An algorithm for controlled text analysis on Wikipedia

No Institute Given

Abstract. While numerous work has examined bias on Wikipedia, most approaches fail to control for possible confounding variables. In this work, given a target corpus for analysis (e.g. biography pages about women), we present a method for constructing a control corpus that matches the target corpus in as many attributes as possible, except the target attribute (e.g. the gender of the subject). This methodology can be used to analyze specific types of bias in Wikipedia articles, for example, gender or racial bias, while minimizing the influence of confounding variables.

1 Introduction

Almost since its inception, Wikipedia has attracted the interest of researchers in various disciplines, including information science, social science, and computer science, because of its unique community and departure from traditional encyclopedias [7, 8]. As a collaborative knowledge platform where anyone can edit pages or add new ones, Wikipedia effectively crowd-sources information. This setup allows for fast and inexpensive dissemination of information, but it risks introducing biases [8]. These biases are problematic - not just because they can influence readers, but also because Wikipedia has become a popular data source for computational models in information retrieval and natural language processing [4, 10, 11], which are prone to absorbing and even amplifying data bias [3, 16, 17].

Prior work has examined possible biases on Wikipedia using a variety of metrics. Most research focuses on gender bias: how do articles about men and women differ? [1, 5, 15]. However, many of these studies draw limited conclusions because of the difficultly in controlling for confounding variables. For example, when computing pointwise mutual information scores or log-likelihood scores to find words that are over-represented in biography articles about men as opposed to women, the most common words consist of sports terms: "football", "footballer", "baseball", "league" [5, 15]. This result is not necessarily indicative of *bias*; we do not suggest that Wikipedia editors omit the football achievements of women. Instead, this imbalance results because, in society and on Wikipedia, there are more male football players than female players.[1] Thus, the difference in occupation, rather than the difference in gender, likely explains this imbalance. While some traits can be accounted for, for example, by including occupation

[1] Whether or not the lack of female football players in society is a sign of bias is beyond the scope of this paper.

as an explanatory variable in a regression model [15], it is difficult to explicitly enumerate all possible confounding variables and limits analysis to particular models, e.g. regression.

In this work, we present a method for identifying a "control" biography page for every page that aligns with a target attribute, where the control page matches as nearly as possible to the target page on all attributes excepted the targeted one. The concept of a control group originates in randomized trials, in which participants in a study are randomly assigned to the "treatment group" or "control group" and the effectiveness of the treatment is measured as the difference in results for the treatment and control groups [12]. In observational studies, when the treatment and outcomes have already occurred, researchers can replicate the conditions of a randomized trial by constructing a treatment group and a control group so that the distribution of covariates is as identical as possible between the two groups for all traits except the target attribute [12]. Then by comparing the constructed treatment and control groups, researchers can isolate the effects of the target attribute from other confounding variables. For example, if our target attribute is gender, the treatment group may consist of Wikipedia biography pages about women and the control group may consist of biography pages about men. The two groups would be constructed so that they have similar distributions of covariates that could be confounding variables, such as age, occupation, age, and nationality.

We present several methods for constructing a control group for a given treatment group and evaluate how well they control for data confounds. Our recommended method uses TF-IDF vectors with a pivot-slope correction [13, 14], where the vectors are constructed from Wikipedia metadata categories in order to identify the best-matching control for each treatment value.

2 Methodology

Given a set of treatment articles $\mathcal{T}$, our goal is to construct a set of control articles $\mathcal{C}$ from a set of candidate controls $\mathcal{A}$, such that $\mathcal{C}$ has a similar covariate distribution as $\mathcal{T}$ for all covariates except the target attribute. For example, $\mathcal{T}$ may be the set of all biography pages about women, and $\mathcal{A}$ maybe be the set of all biography articles about men.

We construct $\mathcal{C}$ using a 1:1 matching algorithm. For each $t \in T$, we identify $c_{best} \in A$ that best matches t and add c_{best} to $\mathcal{C}$. For example, if our target trait is gender, t may be an American female actress born in the 1970s and c_{best} may be an American male actor born in the 1970s.

In order to identify c_{best} for a given t, we leverage the category metadata associated with each article. Wikipedia articles contain category tags that enumerate relevant traits. For example, the page for Steve Jobs includes the categories "Pixar people", "Directors of Apple Inc.", "American people of German descent", etc. While relying on the category tags could introduce some bias, as articles are not always categorized correctly or with equal detail, using this metadata allows us to exclude the target attribute from the matching process.

If our target attribute is gender, we can exclude categories that contain gendered keywords like "Women corporate directors" while including categories like "American corporate directors". We cannot use the article text for matching, as we cannot disambiguate what aspects of the text result from the target attribute and what results from confounding variables.

We use the following four different similarity metrics in order to compare the categories for each candidate $c \in A$ with t. Throughout this section, we use $CAT(c)$ to denote the set of categories associated with c. In all cases, we performed matching with replacement, thus allowing each chosen control to match to multiple treatment values.

Number of Categories In this metric, we choose c_{best} as the person who has the most number of categories in common with t. The intuition behind this metric is that the person with the largest number of overlapping categories has the most in common with t and thus is the best possible match. More formally, we choose

$$c_{best} = \arg\max_{c_i} |CAT(c_i) \cap CAT(t)|$$

Percent Categories One of the drawbacks of matching simply on the raw number of categories is that it favors people who have more categories. For example, a candidate c_i who has 30 categories is more likely to have more categories in common with t than a candidate c_j who only has 10 categories. However, c_i having more categories in common with t does not necessarily mean that c_i is a better match than c_j - it suggests that the article for c_i is better written, rather than suggesting that c_i has more traits in common with t than c_j. Instead of matching on the raw number of categories, we can control this problem by normalizing the number of overlapping categories by the total number of categories in the candidate c_i. Thus, we choose:

$$c_{best} = \arg\max_{c_i} |CAT(c_i) \cap CAT(t)| * \frac{1}{|CAT(c_i)|}$$

TF-IDF Weighting Both of the prior methods assume that all categories are equally meaningful, but this is an oversimplification. For example, a candidate c_i that has the category "American short story writers" in common with t is more likely to be a good match than a candidate that has the category "Living People" in common with t. We adopt a TF-IDF weighting schema from information retrieval to up-weight categories that are less common [13].

We represent each candidate $c_i \in A$ as a sparse category vector. Each element in the vector is a product between the frequency of the category in c_i, e.g. $\frac{1}{|CAT(c_i)|}$ if the category is in c_i and 0 otherwise, and the inverse frequency of the category, e.g. $\frac{1}{|category|}$. Thus, very common categories like "Living People" are down-weighted as compared to more specific categories. We similarly construct a vector representation of t.

We then c_{best} as c_i with the highest cosine similarity between its vector and the vector for t.

Pivot-Slope TF-IDF Weighting TF-IDF weighting and Percent Categories both have a potential problem in that they include the normalization term $\frac{1}{|CAT(c_i)|}$. While this term is intended to normalize for articles having different numbers of categories, in actuality, it over-corrects and causes the algorithm to favor articles with fewer categories. This issue has been observed in information retrieval: using TF-IDF weighting to retrieve relevant documents causes shorter documents to have a higher probability of being retrieved [14].

In order to correct for this, we adopt the pivot-slope normalization mechanism from [14]. In this method, instead of normalizing the TF-IDF term by $|CAT(c_i)|$, the term is normalized with an adjusted value:

$$(1.0 - slope) * pivot + slope * |CAT(c_i)|$$

Following the recommendations in [14], the pivot is set to the average number of categories across all articles in the data set, and the slope is tuned over a development set (described in §5).

3 Data

We gathered a corpus of Wikipedia biography pages by collecting all articles with the category "Living people". We then discarded any articles that had fewer than 2 categories or fewer than 100 tokens in the article. We also discarded any articles marked as stubs, indicated by the presence of a stub category like "Actor stubs".

In matching the articles, we ignore any categories that are focused on traits of the Wikipedia article rather than traits of the person using a manually defined list. The list includes categories containing the words "Use Indian English", "Pages with", "Contains Links", etc.

After filtering, our final data set contain 444,045 biography pages. On average, pages contain 9.3 categories and 628.2 tokens.

4 Experiments

Our ultimate goal in identifying matches is to control for possible confounds in the text. Thus, for a given treatment set of articles, the optimal matching algorithm would produce a matched control set that has identical traits as the treatment set for all attributes, except the one being measured.

In order to assess the performance of each matching metric, we construct a simulated treatment set. We then run each matching algorithm to identify a matched control set, and we use several metrics to examine how closely the control set matches the treatment set. In these simulated treatment sets, we do not fix a target attribute that we mandate differ between the treatment and control sets, and thus we expect a high quality matching algorithm to identify a control set that matches very closely to the treatment set. We use two methods to construct simulated treatment sets:

Random We randomly sample 1000 articles from the corpus.

Category We randomly sample one category that has at least 500 members from the corpus. We then sample 500 articles from the category. In this method, we do not expect there to be any bias towards a single category, since categories are typically very specific. For example, we may sample the category, "Players of American football from Pennsylvania". While we might guess that articles for football players have different characteristics than other articles, we would not expect articles for football players from Pennsylvania to be substantially different than articles for football players from New York or New Jersey. However, this setup does more closely replicate intended analysis setting than random sampling, as we ensure that all people in the treatment group have a common trait.

We then use several metrics to assess how well-matched the treatment and control groups are:

Average bias Standardized bias is the typical method used to evaluate covariate balance in matching methods for causal studies [6]. For a given covariate, the metric is calculated by taking the difference in means between the treatment and control groups and dividing by the standard deviation in the treatment group. In our case, we treat each category as a binary covariate that can be present or absent for each article. We then compute the standardized bias for each category and average across all categories. Since there are some categories that appear only in the treatment group and some that appear only in the control group, we compute this metric in two directions: first for all the categories that appear in the treatment group and second for all the categories that appear in the control group (reported as "Avg. Bias" and "Avg. Bias 2", respectively). A high average bias suggests that the distribution of categories is very different between the treatment and the control.

Number of Categories As discussed in Section 2, one of the concerns with the provided methods is that they may favor articles with more or fewer categories. Thus, we compare the number of categories in the treatment group with the number of categories in the control group by computing Cohen's d. Cohen's d measures effect size as the difference in mean between two groups divided by the pooled standard deviation. A high value indicates that the two groups have different values for this trait.

Text Length The prior two metrics focus on category-level metrics to assess the quality of the match. However, we use categories as a proxy to control for confounds in the text, and thus we ultimately seek to assess how well our matching methods control for differences in the actual article text. We first compare the lengths of the text by computing Cohen's d between the word count of articles in the treatment and control groups.

Polar Log-odds We then compare how different the vocabularies are between the two groups by computing log-odds with a Dirichlet prior [9], which measures to what extent words are overrepresented in one corpus as compared to another. A high log-odds score indicates that a word is much more likely to appear in one corpus than the other. We compute log-odds between the treatment group and control group. We then take the absolute value of all log-odds scores, and compute the mean and standard deviation for the 200 most polar words. If the two groups use similar vocabulary, the polarity of the log-odds scores will be lower.

KL Divergence Finally, rather than examining just word-level differences, we also use topic modelling to examine topic-level differences. We train an LDA model with 100 topics across all articles in the corpus [2]. After running the matching algorithm, we average the topic vector for each article in the control and the treatment group, using 1/1000 additive smoothing to avoid having 0 probabilities for any topic, and then normalize these added vectors into valid probability distributions. Thus, we obtain a topic probability distribution vector for the treatment group and for the control group. We then compute the KL-divergence between these two vectors. Since KL-divergence is not symmetric, we compute it in both the treatment-control and the control-treatment directions (reported as "KL" and "KL 2", respectively).

5 Parameter Tuning

The Pivot-Slope TF-IDF Weighting approach requires setting two parameters, the slope and the pivot, that control the strength of the normalization adjustment. As recommended, we fixed the pivot to be the average of the unadjusted normalization term, in our case, the average number of categories across our data set (9.3) [14]. We then tuned the slope in two ways: first, we sampled a fixed set of 1000 people from the data set, and second, we sampled a fixed set of 10 categories and sampled 500 people from each of these categories during tuning. We then used grid search to select the slope value that minimized the difference between the treatment and control sets using the metrics described in §4. We fixed the slope to 0.3.

6 Results

We evaluate each method using 100 simulations. For each simulation, we randomly draw 1000 people or we randomly draw a category and 500 people from the category in order to build a synthetic treatment group, and then we use the chosen method to identify a matched control group. We report results averaged over the 100 simulations. In addition to the described matching algorithms, we show the results of randomly sampling a control group that has the same number of people as the treatment group.

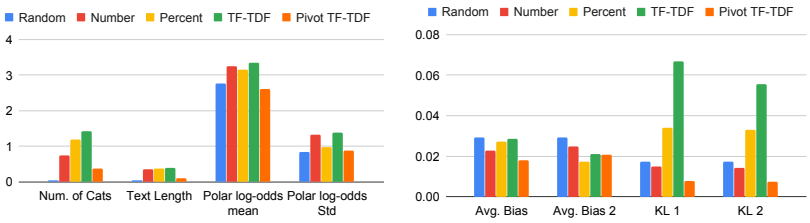

Fig. 1. Evaluation of matching methods using random samples of 1000 people as the treatment group, averaged over 100 simulations. Results are divided into two figures for readability. Pivot-Slope TF-IDF matching performs the best.

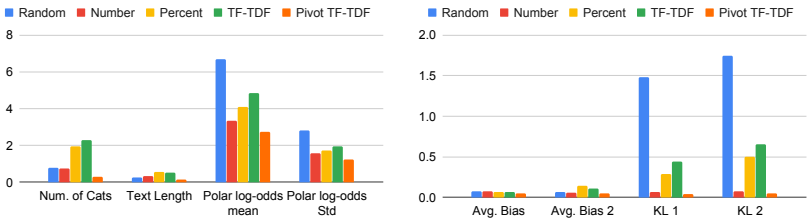

Fig. 2. Evaluation of matching methods by sampling a category and sampling a 500-person treatment group from the category, repeated 100 times.

Figures 1 and 2 report results. All of the evaluation metrics measure differences between the treatment group and the control group, meaning a lower value indicates the control group is a better match. In the Category sampling simulations (Figure 2), which better simulates having a treatment group with a particular trait in common, all matching methods perform better than random sampling, and the Pivot-Slope TF-IDF method performs the best overall. In the Random sampling simulations (Figure 1), random sampling provides a strong baseline. This is unsurprising, as a two randomly chosen groups of 1000 articles are unlikely to differ significantly from each other. Nevertheless, the Pivot-Slope TF-IDF method outperforms random sampling on the text-based metrics (polar log-odds and KL divergence) as well as on average bias.

The Number of Categories (Num. of Cats) and Text Length metrics do show possible biases of these methods. As expected, the Number of Categories, Percent Categories, and TF-IDF Weighting matching methods all exhibit bias towards articles with more or fewer categories, which results in worse performance than random sampling over these 2 metrics in Figures 1 and 2. In the Number of Categories matching method, this effect is positive, indicating that articles in the control group tend to have more categories, while in the Percent Categories and TF-IDF Weighting methods the effect is negative (Figures 1 and 2 report absolute values), indicating that articles in the control group tend to have fewer categories. These differences are also reflected in text length, in that articles

with more categories also tend to be longer. In the Category Evaluation, pivot-slope normalization corrects for this length bias, and demonstrates better-than-random matches. In the Random Evaluation, while the pivot-slope normalization does outperform other metrics, the random method exhibits the least category number and text length bias. However, as mentioned, random sampling is a strong baseline in this setting.

In Tables 1-4, we provide examples of matched controls for a set of sample people. These examples illustrate some of the broader trends of the algorithms. Notably, the Percent Categories and TD-IDF weighting methods strongly prefer short controls with few categories, even when they are not particularly meaningful. Both the Number of Categories and the Pivot TF-IDF methods produce meaningful pairs. However, the TF-IDF weighting upweights more specific categories. The Number of Categories method matches Barak Obama to Michael Moore based on broad categories like "American people of Irish Descent", the TF-IDF weighting matches Barak Obama to Rolan Burris based on more specific categories, like "United States Senators from Illinois". In the Number of Categories method, the match between T-Pain and Pharell Williams seems adept, while in the Pivot-Slope TF-IDF weighting method, Yuna Kim and Mao Asada is a particularly accurate pairing, as these two figure skaters are well-known for their rivalry.

7 Conclusions

We present a method that can be used to control for confounding variables in comparing corpora of Wikipedia articles. While we focus on Wikipedia biography pages, this method could be used to construct controlled corpora for any set of documents with associated metadata categories. Our evaluation metrics suggests that our method successfully constructs well-balanced control corpora. More specifically, our control corpora have more similar covariate distributions, in terms of the covariates they are actively matched on (average bias of categories) as well as possible text-artifacts of those covariates (polar log-odds and topic KL divergence), than random samples. However, our matching methods may exhibit favoritism to short or longer articles, and thus we do not recommend analyzing text length differences in corpora constructed using our approach.

T-Pain, Pharrell Williams
American music industry executives, 21st-century American rappers, American male singers, African-American male singers, Grammy Award winners for rap music, 21st-century American singers, American hip hop singers, African-American record producers, American hip hop record producers, Southern hip hop musicians, African-American male rappers, American contemporary R&B singers
Barack Obama, Michael Moore
Male feminists, HuffPost writers and columnists, 21st-century American male writers, LGBT rights activists from the United States, American gun control activists, American people of English descent, American male non-fiction writers, American political writers, American people of Irish descent, 21st-century American non-fiction writers, 20th-century American male writers, 20th-century American non-fiction writer, American people of Scottish descent
Meryl Streep, Julia Roberts American voice actresses, Best Actress BAFTA Award winners, Actresses of German descent, BAFTA winners (people), Best Drama Actress Golden Globe (film) winners, American people of German descent, American film actresses, American people of English descent, Best Supporting Actress Golden Globe (film) winners, 21st-century American actresses, American people of Irish descent, Actresses of British descent, 20th-century American actresses, American stage actresses, American television actresses, Best Actress Academy Award winners, Outstanding Performance by a Female Actor in a Leading Role Screen, Actors Guild Award winners, Best Musical or Comedy Actress Golden Globe (film) winners
Yuna Kim, Tessa Virtue
Olympic medalists in figure skating, Medalists at the 2010 Winter Olympics, Figure skaters at the 2010 Winter Olympics, World Junior Figure Skating Championships medalists, Medalists at the 2014 Winter Olympics, Figure skaters at the 2014 Winter Olympics, World Figure Skating Championships medalists, Four Continents Figure Skating Championships medalists, Season-end world number one figure skaters)
Amitabh Bachchan, S. P. Balasubrahmanyam
20th-century Indian singers, Filmfare Awards winners, Bollywood playback singers, Recipients of the Padma Shri in arts, Indian male voice actors, 21st-century Indian male actors, Indian male film actors, Indian male film singers, 21st-century Indian singers Indian male singers, 20th-century Indian male actors, Indian television presenters, Recipients of the Padma Bhushan in arts)
Tim Cook, Bob Iger
American chief operating officers, Biography with signature 20th-century American businesspeople, Directors of Apple Inc. 21st-century American businesspeople
Ron Berger (professor), Yukiko Iwai (singer)
1968 births

Table 1. Matches and common categories obtained using the Number of Categories method

T-Pain, Tay Dizm	Musicians from Tallahassee, Florida, Singers from Florida, 21st-century American singers, 21st-century male singers, Jive Records artists, African-American male rappers, RCA Records artists, 21st-century American rappers, Rappers from Florida
Barack Obama, Robert Moffit	American male non-fiction writers, American political writers
Meryl Streep, Meghan Strange	20th-century American actresses, American film actresses, 21st-century American actresses, American television actresses, American voice actresses
Yuna Kim, Park Solhee	1990 births, South Korean writers
Amitabh Bachchan, Kapil Jhaveri	Male actors in Hindi cinema, Indian male film actors
Tim Cook, Joe Fuca	American technology chief executives, 21st-century American businesspeople
Ron Berger (professor), Jean-Christophe Valtat	1968 births

Table 2. Matches and common categories obtained using the Percent Categories method

T-Pain, Vexxed	Twitch streamers
Barack Obama, JJonak	Use American English from August 2018
Meryl Streep, Dorothy Emmerson	American musical theatre actresses, 20th-century American actresses
Yuna Kim, Tommy Chang (martial artist)	South Korean expatriates in Canada
Amitabh Bachchan, Kishore Lulla	Film producers from Mumbai, Hindi film producers
Tim Cook, Stephen Austin (American football)	National Football League executives
Ron Berger (professor), Sheldon Hall (film historian)	Academics of Sheffield Hallam University

Table 3. Matches and common categories obtained using the TF-IDF method

T-Pain, Tay Dizm
21st-century American rappers, Jive Records artists, 21st-century American singers, Rappers from Florida, 21st-century male singers, Musicians from Tallahassee, Florida, Singers from Florida, RCA Records artists, African-American male rappers
Barack Obama, Roland Burris
Illinois Democrats, United States senators from Illinois, African-American United States senators, Democratic Party United States senators, 21st-century American politicians, Politicians from Chicago, African-American people in Illinois politics
Meryl Streep, Joanne Woodward
American stage actresses, Best Miniseries or Television Movie Actress Golden Globe winners, BAFTA winners (people), Outstanding Performance by a Female Actor in a Miniseries or Television Movie Screen Actors Guild Award winners, Best Actress BAFTA Award winners, American film actresses, Kennedy Center honorees, American television actresses, Cannes Film Festival Award for Best Actress winners, 21st-century American actresses, Best Drama Actress Golden Globe (film) winners, Outstanding Performance by a Lead Actress in a Miniseries or Movie Primetime Emmy Award winners, 20th-century American actresses, Best Actress Academy Award winners
Yuna Kim, Mao Asada
Olympic medalists in figure skating, 1990 births, Medalists at the 2010 Winter Olympics, Figure skaters at the 2014 Winter Olympics, World Figure Skating Championships medalists, Four Continents Figure Skating Championships medalists, Season-end world number one figure skaters, Figure skaters at the 2010 Winter Olympics, World Junior Figure Skating Championships medalists
Amitabh Bachchan, Dilip Kumar
Male actors in Hindi cinema, Indian male voice actors, Recipients of the Padma Bhushan in arts, Indian actor-politicians, Indian male film actors, Male actors from Mumbai, 20th-century Indian male actors, Filmfare Awards winners, Recipients of the Padma Vibhushan in arts, Dadasaheb Phalke Award recipients, Biography with signature, Film producers from Mumbai
Tim Cook, Ellen Hancock
American technology chief executives, Apple Inc. executives, IBM employees, American chief operating officers
Ron Berger (professor), Liu Mingkang
Alumni of Cass Business School

Table 4. Matches and common categories obtained using the Pivot-Slope TF-IDF method

References

1. Adams, J., Brückner, H., Naslund, C.: Who counts as a notable sociologist on wikipedia? gender, race, and the "professor test". Socius **5**, 2378023118823946 (2019)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
3. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Advances in neural information processing systems. pp. 4349–4357 (2016)
4. Filatova, E.: Multilingual wikipedia, summarization, and information trustworthiness. In: SIGIR workshop on information access in a multilingual world. vol. 3 (2009)
5. Graells-Garrido, E., Lalmas, M., Menczer, F.: First women, second sex: Gender bias in wikipedia. In: Proceedings of the 26th ACM Conference on Hypertext & Social Media. pp. 165–174 (2015)
6. Harder, V.S., Stuart, E.A., Anthony, J.C.: Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychological methods **15**(3), 234 (2010)
7. Jullien, N.: What We Know About Wikipedia: A Review of the Literature Analyzing the Project(s). (2012), https://hal.archives-ouvertes.fr/hal-00857208
8. Kolbitsch, J., Maurer, H.A.: The transformation of the web: How emerging communities shape the information we consume. J. UCS **12**(2), 187–213 (2006)
9. Monroe, B.L., Colaresi, M.P., Quinn, K.M.: Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis **16**(4), 372–403 (2008)
10. Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E.: A systematic literature review on wikidata. Data Technologies and Applications (2019)
11. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237 (2018)
12. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
13. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management **24**(5), 513–523 (1988)
14. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: ACM SIGIR Forum. vol. 51, pp. 176–184. ACM New York, NY, USA (2017)
15. Wagner, C., Graells-Garrido, E., Garcia, D., Menczer, F.: Women through the glass ceiling: gender asymmetries in wikipedia. EPJ Data Science **5**(1), 5 (2016)
16. Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., Chang, K.W.: Gender bias in contextualized word embeddings. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 629–634 (2019)
17. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2979–2989 (2017)