Diese Arbeit wurde vorgelegt am 15.01.2019
Lehr- und Forschungsgebiet Computational Social Sciences and Humanities
Fakultät für Mathematik, Informatik und Naturwissenschaften
Prof. Dr. Markus Strohmaier

Bachelor Thesis

Concept embeddings for Wikipedia across language editions

vorgelegt von

Felix Ingenerf
Matrikelnummer: 355026
2019-01-15

Erstgutachter:	Prof. Dr. Markus Strohmaier
Zweitgutachter:	Prof. Dr. Stefan Decker
Erstbetreuer:	Dr. Florian Lemmerich
Zweitbetreuer:	Dr. Michael Cochez

Eidesstattliche Versicherung

Felix Ingenerf	355026
Name	Matrikelnummer

Ich versichere hiermit an Eides Statt, dass ich die vorliegende Bachelorarbeit mit dem Titel

Concept embeddings for Wikipedia across language editions

selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Für den Fall, dass die Arbeit zusätzlich auf einem Datenträger eingereicht wird, erkläre ich, dass die schriftliche und die elektronische Form vollständig übereinstimmen. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Aachen, 2019-01-15	
Ort, Datum	Unterschrift

Belehrung:

§ **156 StGB: Falsche Versicherung an Eides Statt**
Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.

§ **161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt**

(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.

(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten dementsprechend.

Die vorstehende Belehrung habe ich zur Kentnis genommen:

Aachen, 2019-01-15	
Ort, Datum	Unterschrift

Contents

1 Introduction

1.1 Motivation

Nowadays many people around the world gather their knowledge and information not anymore from books but form the Internet. Instead of going to a library, they use the Internet and search for a topic and one website which most likely appears is Wikipedia. Wikipedia is a platform in which more or less anybody can provide information in the form of articles in mostly all languages. Only the English version of Wikipedia contains over five million articles, consisting of over 3 billion words. Every day more than 700 new articles are created only in the English edition. There are in total 293 different language editions and if all of these language editions of Wikipedia are added up, it results into 27 billion words in 40 million articles [19]. All the information is covered in multiple articles of various topics, where the requested information can be about a scientific Topic, a historic event or a current politician. In 2017 every second, 6,000 people viewed Wikipedia pages from across the globe. These people live in different parts of the world, speak different languages and belong to different cultures. To be useful for all these people, Wikipedia needs to be available in many different languages. Wikipedia approaches this challenge by establishing a community with editors providing their knowledge in the form of articles. This community is open for everybody and accordingly, virtually any person can write or edit Wikipedia articles. However, not only across one Wikipedia language edition the articles are added by many different people who do not have to be experts or share the same knowledge for the context they are working on, also the authors of one article in different language editions may differ. Therefore, it is not given that the different versions about the same topic contain the same information. This concludes into the question whether it depends on the language a person speaks what information this person gets and if there are specific biases between language editions of Wikipedia with which this person gets confronted when searching for neutral information.
Apart from identifying biases across the Wikipedia language editions, these iden-

tified biases could also help to find cultural biases across countries or regions of the world. Therefore, this thesis aims to find a way to identify biases more easily and to understand why humans behave like they behave in a specific part of the world. These biases could perhaps even explain why certain historic events took place as they did or why there are conflicts between certain groups around the world.

1.2 Research question

The goal of this thesis is to determine whether it is possible to discover differences and similarities between the language editions of Wikipedia. This goal is tackled by first classifying Wikipedia articles and then comparing the different language editions. More precisely possible methods are explored to first compute an embedding for all articles in the different language editions and then by comparing the most similar articles for one article in the different language editions. This approach is chosen to compare the associations of an article in different language editions and in such a way discover possible biases across languages and cultures. With this thesis I tackle the **research question whether it is possible to find cultural differences in Wikipedia with regard to the associations in which articles can be found**. This could be a new way to discover cultural and racial biases between countries or rather languages. In social science it can be a new achievement and help social scientists for example to find new biases or to confirm their earlier hypotheses.

1.3 Methodological approach

Over the last years machine learning algorithms got more and more important and as a computer can usually not handle objects like words adequately, every word requires a representative which was normally just a unique but meaningless id, at least manually supplemented by syntactic and semantic metadata. However, since it would be more useful to have a representative with some meaning which already carries some information about the object, embeddings are considered as such representatives. Embeddings are an approach, first used in natural language processing and they are increasingly utilized in researches about digital humanities and computational social science [1]. It was the goal to get new representatives for words. These representatives are vectors, i.e. points in space. But instead of using a space with one dimension per word, it was the goal to use a space with a much lower dimension. To reduce the dimensionality of the vectors probabilistic models are used to compute these embeddings. Further, analyzing the relation between words is then more efficiently done, due to the new representatives. Since the embeddings are vectors hence points in a multidimensional space, it is trivially done to compare these vectors and obtain the

relation of the words like this. A common example would be that the relation in the space between the embedding of 'King' and 'Man' should be the same than the relation between 'Queen' and 'Woman'. The most methods to compute embeddings are based on deep learning [3] and as an example the neural-networks language-modeling approach *word2vec* is a state-of-the art method to build word embeddings. I make use of this method throughout this work.
For the approach to compute embeddings for words the given information is always some text. Thus, for every word there are different words given, which occur in front or behind that word, so a context is given in which a word appears in. This idea is not feasible if a embedding should be computed for a whole article. Therefore, I use different approaches as the underlying information for the embedding.
Firstly, the link structure is used. That means that a graph is created out of the links which point from one article to another. This allows that the articles themselves are displayed as a node and the links are the edges. Random walks on that graph can then be understood as sentences in a text, where one article occurs before another when it has a link to that one. These walks are then a possible input for *word2vec* by Bengio et al. which produces with this input embeddings for nodes in a graph instead of words. This so called method *node2vec* was described by Grover et al. [7] before.
As a second approach the clickstream of Wikipedia is used, which is quite similar to the first idea but instead of using the whole link-network, this time edges are added for links, which actually were used at least ten times in one month by users of Wikipedia. As a result, the graph gets smaller, but maybe just contains the more relevant associations. That can produce results which show more important differences and biases between the language editions. On this clickstream graph random walks are generated again and these walks are then used as an input for *word2vec* as it is the procedure for the concept with the whole link structure.
As a last approach pre-trained word embeddings by Bojanowski at al. [4] are used. Bojanowski at al. published word embeddings for 294 languages, which are trained on Wikipedia using fastText. The idea with these pre-trained word embeddings is to get all the texts of the articles and then build an average of the embeddings for the words which appear in one article, to get an embedding for that article. That could result in a different outcome since the training of the embedding is based on the actual text, so on the words which describe that topic instead of being based on the link or rather associations to other topics.
With these embeddings it should be possible to do the same comparison as it was done before for words but now for Wikipedia articles. Comparing the embeddings can help computing the similarity of articles and therefore, it can be discovered what articles are closely associated to others. By comparing these associations of an article in the different language editions, possible differences respectively similarities or maybe rather biases would be determined between the different languages and therefore, also between countries or cultures.

1.4 Structure of the work

The following pages start with a brief summary of previous scientific work delivering the fundamental basis for the study which is done in this thesis. In the course of this work the applied methods to tackle the research question are presented while indicating technical challenges which must be faced during the implementation. These methods include approaches to collect the desired datasets from different dumps as well as approaches to train the actual embeddings for all the articles. Furthermore, procedures to compare embeddings are introduced in that part. Finally, the results of the conducted study are presented in different forms. These results are then analyzed and discussed to indicate, how the presented methods approach the stated research question. Subsequently a summary of the presented thesis follows and recommendations for future research are provided.

2 Related work

Already a variety of studies cover aspects of differences on Wikipedia. Hecht et al. [9] showed that there is a high degree of self-focus in Wikipedia editions. Meaning that there is a observable bias towards the knowledge of the editor community. For example geographic articles are substantially more often referenced in a language edition if the respective language is spoken at the location referred to. With these biases in the Wikipedia editions, Laufer et al. [11] investigated similarities in food cultures based on the common links in the respective Wikipedia descriptions. That was done by mining cross-cultural relations from Wikipedia. Apart from studies about similarities in food cultures, Callahan et al. [5] and Aragon et al. [**?**] examined differences in the description of persons. Callahan at al. compared the description of Polish and American persons in Polish and English language editions. That way, they found diverse priorities, e.g. Polish language entries are more likely to include information about professional accomplishments. However, Aragon et al. aimed to understand how social links are recorded across the different language editions and therefore, across cultures. They discovered that similarities exist between distinct groups of language editions, which can help to discover groups of languages which appear across the same cultures. Moreover, Jiang et al. [10] revealed that there is a difference across the multiple Wikipedia communities with regard to the understanding of quality. With this underlying information they analyzed the relationship between similarities in sociocultural factors and the understanding of information quality. Furthermore, Mass et al. [13] developed an open source web tool called *Manypedia* which provided the user with an easy way to compare automatically translated versions of an specific article from different language editions. Further they approached the question, if the different communities for the language editions develop their own diverse Linguistic Point of View. Apart from that, it was approached to develop a multilingual Wikipedia version by Bao et al. [2]. This so-called *Omnipedia* shows the similarities and differences of the diverse language editions and gives information about what facts are unique to a specific language and which are shared across many editions.

Apart from studies about Wikipedia, the work by Bengio et al. [3] also relates to this work. The goal of their work was to to learn the joint probability function of sequences of words in a language. The main issue they faced was the curse of dimensionality, as for example a distribution of 10 consecutive words in a natural language with a vocabulary of the size $100,000$ would have potentially $100,000^{10}-1$ free parameters. The solution Bengio et al. discovered was to learn a distributed representation for words. This way it is possible to reduce the dimension of the embedding dramatically. With these already revealed results, Galke et al. [6] examined practical information retrieval scenarios and the suitability of word embeddings for that. They discovered that word embeddings are a good way to tackle the task of practical information retrieval.
Further, Hamilton et al. [8] examined a similar question related to this work. I compare the same articles in different languages, whereas they compared the same word in different periods of time with regard to their semantics. That was done with the help of word embeddings, which were trained on text corpora from different periods of time, to discover the of change of semantic of words over time.
To present one last work, Sherkat et al. [16] examined if it is possible to build vector embeddings for Wikipedia articles, which is exactly the method I use to compare the language editions. Sherkat et al. successfully trained embeddings for Wikipedia concepts and entities.

3 Methodology

3.1 Examined languages

As it was mentioned before, Wikipedia has 293 different language editions of which seven are chosen for this thesis. For the selection a few different criteria are considered. One criterion is to choose smaller and bigger samples of language editions, meaning to have on the one hand editions like English with over 5 million articles and on the other hand editions like Hindi with just over $120,000$ articles. Furthermore, the geographic area, where a language is spoken, is analyzed, to have samples of languages which are spoken just in a specific region of the world, so only in a limited number of countries, and apart from that also languages spoken in different areas and countries around the globe. Moreover, as another criterion languages are chosen that are spoken in countries that are well distributed around the globe and in which the cultures are different as well, to have a sample of languages which are spoken by groups of people representing a good average of the whole humankind. This way, it should be possible to find on the one hand greater biases, since the cultures of the countries are quite different anyway. On the other hand with the results of these languages, it could be also possible to identify for which language editions there is a greater potential to get good results, so a more significant image of biases.
Under consideration of these criteria I choose Arabic ('ar'), Bengali ('bn'), German ('de'), English ('en'), Hindi ('hi'), Japanese ('ja') and Russian ('ru'). The mentioned country codes, which are also used by Wikipedia for respective language edition, may be used on the following pages instead of the whole name and should refer to that language.

3.2 Used concepts

No. of walks	10
Walk length	80
Dimension	128
Window	10
Min count	0
Iteration	5

Table 3.1: Parameters used for *node2vec* respectively *word2vec*.

Before it is possible to start the actual comparison of the chosen language editions, the embeddings have to be calculated. For the training three different concepts and for each of these different underlying datasets are used. First the link structure, i.e. the links from one page to another, then the clickstream, which means, that just those links are considered which were actually used, and as the last concept the article text are considered. In the following paragraphs, I will introduce each of those concepts in more detail.

Link structure As the first approach, the link structure is used as the input for the *node2vec* algorithm. As mentioned before *node2vec* is more or less a wrapper for *word2vec*. That means, that instead of considering sentences as the input for *word2vec*, simulated random walks on a graph are used as the input.
For this concept all links existing in a Wikipedia article to another article are extracted and then used as edges to construct a graph, where the nodes represent the articles. First I start with the Python package *networkx* as the format to store the graphs, since a *networkx* object can be used as input for an existing implementation of *node2vec*. Unfortunately, because some of the datasets are very large, e.g. the English dataset for the link structure with 5.8 million nodes and over 400 million edges, the *networkx* format is to memory consumptive. Therefore, it is not possible to use the existing implementation of *node2vec*. Consequently I implement a new data format to store the graph as well as a function to simulate random walks on the graphs. The resulting random walks are then used as the input for the *gensim word2vec* implementation [15] to train the embeddings for all articles. The graphs are constructed separately for each language edition and therefore, the embeddings for the different editions are in separate vector spaces. The different parameters to simulate the random walks and for the *word2vec* algorithm are displayed in Table 3.1.

Clickstream The clickstream links are applied as the underlying dataset for the second approach. Accordingly it leads to a similar procedure as for the link structure. First the graphs for different language edition have to be constructed, except this time not every link from one article to another is added as an edge, just the links which were actually used are considered to construct the graph. Like this, the graph is much smaller, but it may contain more concrete biases, since the included links just display what articles are actually viewed after one another. After the construction of the graph, random walks are simulated again

and these are used as the input for the *word2vec* algorithm as before for the link structure concept. The parameters used in this case are the same as for the first method (see Table 3.1).

Article text The last method to compute the embeddings is to use the actual text of the article and then build an average over all the word-embeddings corresponding to the words in the article.
For this concept all the article texts of the different language editions are extracted first. After this, a specific score for every word in every article is calculated. That is done with the help of the *term frequency–inverse document frequency* (*tf-idf*) score. That score shows the value of each word in a article with the help of the inverse proportion of the frequency of the in a particular article to the percentage of documents that word appears in [14]. A higher *tf-idf* score shows a high importance for that specific article, while a rather low score could appear for example in case that this word is a so-called stopword, meaning a word appearing in nearly every article e.g. 'the' or 'a'. This *tf-idf* score can then be used as a weighting score for the words in the article texts. In this method it fulfills the purpose to weight the corresponding embedding of that word and to get an embedding for the whole article by building the sum over all the weighted word embeddings.

3.3 Preparing datasets

For the training of the embeddings for all articles in the chosen languages all the necessary data has to be available. The major part of the data is extracted from different Wikimedia dumps.
First of all some general information about the articles for the different editions is essential. These information contain first of all the titles of each article and the corresponding page id, which varies for the different language editions. Moreover, the Wikidata id is important to extract, because in contrast to the page id, the language editions share the Wikidata id. Therefore, it is fundamental for this work to get this id, since it allows to map the same article in the different languages together to be able to compare these.
Apart from that some mappings have to be created. As the articles are referenced either with their page id or with their title in the different dumps, it is necessary to construct a mapping between the page id and the title. This way, it is possible to match different information for one specific article together. Furthermore, a mapping between the page id and the Wikidata id has to be constructed. That is essential as the information about all articles in one language edition can be used to train the embeddings. Nevertheless it is not possible to compare these embeddings with another language edition without knowing which is the same article in the other edition. Therefore, the mapping between the page id and the Wikidata id is required to match the corresponding article between all the

	Total Links	(1)	(2)	(3)	(4)
Arabic	71,060,941	9.2%	0.9%	4.0%	<0.1%
Bengali	2,456,649	77.6%	8.6%	16.5%	0.0%
German	94,142,228	10.7%	1.5%	5.3%	<0.1%
English	415,309,792	7.8%	2.4%	9.6%	<0.1%
Hindi	6,811,234	14.6%	1.4%	20.9%	0.0%
Japanese	82,584,210	11.5%	0.9%	4.3%	<0.1%
Russian	82,894,501	19.0%	2.4%	12.0%	<0.1%

Table 3.2: Errors during mapping articles to Wikidata id for the links structure dataset.
(1): Mapping error title to page id (target);
(2)/(3): Mapping error page id to Wikidata id (start)/(target);
(4): other Errors

language editions.
The dumps representing a snapshot of the Wikipedia corpus like the titles, page ids or links and which are used for the presented results in this work, are from the 20th September 2018.

Link structure Next to the general information, specific info is needed to indicate the link structure. For that method all the links of every article are required. The required info for this concept can be extracted from a Wikimedia dump providing the page links. After the plain links are extracted, the links have to be mapped to the Wikidata id with the help of the mapping, which was introduced before. Again the linking procedure has to fulfill as well the requirement to match different language editions.
In this mapping process from the title or the page id to the Wikidata id, many errors occurred as shown in Table 3.2. That table displays first the total links, which the dump contained, and further in the other columns the percentage of the different errors. The percentage of other errors is insignificant in this case and therefore, can be ignored. Apart from that, the mapping errors between the title and the page id for the target node of the link seems to be quite high with mostly over 10% for the different editions. Nevertheless, this can be explained by the so called *red links*. These links sometimes appear in articles linking to articles which are not created yet. Therefore, a mapping is not possible for the target node of these links. However, the mapping errors happening while either the start or target node is mapped from the page id to the Wikidata id are critical. For every language combined, the error is nearly over 5% and for more than half of the editions even over 10%. These numbers appear extraordinarily high and result into the assumption that an error in the developed mapping procedure occurred, which could not be solved yet and should be addressed in further research.
This problem could be caused by different aspects. First of all in the process of

	Arabic	Bengali	German	English	Hindi	Japanese	Russian
Pages	611,862	58,128	2,210,098	5,800,041	121,107	1,122,423	1,497,818
Links	71,060,941	2,456,649	94,142,228	415,309,792	6,811,234	82,584,210	82,894,501

Table 3.3: Number of pages and links extracted from the link structure for the different language editions.

	German	English	Japanese	Russian
Pages	565,460	4,894,939	30,484	39,265
Links	4,894,939	28,400,439	248,981	271,758

Table 3.4: Number of pages and links extracted from the clickstream for the different language editions.

extracting the general information like the title, page id and Wikidata id there could be some inaccuracy according to the filtering of the information. For example some Wikidata ids could be missing in the mapping if the specific part of the dump is not recognized as a Wikidata id. Apart from that aspect the mapping errors could also occur because of inconsistencies between different dumps e.g. it could be the case, that in one dump the title of a article is spelled slightly different than in another. If that occurs it would not be possible to fit the gained information of the different dumps together. Nevertheless these are assumptions why these errors may appear, but unfortunately it did not help in the end to fix the mapping and thereby the errors. Hence there is a relatively high amount of data loss and the datasets do not represent the whole Wikipedia corpus for the different language editions anymore.
Nevertheless, the majority of links are extracted and the datasets for the link structure still contain several articles. In Table 3.3 the size of the datasets for the different language editions are shown with the number of pages and links per corpus.

Clickstream For the clickstream method it is a similar procedure as for the link structure apart from the used dumps. In this case just the links which were actually used are needed. For this purpose Wikimedia provides a specific dump which does not display a snapshot of a Wikipedia edition but contains a collection of information over the period of one month. For this method the monthly clickstream dump is used which contains all links that were used at least ten times in that month and states the monthly number of clicks for these link as well. Unfortunately, in 2018, these specific dumps are only available in eleven languages (Chinese, German, English, French, Japanese, Persian, Polish, Portuguese, Russian, Spanish) and therefore, it is only possible to adapt this method for four of the seven chosen language editions. Apart from that, the links have to be mapped to the Wikidata id again. However, this time the sum of all occurring errors is less than 4% for all language editions, which is still reasonable. The statistics of the process for the different languages are displayed in detail in the appendices (see

	Arabic	Bengali	German	English	Hindi	Japanese	Russian
Pages	48,447	13,551	531,362	594,958	18,299	846,594	320,966

Table 3.5: Number of pages extracted from the article text dump for the different language editions.

Tables 1 to 4). Afterwards, the extracted links are used to construct a graph.
After the mapping and the construction of the graphs a different problem occurred. The graph contains just the used links and especially in the Russian and Japanese language edition a fairly low amount of links are used at all. It was difficult to train stable embeddings with such small datasets. As a result the possibility emerged that not only the links from the dump of one month are extracted but also of several months. As Wikimedia started in November 2017 to create these monthly clickstream dumps, it was just possible to join all clickstream links from November 2017 till September 2018, a time-period of eleven months.
After extracting links from eleven instead of one month, the corpora of the four language editions contained more articles, for the absolute numbers see Table 3.4. By comparing the clickstream corpora and the ones from the link structure concept, some aspects are remarkable. To begin the number of pages for the English edition is nearly the same and for the German edition it is only about a quarter. However, both clickstream datasets contain only about 5% of the links compared to the link structure datasets. This is not really surprising as not all links are used. This fact is supporting the intention conducted in this approach as it could help to extract more highly distinct biases. Unfortunately, the corpora of Japanese and Russian still show the same issue as before. In comparison with the link structure datasets, both editions contained more than one million pages and over eighty million links in the link structure method, however, in the clickstream dataset there are less than 3% of the pages and less than 0.5% of the links left. This problem could lead to the case that by comparing these editions, it is not possible to discover distinguished biases.

Article text The underlying dataset for the third approach conducted contains the whole article texts and the word embeddings for the different language editions. For the word embeddings I considered pre-trained word embeddings which were published by Bojanowski et al. [4]. These embeddings are in the dimension 300 and thus the resulting embeddings for the articles have the same dimension. Bojanowski et al. trained word embeddings for several hundred languages on the Wikipedia corpus. Hence they fit good to the data used in this work.
Since the word embeddings are mapped to the actual words, it is also necessary to extract all article texts. As for the other two approaches, Wikimedia dumps contain these information wherefore article texts are extracted from these dumps as well. In Table 3.5 the number of pages in the final datasets are displayed. To be able to connect the different language editions, it is necessary again to map the

	Total Articles	Saved Articles	(1)	(2)
Arabic	525,296	48,447	90.8%	<0.1%
Bengali	59,893	13,551	77.1%	0.2%
German	2,153,920	531,362	75.3%	<0.1%
English	5,126,563	594,958	88.4%	<0.1%
Hindi	97,327	18,299	80.8%	0.4%
Japanese	1,034,271	846,594	15.3%	0.1%
Russian	1,448,868	320,966	77.8%	<0.1%

Table 3.6: Errors during mapping articles to Wikidata id for the article text dataset.
(1): Mapping error title to page id;
(2): Mapping error page id to Wikidata id

articles to the Wikidata id. In this case it worked even more poorly than before for the link structure concept. Unfortunately it was just possible to map about $10 - 20\%$ of the articles to the Wikidata id with the exception of the Japanese edition with a success rat of about 85% for mapping articles. As it is shown in Table 3.6 the most errors occurred while trying to map the title to the page id. Unfortunately, I did not succeed in fixing this problem again and as a result the dataset for this concept is significantly smaller than it could be. That means there is a huge data loss compared to the original Wikipedia corpus which affects the possibility of finding biases.

3.4 Analyzing embeddings

To evaluate the different approaches conducted in this thesis, the different datasets and the embeddings are compared. The standard method to compare embeddings is to analyze relation between the points of the embedding in the vector space. This method cannot be applied for the presented study because the language editions are trained separately and therefore, every language edition is trained in their own vector space. The chosen method to compare a ranking of the most similar articles which means that the association of articles in different languages is analyzed.
The rankings of the most similar articles are highly differential due to the fact that by comparing to editions there is a high quantity of articles which only exist in one of them. Hence, the rankings are adjusted so that a comparison between two languages the corresponding rankings only contain articles which exist in both language editions.

Similarity Scores Two different scores are used to which represent the similarity between two rankings and hence, the similarity for an article of the two corresponding language editions. The two scores are described by Webber et

al. [18]. In their work they established the *ranked-biased overlap* (*RBO*) score. Usually there are three different scores given by *RBO*, a base, a maximum and a extrapolated score. In this work the *RBO* score refers to the extrapolated score, as Webber et al. suggest to use this score if a point estimate is required. The extrapolated score RBO_{EXT} is defined in equation 3.4. The second score which is used throughout this work is the *overlap* score $O_{S,T,d}$, which is part of the extrapolated score and is defined in equation 3.3.
Let S and T be rankings and let S_i be the element at rank i in list S. Further the set of elements from position c to d in list S equals $S_{c:d} = \{S_i : c \leq i \leq d\}$ and thus $S_{:d}$ equals the set of elements from the beginning till the position d. With this set of elements it is possible to calculate the intersection $I_{S,T,d}$ of two rankings for a given depth d:

$$I_{S,T,d} = S_{:d} \cap T_{:d} \tag{3.1}$$

Further let X_k be the length of the intersection for a given depth k.

$$X_{S,T,k} = |I_{S,T,k}| \tag{3.2}$$

With the length of the intersection it is then possible to define the *overlap* score which is defined as the length of the intersection divided by the depth (see equation 3.3). The value of the *overlap* score is between 0 and 1 and with raising value, it indicates a greater similarity of the two rankings.

$$\mathbf{O_{S,T,d}} = \frac{|\mathbf{I_{S,T,d}}|}{\mathbf{d}} = \frac{\mathbf{X_d}}{\mathbf{d}} \tag{3.3}$$

To define the *RBO* score, the *overlap* score is essential as well. That score $RBO_{EXT}(p, k)$ can be seen as the sum over all *overlap* scores O_d with $1 \leq d \leq k$, where all *overlap* scores are weighted with p^d (see equation 3.4). Therefore, the parameter p can be interpreted as the probability of taking rank $i+1$ into account after having examined rank i. In this thesis $p = 0.9$ is used which implies that the first 10 articles of a ranking have 86% of the weight. As well as for the *overlap* score the value of the *RBO* score is between 0 and 1 and with rising value, it indicates a greater similarity of the two rankings.

$$\mathbf{RBO_{EXT}(S, T, p, k)} = \frac{\mathbf{X_k}}{\mathbf{k}} \cdot \mathbf{p^k} + \frac{\mathbf{1 - p}}{\mathbf{p}} \sum_{\mathbf{d=1}}^{\mathbf{k}} \frac{\mathbf{X_d}}{\mathbf{d}} \cdot \mathbf{p^d} \tag{3.4}$$

Compare single articles and topics First of all I compared single articles to investigate, if existing differences between the editions can be detected. This is done by calculating the similarity score for that article. For this, several different articles are chosen by the criteria to have on the one hand controversial articles like "Homosexuality" and "Abortion" and on the other hand some articles with more or less distinct information e.g. about politicians like Donald Trump or

Angela Merkel. However, for these articles the standings in different countries could vary.
Furthermore, I compare topics as well. To be able to do that, sets of elements containing just articles about a specific topic are demanded. This is done by selecting an article which title can already be seen as a topic and then by building a set with all the articles which are linked from the original article. After the set is constructed, the similarity score for every single article in this set is calculated. Then, the similarity score for the whole topic is the average over all scores of the articles in the set. For this method I chose again some controversial articles e.g. "Feminism" or "LGBT". Further some historic events like the first and second world war and apart from that also the articles containing the politics of some countries for example the article "Politics of Japan". At last some articles are chosen which are about a general subject e.g. "Sports" or "Geography".
As a last approach an overall similarity score for the different language editions is calculated. For this purpose the articles with the most pageviews in September 2018 are extracted for every language edition. The similarity score between two language editions is then defined by the average over the similarity scores for the fifty most viewed articles in the one and the other edition whereby just articles in the rankings are considered which exist in both language editions. With this score it is intended to see if there is a general similarity between specific language editions.

Plot 2-dimensional embeddings As it was mentioned in the beginning it is not possible to compare the embeddings of the different language directly. Therefore, the solution is developed to compare the most similar articles. Nevertheless, there is a way to use the actual embeddings to analyze differences in a visual way. For this purpose the embeddings which are based on the link structure are reduced in terms of the dimension from a 128-dimensional space into a 2-dimensional space. Like this, it is possible to plot embeddings of specific articles and analyze the relation between these articles in one language edition. By comparing the plots of two different language editions certain differences can be detected.
For the dimension reduction I used the machine learning algorithm *t-distributed stochastic neighbor embedding* (*t-SNE*). The *t-SNE* algorithm is a technique to visualize high-dimensional data in a low dimensional space [12]. After I started to use *t-SNE* on a single core, the problem arose that it takes too long for such big datasets. Therefore, I use the *multicore t-SNE* implementation by Ulyanov [17] which saved a lot of time in that process.
For this method the most viewed articles in September 2018 are used again, only in this case just the fifteen most viewed per language to keep a good clarity of the plot.

4 Results

In this chapter I present some of the results regarding the similarity score between the language editions. To start of, Figure 4.1 shows an average comparison of all editions computed with the link structure embeddings. The score for a language pair is calculated by averaging the scores of the fifty most viewed articles of the respective language edition and of the other. This way, the resulting scores should show the similarity in general. As it can be examined in Figure 4.1, the similarity scores are fairly even for all language pairs. That implies that the differences between the various language pairs are not highly variegating for articles about general information. However, that does not indicate, that the different editions are still rather different, as the *RBO* score stays below 0.5 and the *overlap* score even below 0.3.

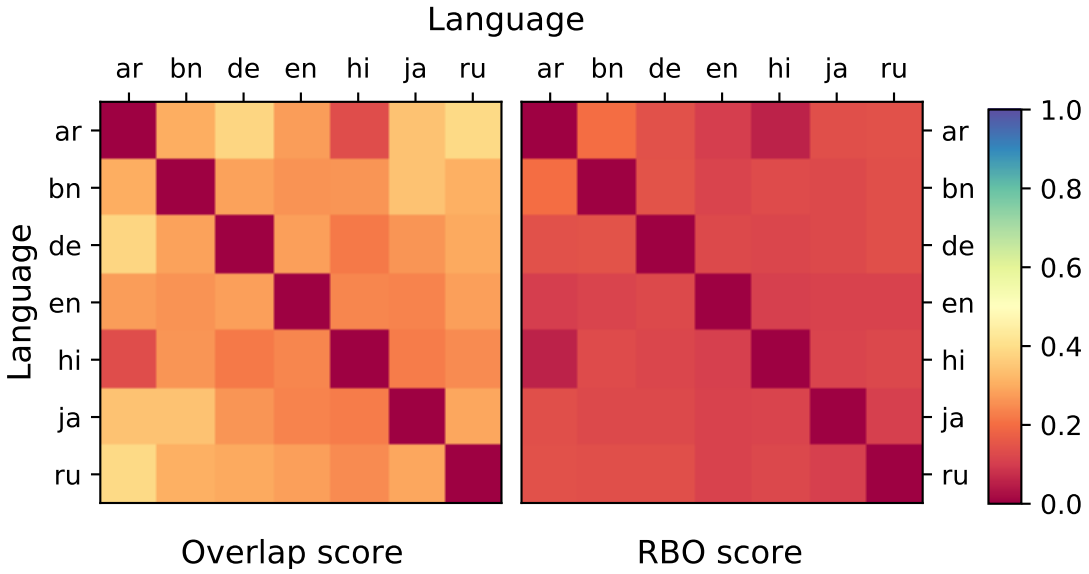

Figure 4.1: Average over computed similarity scores for the most viewed articles for the link structure method.

Furthermore, the average similarity score of all analyzed single articles is presented in Figure 4.2. This plot appears to be different compared to Figure 4.1. That can be explained by the aspect, that the single articles cover more controversial topics, which can lead to a higher variation in the similarity score. In Figure 4.2 it is visible, that German, English, Japanese and Russian seem to result in similar scores. That my be caused by the size of the dataset as the dataset of theses three editions are the largest. Nevertheless, it implies the assumption that there is a greater correlation between the cultures and the general perspective on things. Apart from that, the other language editions seem to be quite different from one another. Especially Hindi has a significant lower score, which may imply, that Hindi is fairly different with regard to culture and the overall viewpoint corresponding to the Wikipedia corpus.

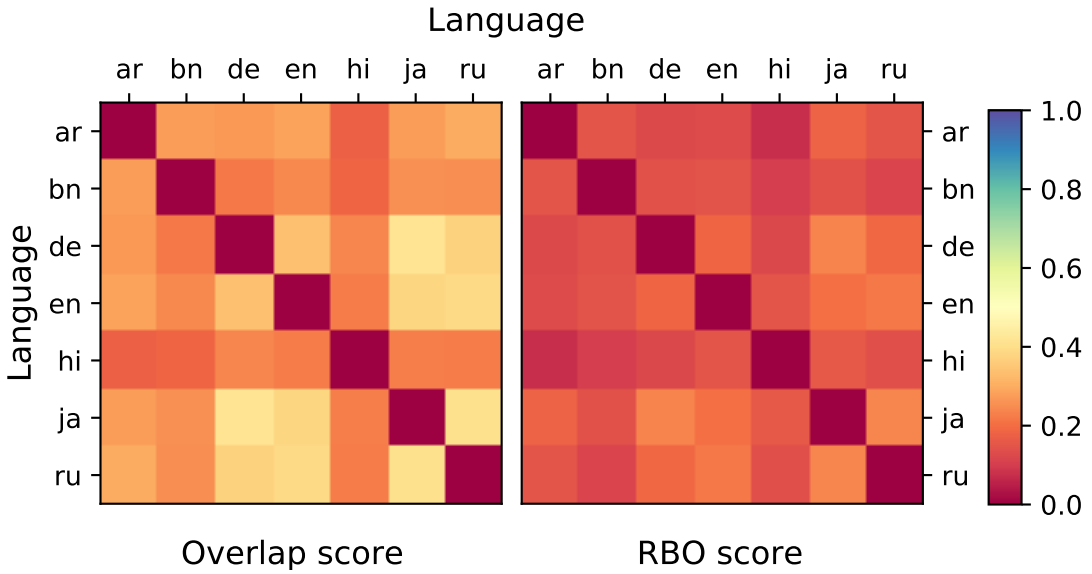

Figure 4.2: Average over computed similarity scores for analyzed single articles for the link structure method.

4.1 Link structure method

After a general comparison, in this section results are presented for the link structure method starting with the comparison of single articles, followed by analyzing topics.

Single article The first presented result is for the article *Homosexuality*. Therefore, the similarity scores are displayed as a heat map in Figure 4.3 and the scores of the comparison between Hindi and the other editions can be examined as a simple line plot in Figure 4.4. In Figure 4.3 it is remarkable that the score for Hindi compared to the other languages is extremely low compared to the other

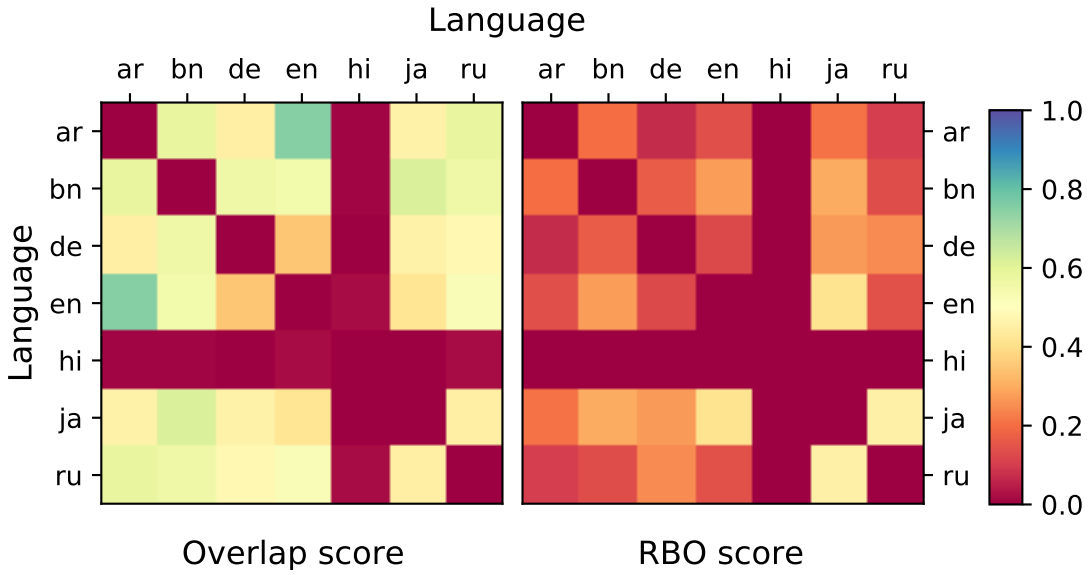

Figure 4.3: Computed similarity scores for the article ***Homosexuality*** for the link structure method.

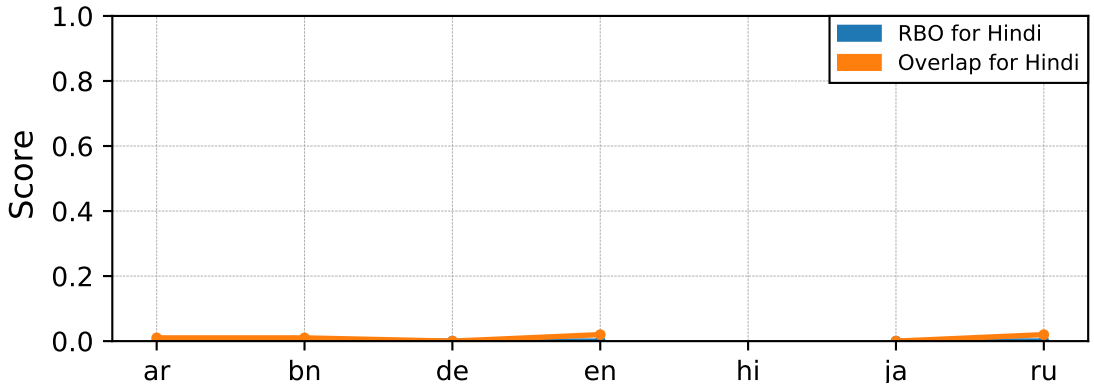

Figure 4.4: Computed similarity scores for the article ***Homosexuality*** for the link structure method. Comparing Hindi to the other languages.

similarity scores. This effect is even more outstanding for the *overlap* score. Furthermore, as presented in Figure 4.4 both similarity scores are close to zero for every comparison between Hindi and another language. That implies that the Hindi Wikipedia article about *Homosexuality* is significantly different associated in comparison to the other editions. This recognition indicates a difference in culture, which would make sense for this specific article because homosexuality was illegal in India until September 2018[1]. Nevertheless, a difference for Arabic compared to the other editions is expected as well since homosexuality is illegal in many Arabian countries. However, this can not be examined in the results.

Next some results for the article *Feminism* are presented in Figure 4.5. In the

[1]https://www.theguardian.com/world/2018/sep/06/indian-supreme-court-decriminalises-homosexuality

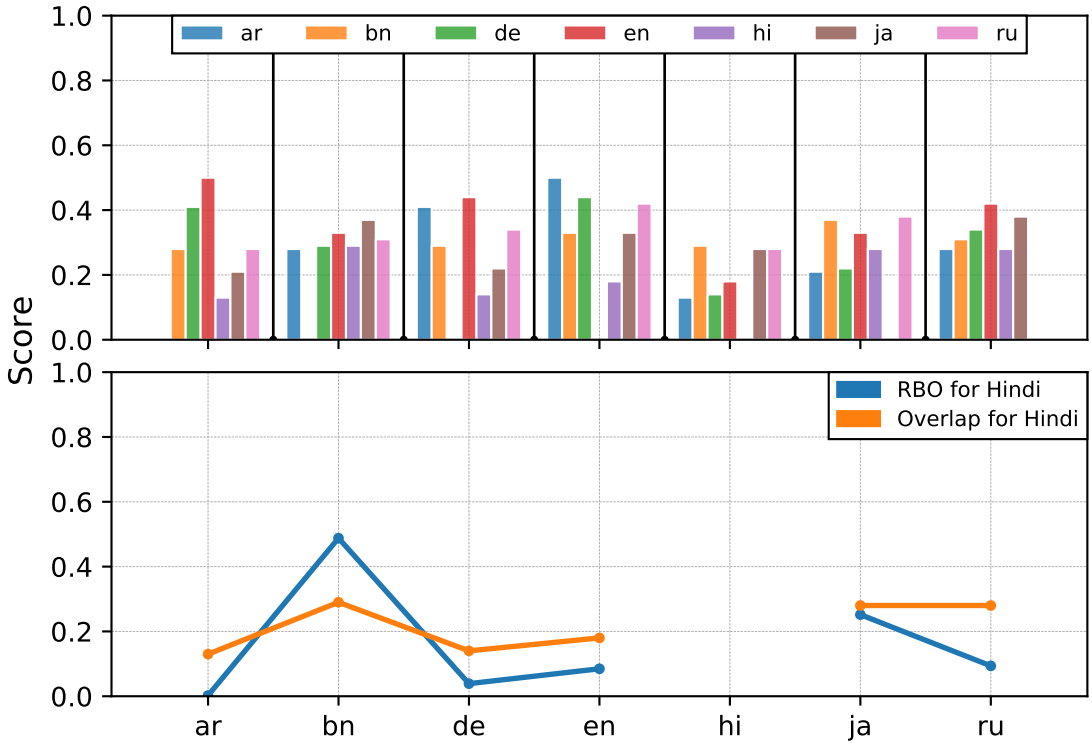

Figure 4.5: Computed similarity scores for the article ***Feminism*** for the link structure method. In the top graph the *overlap* score is displayed and the bottom graph shows both similarity scores for Hindi compared to the other editions.

top graph the *overlap* score for the comparison of every language edition pair is plotted and in the bottom graph both similarity scores for Hindi compared to the other language editions are displayed. The *overlap* score for the comparison with Hindi is a bit lower but not as significant as before. However, in this case the bottom graph is remarkable. It displays both similarity scores and this way it can be observed that these two scores do not have to be proportional to each to other. According that fact, it is important to analyze different scores because they may imply different information. The *overlap* score only represents how many identical articles appear in the given rankings while the *RBO* score also considers the rank of an element in a ranking. Like this, the scores can hold different implications and maybe even a variation of the parameter p in the *RBO* score could benefit into more precise results.

Moreover, the similarity scores for the article *Muslim* are plotted in the Figure 4.6. For this article the *overlap* score for the English edition is relatively lower than for the other editions as it can be examined in the top plot. Further, the *RBO* score is even lower than the *overlap* score in this case (see bottom graph). This outcome is surprising as I assume there should not be a big difference for the article *Muslim* especially for a comparison between for example German and

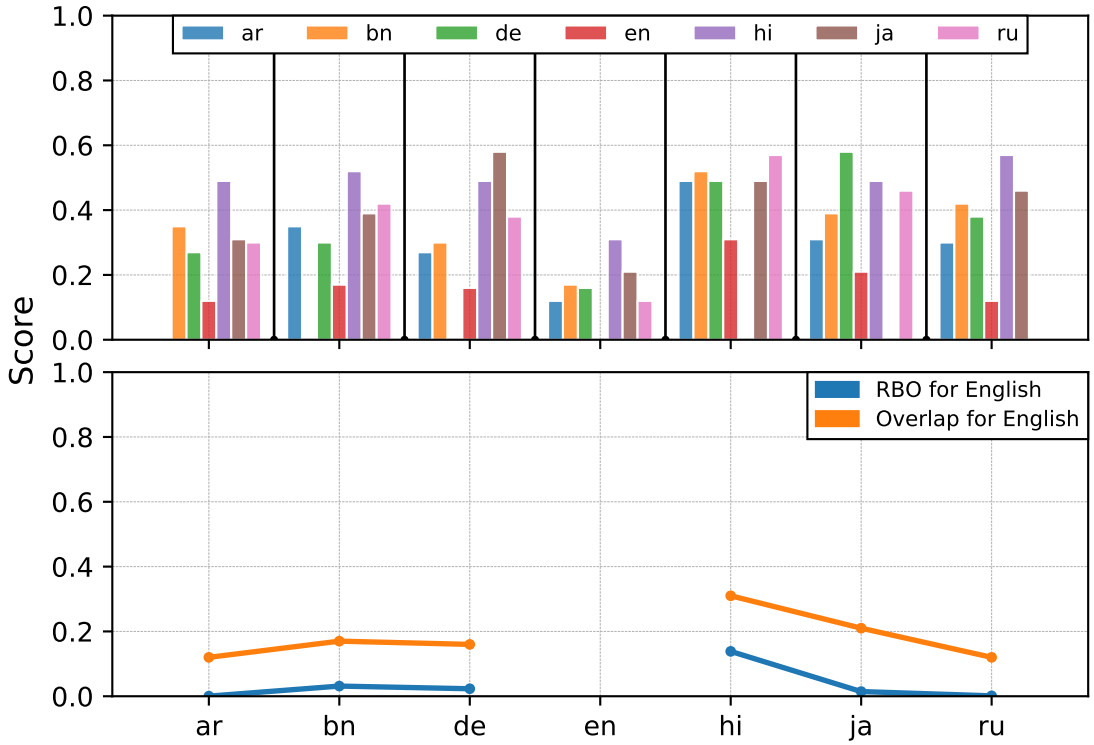

Figure 4.6: Computed similarity scores for the article ***Muslim*** for the link structure method. In the top graph the *overlap* score is displayed and the bottom graph shows both similarity scores for English compared to the other editions.

English, as I would not consider a big difference between German and English in culture or the viewpoint about the topic *Muslim*. However, this result could be explained by the fact, that the English edition is much larger than the others and therefore, different associations could occur.

Finally, as the last result for single articles with the link structure method, the article *Abortion* is considered. As the heat map in Figure 4.7 reveals, the similarity scores for Arabic are conspicuously low, even close to zero for the *RBO* score as it can be observed in Figure 4.8. This exposes a difference in culture which can be discovered outside of Wikipedia as well. This is borne out by the aspect, that abortion is illegal in most Arabian countries and it is rather a taboo topic. For further results of single articles for the link structure method, see the Tables 5a to 22 in the appendices.

Topic Apart from analyzing single articles, topics are compared as well. In Figure 4.9 I present the *RBO* score for the topics *Politics in Japan* and *Politics in the United States*. The corresponding results only represent the languages Arabian, German, English, Japanese and Russian because the two corresponding

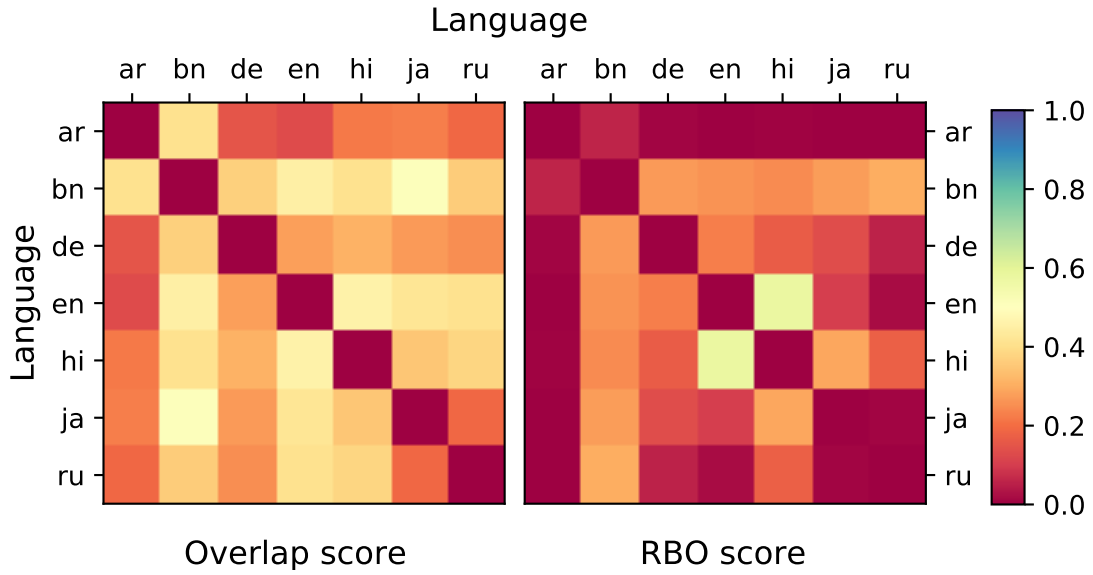

Figure 4.7: Computed similarity scores for the article ***Abortion*** for the link structure method.

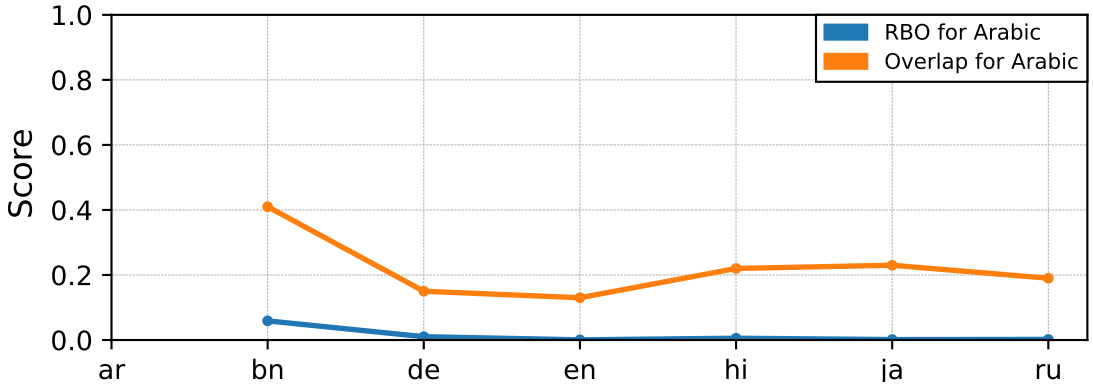

Figure 4.8: Computed similarity scores for the article ***Abortion*** for the link structure method. Comparing Arabic to the other languages.

articles for the topics do not exist in the Bengali or Hindi dataset for the link structure method. What is remarkable about these results is that the *RBO* score for all language edition pairs is fairly even. In contrast I would expect that there is a significant difference for the Japanese edition in the top graph and for the English edition in the bottom one, because I assume that the articles and the corresponding information about the politics in a specific country are more detailed and explicit in the corresponding language edition. Because this difference can not be examined from the results, the question raises why this expectation does not fit the data.

To provide a possible answer to that question, the *RBO* score for the article *Feminism* is displayed in the Figure 4.10. The top plot presents the scores for the single article and the bottom one for the topic of the article *Feminism*. From this

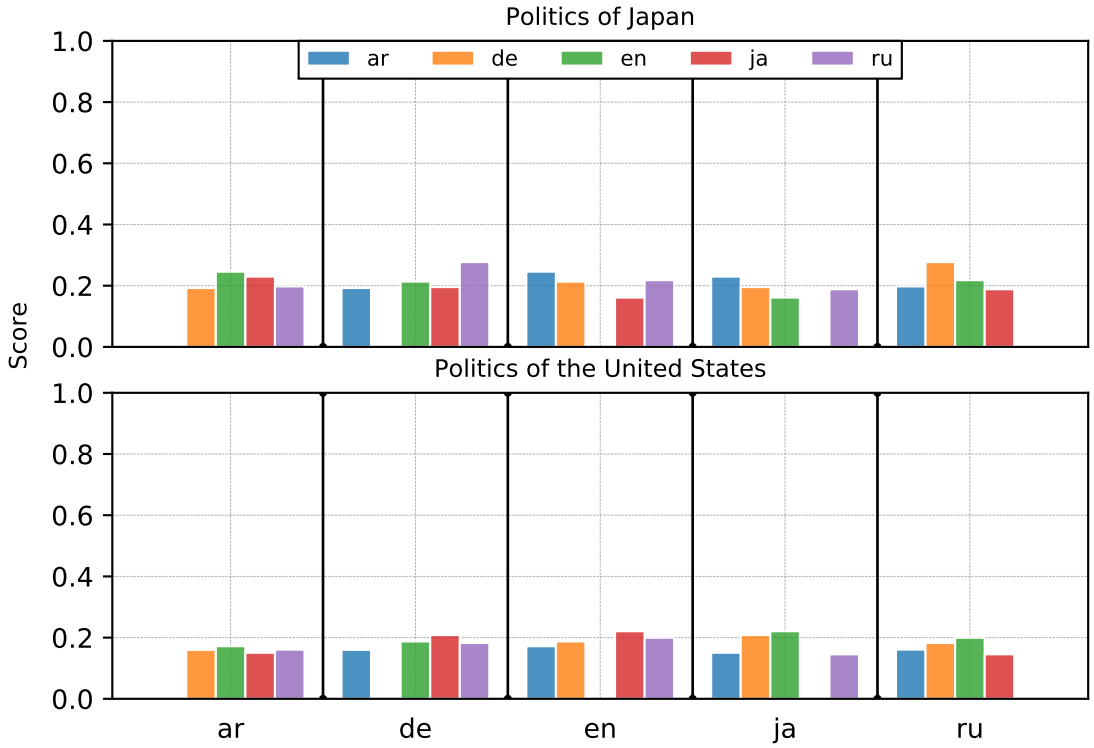

Figure 4.9: Calculated *RBO* score for the two topics ***Politics of Japan*** and ***Politics of the United States*** with the link structure method.

Figure it can be observed, that the scores for the single article differ fairly more whereas the results for the topic are rather even. That aspect of the data can be a result of the fact that due to the averaging over multiple articles, the differences between language editions get evened out. This way, differences discovered for a single article are not as significant for the corresponding topic. Furthermore, that would explain the question from before. Considering that due to more articles, the differences between the language editions get evened out, it does not surprise anymore, that the is not much difference for the Japanese and English edition in Figure 4.9.
More results of other topics for the link structure method are displayed in the appendices in the Tables 23 to 40.

4.2 Clickstream method

In this chapter empirical results as outcomes of the clickstream method are presented, as far as they could be generated.

Unfortunately it is not really possible to compute valuable results out of the clickstream dataset. The data for this method is only available in four of the examined languages and furthermore, the datasets for Japanese and Russian are

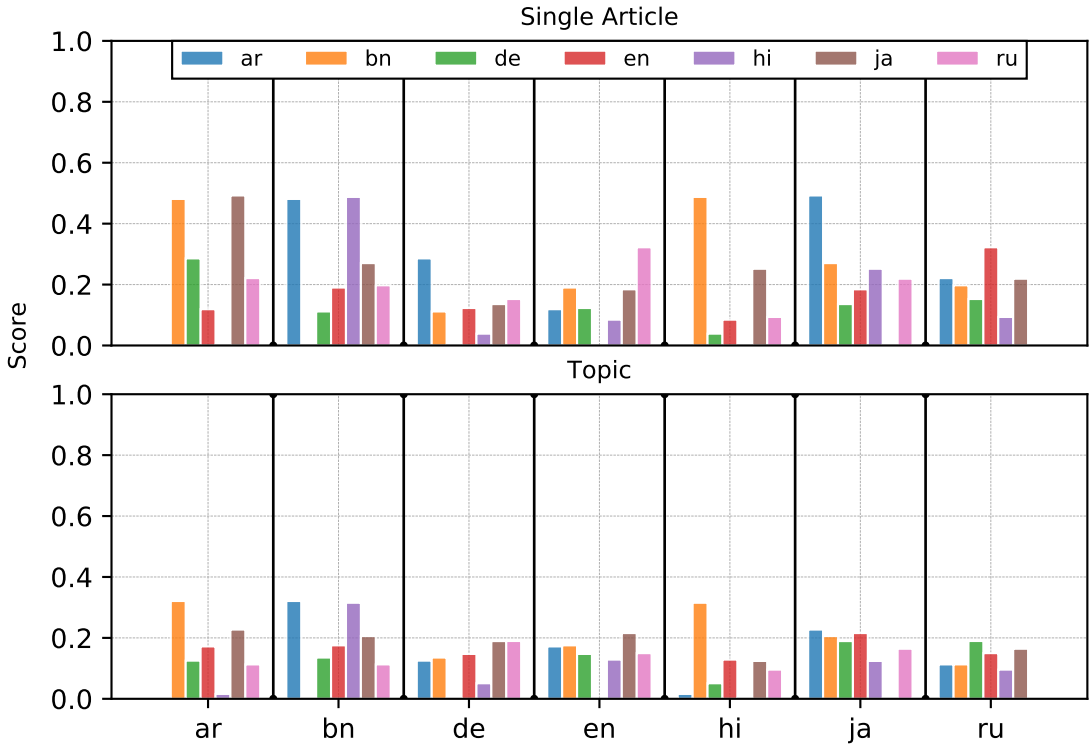

Figure 4.10: *RBO* score for the article ***Feminism*** with link structure embedding. In the top graph showing scores for the single article and in the bottom graph for the topic.

	English	Japanese	Russian
German	0.32	0.00	0.00
English		0.00	0.00
Japanese			0.01

Table 4.1: Overlap score for the article ***Feminism*** for the clickstream method.

fairly small. That is why, many of the analyzed articles do not exist in the dataset of these two languages. As result it is only possible to compute the two similarity scores for the German and English edition. Therefore, there is no comparison with regard to the similarity score of different language edition pairs possible. As a result, it is not likely do detect any biases for these articles.
However, some of the articles exist in all of the four language editions and the

	English	Japanese	Russian
German	0.21	0.00	0.00
English		0.00	0.00
Japanese			0.03

Table 4.2: Overlap score for the article ***Dictatorship*** for the clickstream method.

overlap score of two articles are displayed in the Tables 4.1 and 4.2. Unfortunately even these scores do not present any insights of the differences between the language editions since the score for the comparison with Japanese or Russian is either equal to zero or fairly close to zero. This effect can also be observed for all other results for the clickstream method in which all four language editions appear. Further results are displayed in the appendices in the Tables from 41 to 73.

4.3 Article text method

In the following section I present the results of two different articles for the article text method. On the one hand these results are analyzed separately and on the other hand these results are compared to the equivalent results from the link structure method.

The first results presented in this section are for the article *Feminsim*. The

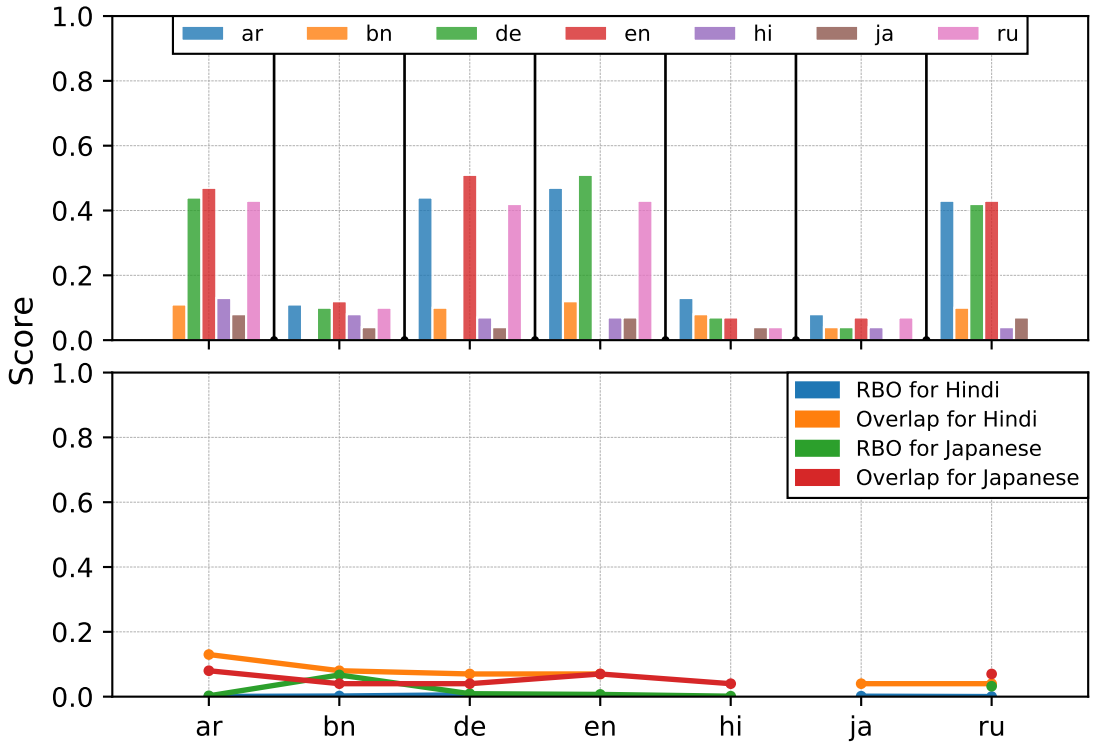

Figure 4.11: Computed similarity scores for the article ***Feminism*** for the article text method. In the top graph the *overlap* score is displayed and the bottom graph shows both similarity scores for Hindi and Japanese compared to the other editions.

corresponding similarity scores are displayed in Figure 4.11 where the top graph presents the *overlap* score and the bottom graph both similarity scores for Hindi

and Japanese compared to the other language editions. The fairly low comparison score for Hindi is outstanding for this article. However, the scores for Japanese and Bengali are equally low for this method. In comparison to the results from the clickstream method, this concludes that the different embeddings from the different show some similarities as the score for Hindi is fairly low in both results.

Furthermore, if the results for the article *Abortion* are analyzed a different out-

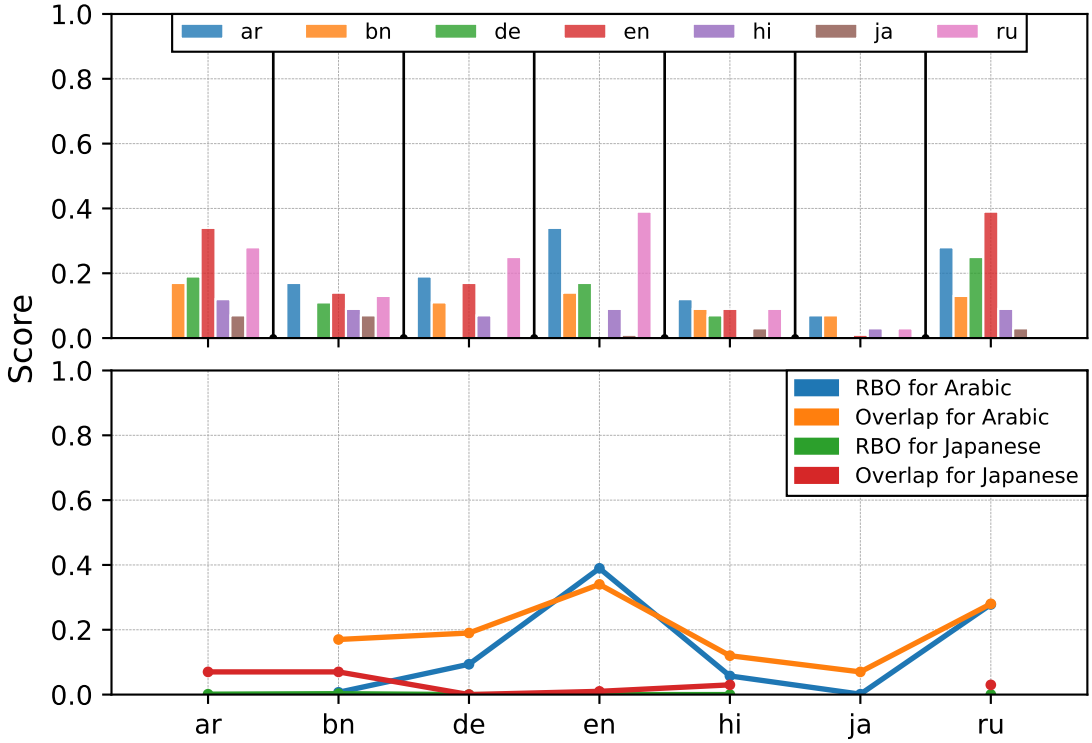

Figure 4.12: Computed similarity scores for the article ***Abortion*** for the article text method. In the top graph the *overlap* score is displayed and the bottom graph shows both similarity scores for Arabic and Japanese compared to the other editions.

come is produced. The similarity score for that article are displayed in Figure 4.12 in the same way as for *Feminsim*. In this case Japanese and Hindi have the lowest similarity scores and as the bottom graph shows, the *RBO* score for Japanese is even lower than the *overlap* score. This is in contrast to the outcome of the link structure method, as in that case the scores for Arabic are the lowest and mostly even close to zero. As it can be examined in the bottom graph in Figure 4.12, the score for Arabic is fairly different for the article text method and is even close to 0.4 for the comparison with the English edition.

These observations can lead to two different conclusion. It can be due to different information and biases which are contained in the specific underlying data. However, it can also reveal that at least one of the methods is inaccurate and produces results of lower quality.

The results for other articles and topics are displayed in the appendices in the Tables 74a to 93.

4.4 2-dimensional plots

As the last part of the result chapter I present some 2-dimensional plots for visualization results. These can be used to analyze the differences between the language editions in a visual way and also to use the actual embeddings for that process instead of the most similar ranking. For these plots I used the embeddings from the link structure method.

The first presented result in this section is displayed in Figure 4.13. For this

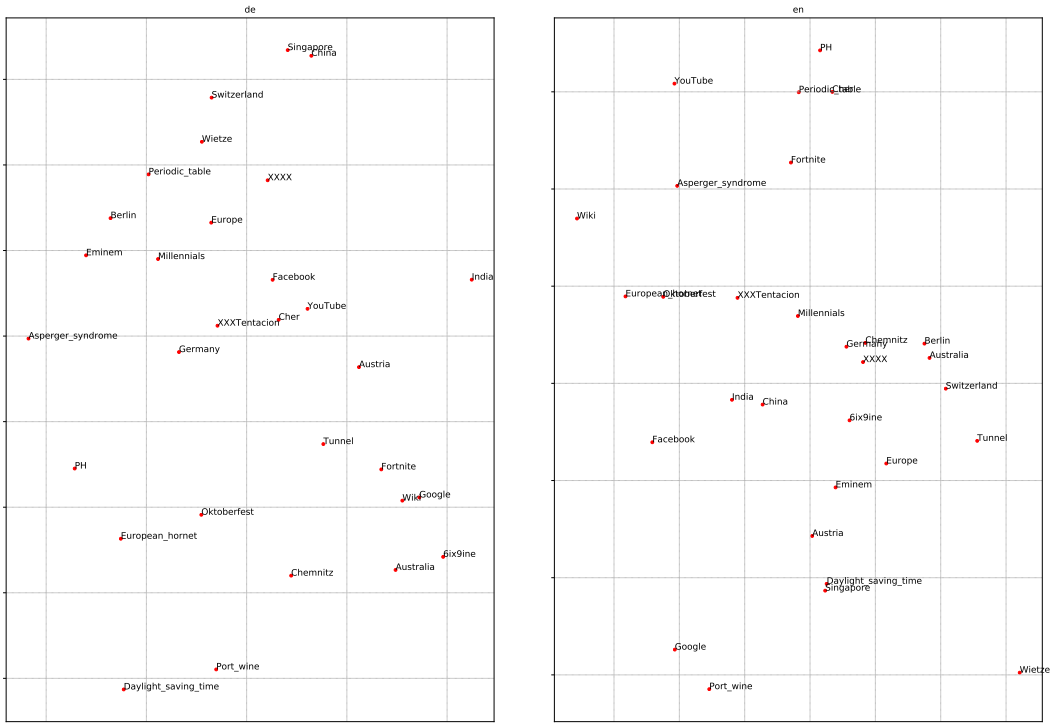

Figure 4.13: 2-dimensional plot of fifteen most viewed articles in German and English.

Figure the fifteen most viewed articles of the English and the German edition are plotted in two dimensions. These plots demonstrate, that the embeddings definitely carry enough information about the different articles, as some obvious relations can be examined with the help of these plots. For the German edition (see the left plot) *Singapore* and *China* are close together in the top of the plot as well as *Facebook* and *Youtube*. Both connections making sense as the articles share common topics. That is also suitable for *Periodic Table* and *PH* in the plot

for the English edition. However, in both plots the points of articles about countries and cities are well scattered over the plot. That aspect seems to be unusual, but could be explained by the fact, that for example the culture of a country is described in the corresponding articles as well and therefore, the associations of different countries with e.g. different culture have fairly different embeddings.

As the last result which is presented in this work, the 2-dimensional plot for the

Figure 4.14: 2-dimensional plot of fifteen most viewed articles in German and Japanese.

comparison between the German and the Japanese edition can be examined in Figure 4.14. The two plots in this Figure are created in the same manner as the first one. As before, some relations can be analyzed in both language plot. On the one hand the points for the articles *Serena Williams* and *Naomi Osaka* are close together in both samples (at the top for he German version, on the right for the Japanese version). That relates well to the reality as both are famous tennis player. On the other hand a connection between *Gin Tama*, *Attack on Titan* and *My Hero Academia* can be examined as these articles are close together on the left in the German scatter plot and in the middle in the Japanese one. Also this aspect relates well to the reality as all three are mangas. However, the point for the article for *Namie Amuro*, who is a Japanese singer, is more outstanding. In the Japanese plot that article is close to the group of mangas, although for the German plot that article is on the complete other side of the plot. This effect

can be explained with the aspect that a Japanese singer is probably described in more detail in the Japanese edition than in the German edition. Therefore, the implied information and also the associations in the two language edition vary presumably in size and detail.
The presented results for the 2-dimensional plot allows to presume that the embeddings for the different language editions represent information and associations about the real world and that is why, the comparison of the 2-dimensional plots can help to detect differences and therefore, biases between language editions.

5 Critical reflection and limitations

Apart from presenting the results, it has to be mentioned that due to the limited time frame some aspects could not be solved and remain for future research. One substantial issue accurred when mapping the articles to the Wikidata id. For the link structure and the article text method a significant data loss is detected. That is why, the results probably do not completely reflect similarities or differences between the original language editions.
Furthermore, the stability of the embeddings can be an issue for this work as it is described in the next section (see 5.1). It can be the case that the embeddings are not trained with enough data and therefore, did not converge. That would lead to unstable embeddings and also to unstable sets of most similar articles.
Moreover, the clickstream method did not produce results which were valuable probably because of the size of the corresponding dataset. As the Japanese and Russian version is really small, it led to similarity scores which did not seem to represent the reality. As the Wikimedia dumps for the clickstream are not available for a longer period of time, it is hard to solve this problem.
After all one of the most important questions is what bias could be extracted from Wikipedia anyway. Certainly it is actually not a bias of a nation or a cultural group. But also and in particular it is a bias which is created by the editors, so a small group which may represent the group of people speaking that language. That aspect was already described by Hecht et al. [9]. However, even if these editors do not represent that group of people well, it is still of high value to detect the containing biases, because all people who read the articles in the end are affected by these biases, since they expect the correctness of the information written in and implied by Wikipedia articles.

5.1 Stability of embeddings

As mentioned before the stability of embeddings can be an issue for the results of this work. Antoniak et al. [1] showed in their work that embeddings are much

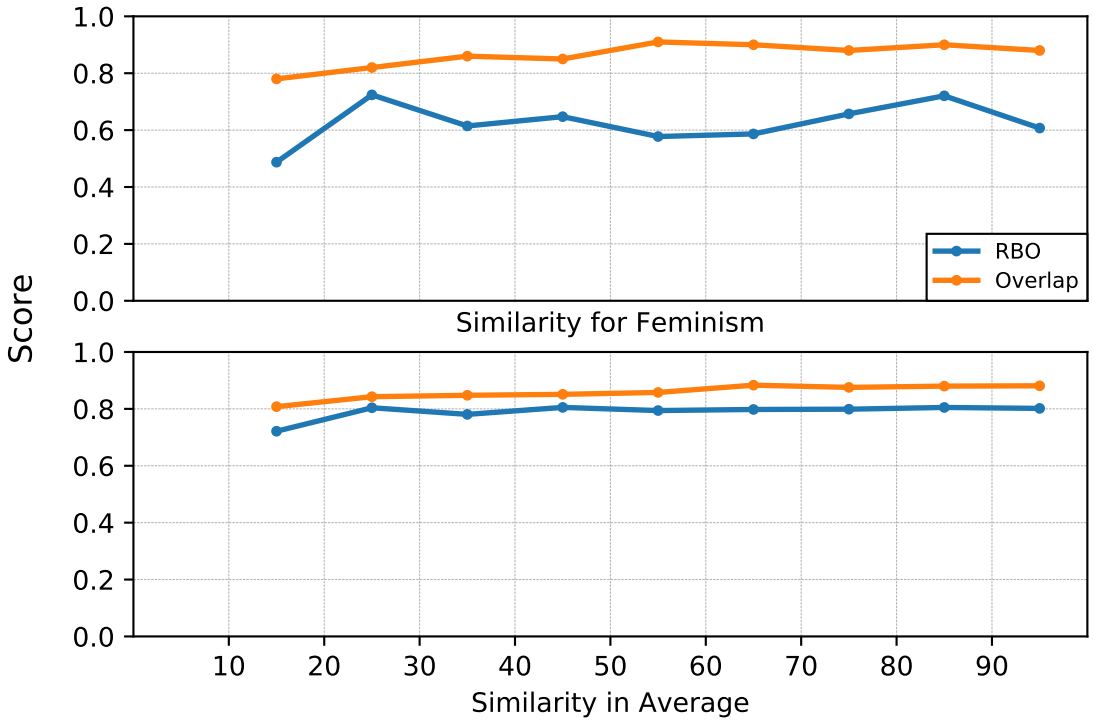

Figure 5.1: Computed similarity score between German link structure embeddings trained with x and $x+10$ walks per node. The top graph shows the similarity score for the article "Feminism" whereas the bottom graph shows the average similarity score for all compared single articles.

more sensitive as it seems to be. This sensitivity is caused by different factors e.g. presence of specific documents, the size of the documents and the size of the corpus. Antoniak et al. showed in their work that embeddings can have a great variation if they are trained on random initializations which leads to instability of the most similar words lists. Since in this work random walks are considered as the initialization, it can result in exactly this instability of most similar rankings. Because of the simulation of random walks, it is not given that each time a embedding is trained a specific document or rather walk is considered. As a solution Antoniak et al. recommend to average over multiple samples to produce more stable most similar rankings and therefore, also more valuable results in the end.
Furthermore, the embedding is expected to converge, as it gets trained with more and more data. If it does converge, it will not change significantly anymore while it is trained with more data. For this work, this aspect is tested for the German link structure model for which the results are displayed in figure 5.1. That results were calculated by training an embedding first with ten walks per node, afterwards add ten more walks per node and so forth. For each embedding

the similarity score was calculated between that embedding and the ones with ten more and ten less walks per node. In figure 5.1 the similarity scores for the article "Feminism" are displayed in the top graph and in the bottom graph average over all calculated similarity scores of the single articles. I did expect that at some stage the score would converge to 1, but the *overlap* score stays stable just below 0.9 and the *RBO* score stays stable just above 0.8. That leads to the conclusion, that the trained embeddings used in this work are not ideal. Maybe the embeddings need even more training. Using the average over multiple samples could be another approach to solve the issue.

6 Summary and future work

6.1 Summary

Is it possible to find cultural differences in Wikipedia with regard to the associations in which articles can be found? That is the research question I tried to answer in this work and therefore, the goal was to determine whether it is possible to discover differences respectively find similarities between the language editions of Wikipedia.
Three different methods are conducted, each with different underlying data. The first method is based on the link structure, meaning every link from one Wikipedia page to another. This link network is then used as the input for the *node2vec* algorithm. The second method is quite similar to the first, however instead of the link structure, this method is based on the clickstream links, so just on the links which were actually used. The third method is based on the actual article text, whereby a article embedding is calculated by building the sum over all word embeddings of the containing words, where the embeddings are weighted with an *tf-idf* score.
Furthermore, information about the different Wikipedia editions is extracted. To begin with some general information like the title, page id and the Wikidata id. Apart from that, specific information for the different methods are extracted as well, e.g. all links or the article texts. All this data is mapped to the corresponding Wikidata id because the different editions share that id. Through this process a lot of errors occurred, wherefore there is some data loss and the results are not calculated on the full Wikipedia corpus.
For the analysis of the embeddings, the lists of most similar articles of one article in different language editions are compared, which means to compare associations instead of comparing the articles directly. The comparison is conducted with the help of two similarity scores, the *overlap* score and the *RBO* score. Both give a score between 0 and 1, where a higher score represents a higher similarity regarding the ranking of most similar articles. Using these similarity scores single

articles as well as topics are compared. As a last analyzing method the vector space of the embeddings was reduced into a 2-dimensional space. By plotting each language, a visual analysis of the embeddings is possible.
Unfortunately the computed embeddings seem to be unstable, which makes the results of the comparison questionable. The results for the different methods are not completely matching with each other. That raises the question if the reason lies in the quality of the embeddings or is due to the different underlying information which may contain different biases. Furthermore, the heavy variation between the language editions regarding size and containing articles is an issue, as it impedes the comparison.
The presented results imply the assumption that biases can be detected in the Wikipedia corpus.For proving this assumption further research tackling the discovered difficulties is needed.

6.2 Future work

During the process of this thesis, many different problems and questions arose. This section contains some of the unanswered questions that may be part of future work.
First of all the mapping to the Wikidata id needs to be fixed, to produce results, which represent the whole Wikipedia corpus. Further some more data for the clickstream method is necessary. More data can be collected with more time or links can be considered as soon as they were clicked at least one time to get a bigger sample. Apart from that aspect, the clickstream method could also be variegated by using a weighted graph with the number of clicks per link as the weight for the edges.
As an approach to make the embeddings more stable, more walks could be simulated for the training. Apart from that an average over multiple samples could solve this issue as well as Antoniak et al. [1] suggested.
Another idea, which arose throughout this work, is to use a multilayer network instead of separate networks for each language edition. Thereby each layer would represent one language edition and these are connected with the help of the Wikidata id. This approach could solve the issue of the high variation of the different language editions. Furthermore, a direct comparison of the embeddings would be possible as well, because then they are in the same vector space.
Moreover, more methods could be implemented by using other embedding algorithms. For example the article text method could also be realized by using *doc2vec*, an algorithm which computes embeddings for documents or in the case of this work for articles.
Finally extending the sample of examined languages, which enables the comparison of even more countries and cultures.

Bibliography

[1] Antoniak, M., Mimno, D.: Evaluating the stability of embedding-based word similarities. Transactions of the Association for Computational Linguistics **6**, 107–119 (2018)

[2] Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., Gergle, D.: Omnipedia: bridging the wikipedia language gap. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1075–1084. ACM (2012)

[3] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of machine learning research **3**(Feb), 1137–1155 (2003)

[4] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)

[5] Callahan, E.S., Herring, S.C.: Cultural bias in wikipedia content on famous persons. Journal of the American society for information science and technology **62**(10), 1899–1915 (2011)

[6] Galke, L., Saleh, A., Scherp, A.: Word embeddings for practical information retrieval. INFORMATIK 2017 (2017)

[7] Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 855–864. ACM (2016)

[8] Hamilton, W.L., Leskovec, J., Jurafsky, D.: Diachronic word embeddings reveal statistical laws of semantic change. arXiv preprint arXiv:1605.09096 (2016)

[9] Hecht, B., Gergle, D.: Measuring self-focus bias in community-maintained knowledge repositories. In: Proceedings of the fourth international conference on Communities and technologies. pp. 11–20. ACM (2009)

[10] Jiang, Y., Bai, W., Zhang, X., Hu, J.: Wikipedia-based information content and semantic similarity computation. Information Processing & Management **53**(1), 248–265 (2017)

[11] Laufer, P., Wagner, C., Flöck, F., Strohmaier, M.: Mining cross-cultural relations from wikipedia: a study of 31 european food cultures. In: Proceedings of the ACM Web Science Conference. p. 3. ACM (2015)

[12] Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)

[13] Massa, P., Scrinzi, F.: Manypedia: Comparing language points of view of wikipedia communities. In: Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration. p. 21. ACM (2012)

[14] Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242, pp. 133–142 (2003)

[15] Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en

[16] Sherkat, E., Milios, E.E.: Vector embedding of wikipedia concepts and entities. In: International Conference on Applications of Natural Language to Information Systems. pp. 418–428. Springer (2017)

[17] Ulyanov, D.: Multicore-tsne. https://github.com/DmitryUlyanov/Multicore-TSNE (2016)

[18] Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Transactions on Information Systems (TOIS) **28**(4), 20 (2010)

[19] Wikipedia: Wikipedia:statistics, https://en.wikipedia.org/wiki/Wikipedia:Statistics

Appendices

1 Clickstream mapping errors

The Tables 1 to 4 show the statistics of the preparing clickstream dataset process. The first column states the month in which the dump was created. The second how many links were successfully extracted from the dump and mapped to the Wikidata id. The column three and four show the mapping errors from the title to the page id, first for the start node and then for the target node. In column five and six the mapping errors from the page id to the Wikidata id are displayed, again first for the start node and then for the target node. After the links are extracted for each month, the resulting datasets are joined together starting with the earliest one. In this process there are always some links added and some already existing. The number of added links can be seen in column seven and the existing links in column eight.

Month	Saved	(1)	(2)	(3)	(4)	Added	Existing
2017-11	2,340,796	2,348	2,678	10,302	10,850	2,340,796	0
2017-12	2,220,195	1,817	2,104	9,321	9,678	448,225	1,771,970
2018-01	2,593,691	1,771	2,063	9,777	9,916	483,601	2,110,090
2018-02	2,310,090	1,372	1,717	7,863	7,969	257,067	2,053,023
2018-03	2,424,799	1,544	1,697	7,161	6,979	234,646	2,190,153
2018-04	2,332,977	1,244	1,373	6,018	6,018	196,193	2,136,784
2018-05	2,430,753	896	1,249	4,987	4,762	218,370	2,212,383
2018-06	2,334,411	722	995	4,015	3,595	169,587	2,164,824
2018-07	2,530,691	567	762	2,980	2,299	202,576	2,328,115
2018-08	2,614,453	207	495	1,545	743	202,231	2,412,222
2018-09	2,458,097	430	1,016	1,211	619	141,647	2,316,450

Table 1: Statistics of the preparing clickstream dataset process for the **German** edition.
(1)/(2): Mapping Error title to page id start/target; (3)/(4):Mapping Error page id to Wikidata id start/target

Month	Saved	(1)	(2)	(3)	(4)	Added	Existing
2017-11	14,916,768	32,720	328,488	193,653	177,477	14,916,661	0
2017-12	14,627,803	32,040	324,711	171,769	155,806	2,444,824	12,182,979
2018-01	15,954,929	33,161	352,992	161,378	144,557	2,220,836	13,734,093
2018-02	14,815,831	29,406	332,629	132,702	119,513	1,231,967	13,583,864
2018-03	15,782,348	33,327	354,489	120,443	109,688	1,293,206	14,489,142
2018-04	16,286,799	35,475	356,985	101,589	92,522	1,280,680	15,006,119
2018-05	16,497,911	37,417	355,274	84,069	73,209	1,178,942	15,318,969
2018-06	16,001,944	36,703	345,828	58,182	53,065	962,750	15,039,194
2018-07	17,047,441	39,475	367,599	35,891	33,638	1,063,957	15,983,484
2018-08	17,435,617	39,780	377,543	12,285	11,403	1,022,944	16,412,673
2018-09	16,894,023	42,458	365,044	8,561	11,926	783,672	16,110,351

Table 2: Statistics of the preparing clickstream dataset process for the **English** edition.
(1)/(2): Mapping Error title to page id start/target; (3)/(4):Mapping Error page id to Wikidata id start/target

Month	Saved	(1)	(2)	(3)	(4)	Added	Existing
2017-11	104,020	148	1,301	390	357	104,020	0
2017-12	98,588	160	1,342	360	323	19,214	79,374
2018-01	106,510	171	1,443	376	306	16,182	90,328
2018-02	96,448	166	1,296	364	272	8,258	88,190
2018-03	102,343	194	1,384	180	240	8,824	93,519
2018-04	112,611	250	1,492	174	226	10,782	101,829
2018-05	128,984	331	1,695	140	214	14,259	114,725
2018-06	124,856	335	1,628	109	160	9,990	114,866
2018-07	128,574	345	1,726	86	91	9,655	118,919
2018-08	133,937	401	1,825	86	64	9,205	124,732
2018-09	168,042	521	1,878	243	189	38,592	129,450

Table 3: Statistics of the preparing clickstream dataset process for the **Japanese** edition.
(1)/(2): Mapping Error title to page id start/target; (3)/(4):Mapping Error page id to Wikidata id start/target

Month	Saved	(1)	(2)	(3)	(4)	Added	Existing
2017-11	152,991	157	473	1,312	1,434	152,991	0
2017-12	152,867	139	360	1,072	1,234	27,068	125,799
2018-01	165,405	123	311	1,048	1,257	22,512	142,893
2018-02	150,272	104	251	874	969	11,141	139,131
2018-03	157,582	108	216	792	924	11,250	146,332
2018-04	151,127	90	223	587	669	9,252	141,875
2018-05	153,209	99	141	424	448	9,160	144,049
2018-06	146,977	86	130	323	319	7,111	139,866
2018-07	153,506	36	65	155	199	7,587	145,919
2018-08	156,846	27	35	76	74	7,505	149,341
2018-09	151,777	37	81	17	36	6,181	145,596

Table 4: Statistics of the preparing clickstream dataset process for the **Russian** edition.
(1)/(2): Mapping Error title to page id start/target; (3)/(4):Mapping Error page id to Wikidata id start/target

2 Results

2.1 Link structure method

	bn	de	en	hi	ja	ru
ar	0.23	0.02	0.01	0.03	0.03	0.08
bn		0.37	0.18	0.20	0.18	0.22
de			0.17	0.15	0.21	0.19
en				0.18	0.17	0.11
hi					0.33	0.12
ja						0.08

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.64	0.38	0.17	0.37	0.14	0.50
bn		0.67	0.65	0.39	0.32	0.64
de			0.27	0.48	0.30	0.41
en				0.40	0.32	0.20
hi					0.23	0.42
ja						0.24

(b) *Overlap* score

Table 5: Similarity score of the single article ***Hiroshima*** (Q34664) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.03	0.25	0.28	0.01	0.50	0.50
bn		0.05	0.06	0.00	0.04	0.05
de			0.26	0.00	0.36	0.24
en				0.00	0.28	0.26
hi					0.00	0.00
ja						0.63

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.14	0.11	0.07	0.19	0.15	0.09
bn		0.08	0.20	0.03	0.12	0.11
de			0.37	0.27	0.60	0.54
en				0.04	0.44	0.52
hi					0.01	0.03
ja						0.56

(b) *Overlap* score

Table 6: Similarity score of the single article ***Winston Churchill*** (Q8016) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.12	0.04	0.04	0.08	0.06	0.09
bn		0.00	0.00	0.00	0.04	0.00
de			0.05	0.04	0.24	0.16
en				0.09	0.12	0.03
hi					0.07	0.13
ja						0.09

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.28	0.25	0.28	0.19	0.29	0.21
bn		0.04	0.06	0.01	0.28	0.03
de			0.37	0.23	0.38	0.34
en				0.24	0.34	0.41
hi					0.23	0.22
ja						0.39

(b) *Overlap* score

Table 7: Similarity score of the single article ***war*** (Q198) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.29	0.28	0.35	0.03	0.31	0.24
bn		0.32	0.41	0.05	0.43	0.29
de			0.25	0.08	0.23	0.11
en				0.05	0.39	0.21
hi					0.05	0.02
ja						0.31

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.25	0.33	0.50	0.31	0.40	0.43
bn		0.26	0.29	0.21	0.30	0.28
de			0.35	0.30	0.35	0.43
en				0.27	0.52	0.55
hi					0.38	0.32
ja						0.50

(b) *Overlap* score

Table 8: Similarity score of the single article ***dictatorship*** (Q317) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.20	0.07	0.14	0.00	0.21	0.11
bn		0.17	0.28	0.00	0.30	0.13
de			0.13	0.00	0.27	0.25
en				0.00	0.41	0.14
hi					0.00	0.00
ja						0.45

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.58	0.45	0.75	0.01	0.46	0.58
bn		0.56	0.55	0.01	0.62	0.56
de			0.35	0.00	0.46	0.48
en				0.02	0.42	0.52
hi					0.00	0.02
ja						0.45

(b) *Overlap* score

Table 9: Similarity score of the single article ***Homosexuality*** (Q6636) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.48	0.29	0.12	0.00	0.49	0.22
bn		0.11	0.19	0.49	0.27	0.20
de			0.12	0.04	0.14	0.15
en				0.08	0.18	0.32
hi					0.25	0.09
ja						0.22

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.28	0.41	0.50	0.13	0.21	0.28
bn		0.29	0.33	0.29	0.37	0.31
de			0.44	0.14	0.22	0.34
en				0.18	0.33	0.42
hi					0.28	0.28
ja						0.38

(b) *Overlap* score

Table 10: Similarity score of the single article ***Feminism*** (Q7252) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.04	0.35	0.03	0.00	0.24	0.04
bn		0.02	0.01	0.00	0.04	0.01
de			0.05	0.00	0.17	0.02
en				0.00	0.02	0.16
hi					0.00	0.00
ja						0.00

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.01	0.11	0.09	0.02	0.17	0.17
bn		0.04	0.04	0.00	0.04	0.04
de			0.26	0.03	0.28	0.23
en				0.01	0.18	0.19
hi					0.03	0.01
ja						0.21

(b) *Overlap* score

Table 11: Similarity score of the single article ***Refugee*** (Q131572) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.11	0.08	0.00	0.11	0.15	0.17
bn		0.02	0.03	0.15	0.08	0.21
de			0.02	0.11	0.35	0.02
en				0.14	0.01	0.00
hi					0.10	0.21
ja						0.15

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.35	0.27	0.12	0.49	0.31	0.30
bn		0.30	0.17	0.52	0.39	0.42
de			0.16	0.49	0.58	0.38
en				0.31	0.21	0.12
hi					0.49	0.57
ja						0.46

(b) *Overlap* score

Table 12: Similarity score of the single article ***Muslim*** (Q47740) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.09	0.06	0.11	0.00	0.06	0.16
bn		0.12	0.13	0.04	0.11	0.02
de			0.48	0.08	0.09	0.38
en				0.04	0.10	0.52
hi					0.08	0.10
ja						0.14

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.08	0.26	0.27	0.08	0.23	0.33
bn		0.10	0.13	0.30	0.15	0.14
de			0.46	0.27	0.48	0.41
en				0.18	0.51	0.41
hi					0.21	0.15
ja						0.40

(b) *Overlap* score

Table 13: Similarity score of the single article ***Nuclear power*** (Q12739) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.09	0.34	0.18	0.21	0.36	0.34
bn		0.39	0.13	0.11	0.13	0.09
de			0.25	0.22	0.30	0.16
en				0.09	0.25	0.34
hi					0.11	0.09
ja						0.33

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.24	0.39	0.45	0.32	0.46	0.50
bn		0.13	0.15	0.15	0.15	0.12
de			0.35	0.41	0.55	0.41
en				0.39	0.35	0.43
hi					0.44	0.45
ja						0.56

(b) *Overlap* score

Table 14: Similarity score of the single article ***Nuclear weapon*** (Q12802) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.06	0.01	0.00	0.01	0.00	0.00
bn		0.27	0.26	0.25	0.28	0.30
de			0.23	0.16	0.14	0.06
en				0.57	0.11	0.02
hi					0.29	0.17
ja						0.01

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.41	0.15	0.13	0.22	0.23	0.19
bn		0.37	0.45	0.41	0.51	0.36
de			0.28	0.31	0.27	0.25
en				0.46	0.42	0.41
hi					0.35	0.38
ja						0.19

(b) *Overlap* score

Table 15: Similarity score of the single article ***Abortion*** (Q8452) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.04	0.04	0.16	0.06	0.03	0.01
bn		0.09	0.09	0.00	0.09	0.09
de			0.05	0.09	0.22	0.23
en				0.08	0.34	0.15
hi					0.08	0.06
ja						0.08

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.05	0.21	0.21	0.12	0.27	0.29
bn		0.08	0.11	0.04	0.10	0.08
de			0.28	0.11	0.37	0.28
en				0.15	0.49	0.51
hi					0.13	0.13
ja						0.44

(b) *Overlap* score

Table 16: Similarity score of the single article ***Capital punishment*** (Q8454) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.10	0.01	0.04	0.00	0.03	0.00
bn		0.14	0.36	0.00	0.11	0.07
de			0.19	0.20	0.25	0.18
en				0.18	0.26	0.13
hi					0.00	0.27
ja						0.14

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.19	0.26	0.18	0.02	0.38	0.17
bn		0.16	0.14	0.00	0.20	0.12
de			0.42	0.05	0.55	0.27
en				0.05	0.38	0.35
hi					0.02	0.05
ja						0.24

(b) *Overlap* score

Table 17: Similarity score of the single article ***Cannabis (drug)*** (Q2845) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.16	0.02	0.06	0.13	0.03	0.01
bn		0.06	0.12	0.31	0.14	0.11
de			0.19	0.06	0.19	0.24
en				0.04	0.05	0.33
hi					0.06	0.05
ja						0.19

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.50	0.19	0.22	0.51	0.21	0.24
bn		0.43	0.43	0.50	0.30	0.49
de			0.26	0.51	0.31	0.34
en				0.52	0.30	0.30
hi					0.46	0.57
ja						0.45

(b) *Overlap* score

Table 18: Similarity score of the single article ***Donald Trump*** (Q22686) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.28	0.19	0.43	0.32	0.35	0.38
bn		0.32	0.22	0.21	0.21	0.11
de			0.11	0.13	0.21	0.09
en				0.35	0.29	0.28
hi					0.36	0.32
ja						0.36

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.21	0.48	0.46	0.10	0.43	0.42
bn		0.23	0.27	0.29	0.23	0.24
de			0.51	0.14	0.52	0.60
en				0.13	0.47	0.55
hi					0.17	0.13
ja						0.60

(b) *Overlap* score

Table 19: Similarity score of the single article ***Kim Jong-un*** (Q56226) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.08	0.04	0.02	0.22	0.06	0.13
bn		0.01	0.01	0.00	0.01	0.01
de			0.19	0.41	0.43	0.37
en				0.38	0.23	0.28
hi					0.37	0.40
ja						0.40

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.33	0.10	0.11	0.05	0.13	0.10
bn		0.08	0.12	0.04	0.14	0.23
de			0.39	0.41	0.56	0.33
en				0.48	0.42	0.45
hi					0.48	0.15
ja						0.35

(b) *Overlap* score

Table 20: Similarity score of the single article ***Vladimir Putin*** (Q7747) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.03	0.18	0.35	0.16	0.33	0.26
bn		0.09	0.06	0.05	0.08	0.14
de			0.43	0.16	0.24	0.26
en				0.26	0.41	0.46
hi					0.29	0.26
ja						0.57

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.17	0.42	0.44	0.03	0.50	0.53
bn		0.14	0.17	0.06	0.30	0.27
de			0.34	0.03	0.49	0.38
en				0.07	0.50	0.48
hi					0.12	0.04
ja						0.64

(b) *Overlap* score

Table 21: Similarity score of the single article ***Angela Merkel*** (Q567) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.28	0.00	0.00	0.01	0.00	0.00
bn		0.00	0.06	0.00	0.01	0.02
de			0.16	0.30	0.23	0.28
en				0.21	0.08	0.20
hi					0.39	0.22
ja						0.20

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.23	0.03	0.15	0.06	0.02	0.04
bn		0.03	0.13	0.11	0.07	0.06
de			0.20	0.14	0.25	0.27
en				0.15	0.27	0.19
hi					0.10	0.14
ja						0.23

(b) *Overlap* score

Table 22: Similarity score of the single article ***Colonialism*** (Q7167) trained with the link structure

	de	en	hi	ja	ru
ar	0.12	0.24	0.20	0.19	0.16
de		0.18	0.22	0.24	0.15
en			0.23	0.21	0.15
hi				0.15	0.02
ja					0.25

(a) *RBO* score

	de	en	hi	ja	ru
ar	0.17	0.39	0.06	0.22	0.31
de		0.23	0.04	0.33	0.28
en			0.07	0.32	0.33
hi				0.04	0.05
ja					0.31

(b) *Overlap* score

Table 23: Similarity score of the topic ***Freedom of the press*** (Q22688) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.29	0.09	0.13	0.08	0.16	0.09
bn		0.21	0.18	0.02	0.20	0.20
de			0.17	0.22	0.19	0.22
en				0.19	0.20	0.15
hi					0.12	0.30
ja						0.18

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.59	0.21	0.36	0.16	0.47	0.28
bn		0.46	0.51	0.23	0.53	0.46
de			0.27	0.18	0.37	0.34
en				0.20	0.40	0.29
hi					0.16	0.18
ja						0.42

(b) *Overlap* score

Table 24: Similarity score of the topic ***Ethics*** (Q9465) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.14	0.09	0.14	0.10	0.15	0.08
bn		0.09	0.12	0.05	0.11	0.05
de			0.18	0.15	0.20	0.18
en				0.14	0.23	0.14
hi					0.15	0.10
ja						0.18

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.27	0.21	0.32	0.21	0.32	0.18
bn		0.26	0.23	0.25	0.29	0.32
de			0.26	0.27	0.35	0.31
en				0.20	0.36	0.30
hi					0.28	0.27
ja						0.34

(b) *Overlap* score

Table 25: Similarity score of the topic ***Politics*** (Q7163) trained with the link structure

	bn	de	en	ja	ru
ar	0.11	0.06	0.19	0.14	0.15
bn		0.17	0.10	0.05	0.31
de			0.15	0.20	0.13
en				0.27	0.19
ja					0.19

(a) *RBO* score

	bn	de	en	ja	ru
ar	0.18	0.19	0.36	0.32	0.30
bn		0.10	0.17	0.20	0.20
de			0.27	0.35	0.23
en				0.44	0.29
ja					0.31

(b) *Overlap* score

Table 26: Similarity score of the topic ***Politics of Germany*** (Q159493) trained with the link structure

	bn	en	hi
ar	0.27	0.21	0.17
bn		0.15	0.21
en			0.13

(a) *RBO* score

	bn	en	hi
ar	0.26	0.28	0.22
bn		0.30	0.30
en			0.15

(b) *Overlap* score

Table 27: Similarity score of the topic ***Politics of India*** (Q1123156) trained with the link structure

	de	en	ja	ru
ar	0.19	0.25	0.23	0.20
de		0.21	0.20	0.28
en			0.16	0.22
ja				0.19

(a) *RBO* score

	de	en	ja	ru
ar	0.26	0.41	0.34	0.33
de		0.39	0.33	0.30
en			0.22	0.32
ja				0.24

(b) *Overlap* score

Table 28: Similarity score of the topic ***Politics of Japan*** (Q865455) trained with the link structure

	de	en	ru
ar	0.18	0.31	0.20
de		0.15	0.13
en			0.18

(a) *RBO* score

	de	en	ru
ar	0.18	0.49	0.40
de		0.28	0.21
en			0.28

(b) *Overlap* score

Table 29: Similarity score of the topic ***Politics of Russia*** (Q1155561) trained with the link structure

	en	hi	ja	ru
de	0.17	0.18	0.25	0.15
en		0.18	0.28	0.20
hi			0.12	0.13
ja				0.22

(a) *RBO* score

	en	hi	ja	ru
de	0.31	0.23	0.40	0.34
en		0.22	0.43	0.36
hi			0.26	0.22
ja				0.41

(b) *Overlap* score

Table 30: Similarity score of the topic ***Politics of the United Kingdom*** (Q678363) trained with the link structure

	de	en	ja	ru
ar	0.16	0.17	0.15	0.16
de		0.19	0.21	0.18
en			0.22	0.20
ja				0.15

(a) *RBO* score

	de	en	ja	ru
ar	0.26	0.28	0.30	0.34
de		0.29	0.37	0.36
en			0.35	0.32
ja				0.32

(b) *Overlap* score

Table 31: Similarity score of the topic ***Politics of the United States*** (Q330963) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.17	0.05	0.10	0.06	0.15	0.18
bn		0.13	0.13	0.13	0.14	0.13
de			0.15	0.20	0.18	0.14
en				0.14	0.17	0.14
hi					0.13	0.01
ja						0.11

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.27	0.26	0.21	0.18	0.32	0.42
bn		0.16	0.25	0.18	0.31	0.29
de			0.21	0.24	0.21	0.16
en				0.27	0.32	0.22
hi					0.19	0.09
ja						0.23

(b) *Overlap* score

Table 32: Similarity score of the topic ***Sport*** (Q349) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.12	0.12	0.19	0.14	0.12	0.14
bn		0.25	0.18	0.19	0.27	0.22
de			0.18	0.23	0.19	0.26
en				0.21	0.19	0.16
hi					0.23	0.26
ja						0.21

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.21	0.25	0.32	0.25	0.32	0.29
bn		0.18	0.20	0.24	0.25	0.21
de			0.28	0.25	0.34	0.30
en				0.27	0.33	0.27
hi					0.33	0.26
ja						0.35

(b) *Overlap* score

Table 33: Similarity score of the topic ***Geography*** (Q1071) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.11	0.08	0.20	0.06	0.22	0.11
bn		0.13	0.15	0.17	0.15	0.15
de			0.13	0.13	0.20	0.21
en				0.10	0.29	0.16
hi					0.09	0.12
ja						0.19

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.19	0.13	0.42	0.11	0.49	0.20
bn		0.14	0.20	0.09	0.22	0.20
de			0.19	0.10	0.20	0.26
en				0.16	0.54	0.24
hi					0.10	0.18
ja						0.27

(b) *Overlap* score

Table 34: Similarity score of the topic ***History*** (Q309) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.13	0.11	0.12	0.10	0.15	0.11
bn		0.07	0.09	0.11	0.17	0.18
de			0.23	0.12	0.26	0.22
en				0.12	0.25	0.20
hi					0.14	0.11
ja						0.24

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.23	0.21	0.29	0.23	0.30	0.27
bn		0.23	0.20	0.29	0.23	0.22
de			0.30	0.19	0.35	0.35
en				0.18	0.38	0.34
hi					0.24	0.20
ja						0.38

(b) *Overlap* score

Table 35: Similarity score of the topic ***World War I*** (Q361) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.19	0.12	0.14	0.14	0.16	0.12
bn		0.17	0.21	0.18	0.17	0.20
de			0.23	0.16	0.23	0.20
en				0.14	0.21	0.17
hi					0.15	0.15
ja						0.22

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.27	0.24	0.26	0.22	0.29	0.27
bn		0.22	0.25	0.22	0.26	0.25
de			0.29	0.25	0.35	0.33
en				0.24	0.31	0.29
hi					0.22	0.22
ja						0.34

(b) *Overlap* score

Table 36: Similarity score of the topic ***World War II*** (Q362) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.15	0.09	0.18	0.10	0.15	0.17
bn		0.15	0.14	0.15	0.16	0.18
de			0.19	0.14	0.25	0.17
en				0.20	0.21	0.19
hi					0.16	0.14
ja						0.23

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.25	0.18	0.36	0.20	0.26	0.38
bn		0.23	0.27	0.27	0.26	0.22
de			0.29	0.22	0.36	0.27
en				0.19	0.32	0.32
hi					0.23	0.21
ja						0.30

(b) *Overlap* score

Table 37: Similarity score of the topic ***Cold War*** (Q8683) trained with the link structure

	en	ja	ru
ar	0.18	0.19	0.22
en		0.18	0.12
ja			0.24

(a) *RBO* score

	en	ja	ru
ar	0.67	0.47	0.57
en		0.38	0.47
ja			0.42

(b) *Overlap* score

Table 38: Similarity score of the topic ***LGBT culture*** (Q51389) trained with the link structure

	bn	de	en	ja	ru
ar	0.16	0.17	0.18	0.17	0.22
bn		0.17	0.13	0.16	0.13
de			0.18	0.26	0.18
en				0.21	0.13
ja					0.24

(a) *RBO* score

	bn	de	en	ja	ru
ar	0.51	0.42	0.64	0.43	0.56
bn		0.50	0.40	0.58	0.47
de			0.39	0.35	0.43
en				0.38	0.44
ja					0.36

(b) *Overlap* score

Table 39: Similarity score of the topic ***LGBT*** (Q17884) trained with the link structure

	bn	de	en	hi	ja	ru
ar	0.32	0.13	0.17	0.02	0.23	0.11
bn		0.14	0.18	0.32	0.21	0.11
de			0.15	0.05	0.19	0.19
en				0.13	0.22	0.15
hi					0.12	0.10
ja						0.16

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.34	0.23	0.41	0.11	0.24	0.24
bn		0.32	0.31	0.34	0.37	0.30
de			0.24	0.16	0.29	0.29
en				0.27	0.32	0.27
hi					0.34	0.24
ja						0.27

(b) *Overlap* score

Table 40: Similarity score of the topic ***Feminism*** (Q7252) trained with the link structure

2.2 Clickstream method

	en	ja	ru
de	0.38	0.01	0.01
en		0.00	0.00
ja			0.00

(a) *RBO* score

	en	ja	ru
de	0.24	0.01	0.07
en		0.00	0.00
ja			0.02

(b) *Overlap* score

Table 41: Similarity score of the single article ***Hiroshima*** (Q34664) trained with the clickstream

	en	ru
de	0.04	0.00
en		0.00

(a) *RBO* score

	en	ru
de	0.39	0.00
en		0.01

(b) *Overlap* score

Table 42: Similarity score of the single article ***Winston Churchill*** (Q8016) trained with the clickstream

	en	ru
de	0.01	0.00
en		0.00

(a) *RBO* score

	en	ru
de	0.10	0.00
en		0.00

(b) *Overlap* score

Table 43: Similarity score of the single article ***war*** (Q198) trained with the clickstream

	en	ja	ru
de	0.32	0.00	0.00
en		0.00	0.00
ja			0.00

(a) *RBO* score

	en	ja	ru
de	0.21	0.00	0.00
en		0.00	0.00
ja			0.03

(b) *Overlap* score

Table 44: Similarity score of the single article ***dictatorship*** (Q317) trained with the clickstream

	en	ru
de	0.16	0.00
en		0.00

(a) *RBO* score

	en	ru
de	0.13	0.00
en		0.00

(b) *Overlap* score

Table 45: Similarity score of the single article ***Homosexuality*** (Q6636) trained with the clickstream

	en	ja	ru
de	0.16	0.00	0.00
en		0.00	0.00
ja			0.00

(a) *RBO* score

	en	ja	ru
de	0.32	0.00	0.00
en		0.00	0.00
ja			0.01

(b) *Overlap* score

Table 46: Similarity score of the single article ***Feminism*** (Q7252) trained with the clickstream

	en
de	0.02

(a) *RBO* score

	en
de	0.03

(b) *Overlap* score

Table 47: Similarity score of the single article ***Refugee*** (Q131572) trained with the clickstream

	en
de	0.00

(a) *RBO* score

	en
de	0.14

(b) *Overlap* score

Table 48: Similarity score of the single article ***Muslim*** (Q47740) trained with the clickstream

	en
de	0.03

(a) *RBO* score

	en
de	0.24

(b) *Overlap* score

Table 49: Similarity score of the single article ***Nuclear power*** (Q12739) trained with the clickstream

	en	ja	ru
de	0.20	0.00	0.00
en		0.00	0.00
ja			0.00

(a) *RBO* score

	en	ja	ru
de	0.32	0.01	0.03
en		0.00	0.02
ja			0.00

(b) *Overlap* score

Table 50: Similarity score of the single article ***Nuclear weapon*** (Q12802) trained with the clickstream

	en	ru
de	0.11	0.00
en		0.00

(a) *RBO* score

	en	ru
de	0.35	0.00
en		0.00

(b) *Overlap* score

Table 51: Similarity score of the single article ***Abortion*** (Q8452) trained with the clickstream

	en
de	0.02

(a) *RBO* score

	en
de	0.16

(b) *Overlap* score

Table 52: Similarity score of the single article ***Capital punishment*** (Q8454) trained with the clickstream

	en	ja	ru
de	0.20	0.00	0.00
en		0.00	0.00
ja			0.00

(a) *RBO* score

	en	ja	ru
de	0.30	0.00	0.00
en		0.00	0.00
ja			0.00

(b) *Overlap* score

Table 53: Similarity score of the single article ***Cannabis (drug)*** (Q2845) trained with the clickstream

	en	ja	ru
de	0.02	0.00	0.00
en		0.00	0.00
ja			0.00

(a) *RBO* score

	en	ja	ru
de	0.18	0.00	0.00
en		0.00	0.02
ja			0.00

(b) *Overlap* score

Table 54: Similarity score of the single article ***Donald Trump*** (Q22686) trained with the clickstream

	en	ru
de	0.17	0.00
en		0.00

(a) *RBO* score

	en	ru
de	0.49	0.00
en		0.00

(b) *Overlap* score

Table 55: Similarity score of the single article ***Kim Jong-un*** (Q56226) trained with the clickstream

	en	ru
de	0.15	0.00
en		0.00

(a) *RBO* score

	en	ru
de	0.25	0.01
en		0.01

(b) *Overlap* score

Table 56: Similarity score of the single article ***Vladimir Putin*** (Q7747) trained with the clickstream

	en	ja	ru
de	0.07	0.00	0.00
en		0.00	0.00
ja			0.00

(a) *RBO* score

	en	ja	ru
de	0.26	0.01	0.02
en		0.01	0.04
ja			0.05

(b) *Overlap* score

Table 57: Similarity score of the single article ***Angela Merkel*** (Q567) trained with the clickstream

	en	ru
de	0.30	0.00
en		0.00

(a) *RBO* score

	en	ru
de	0.09	0.00
en		0.00

(b) *Overlap* score

Table 58: Similarity score of the single article ***Colonialism*** (Q7167) trained with the clickstream

	en
de	0.16

(a) *RBO* score

	en
de	0.07

(b) *Overlap* score

Table 59: Similarity score of the topic ***Freedom of the press*** (Q22688) trained with the clickstream

	en
de	0.15

(a) *RBO* score

	en
de	0.18

(b) *Overlap* score

Table 60: Similarity score of the topic ***Ethics*** (Q9465) trained with the clickstream

	en
de	0.17

(a) *RBO* score

	en
de	0.21

(b) *Overlap* score

Table 61: Similarity score of the topic ***Politics*** (Q7163) trained with the clickstream

	en
de	0.11

(a) *RBO* score

	en
de	0.15

(b) *Overlap* score

Table 62: Similarity score of the topic ***Politics of Germany*** (Q159493) trained with the clickstream

	en
de	0.18

(a) *RBO* score

	en
de	0.23

(b) *Overlap* score

Table 63: Similarity score of the topic ***Politics of Japan*** (Q865455) trained with the clickstream

	en
de	0.14

(a) *RBO* score

	en
de	0.15

(b) *Overlap* score

Table 64: Similarity score of the topic ***Politics of Russia*** (Q1155561) trained with the clickstream

	en
de	0.18

(a) *RBO* score

	en
de	0.18

(b) *Overlap* score

Table 65: Similarity score of the topic ***Politics of the United States*** (Q330963) trained with the clickstream

	en	ja
de	0.10	0.00
en		0.00

(a) *RBO* score

	en	ja
de	0.13	0.00
en		0.00

(b) *Overlap* score

Table 66: Similarity score of the topic ***Sport*** (Q349) trained with the clickstream

	en	ja
de	0.18	0.00
en		0.04

(a) *RBO* score

	en	ja
de	0.16	0.00
en		0.02

(b) *Overlap* score

Table 67: Similarity score of the topic ***Geography*** (Q1071) trained with the clickstream

	en
de	0.13

(a) *RBO* score

	en
de	0.11

(b) *Overlap* score

Table 68: Similarity score of the topic ***History*** (Q309) trained with the clickstream

	en	ru
de	0.15	0.00
en		0.01

(a) *RBO* score

	en	ru
de	0.21	0.01
en		0.01

(b) *Overlap* score

Table 69: Similarity score of the topic ***World War I*** (Q361) trained with the clickstream

	en	ja	ru
de	0.12	0.00	0.00
en		0.00	0.00
ja			0.00

(a) *RBO* score

	en	ja	ru
de	0.19	0.00	0.03
en		0.00	0.04
ja			0.01

(b) *Overlap* score

Table 70: Similarity score of the topic ***World War II*** (Q362) trained with the clickstream

	en
de	0.17

(a) *RBO* score

	en
de	0.24

(b) *Overlap* score

Table 71: Similarity score of the topic ***Cold War*** (Q8683) trained with the clickstream

	en	ja
de	0.16	0.00
en		0.00

(a) *RBO* score

	en	ja
de	0.23	0.00
en		0.01

(b) *Overlap* score

Table 72: Similarity score of the topic ***LGBT*** (Q17884) trained with the clickstream

	en
de	0.16

(a) *RBO* score

	en
de	0.15

(b) *Overlap* score

Table 73: Similarity score of the topic ***Feminism*** (Q7252) trained with the clickstream

2.3 Article text method

	bn	de	en	hi	ja	ru
ar	0.01	0.10	0.13	0.04	0.09	0.16
bn		0.00	0.00	0.01	0.00	0.01
de			0.39	0.00	0.11	0.29
en				0.11	0.13	0.38
hi					0.05	0.11
ja						0.04

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.09	0.07	0.11	0.13	0.03	0.08
bn		0.12	0.10	0.10	0.09	0.18
de			0.50	0.09	0.27	0.24
en				0.07	0.43	0.23
hi					0.06	0.12
ja						0.12

(b) *Overlap* score

Table 74: Similarity score of the single article ***Hiroshima*** (Q34664) trained with the article text

	bn	de	en	hi	ja	ru
ar	0.00	0.02	0.10	0.02	0.00	0.01
bn		0.04	0.00	0.00	0.00	0.00
de			0.25	0.00	0.00	0.00
en				0.00	0.00	0.01
hi					0.00	0.00
ja						0.04

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.07	0.16	0.23	0.10	0.06	0.17
bn		0.06	0.08	0.00	0.02	0.06
de			0.23	0.04	0.06	0.16
en				0.01	0.04	0.24
hi					0.01	0.03
ja						0.07

(b) *Overlap* score

Table 75: Similarity score of the single article ***war*** (Q198) trained with the article text

	bn	de	en	hi	ja	ru
ar	0.10	0.41	0.31	0.17	0.03	0.26
bn		0.04	0.00	0.02	0.01	0.01
de			0.34	0.02	0.02	0.46
en				0.03	0.05	0.44
hi					0.01	0.04
ja						0.04

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.10	0.55	0.61	0.16	0.08	0.57
bn		0.06	0.07	0.07	0.03	0.07
de			0.46	0.12	0.08	0.36
en				0.09	0.09	0.45
hi					0.05	0.09
ja						0.06

(b) *Overlap* score

Table 76: Similarity score of the single article ***dictatorship*** (Q317) trained with the article text

	de	en	hi	ja	ru
bn	0.10	0.09	0.00	0.02	0.07
de		0.45	0.00	0.04	0.43
en			0.00	0.00	0.20
hi				0.00	0.00
ja					0.02

(a) *RBO* score

	de	en	hi	ja	ru
bn	0.10	0.09	0.04	0.06	0.12
de		0.53	0.08	0.10	0.45
en			0.02	0.14	0.52
hi				0.08	0.08
ja					0.16

(b) *Overlap* score

Table 77: Similarity score of the single article ***Homosexuality*** (Q6636) trained with the article text

	bn	de	en	hi	ja	ru
ar	0.01	0.14	0.16	0.00	0.00	0.22
bn		0.12	0.05	0.00	0.07	0.26
de			0.43	0.01	0.01	0.30
en				0.00	0.01	0.30
hi					0.00	0.00
ja						0.03

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.11	0.44	0.47	0.13	0.08	0.43
bn		0.10	0.12	0.08	0.04	0.10
de			0.51	0.07	0.04	0.42
en				0.07	0.07	0.43
hi					0.04	0.04
ja						0.07

(b) *Overlap* score

Table 78: Similarity score of the single article ***Feminism*** (Q7252) trained with the article text

	bn	de	en	hi	ja	ru
ar	0.00	0.13	0.38	0.00	0.01	0.10
bn		0.03	0.01	0.08	0.03	0.04
de			0.18	0.11	0.01	0.14
en				0.12	0.00	0.15
hi					0.00	0.04
ja						0.00

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.04	0.28	0.24	0.00	0.08	0.26
bn		0.10	0.05	0.04	0.09	0.06
de			0.27	0.09	0.11	0.29
en				0.06	0.04	0.24
hi					0.02	0.06
ja						0.04

(b) *Overlap* score

Table 79: Similarity score of the single article ***Refugee*** (Q131572) trained with the article text

	bn	de	en	hi	ja	ru
ar	0.01	0.18	0.31	0.16	0.00	0.17
bn		0.01	0.00	0.01	0.00	0.00
de			0.39	0.21	0.00	0.15
en				0.17	0.00	0.31
hi					0.00	0.21
ja						0.00

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.04	0.23	0.20	0.16	0.04	0.18
bn		0.09	0.05	0.05	0.02	0.04
de			0.44	0.27	0.01	0.49
en				0.30	0.01	0.44
hi					0.05	0.29
ja						0.03

(b) *Overlap* score

Table 80: Similarity score of the single article ***Muslim*** (Q47740) trained with the article text

	ja
de	0.10

(a) *RBO* score

	ja
de	0.20

(b) *Overlap* score

Table 81: Similarity score of the single article ***Nuclear power*** (Q12739) trained with the article text

	ja
de	0.16

(a) *RBO* score

	ja
de	0.07

(b) *Overlap* score

Table 82: Similarity score of the single article ***Nuclear weapon*** (Q12802) trained with the article text

	bn	de	en	hi	ja	ru
ar	0.01	0.09	0.39	0.06	0.00	0.28
bn		0.19	0.13	0.00	0.00	0.01
de			0.07	0.01	0.00	0.28
en				0.09	0.00	0.28
hi					0.00	0.19
ja						0.00

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.17	0.19	0.34	0.12	0.07	0.28
bn		0.11	0.14	0.09	0.07	0.13
de			0.17	0.07	0.00	0.25
en				0.09	0.01	0.39
hi					0.03	0.09
ja						0.03

(b) *Overlap* score

Table 83: Similarity score of the single article ***Abortion*** (Q8452) trained with the article text

	de	hi	ja
bn	0.00	0.00	0.00
de		0.03	0.11
hi			0.00

(a) *RBO* score

	de	hi	ja
bn	0.07	0.09	0.04
de		0.03	0.07
hi			0.05

(b) *Overlap* score

Table 84: Similarity score of the single article ***Capital punishment*** (Q8454) trained with the article text

	hi	ja	ru
ar	0.01	0.01	0.28
hi		0.03	0.00
ja			0.01

(a) *RBO* score

	hi	ja	ru
ar	0.11	0.02	0.24
hi		0.09	0.05
ja			0.06

(b) *Overlap* score

Table 85: Similarity score of the single article ***Cannabis (drug)*** (Q2845) trained with the article text

	bn	de	en	hi	ja	ru
ar	0.01	0.00	0.00	0.02	0.00	0.02
bn		0.00	0.00	0.03	0.00	0.00
de			0.46	0.11	0.00	0.34
en				0.11	0.01	0.32
hi					0.02	0.26
ja						0.05

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.11	0.02	0.02	0.13	0.04	0.09
bn		0.05	0.04	0.06	0.01	0.05
de			0.35	0.05	0.03	0.17
en				0.08	0.05	0.26
hi					0.03	0.10
ja						0.05

(b) *Overlap* score

Table 86: Similarity score of the single article ***Colonialism*** (Q7167) trained with the article text

	bn	de	en	hi	ja	ru
ar	0.02	0.17	0.23	0.04	0.05	0.15
bn		0.05	0.03	0.00	0.00	0.02
de			0.25	0.01	0.07	0.19
en				0.04	0.04	0.20
hi					0.03	0.00
ja						0.06

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.09	0.31	0.36	0.08	0.10	0.32
bn		0.09	0.10	0.05	0.05	0.11
de			0.41	0.05	0.10	0.34
en				0.10	0.08	0.37
hi					0.05	0.06
ja						0.10

(b) *Overlap* score

Table 87: Similarity score of the topic ***Ethics*** (Q9465) trained with the article text

	bn	de	en	hi	ja	ru
ar	0.08	0.28	0.28	0.06	0.04	0.24
bn		0.05	0.03	0.08	0.03	0.15
de			0.25	0.08	0.05	0.24
en				0.04	0.05	0.23
hi					0.05	0.09
ja						0.10

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.07	0.38	0.37	0.12	0.10	0.41
bn		0.10	0.07	0.07	0.05	0.10
de			0.38	0.10	0.08	0.41
en				0.07	0.07	0.45
hi					0.05	0.11
ja						0.10

(b) *Overlap* score

Table 88: Similarity score of the topic ***Politics*** (Q7163) trained with the article text

	bn	de	en	hi	ja	ru
ar	0.00	0.08	0.12	0.00	0.06	0.14
bn		0.02	0.06	0.00	0.03	0.08
de			0.17	0.00	0.06	0.19
en				0.00	0.06	0.20
hi					0.00	0.00
ja						0.10

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.07	0.21	0.23	0.08	0.09	0.23
bn		0.07	0.07	0.06	0.05	0.07
de			0.22	0.04	0.10	0.34
en				0.05	0.08	0.40
hi					0.04	0.04
ja						0.12

(b) *Overlap* score

Table 89: Similarity score of the topic ***Sport*** (Q349) trained with the article text

	bn	de	en	hi	ja	ru
ar	0.13	0.31	0.36	0.06	0.07	0.26
bn		0.06	0.07	0.03	0.02	0.07
de			0.21	0.06	0.03	0.16
en				0.04	0.04	0.25
hi					0.03	0.02
ja						0.03

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.10	0.28	0.36	0.10	0.08	0.31
bn		0.08	0.08	0.07	0.06	0.08
de			0.27	0.09	0.05	0.24
en				0.07	0.07	0.35
hi					0.05	0.09
ja						0.06

(b) *Overlap* score

Table 90: Similarity score of the topic ***Geography*** (Q1071) trained with the article text

	bn	de	en	hi	ja	ru
ar	0.03	0.10	0.26	0.10	0.07	0.34
bn		0.01	0.02	0.01	0.02	0.01
de			0.16	0.00	0.03	0.15
en				0.04	0.05	0.18
hi					0.02	0.05
ja						0.03

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.10	0.23	0.34	0.14	0.09	0.36
bn		0.06	0.07	0.11	0.05	0.09
de			0.21	0.04	0.06	0.22
en				0.08	0.06	0.32
hi					0.11	0.11
ja						0.08

(b) *Overlap* score

Table 91: Similarity score of the topic ***History*** (Q309) trained with the article text

	de	en	ja	ru
bn	0.07	0.06	0.01	0.06
de		0.39	0.02	0.29
en			0.01	0.31
ja				0.07

(a) *RBO* score

	de	en	ja	ru
bn	0.11	0.09	0.06	0.11
de		0.52	0.05	0.39
en			0.05	0.46
ja				0.05

(b) *Overlap* score

Table 92: Similarity score of the topic ***LGBT*** (Q17884) trained with the article text

	bn	de	en	hi	ja	ru
ar	0.02	0.18	0.16	0.00	0.02	0.18
bn		0.06	0.04	0.03	0.01	0.04
de			0.29	0.04	0.01	0.16
en				0.05	0.04	0.26
hi					0.05	0.06
ja						0.07

(a) *RBO* score

	bn	de	en	hi	ja	ru
ar	0.09	0.31	0.31	0.13	0.11	0.35
bn		0.09	0.08	0.07	0.06	0.10
de			0.39	0.10	0.06	0.28
en				0.09	0.07	0.35
hi					0.06	0.08
ja						0.09

(b) *Overlap* score

Table 93: Similarity score of the topic ***Feminism*** (Q7252) trained with the article text