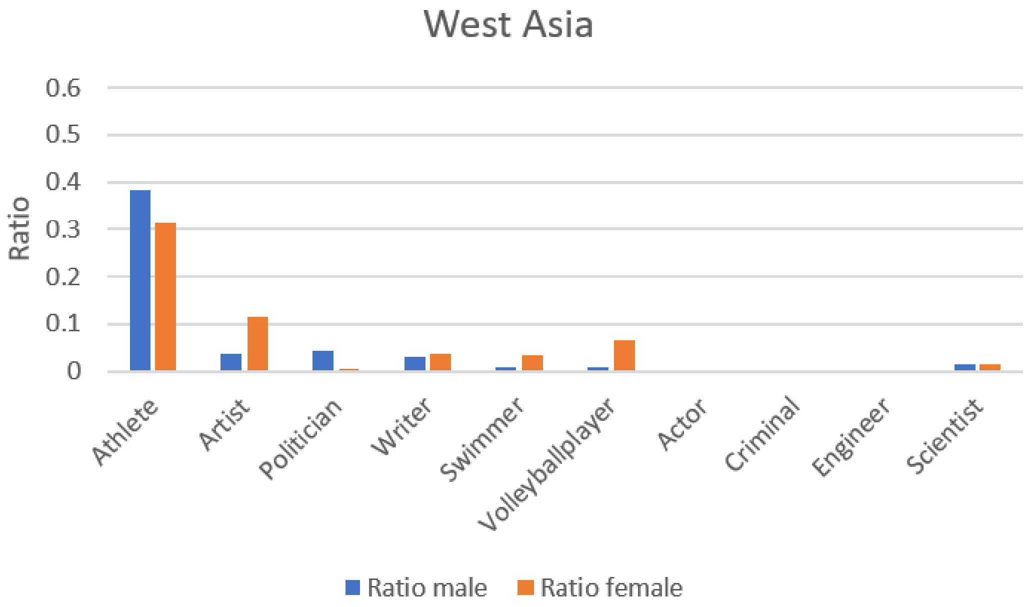


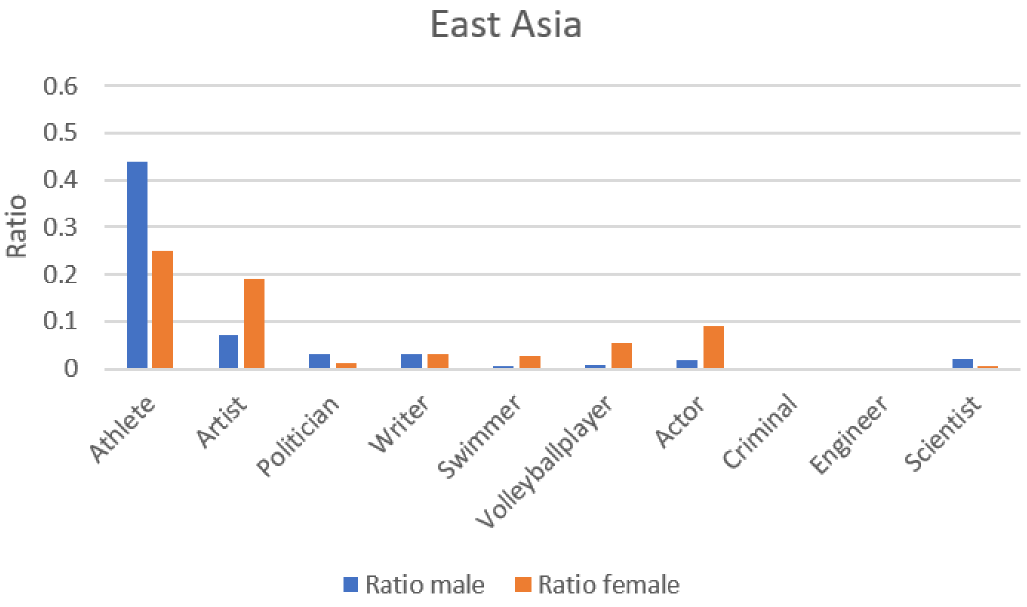
Figure 4.9: DBpedia: Ratios by region.

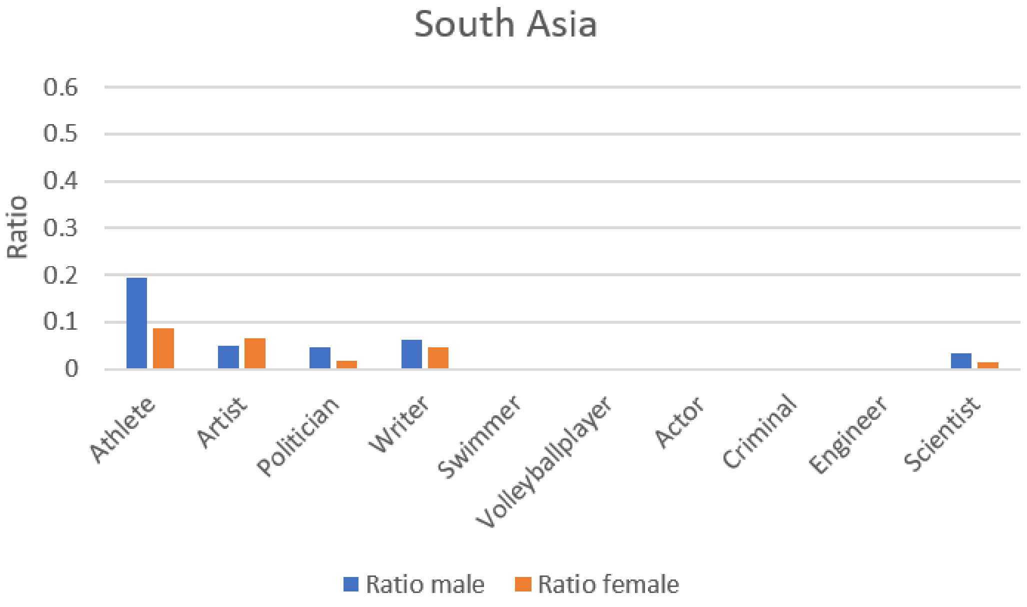

South Europe
Ratio
0.6
0.5
0.4
0.3
0.2
0.1
0
Athlete
Artist
Politician
Writer
Swimmer
Volleyballplayer
Actor
Criminal
Engineer
Scientist
Ratio male
Ratio female

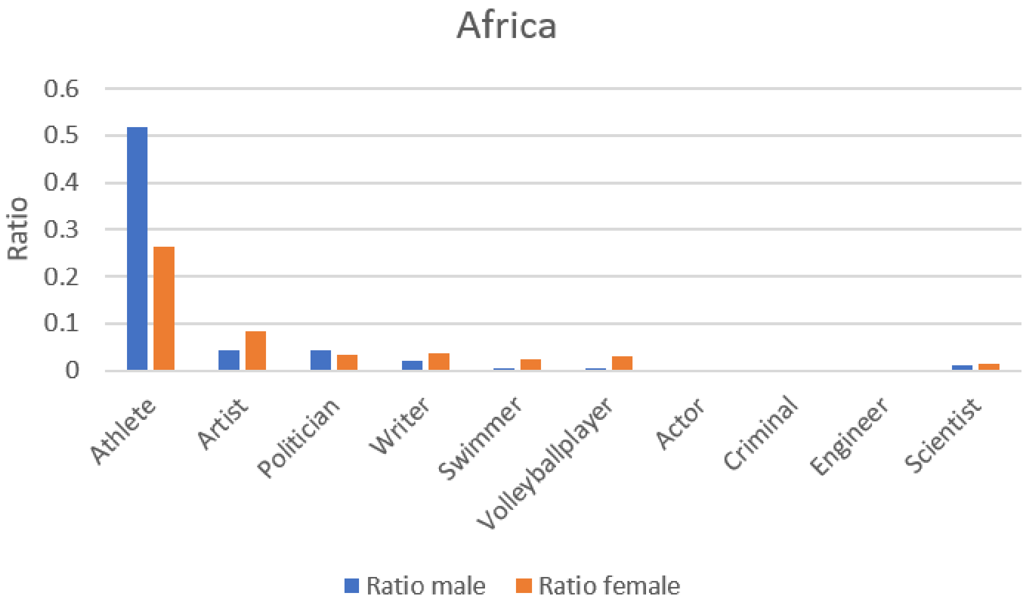

North America
Ratio
0.6
0.5
0.4
0.3
0.2
0.1
0
Athlete
Artist
Politician
Writer
Swimmer
Volleyballplayer
Actor
Criminal
Engineer
Scientist
Ratio male
Ratio female

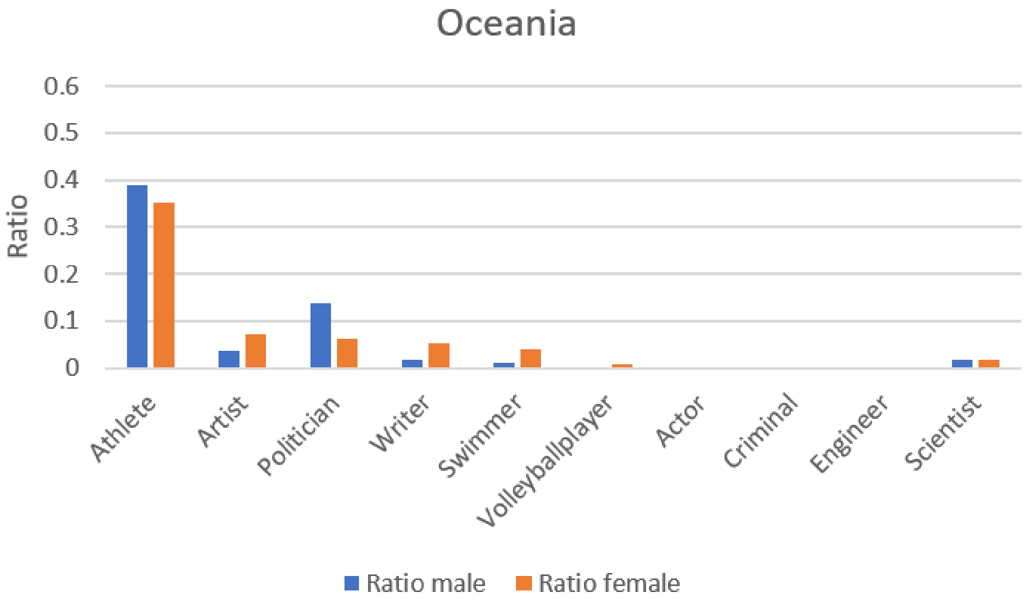

Latin America
Ratio
0.6
0.5
0.4
0.3
0.2
0.1
0
Athlete
Artist
Politician
Writer
Swimmer
Volleyballplayer
Actor
Criminal
Engineer
Scientist
■ Ratio male
■ Ratio female


West Asia
Ratio
0.6
0.5
0.4
0.3
0.2
0.1
0
Athlete
Artist
Politician
Writer
Swimmer
Volleyballplayer
Actor
Criminal
Engineer
Scientist
■ Ratio male
■ Ratio female


East Asia
Ratio
0.6
0.5
0.4
0.3
0.2
0.1
0
Athlete
Artist
Politician
Writer
Swimmer
Volleyballplayer
Actor
Criminal
Engineer
Scientist
Ratio male
Ratio female


South Asia
Ratio
0.6
0.5
0.4
0.3
0.2
0.1
0
Athlete
Artist
Politician
Writer
Swimmer
Volleyballplayer
Actor
Criminal
Engineer
Scientist
Ratio male
Ratio female


Africa
Ratio
0.6
0.5
0.4
0.3
0.2
0.1
0
Athlete
Artist
Politician
Writer
Swimmer
Volleyballplayer
Actor
Criminal
Engineer
Scientist
Ratio male
Ratio female


Oceania
Ratio
0.6
0.5
0.4
0.3
0.2
0.1
0
Athlete
Artist
Politician
Writer
Swimmer
Volleyballplayer
Actor
Criminal
Engineer
Scientist
Ratio male
Ratio female

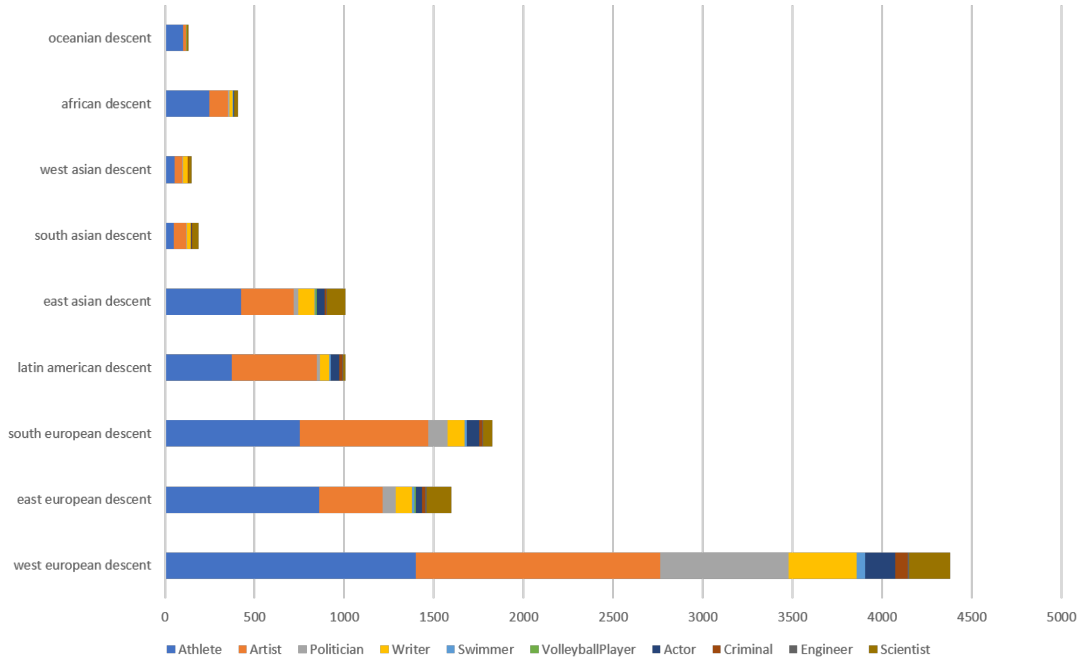


Figure 4.10: DBpedia: Number of paths for americans of foreign descent grouped by region.

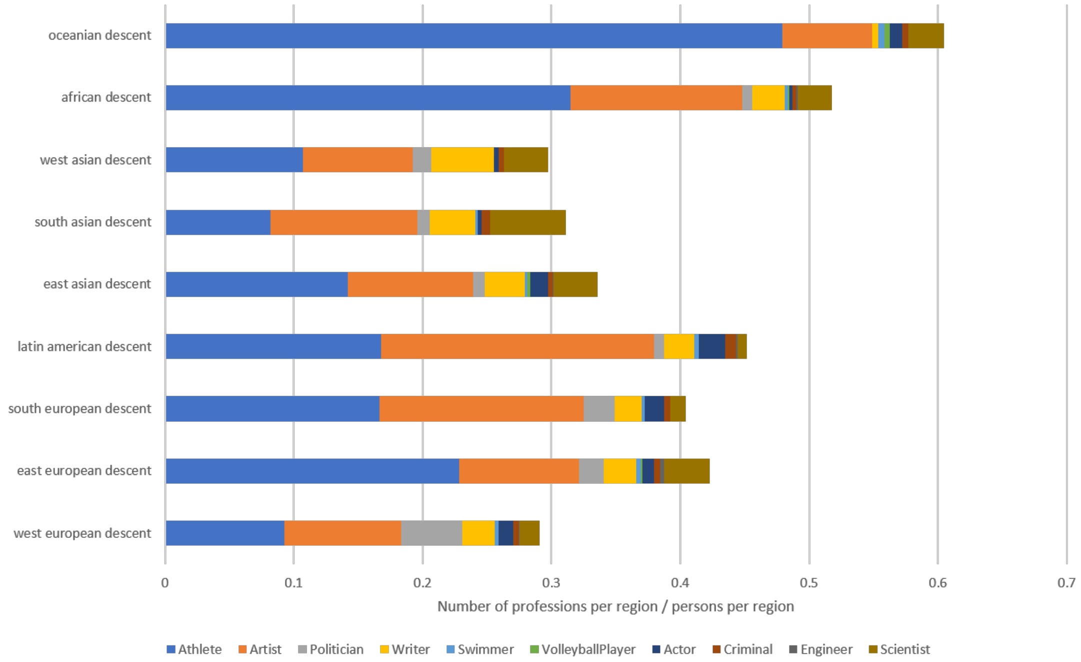


Figure 4.11: DBpedia: Ratio for americans of foreign descent grouped by region.

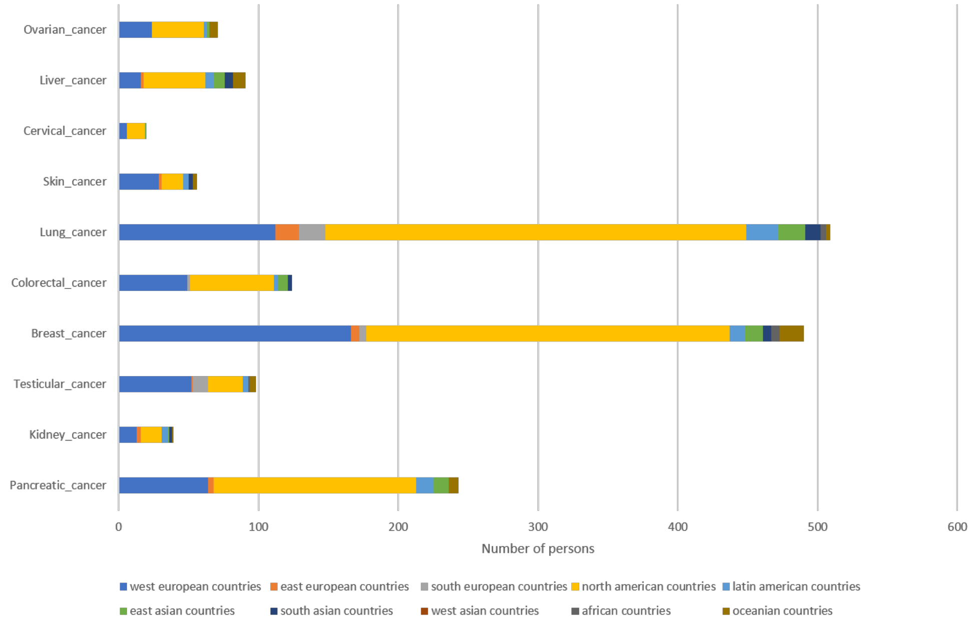


Figure 4.12: DBpedia: Total of cancer diseases by region.

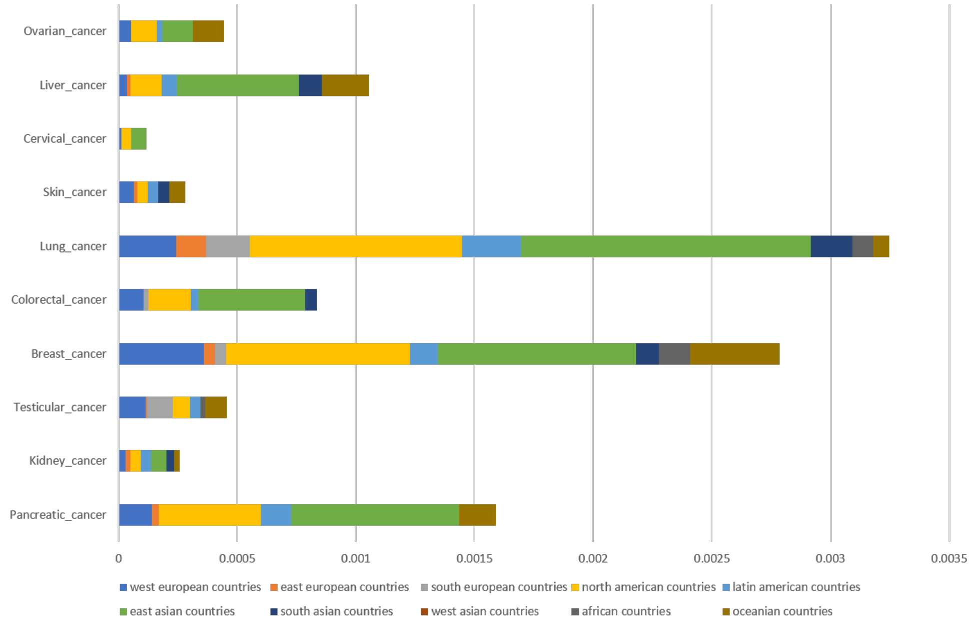


Figure 4.13: DBpedia: Ratio of cancer diseases per region.

Table 4.3: DBpedia: Number cancer diseases per region and gender

		Cancer	Pancreatic cancer	Kidney cancer	Testicular cancer	Breast cancer	Colorectal cancer	Lung cancer	Skin cancer	Cervical cancer	Liver cancer	Ovarian cancer
west europes	male	327	43	13	52	0	32	87	11	0	15	0
	female	144	21	0	0	166	17	25	18	6	1	24
	total	471	64	13	52	166	49	112	29	6	16	24
east europe	male	27	4	3	1	0	0	10	0	0	2	0
	female	4	0	0	0	6	0	7	2	0	0	0
	total	31	4	3	1	6	0	17	2	0	2	0
south europe	male	20	0	0	11	0	2	16	0	0	0	0
	female	30	0	0	0	5	0	3	0	0	0	0
	total	50	0	0	11	5	2	19	0	0	0	0
north america	male	579	110	12	25	7	44	206	9	0	26	0
	female	196	35	3	0	253	16	95	6	13	18	37
	total	775	145	15	25	260	60	301	15	13	44	37
latin america	male	30	12	4	4	0	3	16	4	0	6	0
	female	16	0	0	0	11	0	7	0	0	0	2
	total	46	12	4	4	11	3	23	4	0	6	2
east asia	male	15	3	1	0	0	1	14	0	0	6	0
	female	4	8	0	0	13	6	5	0	1	2	2
	total	19	11	1	0	13	7	19	0	1	8	2
south asia	male	32	0	2	0	0	3	11	3	0	6	0
	female	12	0	0	0	6	0	0	0	0	0	0
	total	44	0	2	0	6	3	11	3	0	6	0
west asia	male	2	0	0	0	0	0	0	0	0	0	0
	female	3	0	0	0	0	0	0	0	0	0	0
	total	5	0	0	0	0	0	0	0	0	0	0
africa	male	11	0	0	1	0	0	4	0	0	0	0
	female	0	0	0	0	6	0	0	0	0	0	0
	total	11	0	0	1	6	0	4	0	0	0	0
oceania	male	38	7	1	4	0	0	3	3	0	6	0
	female	13	0	0	0	17	0	0	0	0	3	6
	total	51	7	1	4	17	0	3	3	0	9	6

4.3 Yago: Detecting Biases using Embeddings

Deepwalk provides the possibility to change certain parameters. It's possible to set the maximal walk length, the number of walks per node or a restart probability for the random walks. This probability (not to be confused with the restart probability of section 3.6), is zero per default. From our results of the enumerations, we know that the distance from a gender to a profession is in most cases 3 steps or less. Hence, we made experiments with a walk length of 3 or 4. Due to the fact, that our graph has a very high number of nodes, we perform up to a maximum of 100 walks per node.

To compare the embeddings, we compute the cosine similarity between the vector for a gender and a profession. Results close to 1.0 stand for a high similarity, whereas values tending towards zero is indicating a low similarity. A big difference of the similarities from male or female to a profession indicates a bias toward one of the genders (see Table 4.4). The differences are plotted in Figure 4.14. The different colours are representing walks with varying parameter like walk length and walks per node. Professions like Mathematician or Engineer show a higher similarity to male, whereas Dancer or Volleyball Player are closer to female.

Changing the parameters like the walk length or number of walks per node, has only a limited impact on the results. Even with the same settings and parameters, there are small differences regarding the results. This is due to the nature of random walks used as input for Word2Vec. The biases are present with every set of parameters we tried.

Such embeddings have the advantage, that there is no need of a deeper knowledge of the dataset. You have to build a graph using the desired data and perform the random walks. You can easily compare nodes with each other using the cosine similarity. A drawback of this method is, that there is no possibility to reveal regional cultural differences like we did with the other two methods. The results are always produced in the context of all nodes in the graph.

Table 4.4: Cosine similarity of word embeddings with walk length 3 and 50 walks per node

	male	female	male - female
<wordnet_mathematician_110301261>	0.536335302	0.162627634	0.373707668
<wordnet_engineer_109615807>	0.695028506	0.226293437	0.46873507
<wordnet_dancer_109989502>	0.383405272	0.765057006	-0.381651734
<wordnet_computer_scientist_109951070>	0.488757432	0.434029309	0.054728123
<wordnet_chemist_109913824>	0.482120146	0.265634262	0.216485884
<wordnet_biologist_109855630>	0.476915872	0.332409343	0.144506529
<wordnet_artist_109812338>	0.524696209	0.538441455	-0.013745246
<wordnet_writer_110794014>	0.460473235	0.630452873	-0.169979638
<wordnet_politician_110451263>	0.509161413	0.247779051	0.261382362
<wordnet_poet_110444194>	0.397299461	0.498857017	-0.101557556
<wordnet_psychologist_110488865>	0.393059417	0.438308634	-0.045249217
<wordnet_physicist_110428004>	0.450851852	0.353285701	0.09756615
<wordnet_volleyball_player_110759047>	0.335864383	0.627362753	-0.29149837
<wordnet_swimmer_110683349>	0.461678351	0.612697959	-0.151019608

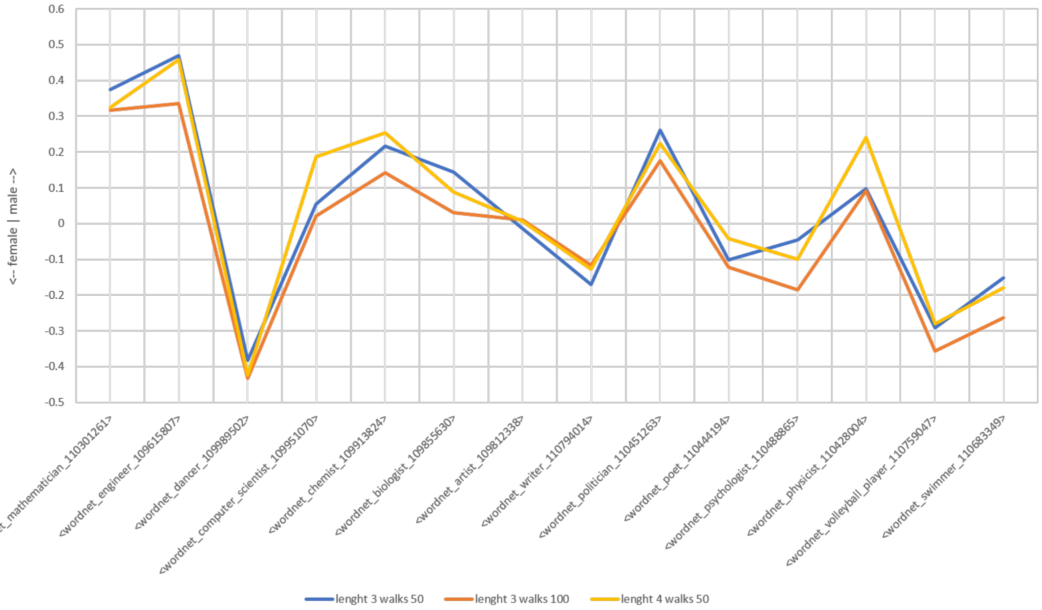


Figure 4.14: Difference of Cosine Similarities of vectors representing a gender, respectively a profession. Negative values are indicating a female bias, positive a male bias.

4.4 Yago: Mitigating Biases with Embeddings

Optimally the cosine similarity of the vectors representing male to one representing a profession is close to the cosine similarity of female and the same profession. Our research revealed big differences regarding this, hence we want to use the method described in Section 3.6 to de-bias the outcome. We have the possibility to adjust the following parameters:

- walk length
- number of walks per node
- restart probability $alpha = \frac{1}{degree^x}$
- window size of skip gram model

We have seen in previous tests, that the walk length and the number of walks only have a small influence on the outcome of the experiment. Unfortunately, also the other two parameters did not lead to the desired results. We tried to use the modified random walk algorithm to de-bias the vectors for the professions 'Engineer' and 'Dancer'. Even with a restart probability of 100%, the cosine similarity's difference between the vectors remained on a similar level (see Figure 4.15). Changing the window size didn't cause significant different results. The different colours are representing walks with varying parameter like walk length, walks per node, restart probability and window size.

The presented example graph (see Figure 3.6) is based on a simplified model. Our graph is more complex and the random walks can lead to a complete other end node, and not necessarily to a gender. In our tests, only 8-10% of the random walks starting from the node 'Engineer' reached the end node 'male'. For the node 'Dancer', it was even lower. We know from our research, that the dataset contains significantly fewer females than males. This means, the probability of reaching the opposite gender (e.g. 'Engineer' ⟶ 'female') is once again much lower. Because we start walks from every node in the graph, we have a total of several millions of walks. Every node is present at least once and the Neural Network of Word2Vec is capable of building similar embeddings, even if we manipulate the input by erasing every walk representing a bias (e.g. 'Engineer' ⟶ 'male').

4.5 Assessment of the Methods

Each of the methods we used to detect biases in the knowledge bases has its advantages and disadvantages. There are differences regarding the required time to retrieve the desired information, a deeper understanding of the structure and the data of the KB, the quality and reliability of the results.

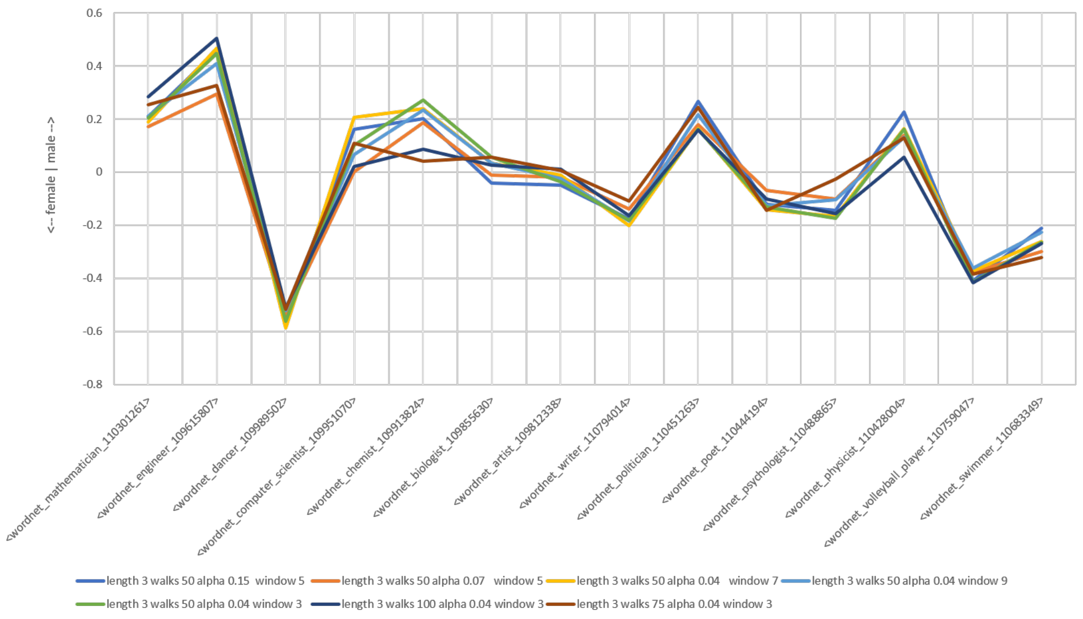


Figure 4.15: Difference of Cosine Similarities of vectors representing a gender, respectively a profession with walk eraser. Alpha values are referring to x in the formula of the restart probability

Enumeration

- *Advantages:* The only prerequisite to perform enumerations is a data dump of the knowledge base. It's possible to retrieve additional information for a person like geographical or temporal data.

- *Disadvantages:* Searching all paths from one node to another is very time consuming, especially for a graph of the size of several millions of nodes and edges. The runtime increases with a higher path length or with a higher number of professions. The resulting number of retrieved paths is not corresponding with the actual number of persons in the dataset. Additional information like the geographical or temporal data is relying on nodes we find within the paths. The quality of the data depends on the completeness of such intermediate nodes. In other words, every relevant person should be connected with a node representing the geographical information we need.

SPARQL

- *Advantages:* Compared to the other two methods, there is no need of downloading the dataset (at least for the KB DBpedia). There is a public SPARQL endpoint. The ontology is accessible on the web page of DBpedia, what makes it comfortable to explore its structure and find the desired classes and attributes. Adding new professions or countries can be done easily by extending the query. This method is by far the fastest of the three variants.

- *Disadvantages:* The quality of the results is depending on the manually edited Wikipedia articles. We found several attributes describing basically the same thing (e.g. the death cause or the profession of a person). This inconsistent usage of such attributes increase the complexity of the queries.

Embeddings

- *Advantages:* As with enumeration, a data dump of the knowledge base is needed to build the graph, but no deeper knowledge of the structure of the data is required, because the walks are produced randomly. With the resulting embeddings, you can compare every node with each other by computing the cosine similarity.
- *Disadvantages:* With this method, it is not possible to analyse further information like cultural differences based on regions. Building the random walks and especially the training of the neural network takes a lot of time for a graph of this size.

Table 4.5: Overview: Evaluation of the three methods for detecting biases

	Enumeration	SPARQL	Embeddings
Runtime	very high	low	high
Required knowledge of dataset	high	medium/high	low
Quality of results	low/medium	medium/high	medium/high
Flexibility	low	high	high

If a SPARQL endpoint of the KB is available, this is the fastest way to find biases. Like that, it is also possible to refine the search and investigate other properties like regional or temporal information. Otherwise, the choice falls on one of the remaining two methods. If no further regional or temporal information has to be evaluated, embeddings are easier to use and more flexible compared to enumerations.

5

Conclusions and Future Work

We introduced three different method to detect social biases in knowledge bases and examined differences regarding genders, professions and regional as well as cultural influences in DBpedia and YAGO. Every of these three methods has it own advantages and disadvantages but they all revealed severe social biases in the datasets.

The number of males is significantly higher compared to the number of females in both evaluated KBs. This difference is more pronounced in less developed countries, but still existent in every examined region. The largest share of persons in the KB are from European countries, especially from Western Europe, despite the much higher population in Asia or Africa. This distribution contrasts clearly with the demographic population in reality. In this thesis, we evaluated the English datasets, what could have an impact on the results. Nevertheless, other aspects can have an influence on the data as well, like the technological development of a country, the internet usage of its population, potential censorship, the popularity of Wikipedia or different cultural aspects, as the source data in Wikipedia is created and edited manually by humans.

If we take a further look at the proportional share of females or males in particular professions, we can find regional differences. Professions with a mathematical background like Engineer or Mathematician are dominated by males. Other professions like Artist or Dancer have a higher share of females in most regions. Still, males outnumber females with only a few exceptions. Historically, the difference between the genders was even bigger as we could see for the professions Mathematician and Poet in the dataset of YAGO. For several centuries, not a single female mathematician is present in the dataset.

You can argue, that certain social biases, like the proportion of males to females in a profession is only reflecting the status of the real world. On the other hand, we discovered social biases, which are obviously not portraying the status of the world. Occupations with a high public attention are clearly overrepresented in both KBs. A common Engineer usually doesn't have an article in Wikipedia and is consequently not mapped in the KB, whereas an Actor, Athletes or Artist is present more likely. Another distortion is the very big influence of Western European countries.

To determine, if such a bias is reflecting a condition in the real world or is only present in the KB is difficult. The users of the KBs must be aware, that such biases are existing and they have to handle it, if they have a potentially harmful influence on their activities and results. Mitigating such biases in the KB would be possible by adding artificial data (nodes and/or edges) or deleting existing entries, but this would lead either to an information loss or the problem, how to treat such artificial data. We tried to 'de-biase' embeddings by manipulating the input for the neural network, i.e. the random walks. These changes didn't show the desired effect.

Our focus was on DBpedia and YAGO, but there are other big knowledge bases. Since we found severe evidence for social biases in both KBs, an examination of further KBs would be recommended. In this thesis, we are focussing on the detection of social biases. The results are restricted on the English datasets of DBpedia and YAGO. We investigated a set of professions with different background like science, arts and sports. The cause of the detected biases is not investigated systematically, just as the impact of biases on systems using the KBs. Furthermore, we compare different regions with each other. Those regions are only a rough segmentation based on the best of our knowledge and could be done in many different ways.

Many different combinations of regions and professions have been left for the future due to lack of time. Future work could concern deeper statistical analysis of the results. Another interesting topic is comparing datasets of the KBs in different languages, as we only worked with the English version of DBpedia and YAGO. The biases in the KBs could possibly be a representation of a bias existing in reality. Comparing our data with statistical data from different countries regarding the proportion of males and females practising a profession, could reveal KB specific biases. This can help to distinguish between biases existing in the real world and such only existing in the KBs.

The impact of the detected biases on the programs using this data is also an interesting topic. If the biases have a harmful influence on those programs, a way to mitigate it should be developed.

References

[Allemani et al., 2015] Allemani, C., Weir, H. K., Carreira, H., Harewood, R., Spika, D., Wang, X.-S., Bannon, F., Ahn, J. V., Johnson, C. J., Bonaventure, A., et al. (2015). Global surveillance of cancer survival 1995–2009: analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (concord-2). *The Lancet*, 385(9972):977–1010.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific american*, 284(5):28–37.

[Caliskan et al., 2017] Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

[Geopy, 2015] Geopy (2015). *https://github.com/geopy/geopy/tree/1.11.0*. [Online; accessed 07.11.2017].

[Greenwald et al., 1998] Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

[Grieco and Cassidy, 2001] Grieco, E. M. and Cassidy, R. C. (2001). *Overview of race and Hispanic origin, 2000*, volume 8. US Department of Commerce, Economics and Statistics Administration, US Census Bureau.

[Hill et al., 2010] Hill, C., Corbett, C., and St Rose, A. (2010). *Why so few? Women in science, technology, engineering, and mathematics.* ERIC.

[Koehn, 2009] Koehn, P. (2009). *Statistical machine translation.* Cambridge University Press.

[Lavergne and Mullainathan, 2004] Lavergne, M. and Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *The American Economic Review*, 94(4):991–1013.

[Lehmann et al., 2015] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C.

(2015). Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

[Leondes, 2002] Leondes, C. T. (2002). *Expert systems: the technology of knowledge management and decision making for the 21st century.* Academic Press, Los Angeles.

[Mahdisoltani et al., 2014] Mahdisoltani, F., Biega, J., and Suchanek, F. (2014). Yago3: A knowledge base from multilingual wikipedias.

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

[Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Mitchell et al., 2015] Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Rivard, K., Mohammad, T., Nakashole, N., Platanios, E. A., Ritter, A., Samadi, M., Settles, B., Wang, R., and Welling, J. (2015). Never-ending learning.

[Nosek et al., 2002] Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002). Math= male, me= female, therefore math$\neq$ me. *Journal of personality and social psychology*, 83(1):44.

[Perozzi et al., 2014] Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 701–710, New York, NY, USA. ACM.

[Ristoski and Paulheim, 2016] Ristoski, P. and Paulheim, H. (2016). Rdf2vec: Rdf graph embeddings for data mining. In *International Semantic Web Conference*, pages 498–514. Springer.

[Rumelhart et al., 1988] Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1.

[Sedgewick, 2001] Sedgewick, R. (2001). *Algorithms in C, Part 5: Graph Algorithms.* Addison Wesley Professional.

[Union, 2017] Union, I. T. (2017). The world in 2017: Ict facts and figures. *International Telecommunication Union.*

[W3C, 2014] W3C (2014). *https://www.w3.org/RDF/.* [Online; accessed 31.12.2017].

[Warf, 2011] Warf, B. (2011). Geographies of global internet censorship. *GeoJournal*, 76(1):1–23.

[Yago, 2017] Yago (2017). Yago homepage. [Online; accessed 05.12.2017].

[Zaveri et al., 2013] Zaveri, A., Kontokostas, D., Sherif, M. A., Bühmann, L., Morsey, M., Auer, S., and Lehmann, J. (2013). User-driven quality evaluation of dbpedia. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 97–104. ACM.

A

Appendix

A.1 Yago enumeration

Table A.1: Timeline for mathematicians and poets

	Mathematicians		Poets	
	male	female	male	female
5th-century	9	0	36	0
6th-century	24	0	72	0
7th-century	12	0	45	0
8th-century	18	0	24	2
9th-century	63	0	72	3
10th-century	78	0	117	2
11th-century	54	0	72	0
12th-century	42	0	144	2
13th-century	72	0	132	3
14th-century	66	0	42	0
15th-century	84	0	51	2
16th-century	381	0	105	2
17th-century	573	0	99	2
18th-century	618	5	48	5
19th-century	1905	12	1668	58
20th-century	8820	173	4470	260

Table A.2: Yago: Total number of paths per region

	male	female
West Europe	32263	23152
East Europe	11596	7832
South Europe	9523	3688
Nord America	14868	3195
Latin America	7580	992
East Asia	4782	1203
South Asia	3835	556
West Asia	2205	233
Africa	6416	760
Ozeania	2895	443
Total	95963	42054

Table A.3: Paths per region and profession. Ratio = Number of pahts divided by total number of paths from A.2

West Europe	male	female	male ratio	female ratio
Mathematicians	2313	108	0.071692031	0.004664824
Computer Scientists	701	38	0.021727676	0.001641327
Physicists	2349	104	0.07280786	0.004492053
Chemists	1764	64	0.054675635	0.00276434
Engineers	2137	23	0.066236866	0.000993435
Psychologists	600	152	0.018597155	0.006565308
Poets	4131	473	0.12804141	0.0204302
Writers	6973	1538	0.216129932	0.066430546
Dancers	118	115	0.00365744	0.004967173
Artists	2439	637	0.075597434	0.027513822
Biologists	975	114	0.030220376	0.004923981
Swimmer	636	1362	0.019712984	0.058828611
Volleyball Player	152	94	0.004711279	0.004060124

East Europe	male	female	male ratio	female ratio
Mathematicians	885	38	0.07631942	0.00485189
Computer Scientists	119	8	0.010262159	0.00102145
Physicists	599	14	0.051655743	0.001787538
Chemists	221	15	0.019058296	0.00191522
Engineers	584	14	0.050362194	0.001787538
Psychologists	118	20	0.010175923	0.002553626
Poets	1692	216	0.145912384	0.027579162
Writers	2665	331	0.229820628	0.042262513
Dancers	31	26	0.002673336	0.003319714

Artists	832	217	0.071748879	0.027706844
Biologists	177	17	0.015263884	0.002170582
Swimmer	260	495	0.022421525	0.063202247
Volleyball Player	334	110	0.028803036	0.014044944

South Europe	male	female	male ratio	female ratio
Mathematicians	353	11	0.037068151	0.002982646
Computer Scientists	62	7	0.006510553	0.001898048
Physicists	242	19	0.02541216	0.005151844
Chemists	59	1	0.006195527	0.00027115
Engineers	291	6	0.030557597	0.001626898
Psychologists	23	6	0.002415205	0.001626898
Poets	839	51	0.088102489	0.013828633
Writers	734	106	0.077076552	0.028741866
Dancers	27	43	0.002835241	0.011659436
Artists	401	113	0.042108579	0.030639913
Biologists	65	11	0.00682558	0.002982646
Swimmer	136	231	0.014281214	0.062635575
Volleyball Player	109	41	0.011445973	0.011117137

Nord America	male	female	male ratio	female ratio
Mathematicians	866	79	0.058245897	0.024726135
Computer Scientists	1340	153	0.090126446	0.047887324
Physicists	1758	111	0.118240517	0.034741784
Chemists	934	106	0.062819478	0.033176839
Engineers	1353	61	0.091000807	0.019092332
Psychologists	1208	397	0.081248319	0.124256651
Poets	413	166	0.027777778	0.051956182
Writers	2137	685	0.143731504	0.214397496
Dancers	135	195	0.009079903	0.061032864
Artists	1535	525	0.103241862	0.164319249
Biologists	732	138	0.049233253	0.043192488
Swimmer	311	540	0.020917407	0.169014085
Volleyball Player	75	68	0.005044391	0.021283255

Latin America	male	female	male ratio	female ratio
Mathematicians	75	1	0.009894459	0.001008065
Computer Scientists	30	0	0.003957784	0
Physicists	95	7	0.012532982	0.007056452
Chemists	37	0	0.004881266	0
Engineers	265	16	0.034960422	0.016129032
Psychologists	12	14	0.001583113	0.014112903
Poets	730	191	0.096306069	0.192540323

Writers	1290	207	0.170184697	0.208669355
Dancers	25	61	0.003298153	0.061491935
Artists	468	130	0.061741425	0.131048387
Biologists	68	18	0.008970976	0.018145161
Swimmer	176	261	0.023218997	0.263104839
Volleyball Player	141	121	0.018601583	0.121975806

East Asia	male	female	male ratio	female ratio
Mathematicians	149	8	0.031158511	0.006650042
Computer Scientists	104	5	0.021748223	0.004156276
Physicists	155	6	0.032413216	0.004987531
Chemists	68	11	0.014219992	0.009143807
Engineers	196	8	0.040987035	0.006650042
Psychologists	12	2	0.00250941	0.00166251
Poets	599	93	0.125261397	0.077306733
Writers	1211	288	0.253241322	0.239401496
Dancers	54	44	0.011292346	0.036575229
Artists	325	106	0.067963195	0.088113051
Biologists	81	5	0.016938519	0.004156276
Swimmer	123	198	0.025721455	0.164588529
Volleyball Player	53	131	0.011083229	0.108894431

South Asia	male	female	male ratio	female ratio
Mathematicians	150	9	0.039113429	0.01618705
Computer Scientists	82	7	0.021382008	0.012589928
Physicists	144	9	0.037548892	0.01618705
Chemists	29	1	0.00756193	0.001798561
Engineers	138	3	0.035984355	0.005395683
Psychologists	18	6	0.004693611	0.010791367
Poets	560	75	0.146023468	0.134892086
Writers	1065	163	0.277705346	0.293165468
Dancers	11	2	0.002868318	0.003597122
Artists	114	49	0.029726206	0.088129496
Biologists	22	7	0.005736636	0.012589928
Swimmer	12	21	0.003129074	0.037769784
Volleyball Player	16	2	0.004172099	0.003597122

West Asia	male	female	male ratio	female ratio
Mathematicians	24	2	0.010884354	0.008583691
Computer Scientists	0	0	0	0
Physicists	4	2	0.001814059	0.008583691
Chemists	0	1	0	0.004291845
Engineers	86	4	0.039002268	0.017167382

Psychologists	0	0	0	0
Poets	261	33	0.118367347	0.141630901
Writers	469	73	0.212698413	0.313304721
Dancers	1	2	0.000453515	0.008583691
Artists	127	54	0.057596372	0.231759657
Biologists	0	0	0	0
Swimmer	48	57	0.021768707	0.244635193
Volleyball Player	6	12	0.002721088	0.051502146

Africa	male	female	male ratio	female ratio
Mathematicians	50	4	0.007793017	0.005263158
Computer Scientists	5	1	0.000779302	0.001315789
Physicists	24	4	0.003740648	0.005263158
Chemists	14	0	0.002182045	0
Engineers	181	7	0.028210723	0.009210526
Psychologists	4	4	0.000623441	0.005263158
Poets	377	57	0.058759352	0.075
Writers	964	209	0.150249377	0.275
Dancers	4	7	0.000623441	0.009210526
Artists	293	95	0.045667082	0.125
Biologists	15	3	0.002337905	0.003947368
Swimmer	69	129	0.010754364	0.169736842
Volleyball Player	15	15	0.002337905	0.019736842

Oceania	male	female	male ratio	female ratio
Mathematicians	93	8	0.032124352	0.018058691
Computer Scientists	62	1	0.021416235	0.002257336
Physicists	111	8	0.038341969	0.018058691
Chemists	82	5	0.028324698	0.011286682
Engineers	209	1	0.072193437	0.002257336
Psychologists	32	13	0.011053541	0.029345372
Poets	243	86	0.083937824	0.194130926
Writers	410	174	0.141623489	0.392776524
Dancers	6	5	0.002072539	0.011286682
Artists	248	61	0.08566494	0.137697517
Biologists	64	13	0.022107081	0.029345372
Swimmer	135	249	0.046632124	0.562076749
Volleyball Player	13	4	0.004490501	0.009029345

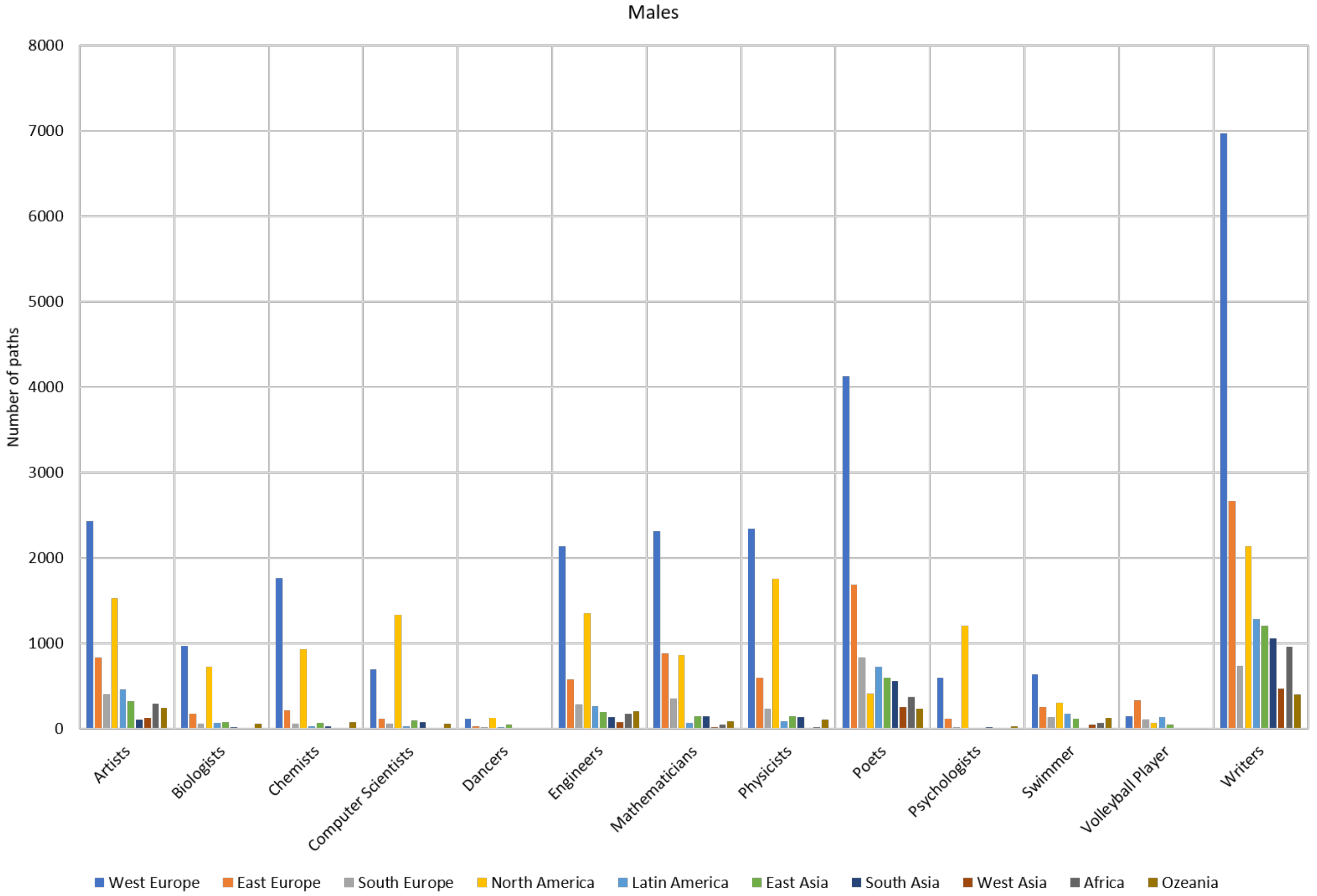


Figure A.1: YAGO: Number of males per region and profession.

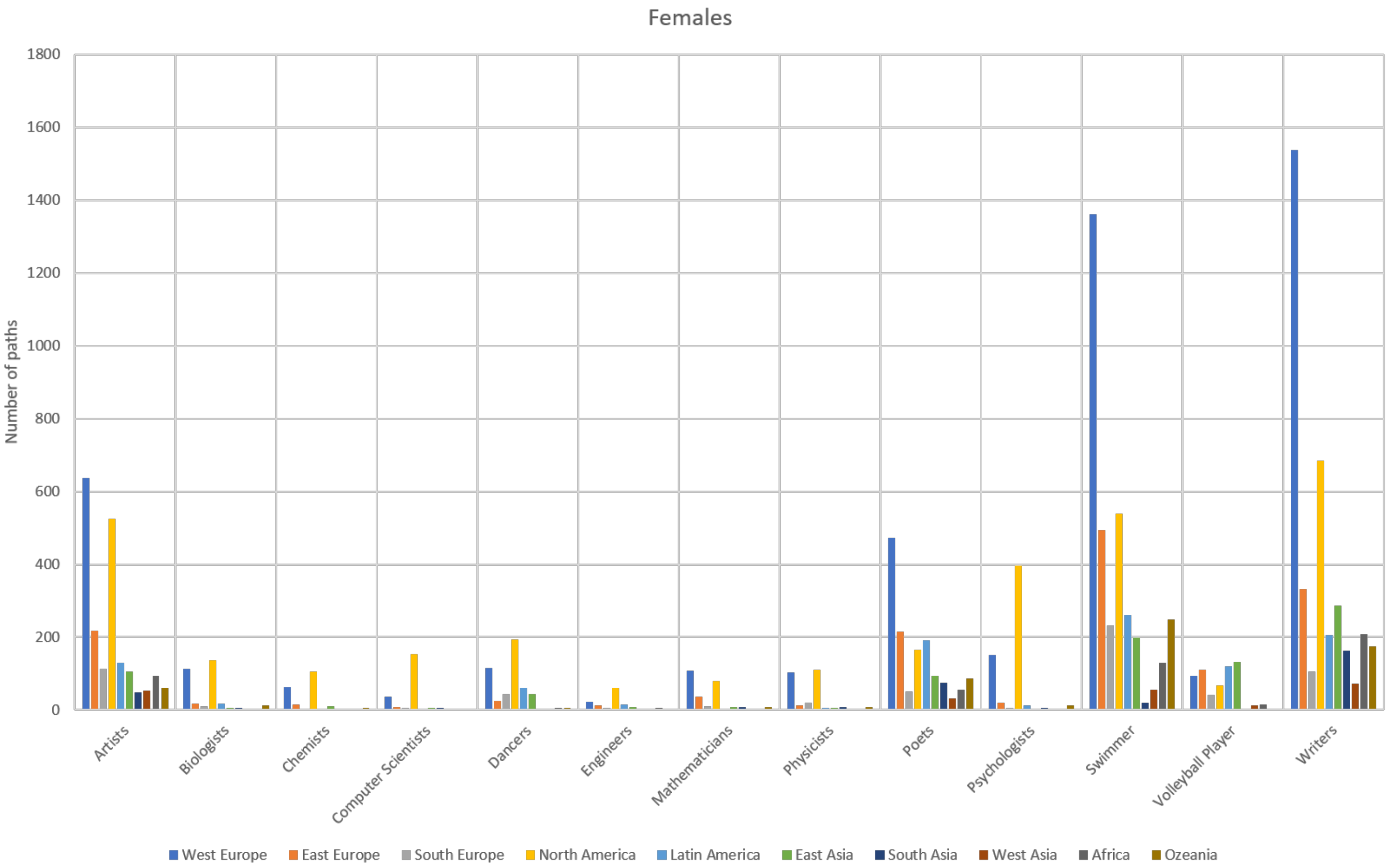


Figure A.2: YAGO: Number of females per region and profession.

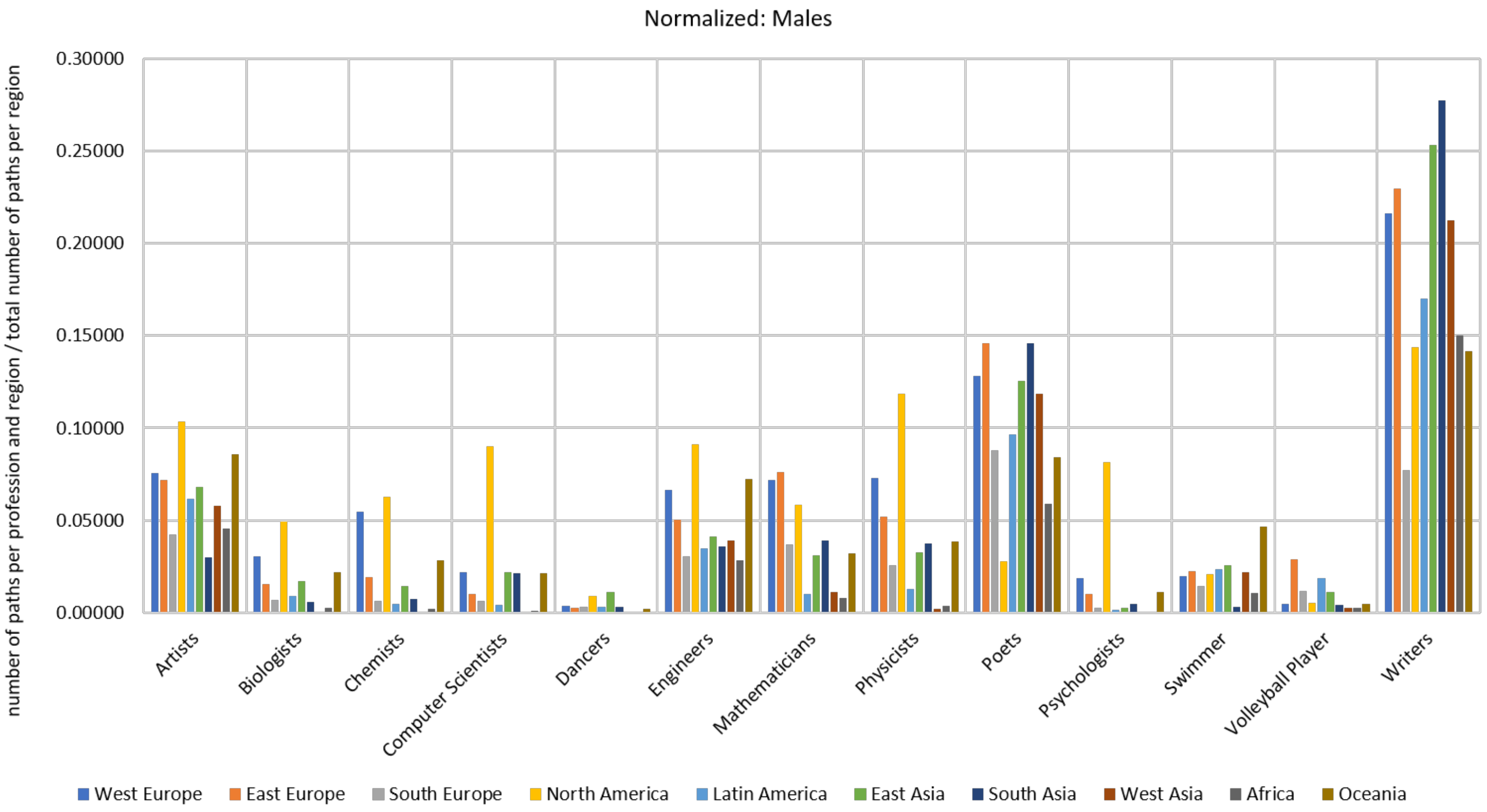


Figure A.3: YAGO: Ratio for males per region and profession.

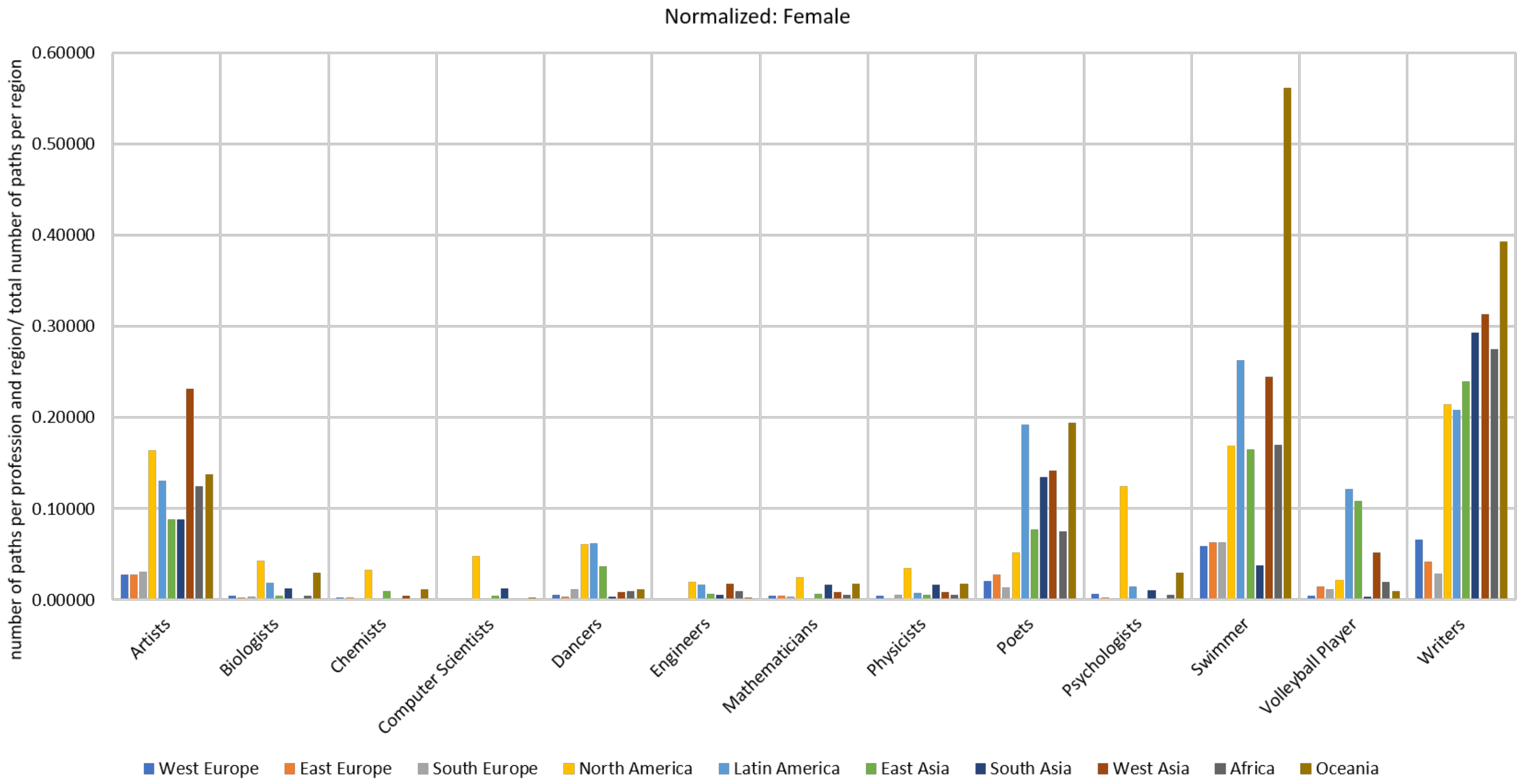


Figure A.4: YAGO: Ratio for females per region and profession.

A.2 DBpedia: SPRQL query results

Table A.4: DBpedia: Total number of persons per region

	male	female
West Europe	392836	67086
East Europe	113091	22335
South Europe	88978	14117
North America	281127	55387
Latin America	77649	15689
West Asia	13528	2015
East Asia	47937	14341
South Asia	56046	11896
Africa	39378	6608
Oceania	36688	8488
Total	1147258	217962

Table A.5: Results of section 3.4.1, professions modelled as subclasses

West Europe	male	female	Ratio male	Ratio female
Athlete	131962	12713	0.335921351	0.189503026
Artist	24755	7424	0.063016119	0.110663924
Politician	9767	1437	0.024862793	0.021420267
Writer	9678	3462	0.024636235	0.051605402
Swimmer	1406	1377	0.003579102	0.020525892
Volleyball Player	850	627	0.002163753	0.009346212
Actor	3010	2109	0.007662231	0.03143726
Criminal	371	47	0.000944414	0.000700593
Engineer	971	4	0.002471769	5.9625E-05
Scientist	16703	1328	0.042519016	0.019795486

East Europe	male	female	Ratio male	Ratio female
Athlete	46679	9451	0.412756099	0.423147526
Artist	4651	1709	0.041126173	0.076516678
Politician	3307	342	0.029241938	0.01531229
Writer	3709	670	0.032796597	0.029997761
Swimmer	961	664	0.008497582	0.029729125
Volleyball Player	1317	1257	0.011645489	0.056279382
Actor	31	85	0.000274116	0.003805686
Criminal	106	10	0.000937298	0.000447728
Engineer	90	2	0.000795819	8.95456E-05
Scientist	4221	268	0.037323925	0.011999105

South Europe	male	female	Ratio male	Ratio female
Athlete	26010	3127	0.29231945	0.221505986
Artist	6440	933	0.072377442	0.066090529
Politician	1568	117	0.017622334	0.00828788
Writer	1958	397	0.02200544	0.028122122
Swimmer	399	261	0.004484255	0.018488347
Volleyball Player	545	334	0.00612511	0.023659418
Actor	51	74	0.000573175	0.005241907
Criminal	196	11	0.002202792	0.000779202
Engineer	51	0	0.000573175	0
Scientist	1914	149	0.021510935	0.01055465

North America	male	female	Ratio male	Ratio female
Athlete	77710	7169	0.276423111	0.129434705
Artist	22259	6819	0.079177738	0.123115533
Politician	19743	1022	0.070228046	0.018451983
Writer	9010	4103	0.032049572	0.074078755
Swimmer	517	565	0.001839026	0.01020095
Volleyball Player	301	349	0.00107069	0.006301118
Actor	377	644	0.001341031	0.011627277
Criminal	879	131	0.003126701	0.002365176
Engineer	249	5	0.000885721	9.02739E-05
Scientist	10836	1598	0.038544857	0.028851536

Latin America	male	female	Ratio male	Ratio female
Athlete	39353	3658	0.506806269	0.233156989
Artist	383	1165	0.004932452	0.074255848
Politician	3410	220	0.043915569	0.014022564
Writer	1797	450	0.023142603	0.028682516
Swimmer	733	399	0.009439916	0.025431831
Volleyball Player	782	896	0.01007096	0.057110077
Actor	26	66	0.00033484	0.004206769
Criminal	131	9	0.001687079	0.00057365
Engineer	31	0	0.000399232	0
Scientist	900	138	0.011590619	0.008795972

West Asia	male	female	Ratio male	Ratio female
Athlete	5180	634	0.382909521	0.314640199
Artist	499	235	0.036886458	0.11662531
Politician	596	12	0.044056771	0.005955335
Writer	427	75	0.031564163	0.037220844
Swimmer	114	67	0.008426966	0.03325062
Volleyball Player	105	133	0.007761679	0.066004963

Actor	2	4	0.000147842	0.001985112
Criminal	15	0	0.001108811	0
Engineer	2	0	0.000147842	0
Scientist	215	27	0.015892963	0.013399504

East Asia	male	female	Ratio male	Ratio female
Athlete	21104	3596	0.440244488	0.250749599
Artist	3380	2755	0.07050921	0.192106548
Politician	1408	162	0.029371884	0.011296283
Writer	1401	447	0.029225859	0.031169375
Swimmer	304	371	0.006341657	0.025869884
Volleyball Player	392	793	0.0081774	0.055296004
Actor	876	1299	0.018273985	0.090579457
Criminal	61	5	0.001272503	0.000348651
Engineer	22	0	0.000458936	0
Scientist	1010	75	0.02106932	0.005229761

South Asia	male	female	Ratio male	Ratio female
Athlete	10808	1038	0.192841594	0.087256221
Artist	2705	784	0.048263926	0.065904506
Politician	2575	220	0.045944403	0.018493611
Writer	3483	546	0.062145381	0.045897781
Swimmer	61	30	0.001088392	0.002521856
Volleyball Player	162	31	0.002890483	0.002605918
Actor	29	18	0.000517432	0.001513114
Criminal	72	8	0.001284659	0.000672495
Engineer	15	0	0.000267637	0
Scientist	1888	175	0.033686615	0.014710827

Africa	male	female	Ratio male	Ratio female
Athlete	20358	1747	0.516989182	0.264376513
Artist	1682	558	0.042714206	0.084443099
Politician	1707	220	0.043349078	0.033292978
Writer	772	242	0.019604856	0.036622276
Swimmer	183	169	0.004647265	0.025575061
Volleyball Player	239	195	0.006069379	0.029509685
Actor	26	17	0.000660267	0.002572639
Criminal	43	5	0.00109198	0.000756659
Engineer	10	2	0.000253949	0.000302663
Scientist	493	101	0.012519681	0.015284504

Oceania	male	female	Ratio male	Ratio female
Athlete	14308	2998	0.389991278	0.353204524

Artist	1385	610	0.037750763	0.071866164
Politician	5072	515	0.138246838	0.060673893
Writer	653	439	0.017798735	0.051720075
Swimmer	389	336	0.010602922	0.039585297
Volleyball Player	63	63	0.001717183	0.007422243
Actor	15	18	0.000408853	0.002120641
Criminal	98	12	0.002671173	0.001413761
Engineer	48	0	0.00130833	0
Scientist	692	152	0.018861753	0.017907634

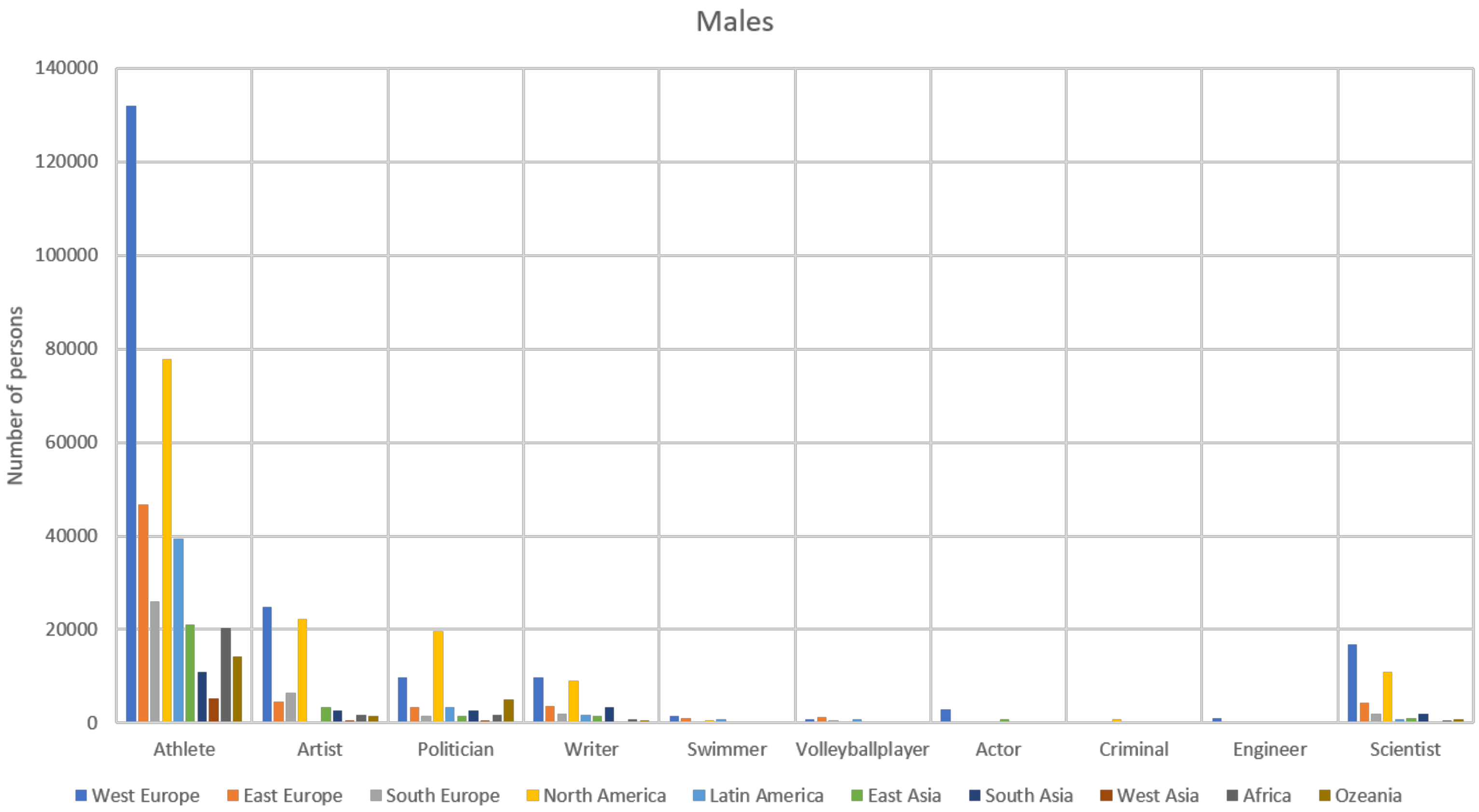


Figure A.5: DBpedia: Number of males per region and profession.

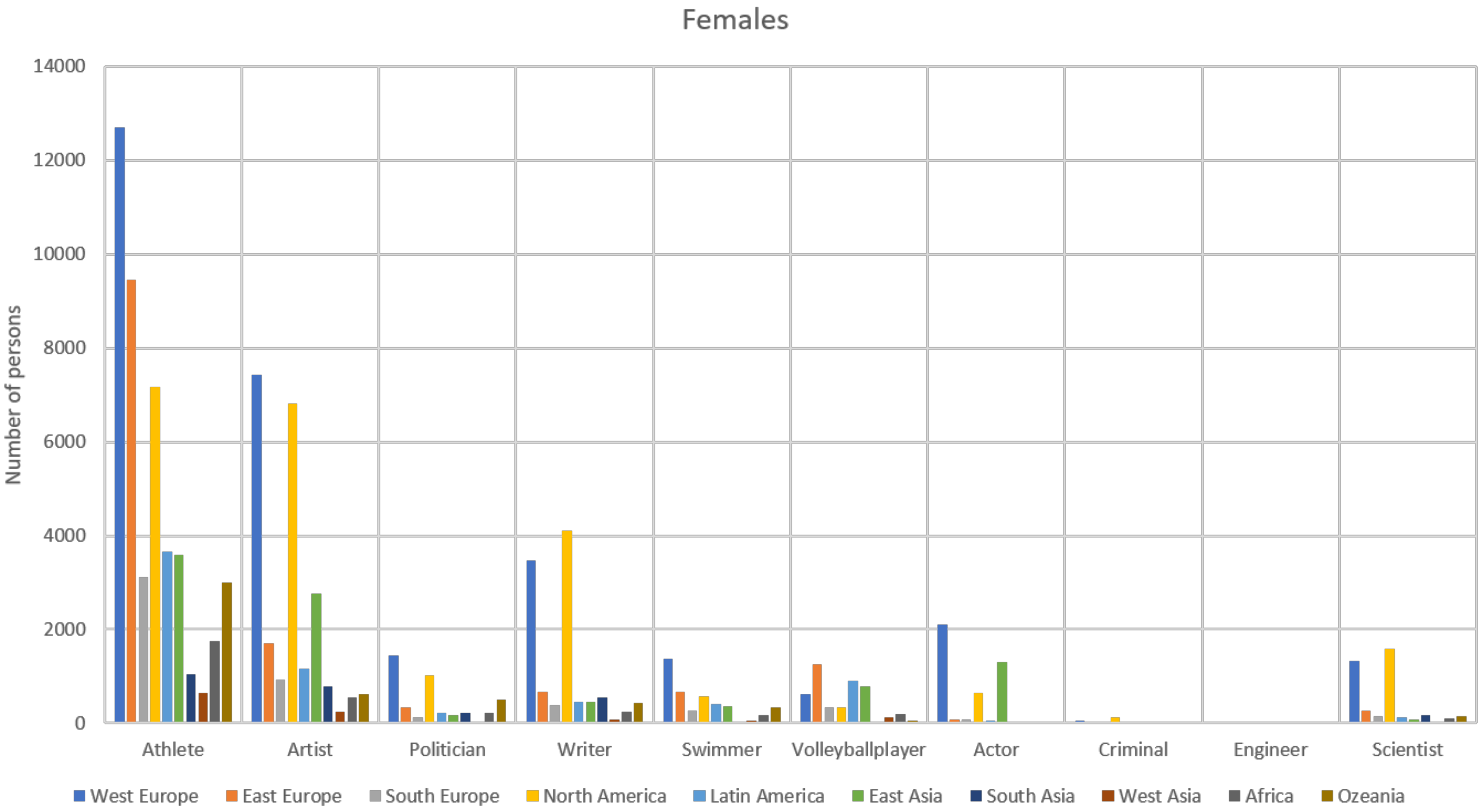


Figure A.6: DBpedia: Number of females per region and profession.

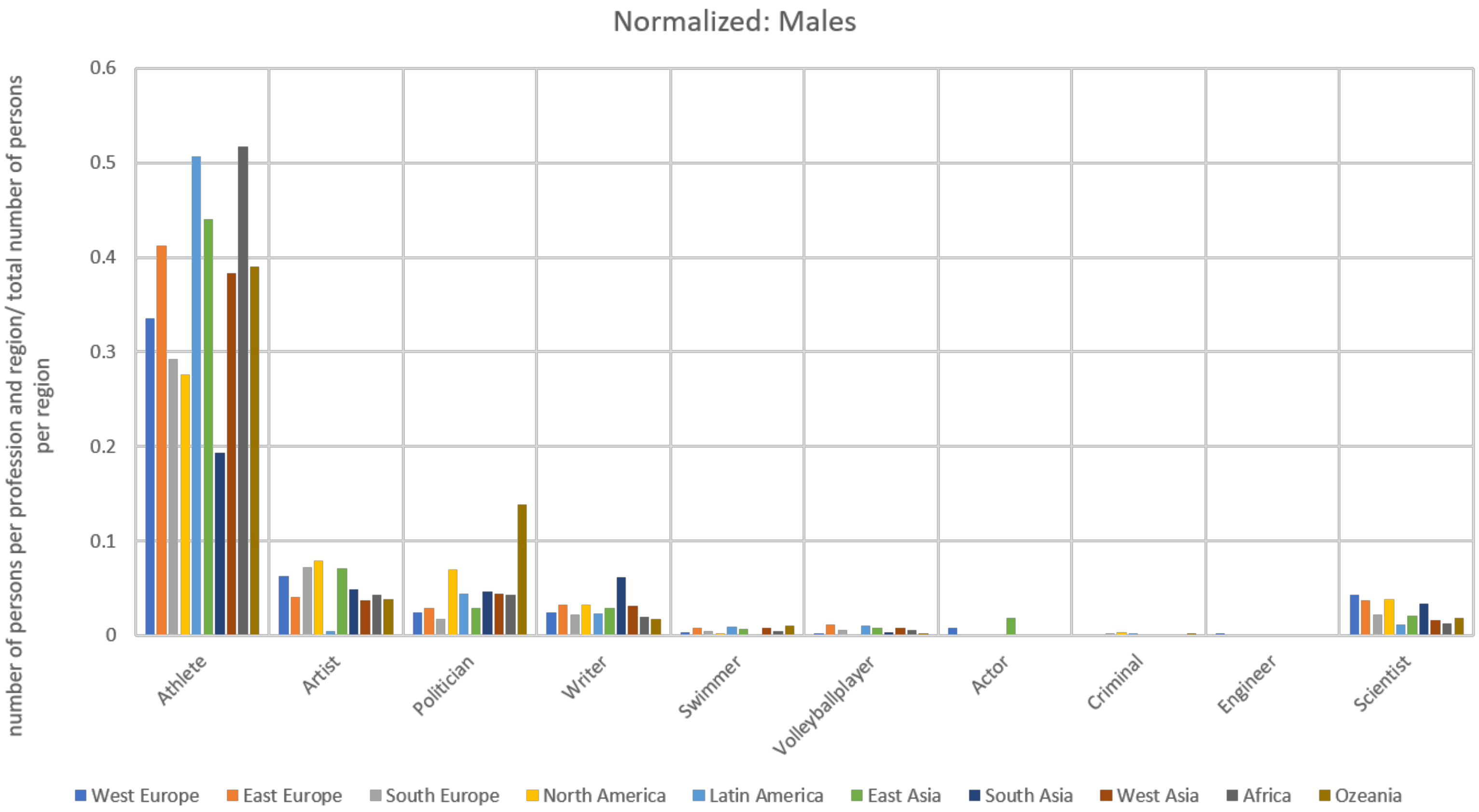


Figure A.7: DBpedia: Ratio for males per region and profession.

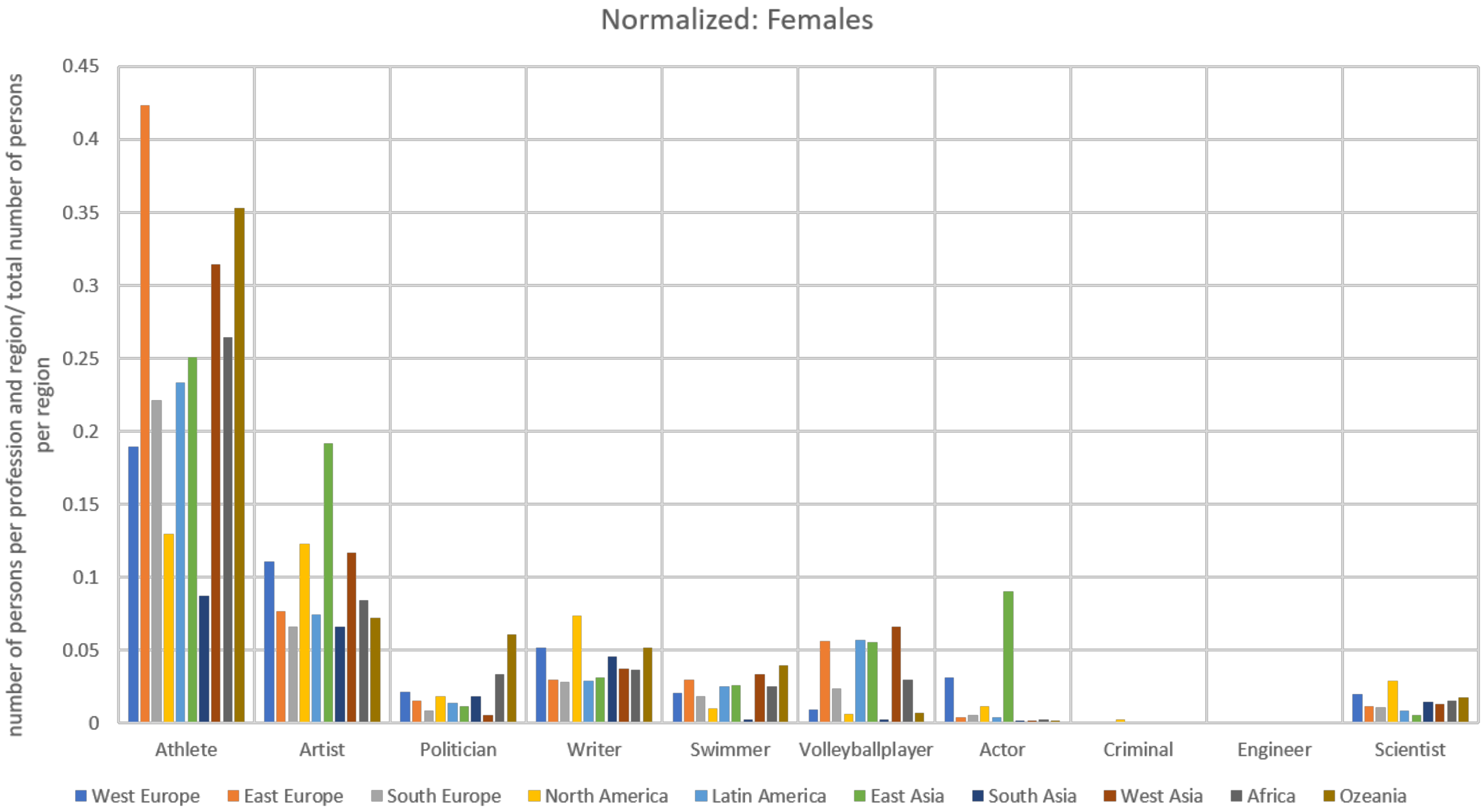


Figure A.8: YAGO: Ratio for females per region and profession.

List of Figures

List of Tables