Mining cultural insights from online texts

Big Data Science @ SDSU
March 7, 2015

Rob Malouf
Department of Linguistics and Asian/Middle Eastern Languages
“According to Computer World, unstructured information may account for more than 70% to 80% of all data in organizations. These data, which mostly originate from social media, constitute 80% of the data worldwide and account for 90% of Big Data.”


Figure 2: Dependency tree voor de zin *Kim wil weten of Anne komt*
Computational linguistics

Natural Language Processing: Make information contained linguistically structured data available for further processing

Deep analysis vs. shallow analysis

- Quality of results
- Implementation difficulty
- Scalability
Meaning

Shallow methods scale to billions or trillions of words (“There’s no data like more data!”)

Start from scratch and bootstrap linguistic knowledge

  Word meanings
  Phrase types
  Constructions

After linguistic patterns are established, we can extract real-world knowledge from texts
“You shall know a word by the company it keeps.”

(J.R. Firth, 1957)

(a) Typical term-document incidence matrix $C$ ($C_{ij} = n \leftrightarrow$ document $D_j$ contains term $W_i$ exactly $n$ times)

(b) Typical term-term similarity matrix $R$

\[
\begin{pmatrix}
R_{i1} & R_{i2} & \cdots & R_{in}
\end{pmatrix} = R
\]

\[
\left( R_{ij} = R_{ji} = \sum_{k=1}^{m} C_{ik} C_{kj} / \sqrt{ \left( \sum_{k=1}^{m} (C_{ik})^2 \right) \left( \sum_{k=1}^{m} (C_{kj})^2 \right) } \right)
\]

**Fig. 2.** Matrices used for the generation of term associations.
Reduced singular value decomposition of the term x document matrix, X. Where:

- T has orthogonal, unit-length columns ($T^T T = I$)
- D has orthogonal, unit-length columns ($D^T D = I$)
- S is the diagonal matrix of singular values
- $t$ is the number of rows of X
- $d$ is the number of columns of X
- $m$ is the rank of X ($\leq \min(t,d)$)
- $k$ is the chosen number of dimensions in the reduced model ($k \leq m$)

FIG. 3. Schematic of the reduced Singular Value Decomposition (SVD) of a term by document matrix. The original term by document matrix is approximated using the $k$ largest singular values and their corresponding singular vectors.

Vector Space Models

Vector Space Models are one way to operationalize Firth’s distributional notion of meaning.

Results from shallow methods can only be as good as the input (representativeness).

Corpus of reading material for K-12 students

**bicycle:** pedals, handlebars, bicycles, pedaling, bike, starley, highwheeler, boneshaker, mede, lallement, gearwheels, gearwheel, drais, bikers, bikes, wheels, wheel, bicycled, pedal

**patriot:** 1775, patriots, lexington, concord, loyalist, loyalists, 1777, bunker, minutemen, hancock, 1776, redcoats, ticonderoga, sniping, framingham, edgel, revere, cornwallis, saratoga
Malouf, Edwards, Perez Ruiz, Richette, Southam, and DiChiara. “A computational lexical analysis of the language commonly used to describe gout.” (submitted)
This Is the Williamsburg of Your City: A Map of Hip America

Max Read

Filed to: WILLIAMSBURG  1/29/14 11:30am
310,000 posts
35 million words

San Diego & San Francisco forums

tokenization tagging

term extraction

term/document matrix

latent semantic model

term similarities

north park
real estate
young kids
cal state san marcos
san diego unified
school district
north park

north park (0.000) south park (0.054) university heights (0.055) normal heights (0.096) golden hill (0.128) hillcrest (0.147) kensington (0.149) mission hills (0.177) adams ave (0.191) hipster (0.201) np (0.206) bankers hill (0.234) morley field (0.240) adams avenue (0.322) other neighborhoods (0.353) banker (0.360) craftsman (0.367) burlingame (0.369) park west (0.372) adams (0.373) gentrified (0.381) nh (0.384) flight path (0.385) coffee shops (0.385) funky (0.402) artsy (0.418) cottage (0.431) neighborhoods (0.438) little italy (0.444) mewzikguy (0.454) housing stock (0.456) iffy (0.458) walkable (0.459) gritty (0.459) bungalow (0.462) damon (0.463) urban (0.468) hip (0.482) main drag (0.489) hoods (0.492) hillcrest area (0.494) university avenue (0.498) uh (0.502) neighborhood (0.505) kettlepot (0.506) university ave (0.508) hipsters (0.521) mansions (0.529) apartment buildings (0.532) pubs (0.538) charm (0.541) sherman heights (0.548) park blvd (0.552) trendy (0.556) great neighborhood (0.562) talmadge (0.569) antique (0.569) univ (0.574) walkability (0.576) great areas (0.581) character (0.586) balboa park (0.590) parts (0.595) eclectic (0.606) gay (0.607) bars (0.607) sp (0.608) ocean beach (0.609) tattoo (0.613) urban areas (0.618) pricier (0.618) shops (0.621) gentrification (0.626) counts (0.627) sketchy (0.628) cottages (0.630) particularly (0.634) coffee shop (0.634) beach communities (0.635) upscale (0.636) blocks (0.638) northpark (0.645) heights (0.645) cortez hill (0.648) central sd (0.650) urban neighborhoods (0.650) congested (0.652) bungalows (0.654) small city (0.655) vibe (0.662) charming (0.664) neighboring (0.666) reached (0.670) hood (0.672) northern part (0.673) hill (0.673) urban core (0.674) downside (0.675) rental budget (0.675) central san diego (0.677)
north park

hipster (0.201) craftsman (0.367) gentrified (0.381) flight path (0.385) coffee shops (0.385) funky (0.402) artsy (0.418) cottage (0.431) housing stock (0.456) iffy (0.458) walkable (0.459) gritty (0.459) bungalow (0.462) urban (0.468) hip (0.482) main drag (0.489) hipsters (0.521) mansions (0.529) apartment buildings (0.532) pubs (0.538) charm (0.541) trendy (0.556) great neighborhood (0.562) antique (0.569) walkability (0.576) great areas (0.581) character (0.586) parts (0.595) eclectic (0.606) gay (0.607) bars (0.607) tattoo (0.613) urban areas (0.618) pricier (0.618) shops (0.621) gentrification (0.626) counts (0.627) sketchy (0.628) cottages (0.630) particularly (0.634) coffee shop (0.634) beach communities (0.635) upscale (0.636) blocks (0.638) urban neighborhoods (0.650) congested (0.652) bungalows (0.654) small city (0.655) vibe (0.662) charming (0.664) neighboring (0.666) reached (0.670) northern part (0.673) hill (0.673) urban core (0.674) downside (0.675) rental budget (0.675)
clairemont mesa

centrally (0.389) mesa college (0.459) apartment complexes (0.472) shopping centers (0.493) single family homes (0.512) supermarkets (0.512) located (0.519) supermarket (0.523) easy access (0.528) shopping malls (0.551) near (0.561) zip (0.562) branch (0.572) min drive (0.582) albertsons (0.608) apartments (0.609) good neighborhoods (0.614) branches (0.617) military housing (0.618) home depot (0.619) close (0.638) only one (0.640) classifieds (0.640) quiet (0.650) campus (0.656) henry (0.658) short commute (0.658) nasty (0.661) newly (0.661) complex (0.662) rush hour (0.663) short drive (0.663) repair (0.663) shopping center (0.665) pricey (0.669) condo complex (0.671) apts (0.671) item (0.680) recommended (0.682) centers (0.682) congestion (0.683) roommate (0.685) nearby (0.685) stores (0.686) parkway (0.688) ins (0.690) hey everyone (0.690) good area (0.691) pricier (0.691) mcdonalds (0.693) nicest (0.693)
kearny mesa

chinese (0.247) korean (0.275) seafood (0.282) cuisine (0.287) authentic (0.313) sushi (0.316) convoy (0.325) japanese (0.325) thai (0.330) sandwiches (0.354) sauce (0.368) vietnamese (0.377) tasty (0.386) menu (0.389) good food (0.394) chef (0.397) hole (0.398) bread (0.415) cook (0.421) asian (0.428) cafe (0.431) food (0.435) grill (0.437) fresh (0.439) taco (0.444) deli (0.451) fries (0.462) mexican food (0.465) yelp (0.466) steak (0.472) restaurant (0.477) italian (0.478) burrito (0.481) wall (0.488) burger (0.491) ethnic (0.493) great food (0.495) cooked (0.495) french (0.498) strip mall (0.499) delicious (0.500) rolls (0.501) fried (0.502) beef (0.502) bacon (0.505) gems (0.510) cooking (0.511) asada (0.512) indian (0.514) filipino (0.515) egg (0.518) supermarkets (0.519) meat (0.520) eat (0.521) flavor (0.532) sandwich (0.534) chocolate (0.536) tacos (0.538) carne (0.540) burgers (0.543) dishes (0.546) fish (0.550) henry (0.552) portions (0.552) reservations (0.553) taco shop (0.553) lunch (0.554) rave (0.559) variety (0.560) el (0.560) eaten (0.560) pub (0.561) gourmet (0.562) chips (0.562) bakery (0.563) disappointed (0.564) chain (0.567) eating (0.567) mediocre (0.568) good places (0.570) try (0.572) breakfast (0.573) german (0.576) salad (0.581)
<table>
<thead>
<tr>
<th>Location</th>
<th>San Diego</th>
<th>San Francisco</th>
</tr>
</thead>
<tbody>
<tr>
<td>hipster</td>
<td>south park</td>
<td>mission</td>
</tr>
<tr>
<td>dim sum</td>
<td>kearny mesa</td>
<td>chinatown</td>
</tr>
<tr>
<td>affordable</td>
<td>(temecula)</td>
<td>gilroy</td>
</tr>
<tr>
<td>pricey</td>
<td>del mar</td>
<td>pleasant hill</td>
</tr>
<tr>
<td>dangerous</td>
<td>logan heights</td>
<td>tenderloin</td>
</tr>
<tr>
<td>tourist</td>
<td>old town</td>
<td>union square</td>
</tr>
<tr>
<td>condos</td>
<td>downtown</td>
<td>soma</td>
</tr>
<tr>
<td>gay</td>
<td>hillcrest</td>
<td>(castro)</td>
</tr>
<tr>
<td>good schools</td>
<td>carmel mountain ranch</td>
<td>san ramon</td>
</tr>
<tr>
<td>hiking</td>
<td>ocean beach</td>
<td>presidio</td>
</tr>
<tr>
<td>white collar</td>
<td>sorrento valley</td>
<td>walnut creek</td>
</tr>
<tr>
<td>flight path</td>
<td>bankers hill</td>
<td>(burlingame)</td>
</tr>
</tbody>
</table>
Prospects

Big data linguistic techniques applied to broad spectrum texts allow us to extract real-world intelligence from ‘unstructured’ data

When applied to more focused corpora, they yield insights about speakers that would not be accessible via traditional qualitative methods

Hybrid quantitative / qualitative methods