Name Embeddings and Online News Analysis

Speaker: Junting Ye
Department: Computer Science
Advisor: Prof. Steven Skiena
Outline

● **Overview**

● **Name Embeddings**
  ○ Nationality Classification
  ○ Ethnicity & Gender Embeddings

● **Quality Analysis of News and Social Media**
  ○ Motivation
  ○ *MediaRank* Overview
  ○ Progress

● **Future Work**
## Overview

### News Analysis

### Name Embeddings

### Opinion Spam Detection

### Others
Outline

● Overview

● Name Embeddings
  ○ Nationality Classification
  ○ Ethnicity & Gender Embeddings

● Quality Analysis of News and Social Media
  ○ Motivation
  ○ MediaRank Overview
  ○ Progress

● Future Work
Name Embeddings

Using our machine learning algorithm, each name part (first or last name) is represented by a 100-dimension vector (i.e. embedding).

When projecting 100-dimension to 2-dimension:
### Name Embeddings

<table>
<thead>
<tr>
<th>Input (examples)</th>
<th>Gerda_Zavada@ Roxana Carmen, Adina Margine, Radoi Seicaru, Drînd Ramona,…</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chilap_ja@ leung Ja, Chow Iris, Ken Ja, Betty Cheung, Chan Stone, Donna Tang, …</td>
</tr>
<tr>
<td></td>
<td>balbirsingh@ Krishan Singh, Neeraj Kumar, Pankaj Bawa, Vijay Kumar, …</td>
</tr>
</tbody>
</table>

| Objective Function (negative sampling) | \[
\log \sigma(w_O^T v_I) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[ \log \sigma(-v_i^T v_I) \right]
\] |

<table>
<thead>
<tr>
<th>Labels</th>
<th>Positive: name part pairs in the same list</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative: random name part pairs</td>
</tr>
</tbody>
</table>

| Output | Distributed representation of name parts |
NamePrism: A nationality classifier

Our API* has been supporting 100+ research projects from social science, economics, etc..

<table>
<thead>
<tr>
<th>Research Project Goal</th>
<th>Research Group</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>“working on racial representation in historical bureaucracies”</td>
<td>Haas School of Business, UC Berkley</td>
<td>U.S.</td>
</tr>
<tr>
<td>“determine if ethnic group size impacts national cabinet diversity”</td>
<td>Department of Political Science, Washington University in St. Louis</td>
<td>U.S.</td>
</tr>
<tr>
<td>“promote the contributions of Iranian Americans to members with-in and outside of the Iranian community living in America.”</td>
<td>Iranian Americans’ Contributions Project</td>
<td>U.S.</td>
</tr>
<tr>
<td>“determine if ethnicity plays a part/plays no part in whether a written evidence submitted to a Parliamentary Inquiry is accepted or rejected”</td>
<td>Parliamentary Digital Service</td>
<td>UK</td>
</tr>
<tr>
<td>“working on a study on the network effects for long term unemployed”</td>
<td>German Institute for Employment Research</td>
<td>Germany</td>
</tr>
<tr>
<td>“unveiling the origins of French citizens in order to study discrimination in several areas of the French society”</td>
<td>Laboratoire Interdisciplinaire Sciences Innovations Sociétés (LISIS)</td>
<td>French</td>
</tr>
<tr>
<td>“Investigate whether hosts on Airbnb get discriminated based on their ethnicity”</td>
<td>Stockholm School of Economics</td>
<td>Sweden</td>
</tr>
</tbody>
</table>

Media Coverage

- WIRED Magazine;
- Irish Tech News;
- TyN Magazine;
- 24 Heures;
- ….

*: www.name-prism.com
## Gender & Ethnicity Classification

<table>
<thead>
<tr>
<th>Embedding</th>
<th>White</th>
<th>Black</th>
<th>API</th>
<th>Hisp.</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retweet</td>
<td>0.92</td>
<td>0.20</td>
<td>0.57</td>
<td>0.64</td>
<td>0.58</td>
</tr>
<tr>
<td>Mention</td>
<td>0.93</td>
<td>0.22</td>
<td>0.61</td>
<td>0.71</td>
<td>0.62</td>
</tr>
<tr>
<td>Follower</td>
<td>0.94</td>
<td>0.31</td>
<td>0.77</td>
<td>0.86</td>
<td>0.72</td>
</tr>
<tr>
<td>Followee</td>
<td>0.92</td>
<td>0.27</td>
<td>0.72</td>
<td>0.81</td>
<td>0.68</td>
</tr>
<tr>
<td>Followee*</td>
<td>0.94</td>
<td>0.31</td>
<td>0.77</td>
<td>0.84</td>
<td>0.72</td>
</tr>
<tr>
<td>Friends</td>
<td>0.93</td>
<td>0.28</td>
<td>0.74</td>
<td>0.81</td>
<td>0.69</td>
</tr>
<tr>
<td>NonFriends</td>
<td>0.92</td>
<td>0.26</td>
<td>0.71</td>
<td>0.82</td>
<td>0.68</td>
</tr>
<tr>
<td>Aggregated</td>
<td>0.93</td>
<td>0.32</td>
<td>0.76</td>
<td>0.83</td>
<td>0.71</td>
</tr>
<tr>
<td>Aggregated*</td>
<td>0.94</td>
<td>0.33</td>
<td>0.79</td>
<td>0.86</td>
<td>0.73</td>
</tr>
<tr>
<td>Email</td>
<td>0.96</td>
<td>0.47</td>
<td>0.83</td>
<td>0.87</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Table 3: Ratios of same-ethnicity names among nearest neighbor (i.e. $k = 1$). *Aggregated* gets promising performance. It achieves comparable performance on *White* and *Hispanic*. *Black* names are harder task because they only take up 3.5% of all labels. (API: Asian and Pacific Islander)

Figure 2: Ratio of same-gender names among top $k$ nearest neighbors ($k \in [1, 10, 50, 100]$). *Mention* performs the best (avg. on female: 0.94, male: 0.74). *Aggregated* outperforms *Email* (female (avg.): 0.94 vs. 0.91, male: 0.67 vs. 0.59). Performance of male names are generally lower than female, because there are far less male names (29% vs. 71%).
Outline

● Overview

● Name Embeddings
  ○ Nationality Classification
  ○ Ethnicity & Gender Embeddings

● Quality Analysis of News and Social Media
  ○ Motivation
  ○ MediaRank Overview
  ○ Progress

● Future Work
Quality Analysis of News and Social Media

Motivation

- **Fake news** went viral in 2016 election
  - Pizzagate of Hillary Clinton
  - Pope endorse Donald Trump
  - ISIS leader calls for American Muslim voters to support Hillary Clinton
  - Donald Trump sent his own plane to transport 200 stranded marines in 1991
  - ...

- **Impact** of fake news on social media
  - 62% U.S. adults get news on social media in 2016 [1]
  - 15% recall seeing fake news headlines [1]
  - Popular fake news shared more times and faster on Facebook than mainstream news [2]


[2]: [S. Vosoughi, Science, 2017]
MediaRank

Figure 1: Four major components of MediaRank system.
MediaRank: System Overview

- **AWS**: Website server for UI
- **Master server with 50TB storage**
- **Cluster of 85 workers**
- **OpenStack for virtualization; Ansible for cluster management**
- **Celery for distributed task management**

### Table 1: Sizes of datasets from last 6 months and they are growing on daily basis.

<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Html</th>
<th>T_Post</th>
<th>T_User</th>
<th>FB_Post</th>
<th>FB_Comment</th>
<th>FB_Like</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td></td>
<td>71K</td>
<td>224M</td>
<td>375M</td>
<td>30M</td>
<td>51M</td>
<td>714M</td>
</tr>
<tr>
<td>Daily</td>
<td>0.15K</td>
<td>1.1M</td>
<td>2.5M</td>
<td>10K</td>
<td>150K</td>
<td>2.1M</td>
<td>–</td>
</tr>
</tbody>
</table>

*T*: Twitter, *FB*: Facebook.
MediaRank: News Analysis

- Independent Signals
  - Social Media
  - Monetization
  - Political Bias
  - Quality of the Coverage
  - Duplicate Articles
  - Popularity
  - Readability

- Relations
  - Hyperlinks
  - Common News Reader
Timeline for Following Year

- Aug. 2018 ~ Dec. 2018:
  - Investigating political bias and monetization;
  - Leading a team of two PhD and three master students on computing remaining signals and building reliable system;

- Jan. 2018 ~ May. 2018:
  - Modeling heterogeneous signals;
  - Publishing papers and defend thesis;
Q & A