# Narrative Bias In Wikipedia: Different Links, Different Stories

Raine Hoover (raine@stanford.edu)
Stanford University, December 12, 2016

## Theory

Language (narrative) determines what concepts are used to explain a topic on Wikipedia; a narrative is defined by the concepts it chooses to include or exclude.

## Experiments And Hypotheses

(1) Given the concept graph for all languages….

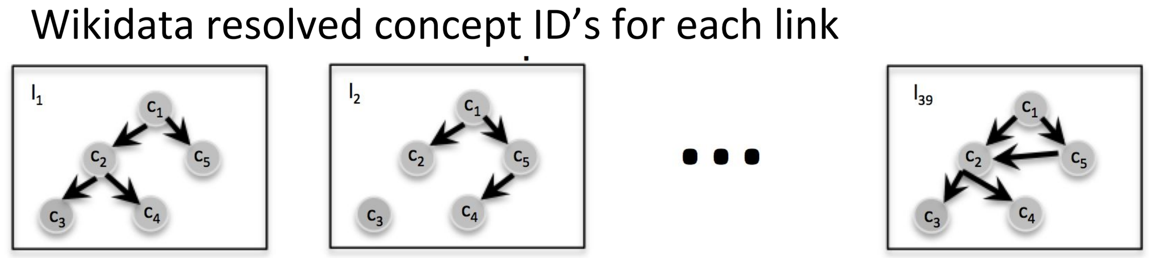
- Clustering
  - Running pagerank on the different concept networks in different languages and then clustering those pageranks will show geopolitical alignments across narratives.

(2) Given articles written in Hebrew or Arabic, and the concept links that each article does or does not contain…

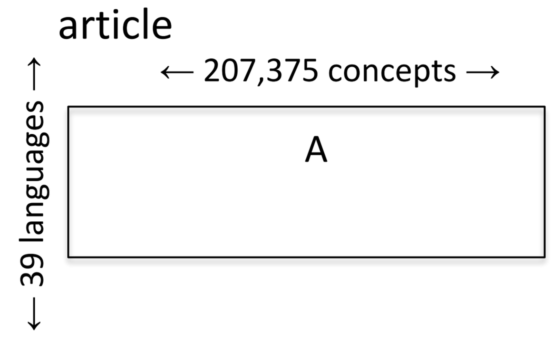
- (a) Logistic Regression:
  - We can predict which language an article is written in based solely on the concepts it does or does not contain
- (b) SVD:
  - One of the main principal components of the articles' concept links will be the language in which the article is written

## Data Set And Experiment Setup

- Wikipedia links from Wikipedia.org
- Wikidata resolved concept ID's for each link

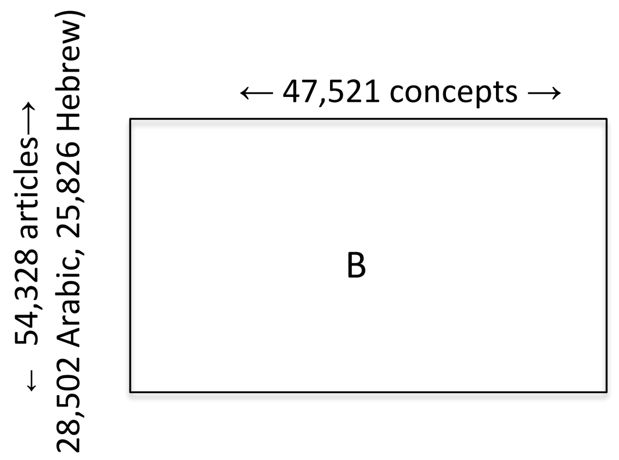


(1) Clustering setup
- 2-hop depth first search from seed article 'Arab-Israeli Conflict' in each language
- Calculate personalized pagerank with a 1 for the seed article




$a_{ij}$ = pagerank for concept i in language j's network

(2) Logistic Regression and SVD setup
- Combined adjacency matrix for Arabic and Israeli graphs




$b_{ij}$ = if article i cites concept j

## Method of Analysis

(1) Clustering
- Hierarchical agglomerative clustering
- Using a distance metric, assign each element (or group of elements) to closest cluster until there is only one cluster
  - i.e. Euclidean metric (I used Euclidean and cosine)[1] :

$$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

(2) (a) Logistic Regression
- Model how features interact to predict label as coefficients on features, θ.
  - Assign label according to [2]:

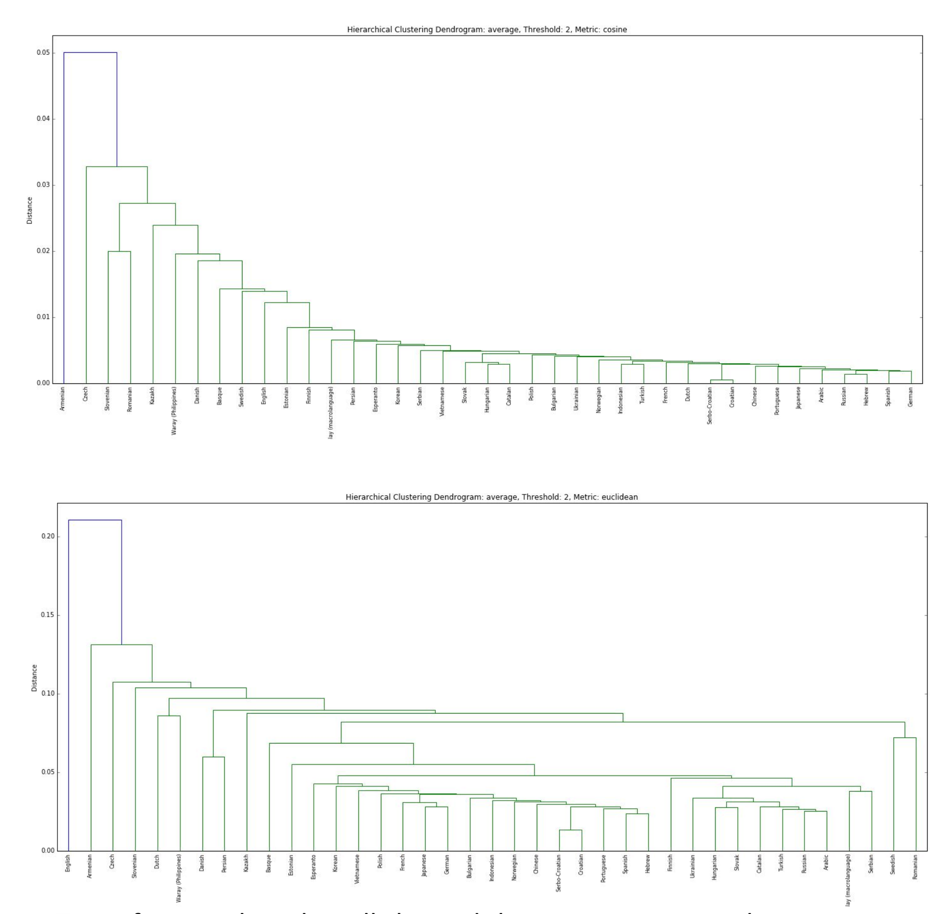
$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

- Split the data into train and test sets randomly.
- Learn θ in order to maximize likelihood of training data.
- Predict on test data.

(2) (b) SVD
- Find the principal components of the data (the axes that explain the most variance across the data) using the formula [3]:

$$\mathbf{X} = \mathbf{UDV}^\intercal$$

## (1) Clustering Results



Two of many plots that all showed the same pattern: a degenerate clustering.

The subspace of all languages' concepts was too large for any clustering pattern to emerge. In many of the plots, Hebrew and Arabic, the two languages most likely to diverge in narrative, were the closest together.

On hand inspection of the weights in Arabic and Hebrew's pageranks, however, there was an interesting pattern for some of the more controversial topics; this implied that there was a difference in link structure, at least between these two languages.

## (2) (a) Logistic Regression Results

| | Full data | Just intersect |
|---|---|---|
| Accuracy | 0.95325 | 0.93377 |

**Coefficients**

Salient categories for the concepts that were heavily weighted towards Hebrew:

- Dates and Years
  - "October 12", "1980s", "2000s", "1967", "1998"
- Loaded Pro-Israel terms
  - "Land of Israel", "Aliyah", "State of Palestine", "Canaan"
- Arab aggression
  - "Arab League Boycott of Israel", "Palestinian political violence", "Palestinian stone-throwing"
- Nuclear threat
  - "Plutonium", "Nuclear reactor"
- Cold War
  - "Space Race", "Berlin Wall"
- Peace Negotiations/Goodwill
  - "Two-state solution", "Israeli Disengagement from Gaza"

Salient categories for the concepts that were heavily weighted towards Arabic:

- Religion
  - "Hebrew calendar", "Book of Ruth", "Solomon", "Jewish philosophy", "Islam", "Christianity", "Judaism"
- Israeli Aggression
  - "1982 Lebanon War", "1948 Palestinian Exodus"

**Confusion Matrices**

Predicted (columns), Actual (rows)

| ALL | 'ar' | 'he' |
|---|---|---|
| 'ar' | 13,715 | 459 |
| 'he' | 811 | 12,179 |

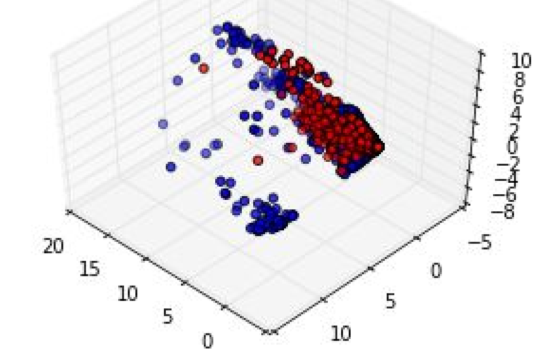
Predicted (columns), Actual (rows)

| INTERSECT | 'ar' | 'he' |
|---|---|---|
| 'ar' | 13,755 | 500 |
| 'he' | 1,299 | 11,610 |

Performance is best for Arabic articles, perhaps because there are ~3000 more of them in the dataset.

An in-depth error analysis is yet to be conducted.

## (2) (b) SVD Results



Plot of data across 3 principal components
(Arabic is blue, Hebrew is red)

Percentage of variance explained by each component:

| 0 | 1 | 2 |
|---|---|---|
| 68.07925 | 18.24344 | 13.67731 |

On manually inspecting the components (the first three vectors of U and V), the first component appears to be topic.

For all the components of U, there is at most a 20% correlation between the coefficients of U for that component and the language. This suggests that there is not a strong argument for language being a key factor in explaining variance across link structures.

## Discussion and Future Work

Overall, the experiment yielded mixed results: clustering and SVD did not yield obvious endorsements of the theory, but the high performance of the logistic regression classifier suggests that indeed we can define a particular language's narrative based on the concepts an article uses to explain a topic.

- **(1) Clustering**:
  - It is clear that there is a good deal of noise when it comes to concepts that are included but are not directly related to the topic of interest. Without these irrelevant articles, some patterns might emerge between the pagerank vectors, as they would be in a lower dimensional space.
- **(2) (a) Logistic Regression**:
  - Error analysis
    - Digging more deeply into what articles the classifier did or did not classify correctly would illuminate what the coefficients are not capturing
    - Investigating what it came close to classifying incorrectly (probabilities close to .5) would illuminate what the classifier is seeing as a "neutral" article. Perhaps we could use this result in deciding what is an "irrelevant" concept or article.
- **(2) (b) SVD**
  - Normalization
    - In doing SVD, no normalization on the adjacency matrix was performed. Results could be skewed by the number of links an article contains, or the number of articles that link to a concept.
  - Inspecting other components
    - The first component was topic, as evidenced by manual categorization of the articles-- a more detailed manual inspection of the other components is needed before their meaning can be interpreted.
  - Eliminating irrelevant concepts and articles.

## References

[1] "Distance -- from Wolfram MathWorld", *Mathworld.wolfram.com*, 2016. [Online]. Available: http://mathworld.wolfram.com/Distance.html. [Accessed: 13- Dec- 2016]
[2] "CS229 Notes 1: Supervised Learning". *cs229.stanford.edu*. N.p., 2016. Web. 13 Dec. 2016.
[3] Niculae, Vlad, et al. "Quotus: The structure of political media coverage as revealed by quoting patterns." *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015.