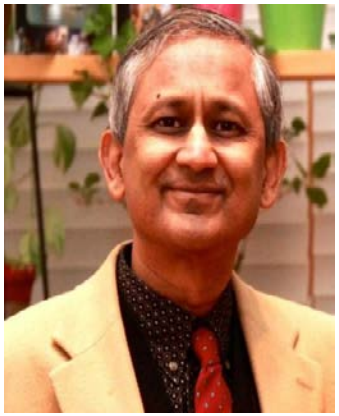
Knowledge Enabled Location Prediction of Twitter Users



Revathy Krishnamurthy
revathy@knoesis.org

Pavan Kapanipathi
pavan@knoesis.org

Amit Sheth
amit@knoesis.org

Krishnaprasad Thirunarayan
tkprasad@knoesis.org

Kno.e.sis: Ohio Center of Excellence in Knowledge Enabled Computing
http://knoesis.org

City of Lights

City of Lights

Background Knowledge can improve a machine's ability to interpret text

BACKGROUND KNOWLEDGE

Geographic footprint of a Twitter user

WHY IS LOCATION IMPORTANT?

News Recommender Systems

Recommended for you

Beavercreek preschool to open in 2015
By Sharon D. Boykin

A $5.1 million preschool in Beavercreek city Schools district will help accommodate a growing of student population and reduce overcrowding, according to school officials.

Ohio's health exchange to include more competition
By Randy Tucker

It was just a year ago that the insurance industry fretted over potential loses from the new insurance market created by Affordable Care Act.

WHY IS LOCATION IMPORTANT?

News Recommender Systems

Recommended for you

Beavercreek preschool to open in 2015
By Sharon D. Boykin

A $5.1 million preschool in Beavercreek city Schools district will help accommodate a growing of student population and reduce overcrowding, according to school officials.

Ohio's health exchange to include more competition
By Randy Tucker

It was just a year ago that the insurance industry fretted over potential loses from the new insurance market created by Affordable Care Act.

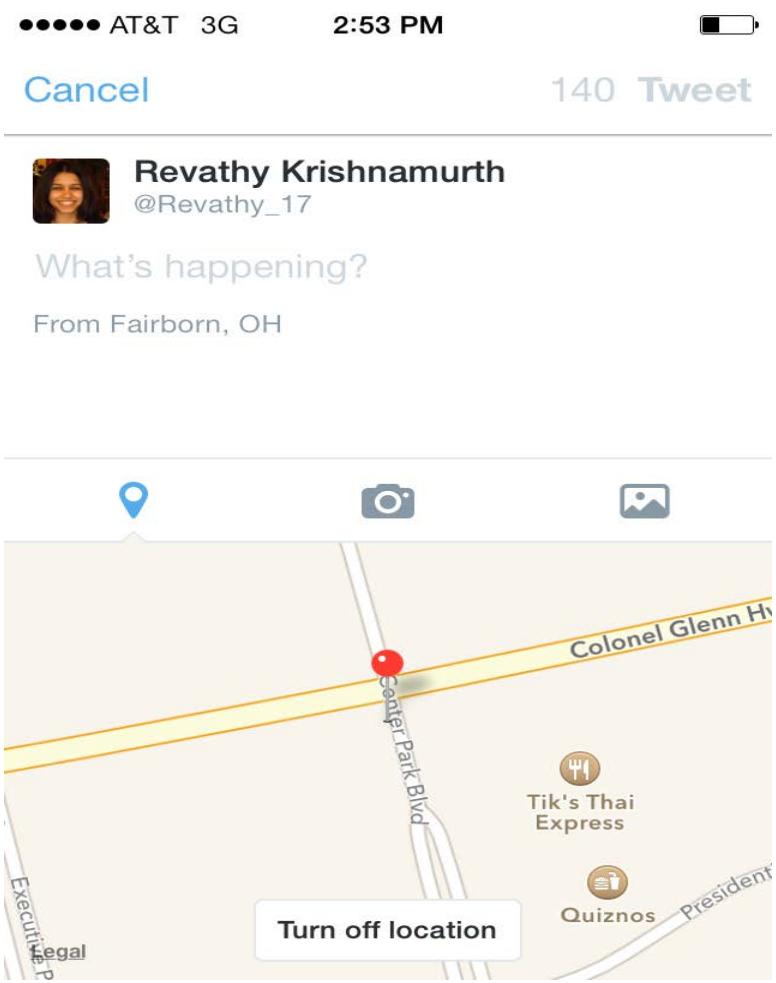
Other applications

- Targeted advertising
- Opinion Analysis
- Disaster Response
- Location Based Services

LOCATION PUBLISHED BY USER

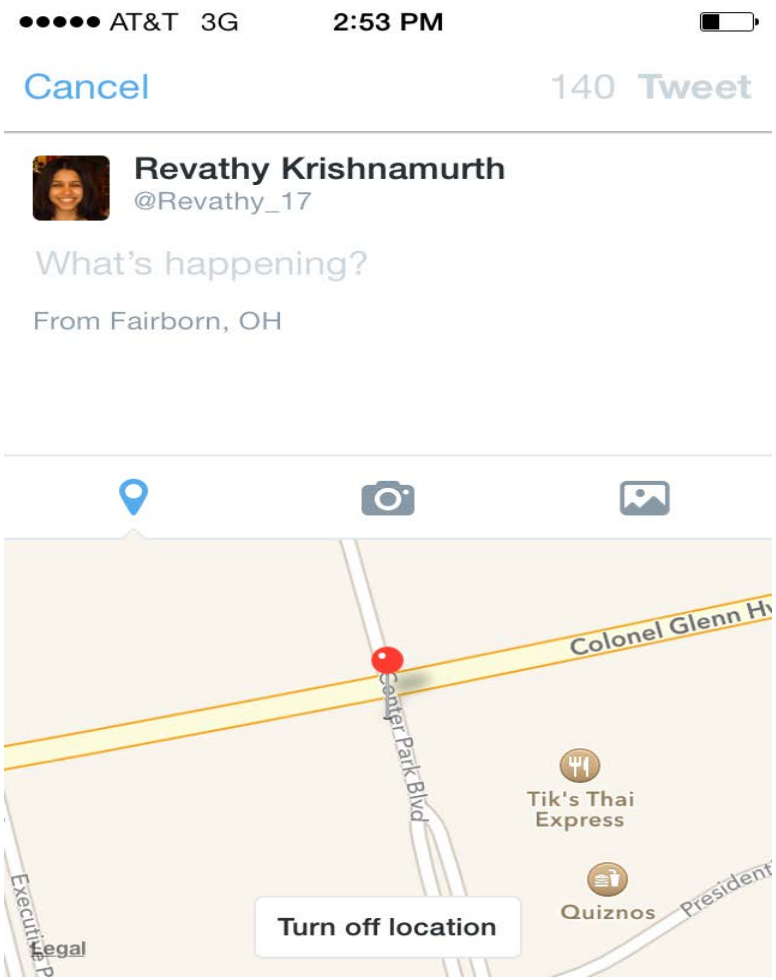
LOCATION PUBLISHED BY USER

Geo-tagged Tweets

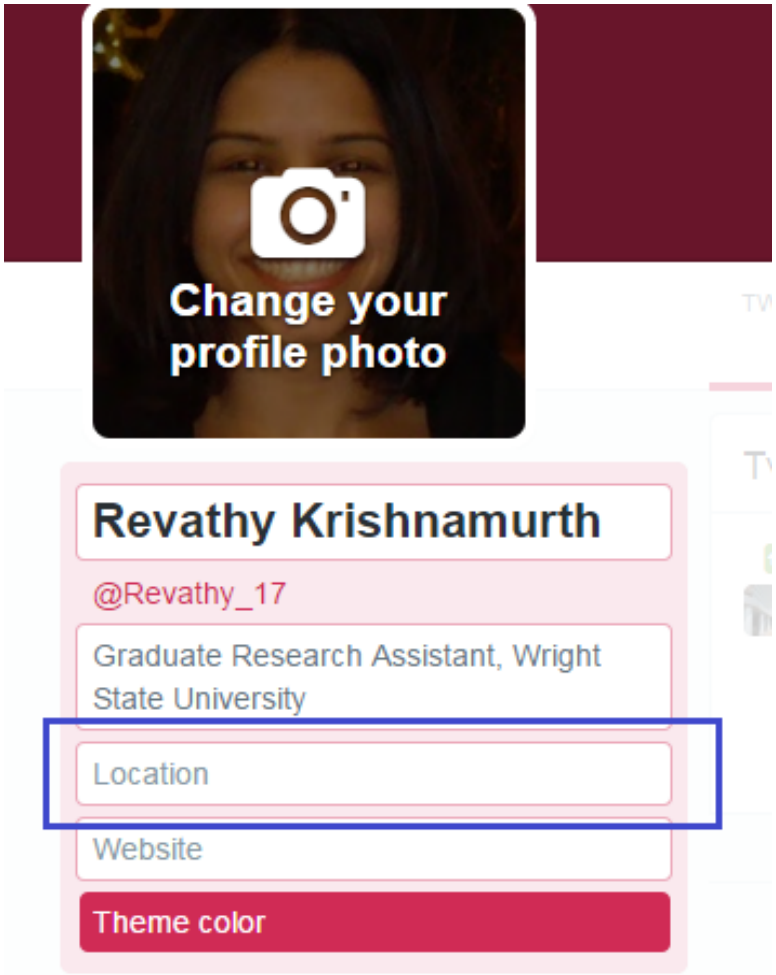


LOCATION PUBLISHED BY USER

Geo-tagged Tweets

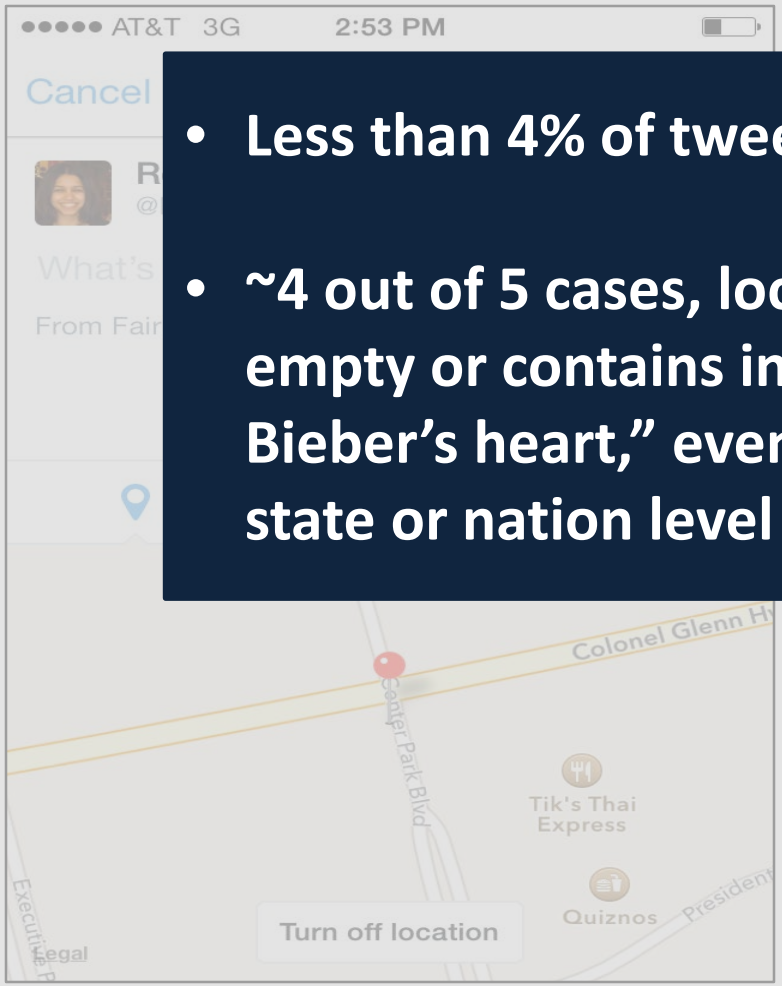


Profile Information

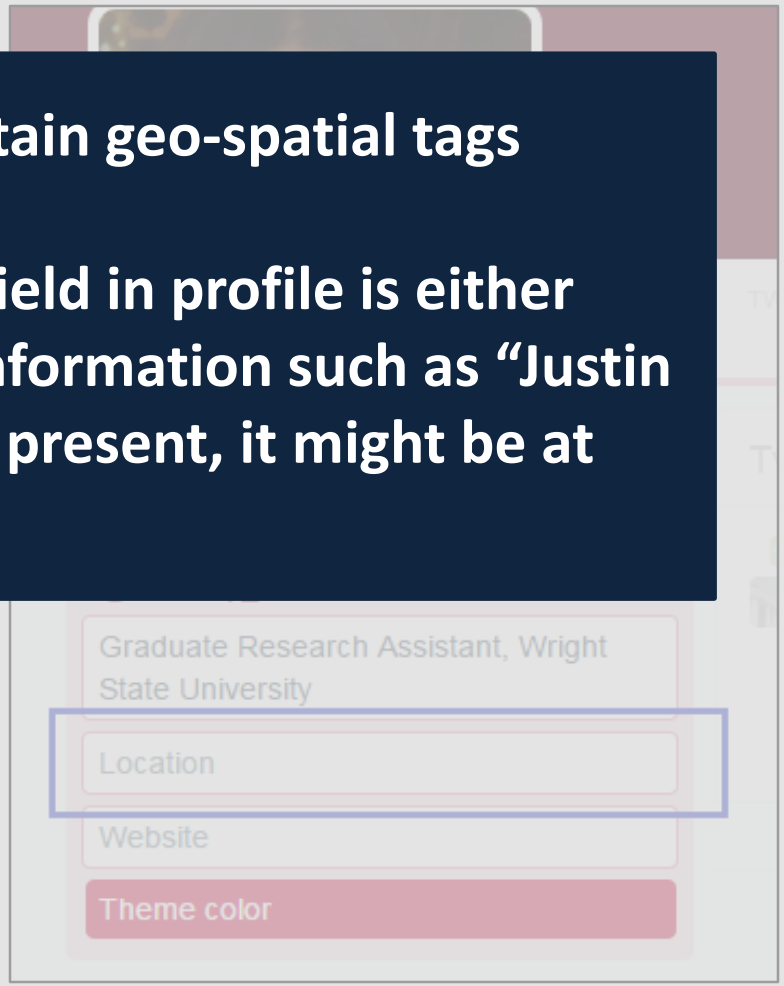


LOCATION PUBLISHED BY USER

Geo-tagged Tweets



Profile Information

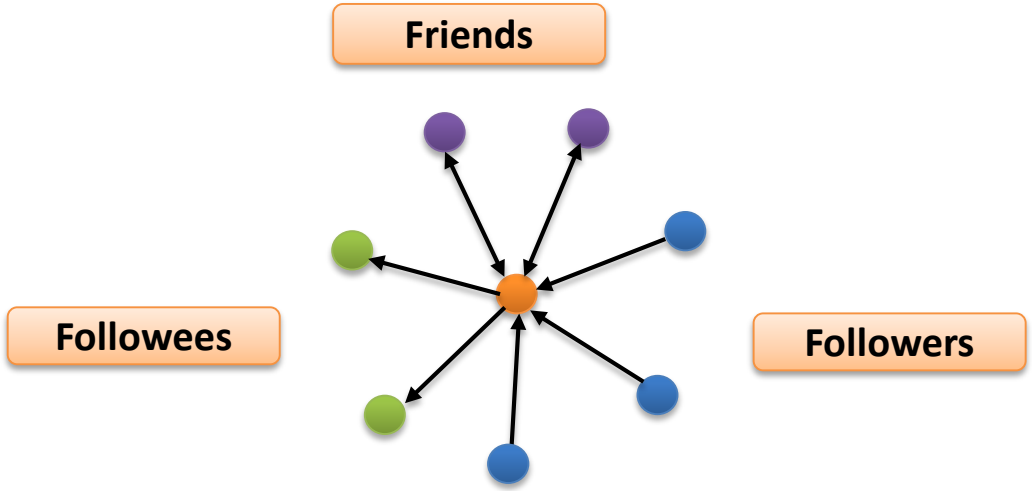


- Less than 4% of tweets contain geo-spatial tags
- ~4 out of 5 cases, location field in profile is either empty or contains invalid information such as “Justin Bieber’s heart,” even when present, it might be at state or nation level

LOCATION INFERENCE

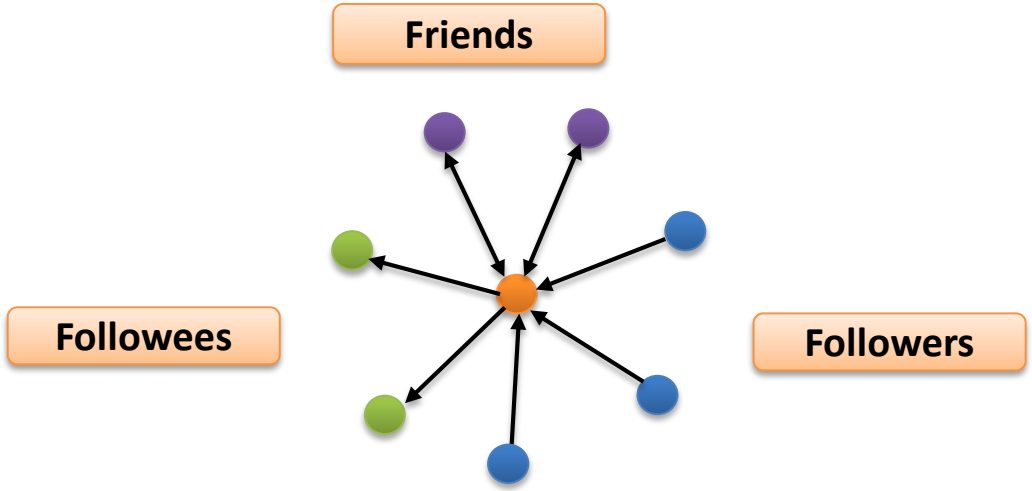
LOCATION INFERENCE

Network based



LOCATION INFERENCE

Network based



Content based

Just drove around Golden Gate Park two times trying to get in

Cleveland Browns confuse me. When I give up on them, they actually show up to play.

CONTENT BASED APPROACHES

Just drove around Golden Gate Park two times trying to get in

Cleveland Browns confuse me. When I give up on them, they actually show up to play.

CONTENT BASED APPROACHES

Just drove around Golden Gate Park two times trying to get in

Cleveland Browns confuse me. When I give up on them, they actually show up to play.

Geographic location of a user influences the contents of their tweets

CONTENT BASED APPROACHES

Just drove around Golden Gate Park two times trying to get in

Cleveland Browns confuse me. When I give up on them, they actually show up to play.

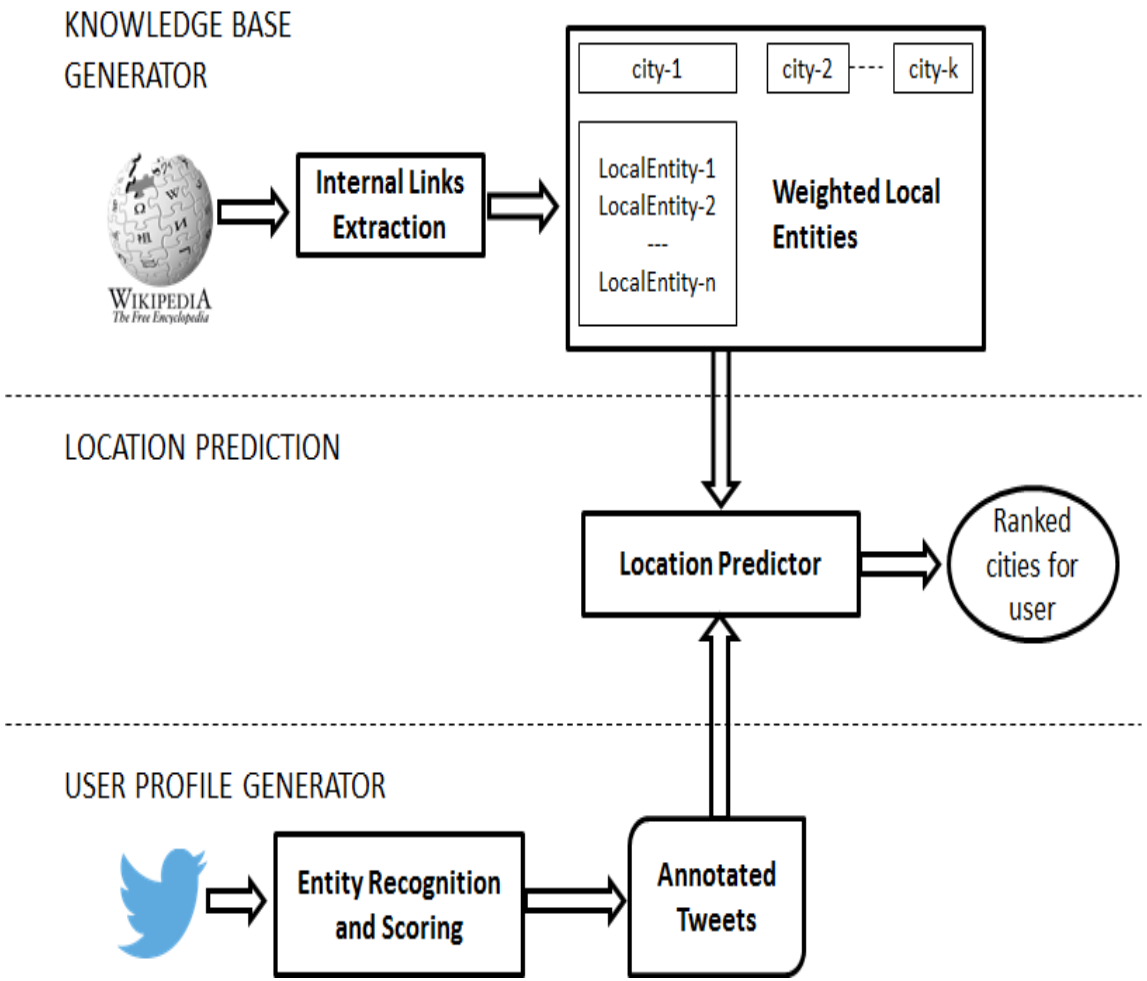
Geographic location of a user influences the contents of their tweets

- Supervised Approaches
  - Probabilistic Models – (Cheng, Caverlee, and Lee, 2010)
  - Cascading Topic Models – (Eisenstein, Connor, Smith, and Xing, 2010)
  - Gaussian Mixture Model – (Chang, Lee, Eltaher, and Lee, 2012)
  - Language Models – (Doran, Gokhale, and Dagnino, 2014)
  - Ensemble of Statistical and Heuristic Classifiers – (Mahmud, Nichols, and Drews, 2014)

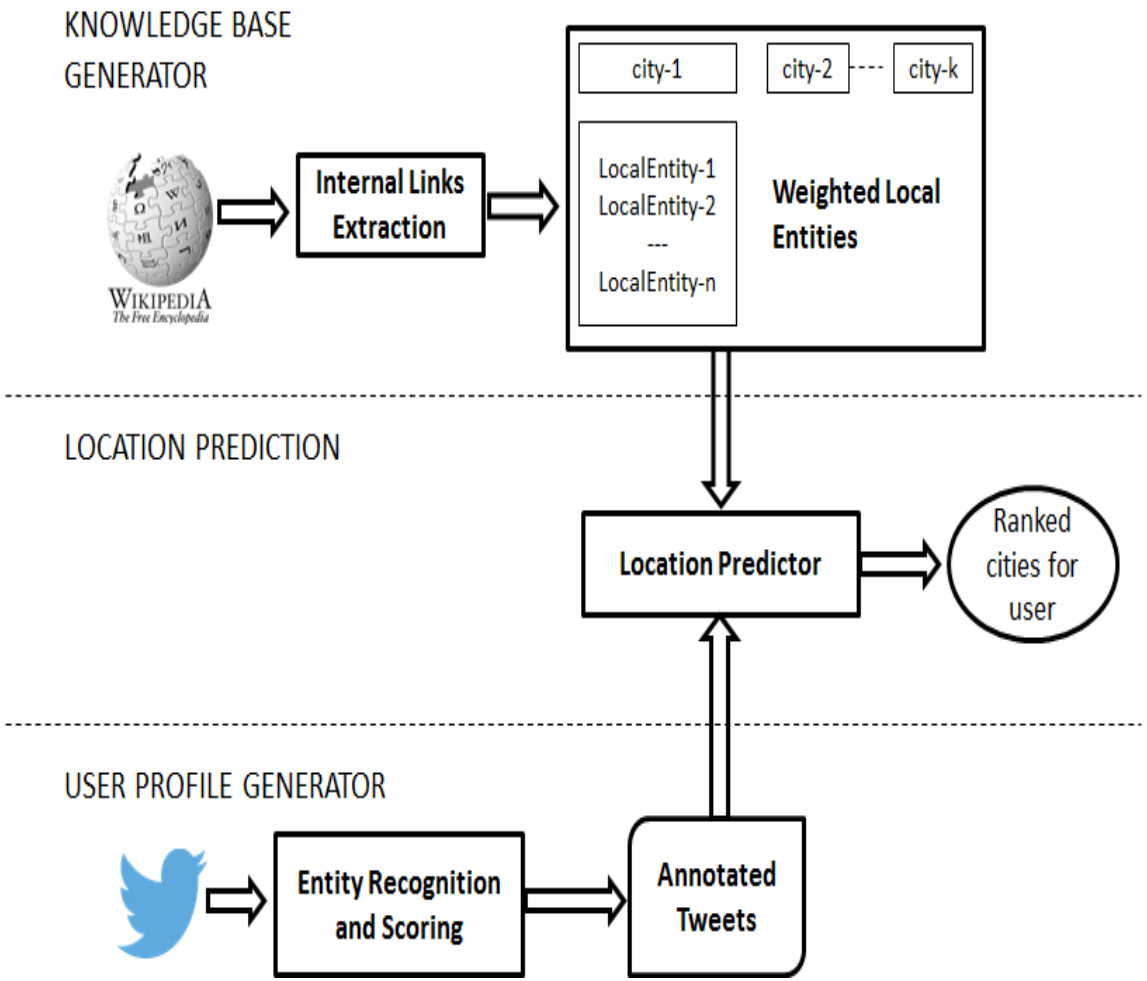
PROBLEM STATEMENT

Predict the location of a Twitter user based on their tweets, by exploiting Wikipedia to create a location specific knowledgebase

KNOWLEDGE-BASE ENABLED APPROACH

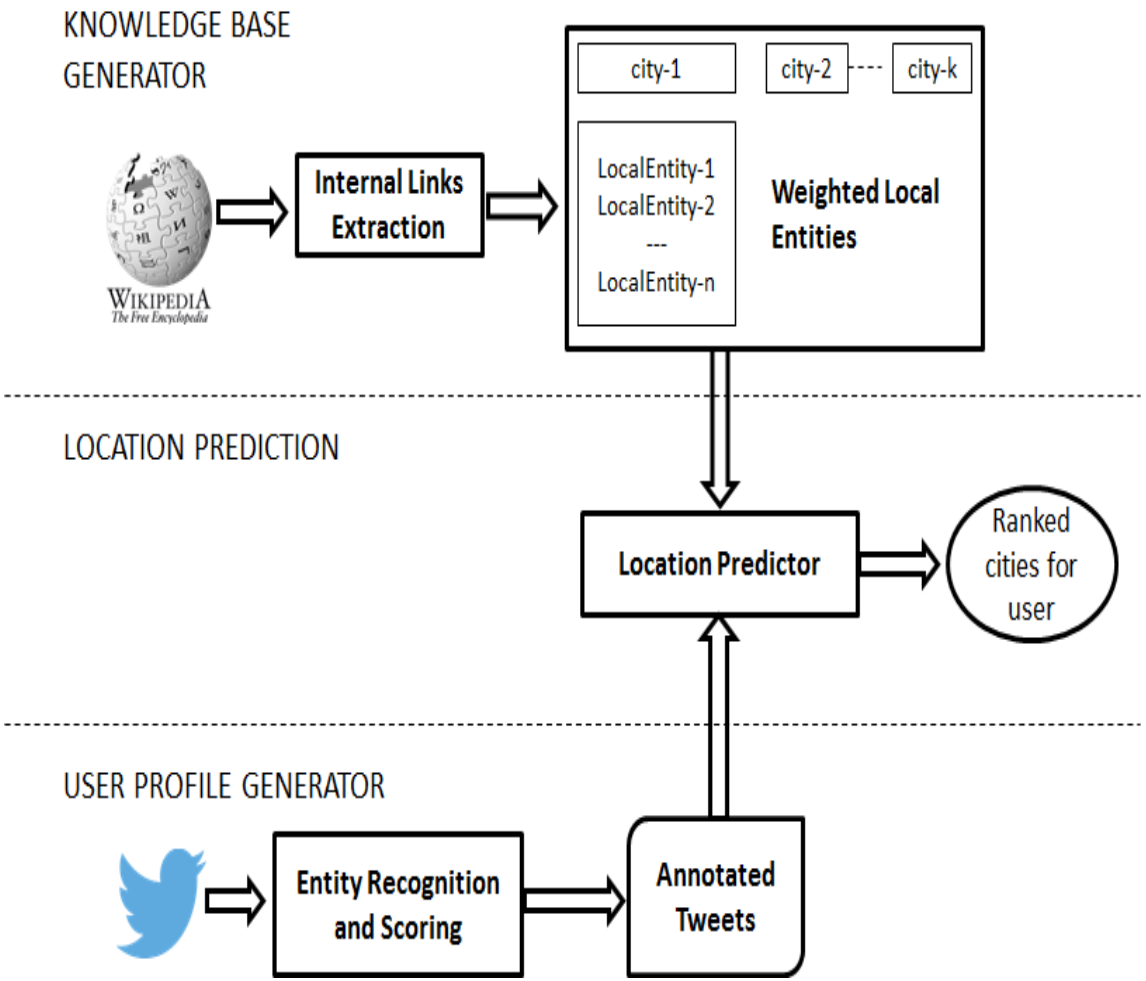


KNOWLEDGE-BASE ENABLED APPROACH



San Francisco:
Golden Gate Bridge,
San Francisco 49ers,
San Francisco Chronicle …

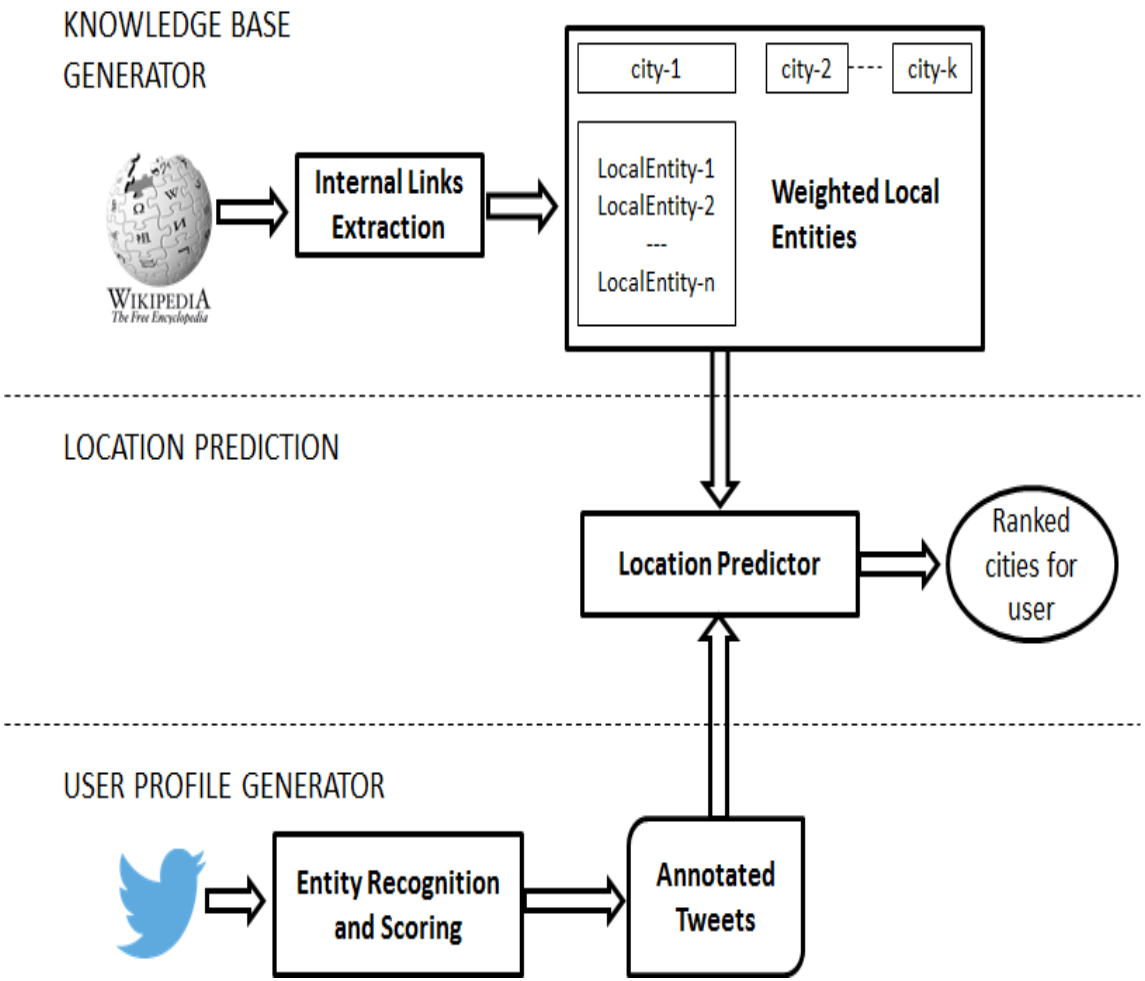
KNOWLEDGE-BASE ENABLED APPROACH



San Francisco:
Golden Gate Bridge,
San Francisco 49ers,
San Francisco Chronicle …

Entity	Count
Golden Gate Bridge	4
San Francisco 49ers	2
San Francisco Chronicle	1

KNOWLEDGE-BASE ENABLED APPROACH

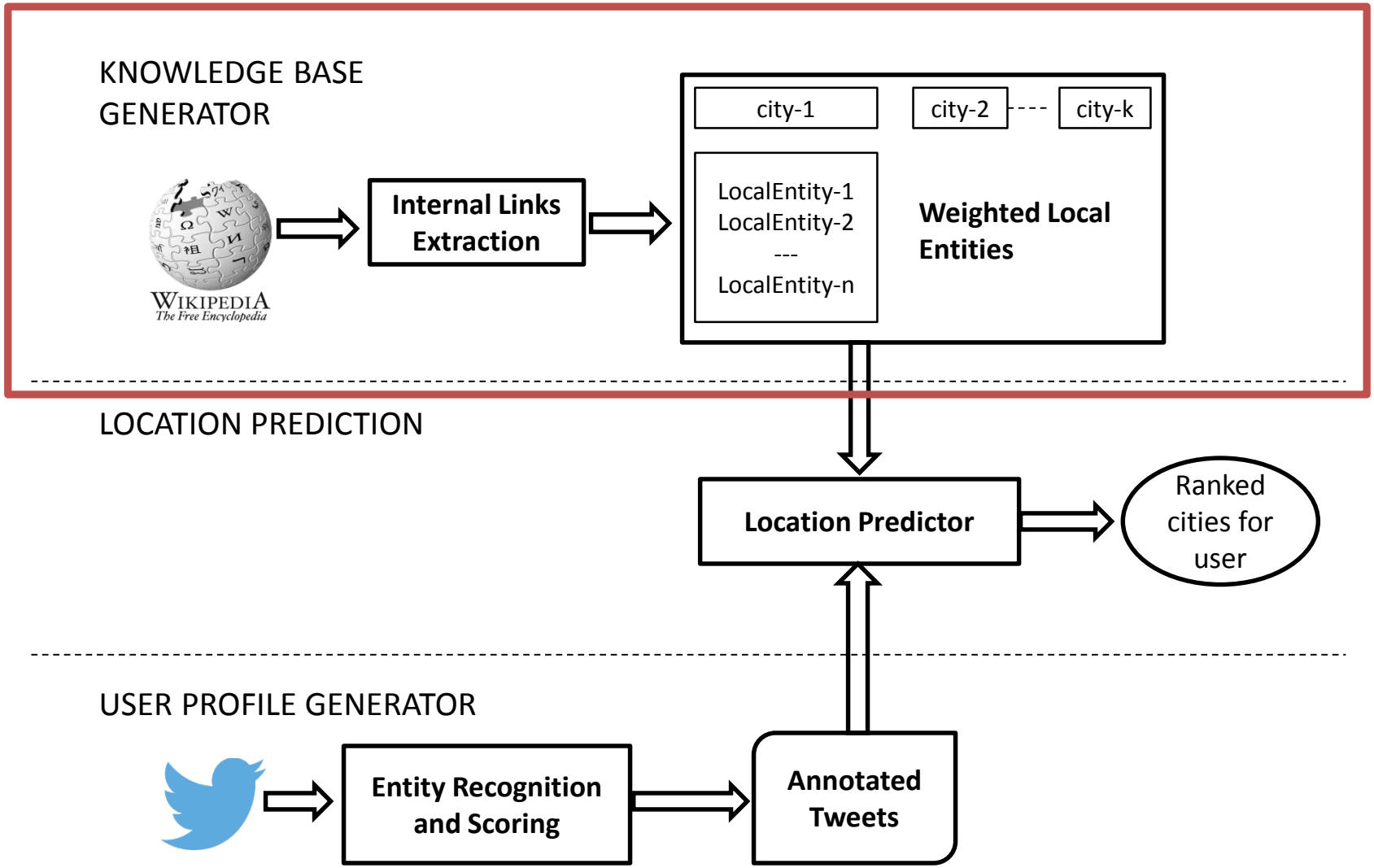


San Francisco:
Golden Gate Bridge,
San Francisco 49ers,
San Francisco Chronicle …

Top-k predictions:
San Francisco
Oakland
Palo Alto

Entity	Count
Golden Gate Bridge	4
San Francisco 49ers	2
San Francisco Chronicle	1

KNOWLEDGE-BASE ENABLED APPROACH



LOCAL ENTITIES

LOCAL ENTITIES

SAN FRANCISCO

NEW YORK CITY

HOUSTON

METRORail

WIKIPEDIA

WIKIPEDIA

- Collaborative encyclopedia

WIKIPEDIA

- Collaborative encyclopedia
- As of 2014, English Wikipedia has 4.6 million articles, 18 billion pages views and 500 million unique visitors per month.

WIKIPEDIA

- Collaborative encyclopedia
- As of 2014, English Wikipedia has 4.6 million articles, 18 billion pages views and 500 million unique visitors per month.
- **Category Structure**

WIKIPEDIA

- Collaborative encyclopedia
- As of 2014, English Wikipedia has 4.6 million articles, 18 billion pages views and 500 million unique visitors per month.
- **Category Structure**
  - Used for document clustering, tweet classification, personalization systems etc.

WIKIPEDIA

- Collaborative encyclopedia
- As of 2014, English Wikipedia has 4.6 million articles, 18 billion pages views and 500 million unique visitors per month.
- **Category Structure**
  - Used for document clustering, tweet classification, personalization systems etc.
- **Link Structure**

WIKIPEDIA

- Collaborative encyclopedia
- As of 2014, English Wikipedia has 4.6 million articles, 18 billion pages views and 500 million unique visitors per month.
- **Category Structure**
  - Used for document clustering, tweet classification, personalization systems etc.
- **Link Structure**
  - Used for word sense disambiguation, semantic relatedness between terms etc.

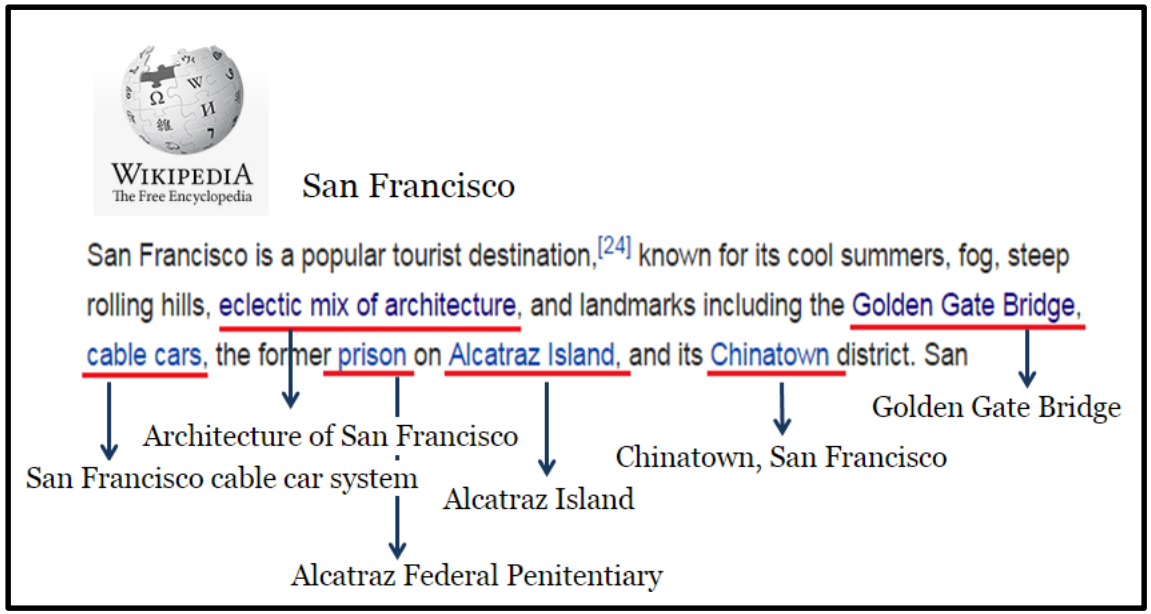
LOCAL ENTITIES

LOCAL ENTITIES

- We consider the internal links of location pages as *Local Entities* of the city

LOCAL ENTITIES

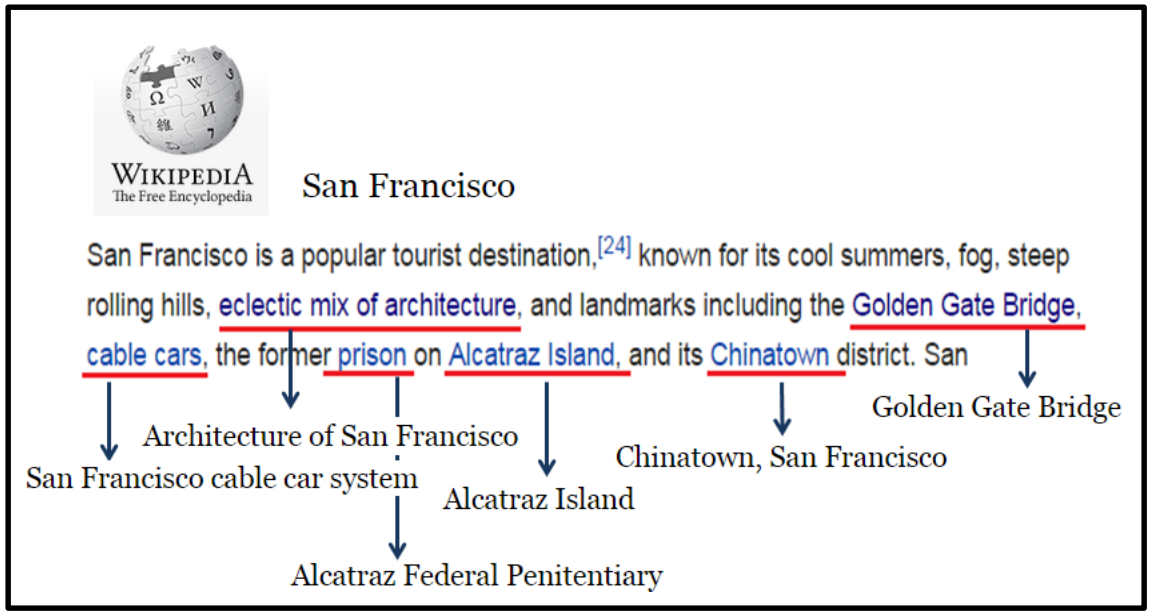
- We consider the internal links of location pages as *Local Entities* of the city



Local Entities of San Francisco

LOCAL ENTITIES

- We consider the internal links of location pages as *Local Entities* of the city



Local Entities of San Francisco

- While a city does not contain link to itself, we use the city as a local entity

ARE ALL ENTITIES EQUALLY LOCAL?

San Francisco

Media [edit]

Main article: Media in San Francisco

The major daily newspaper in San Francisco is the *San Francisco Chronicle*, which is currently Northern California's most widely circulated newspaper.[242] The Chronicle is most famous for a former columnist, the late Herb Caen, whose daily musings attracted critical acclaim and represented the "voice of San Francisco". The *San Francisco Examiner*, once the cornerstone of William Randolph Hearst's media empire and the home of Ambrose Bierce, declined in circulation over the years and now takes the form of a free daily tabloid, under new ownership.[243][244] *Sing Tao Daily* claims to be the largest of several Chinese language dailies that serve the Bay Area.[245] Alternative weekly newspapers include the *San Francisco Bay Guardian* and *SF Weekly*. *San Francisco Magazine* and *7x7* are major glossy magazines about San Francisco. The national newsmagazine *Mother Jones* is also based in San Francisco.

The San Francisco Bay Area is the sixth-largest TV market[246] and the fourth-largest radio market[247] in the U.S. The city's oldest radio station, KCBS (AM), began as an experimental station in San Jose in 1909, before the beginning of commercial broadcasting. KALW was the city's first FM radio station when it signed on the air in 1941. The city's first television station was KPIX, which began broadcasting in 1948.

All major U.S. television networks have affiliates serving the region, with most of them based in the city. CNN, MSNBC, BBC, Al Jazeera America, Russia Today, and CCTV America also have regional news bureaus in San Francisco. Bloomberg West was launched in 2011 from a studio on the Embarcadero. ESPN uses the local ABC studio for their broadcasting. The regional sports network, Comcast SportsNet Bay Area and it's sister station Comcast SportsNet California, are both located in San Francisco. The Pac-12 Network is also based in San Francisco.

ARE ALL ENTITIES EQUALLY LOCAL?

San Francisco

Media [edit]

Main article: Media in San Francisco

San Francisco Chronicle

The major daily newspaper in San Francisco is the San Francisco Chronicle, which is currently Northern California's most widely circulated newspaper.[242] The Chronicle is most famous for a former columnist, the late Herb Caen, whose daily musings attracted critical acclaim and represented the "voice of San Francisco". The *San Francisco Examiner*, once the cornerstone of William Randolph Hearst's media empire and the home of Ambrose Bierce, declined in circulation over the years and now takes the form of a free daily tabloid, under new ownership.[243][244] *Sing Tao Daily* claims to be the largest of several Chinese language dailies that serve the Bay Area.[245] Alternative weekly newspapers are the *San Francisco Bay Guardian* and *SF Weekly*. *San Francisco Magazine* and *7x7* are major glossy magazines about San Francisco. The national newsmagazine *Mother Jones* is also based in San Francisco.

SF Weekly **San Francisco Examiner**

The San Francisco Bay Area is the sixth-largest TV market[246] and the fourth-largest radio market[247] in the U.S. The city's oldest radio station, KCBS (AM), began as an experimental station in San Jose in 1909, before the beginning of commercial broadcasting. KALW was the city's first FM radio station when it signed on the air in 1941. The city's first television station was KPIX, which began broadcasting in 1948.

CNN **BBC**

All major U.S. television networks have affiliates serving the region, with most of them based in the city. CNN, MSNBC, BBC, Al Jazeera America, Fuji Sky Tokyo and CCTV America also have regional news bureaus in San Francisco. Bloomberg West was launched in 2011 from a studio on the Embarcadero. ESPN uses the local ABC studio for their broadcasting. The regional sports network, Comcast SportsNet Bay Area and it's sister station Comcast SportsNet California, are both located in San Francisco. The Pac-12 Network is also based in San Francisco.

Al Jazeera America

LOCALNESS MEASURE OF ENTITIES

$$PMI(le, c) = log_2 \frac{P(le,c)}{P(le).P(c)}$$

LOCALNESS MEASURE OF ENTITIES

Association-based Measure

- **Pointwise Mutual Information** – standard measure of association between two variables

- Assumption is that higher is the localness of an entity with respect to the city, higher will be the statistical dependence between them

- Computed as:

$$PMI(le, c) = log_2 \frac{P(le,c)}{P(le).P(c)}$$

where *le* is the local entity, *c* is the city, *P(le,c)* is the joint probability of occurrence of the city and the local entity in the Wikipedia dump, *P(e)* and *P(c)* are the individual probability of occurrence of the local entity and city respectively.

LOCALNESS MEASURE OF ENTITIES

Graph-based Measure

LOCALNESS MEASURE OF ENTITIES

Graph-based Measure

Boston

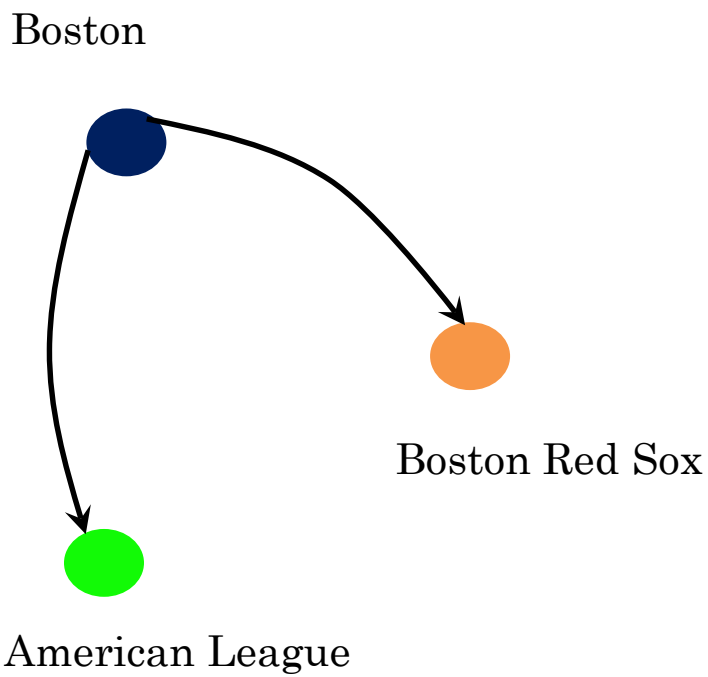
The Boston Red Sox, a founding member of the American League of Major League Baseball in 1901..

LOCALNESS MEASURE OF ENTITIES

Graph-based Measure

Boston

The Boston Red Sox, a founding member of the American League of Major League Baseball in 1901..



LOCALNESS MEASURE OF ENTITIES

Graph-based Measure

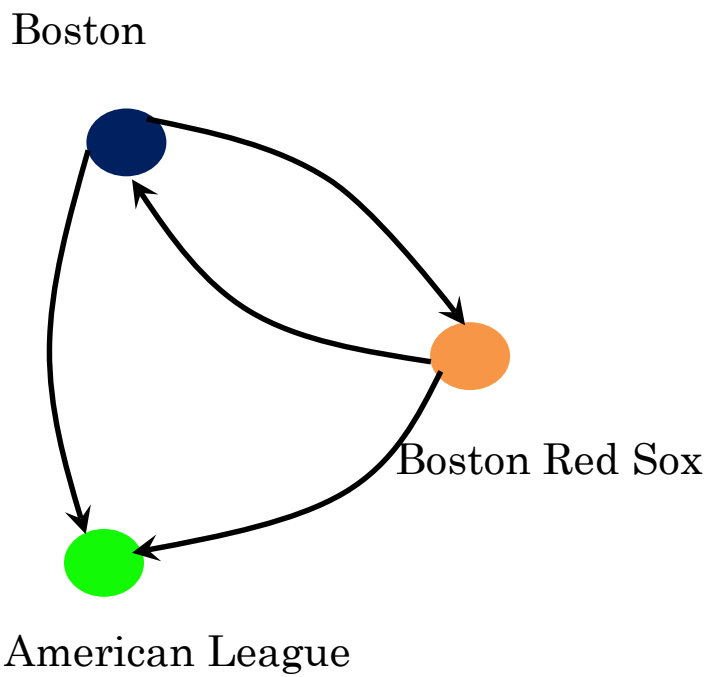
Boston

The Boston Red Sox, a founding member of the American League of Major League Baseball in 1901..

Boston Red Sox

The Boston Red Sox are an American professional baseball team based in Boston, Massachusetts ..

They are members of American League (AL)



LOCALNESS MEASURE OF ENTITIES

Graph-based Measure

- **Betweenness Centrality (BC)** – Measures the importance of a node relative to the rest of the nodes in the graph

- A high BC score of a vertex in a graph indicates that it lies on considerable fraction of shortest path connecting others

- Computed as:

$$C_B(le, c) = \sum_{le_i \neq le \neq le_j} \frac{\sigma_{le_i le_j}(le)}{\sigma_{le_i le_j}}$$

where le_i, le_j, le are local entities of c, $\sigma_{le_i le_j}$ represents the total number of shortest paths from le_i *to* le_j

LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure

LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure

San Francisco

Golden Gate Bridge

LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure

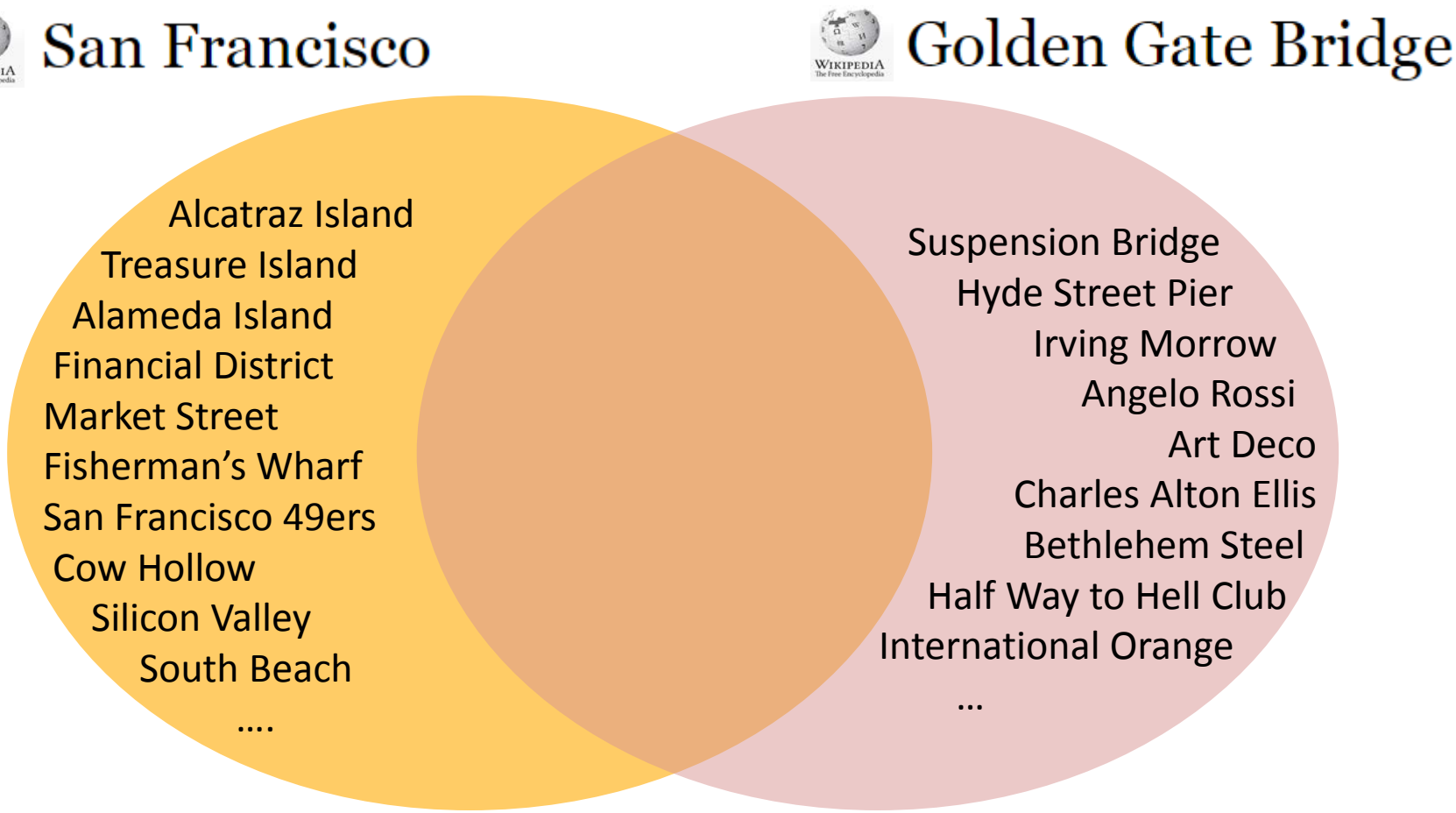
San Francisco

Golden Gate Bridge

Alcatraz Island
Treasure Island
Alameda Island
Financial District
Market Street
Fisherman's Wharf
San Francisco 49ers
Cow Hollow
Silicon Valley
South Beach
....

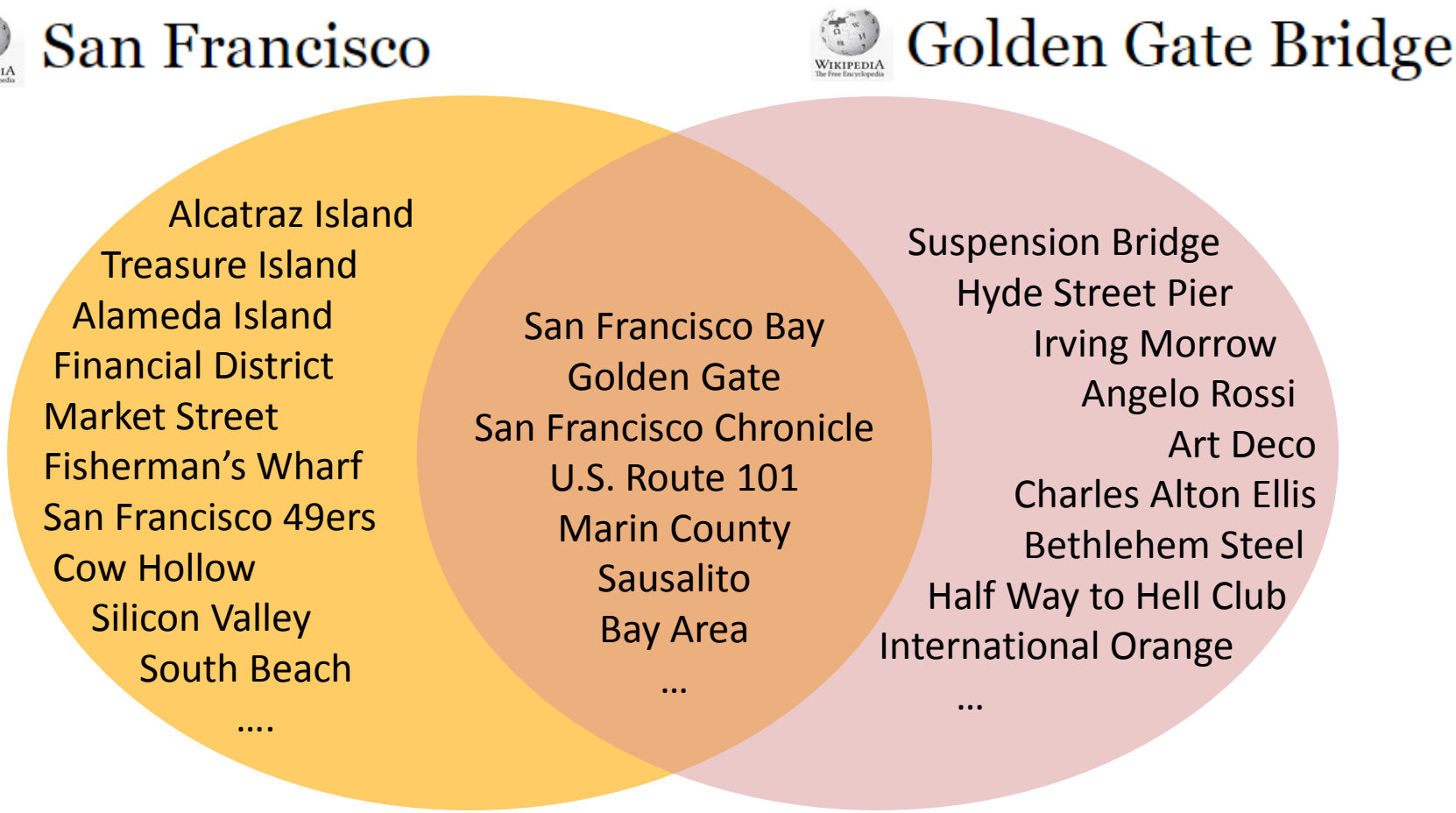
LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure



LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure



LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure

$$jaccard(le, c) = \frac{|IL(c) \cap IL(le)|}{|IL(c) \cup IL(le)|}$$

LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure

- Measures the relatedness between concepts with the intuition that related concepts are connected to similar entities

$$jaccard(le, c) = \frac{|IL(c) \cap IL(le)|}{|IL(c) \cup IL(le)|}$$

LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure

- Measures the relatedness between concepts with the intuition that related concepts are connected to similar entities

- **Jaccard Index:** Overlap between two sets

$$jaccard(le, c) = \frac{|IL(c) \cap IL(le)|}{|IL(c) \cup IL(le)|}$$

LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure

- Measures the relatedness between concepts with the intuition that related concepts are connected to similar entities

- **Jaccard Index:** Overlap between two sets

$$jaccard(le, c) = \frac{|IL(c) \cap IL(le)|}{|IL(c) \cup IL(le)|}$$

LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure

- Measures the relatedness between concepts with the intuition that related concepts are connected to similar entities

- **Jaccard Index:** Overlap between two sets

$$jaccard(le, c) = \frac{|IL(c) \cap IL(le)|}{|IL(c) \cup IL(le)|}$$

Where *IL(c)* and *IL(e)* and are the internal links found in the Wikipedia page of the city *c* and the local entity *le*.

LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure

$$i(le, c) = \frac{|IL(c) \cap IL(le)|}{|IL(c) \cap IL(le)| + \alpha|IL(c) - IL(le)| + \beta|IL(le) - IL(c)|}$$

$L(c)$ and $IL(le)$ are the internal links found in the Wikipedia page of the city c and the local entity le

- = 0 and β = 1

LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure

- **Tversky Index**: Asymmetric similarity measure between two sets

$$i(le, c) = \frac{|IL(c) \cap IL(le)|}{|IL(c) \cap IL(le)| + \alpha|IL(c) - IL(le)| + \beta|IL(le) - IL(c)|}$$

$L(c)$ and $IL(le)$ are the internal links found in the Wikipedia page of the city c and the local entity le

- = 0 and β = 1

LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure

- $cc\ le, c = |IL\ c \cap IL\ le|\ \ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le + \beta |IL\ le - IL\ c|\ |IILL\ c\ cc\ c \cap IILL\ le\ llee\ le|\ |IL\ c \cap IL\ le|\ \ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le + \beta |IL\ le - IL\ c|\ \ IL\ c \cap IL\ le\ \ IILL\ c\ cc\ c \cap IILL\ le\ llee\ le\ \ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le\ \ IILL\ c\ cc\ c - IILL\ le\ llee\ le\ \ IL\ c - IL\ le + \beta |IILL\ le\ llee\ le - IILL\ c\ cc\ c|\ |IL\ c \cap IL\ le|\ \ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le + \beta |IL\ le - IL\ c|$
- **Tversky Index**: Asymmetric similarity measure between two sets

$$ti\ le, c = \frac{|IL(c) \cap IL(le)|}{|IL(c) \cap IL(le)| + \alpha |IL(c) - IL(le)| + \beta |IL(le) - IL(c)|}$$

$L(c)$ and $IL(le)$ are the internal links found in the Wikipedia page of the city c and the local entity le

LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure

- cc) and $IILL\ le\ llee\ le$ are the internal links found in the Wikipedia page of the city cc and the local entity $llee$
- $cc\ le, c = |IL\ c \cap IL\ le|\ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le + \beta |IL\ le - IL\ c|\ |IILL\ c\ cc\ c \cap IILL\ le\ llee\ le|\ |IL\ c \cap IL\ le|\ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le + \beta |IL\ le - IL\ c|\ IL\ c \cap IL\ le\ IILL\ c\ cc\ c \cap IILL\ l\ e\ llee\ le\ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le\ IILL\ c\ cc\ c - IILL\ le\ llee\ le\ IL\ c - IL\ le + \beta |IILL\ le\ llee\ le - IILL\ c\ cc\ c|\ |IL\ c \cap IL\ le|\ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le + \beta |IL\ le - IL\ c|$
- **Tversky Index**: Asymmetric similarity measure between two sets

$$ti\ le, c = \frac{|IL(c) \cap IL(le)|}{|IL(c) \cap IL(le)| + \alpha |IL(c) - IL(le)| + \beta |IL(le) - IL(c)|}$$

Where $IL(c)$ and $IL(le)$ are the internal links found in the Wikipedia page of the city c and the local entity le

LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure

- = 0 and β = 1
- cc) and $IILL\ le\ llee\ le$ are the internal links found in the Wikipedia page of the city cc and the local entity $llee$
- $cc\ le, c = |IL\ c \cap IL\ le|\ \ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le + \beta|IL\ le - IL\ c|\ |IILL\ c\ cc\ c \cap IILL\ le\ llee\ le|\ |IL\ c \cap IL\ le|\ \ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le + \beta|IL\ le - IL\ c|\ \ IL\ c \cap IL\ le\ \ IILL\ c\ cc\ c \cap IILL\ l\ e\ llee\ le\ \ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le\ \ IILL\ c\ cc\ c - IILL\ le\ llee\ le\ \ IL\ c - IL\ le + \beta|IILL\ le\ llee\ le - IILL\ c\ cc\ c|\ |IL\ c \cap IL\ le|\ \ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le + \beta|IL\ le - IL\ c|$
- **Tversky Index**: Asymmetric similarity measure between two sets

$$ti\ le, c = \frac{|IL(c) \cap IL(le)|}{|IL(c) \cap IL(le)| + \alpha|IL(c) - IL(le)| + \beta|IL(le) - IL(c)|}$$

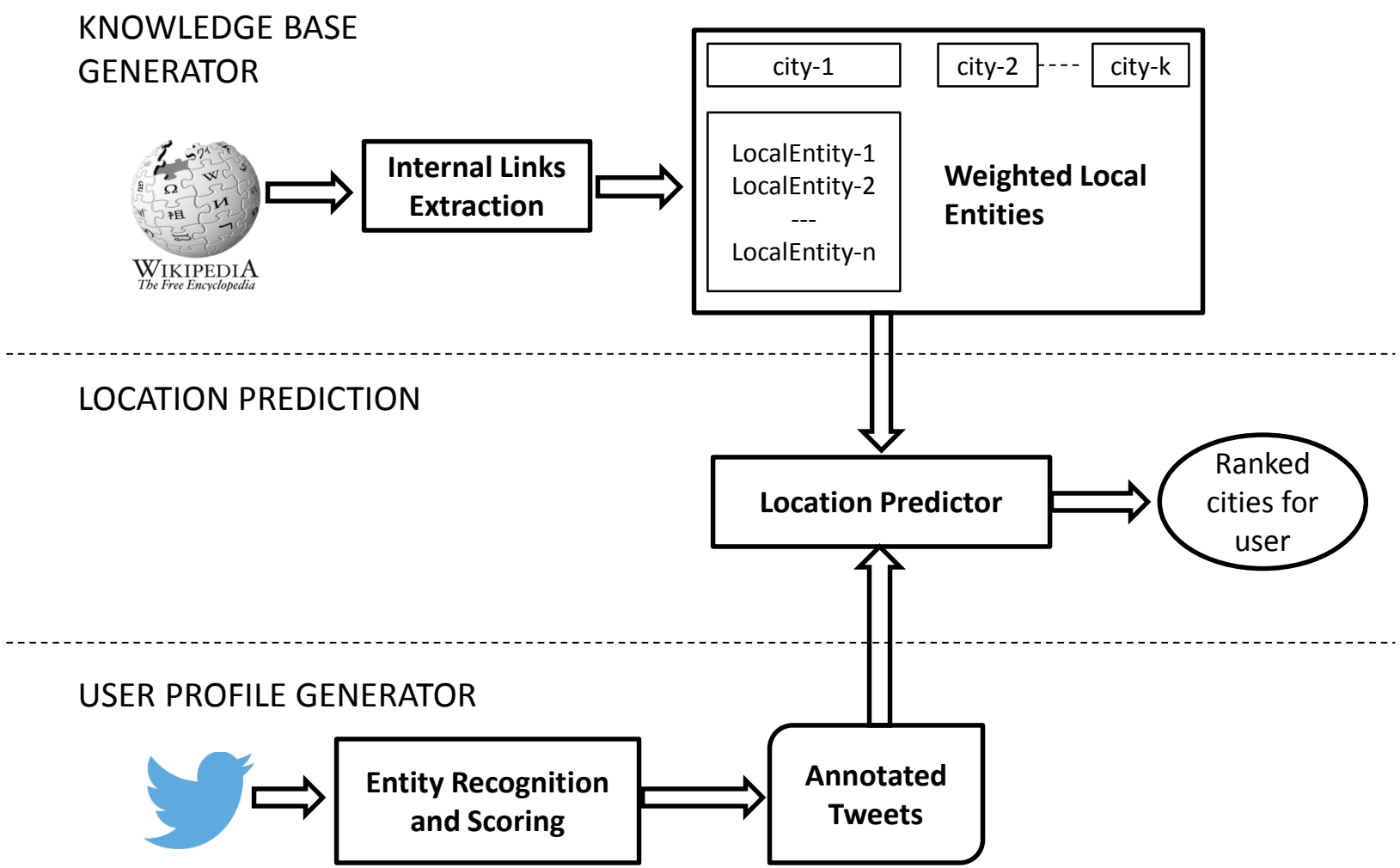
LOCALNESS MEASURE OF ENTITIES

Semantic Overlap Measure

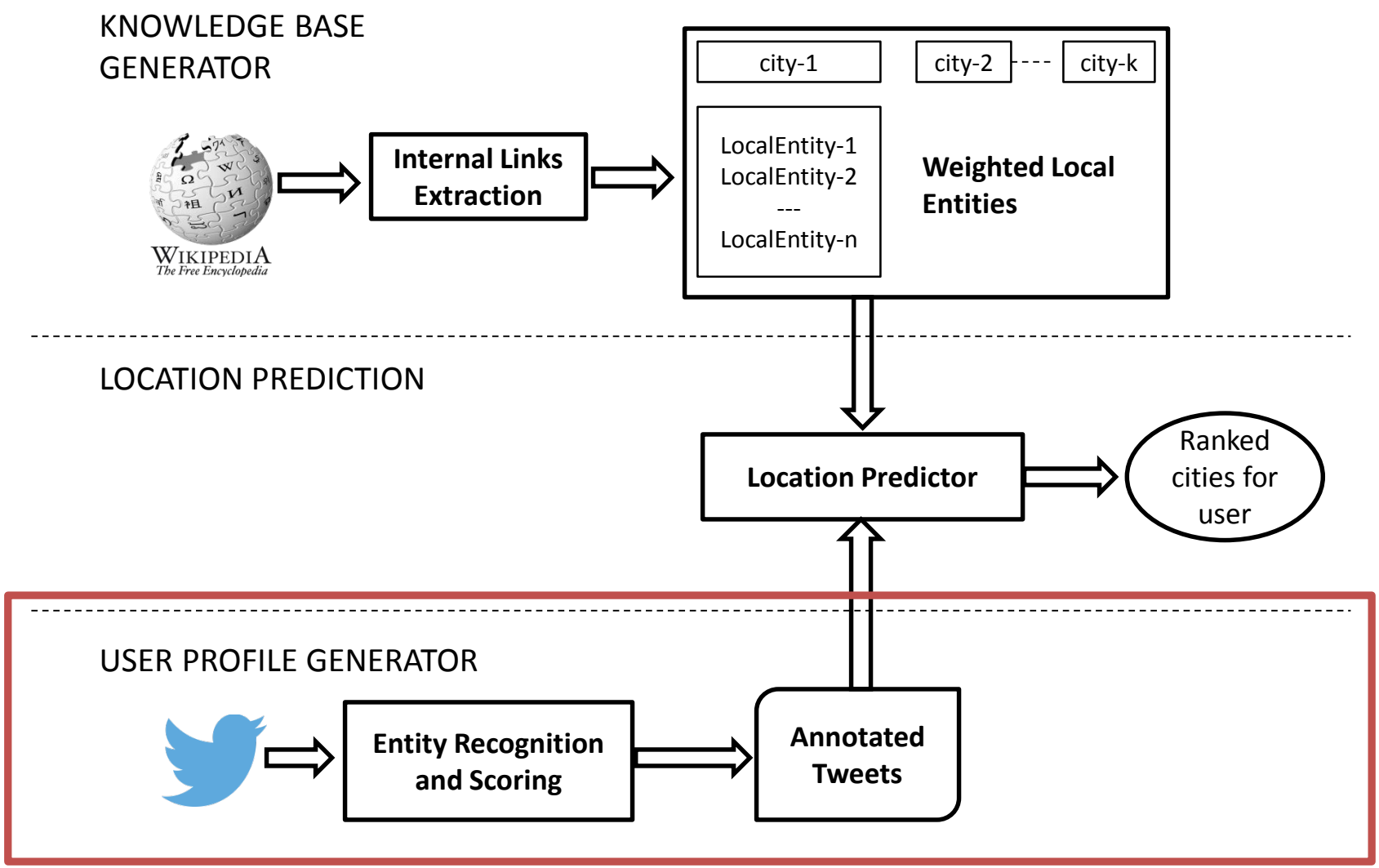
- = 0 and β = 1
- cc) and $IILL\ le\ llee\ le$ are the internal links found in the Wikipedia page of the city cc and the local entity $llee$
- $cc\ le, c = |IL\ c \cap IL\ le|\ \ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le + \beta|IL\ le - IL\ c|\ |IILL\ c\ cc\ c \cap IILL\ le\ llee\ le|\ |IL\ c \cap IL\ le|\ \ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le + \beta|IL\ le - IL\ c|\ \ IL\ c \cap IL\ le\ \ IILL\ c\ cc\ c \cap IILL\ l\ e\ llee\ le\ \ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le\ \ IILL\ c\ cc\ c - IILL\ le\ llee\ le\ \ IL\ c - IL\ le + \beta|IILL\ le\ llee\ le - IILL\ c\ cc\ c|\ |IL\ c \cap IL\ le|\ \ IL\ c \cap IL\ le + \alpha\ IL\ c - IL\ le + \beta|IL\ le - IL\ c|$
- **Tversky Index**: Asymmetric similarity measure between two sets

$$ti\ le, c = \frac{|IL(c) \cap IL(le)|}{|IL(c) \cap IL(le)| + \alpha|IL(c) - IL(le)| + \beta|IL(le) - IL(c)|}$$

KNOWLEDGE-BASE ENABLED APPROACH



KNOWLEDGE-BASE ENABLED APPROACH



CREATION OF USER PROFILE

CREATION OF USER PROFILE

Just drove around Golden Gate Park trying to get in.

CREATION OF USER PROFILE

Just drove around Golden Gate Park trying to get in.

Step 1: Entity Linking

CREATION OF USER PROFILE

Just drove around Golden Gate Park trying to get in.

Step 1: Entity Linking

Golden Gate Park

CREATION OF USER PROFILE

Just drove around Golden Gate Park trying to get in.

Step 1: Entity Linking

Golden Gate Park

CREATION OF USER PROFILE

Just drove around Golden Gate Park trying to get in.

Step 1: Entity Linking

Wikipedia
The Free Encyclopedia

Golden Gate Park

We use Zemanta for Entity Linking

CREATION OF USER PROFILE

Just drove around Golden Gate Park trying to get in.

Step 1: Entity Linking

Golden Gate Park

We use Zemanta for Entity Linking

CREATION OF USER PROFILE

Just drove around Golden Gate Park trying to get in.

Step 1: Entity Linking

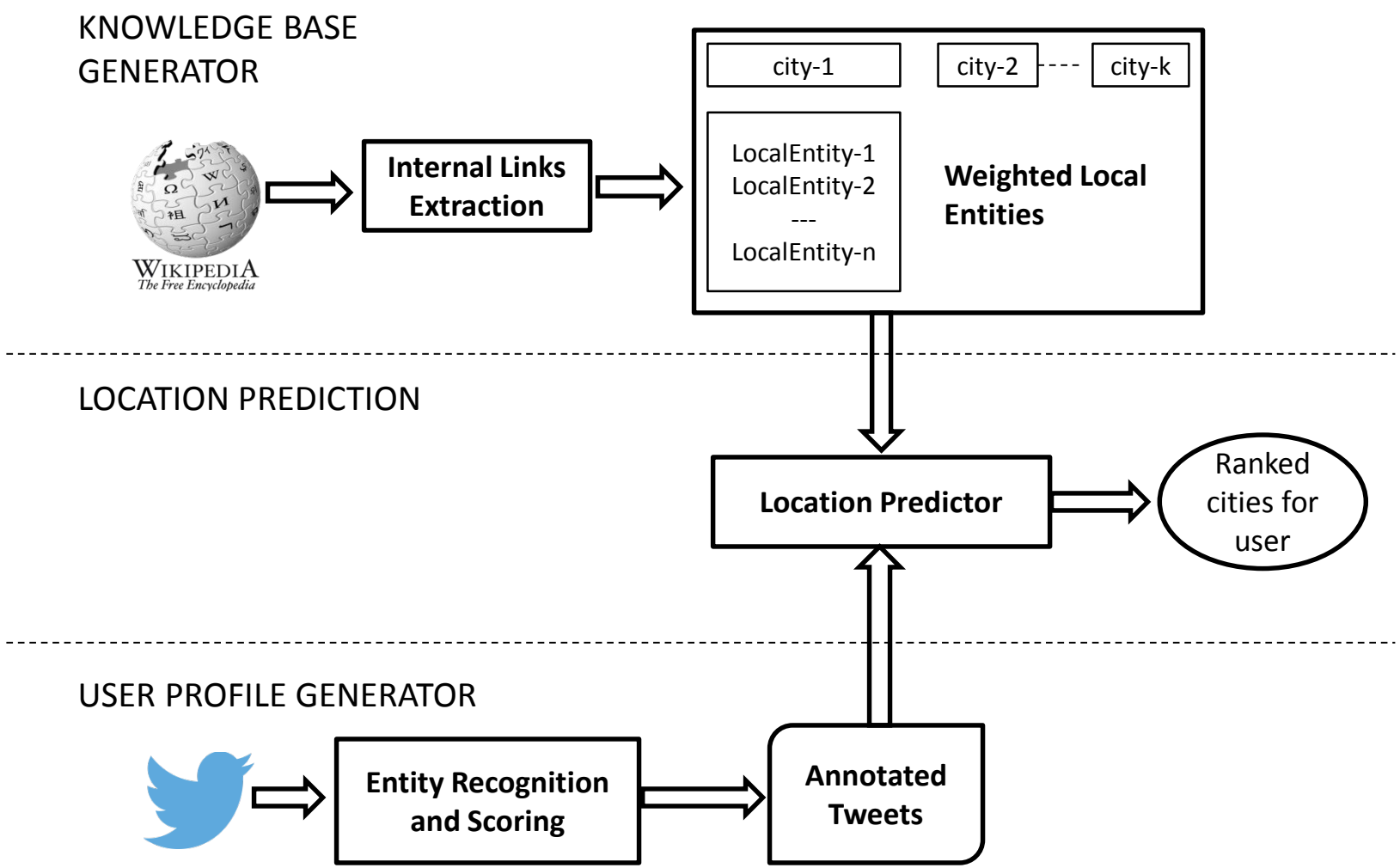
Golden Gate Park

We use Zemanta for Entity Linking

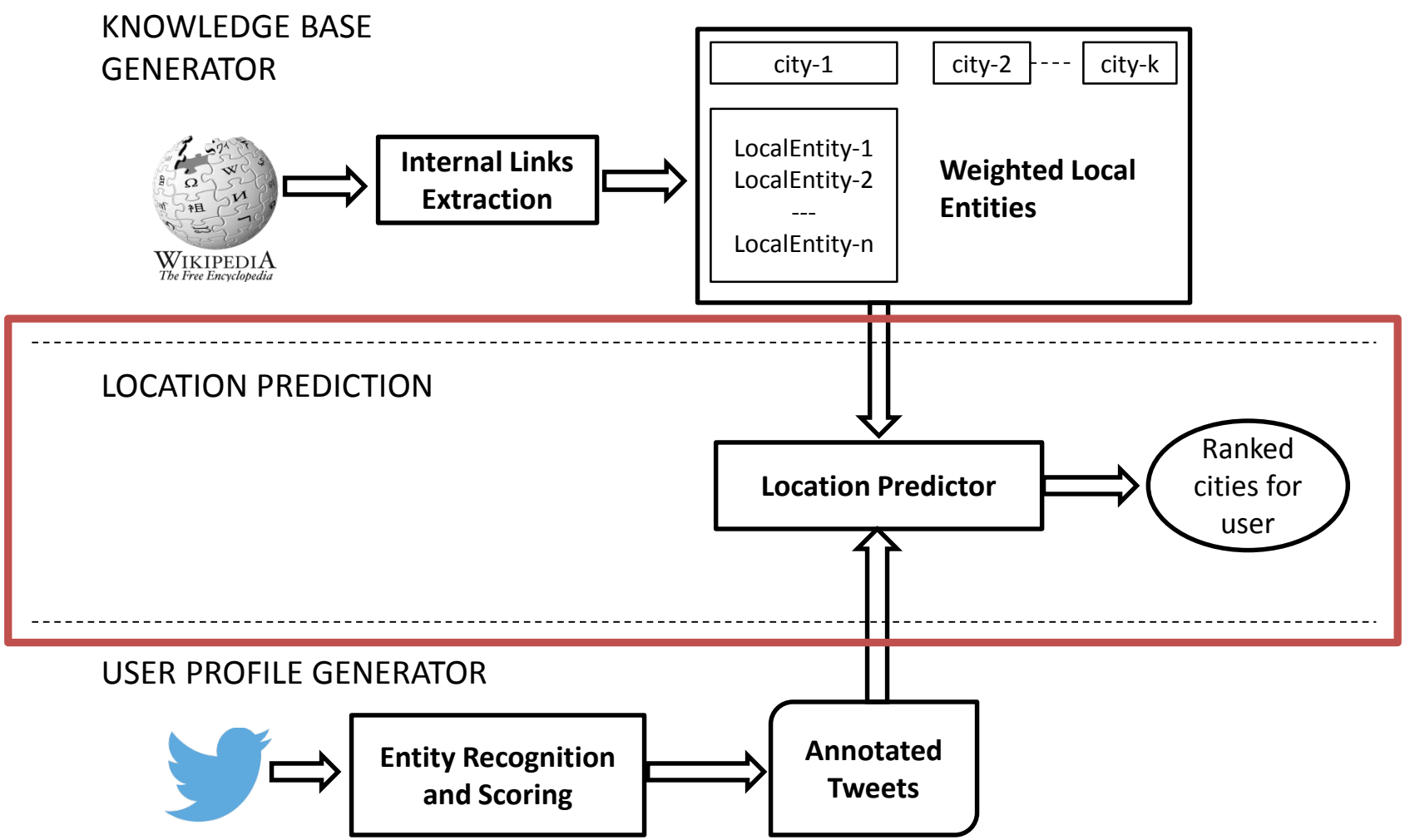
Step 2: Entity Scoring

Entity	Count
Golden Gate Bridge	4
San Francisco 49ers	2
San Francisco Chronicle	1

KNOWLEDGE-BASE ENABLED APPROACH



KNOWLEDGE-BASE ENABLED APPROACH



LOCATION PREDICTION

LOCATION PREDICTION

- Compute an aggregate score for each city whose local entities are found in a user's tweets

LOCATION PREDICTION

- Compute an aggregate score for each city whose local entities are found in a user's tweets

LOCATION PREDICTION

- $cc\ u,c = \ e \in LEcu\ \ locl\ c, e \times se\ ee \in LLEEccuu\ e \in LEcu\ \ locl\ c, e \times se\ \ e \in LEcu\ \ locl\ c, e \times se\ lloo ccll\ c, e\ cc, ee\ c, e \times ssee\ e \in LEcu\ \ locl\ c, e \times se$
- Compute an aggregate score for each city whose local entities are found in a user's tweets

$$locScore\ u, c\ \ = \sum_{e \in LEcu} locl(c, e) \times se$$

LOCATION PREDICTION

- $cc\ u,c = \ e \in LEcu\ \ locl\ c, e \times se\ ee \in LLEEccuu\ e \in LEcu\ \ locl\ c, e \times se\ \ e \in LEcu\ \ locl\ c, e \times se\ llooccll\ c, e\ cc, ee\ c, e \times ssee\ e \in LEcu\ \ locl\ c, e \times se$
- Compute an aggregate score for each city whose local entities are found in a user's tweets

LOCATION PREDICTION

- *ee,cc) is the localness measure of the entity ee with respect to city cc*
- *cc u,c = e∈LEcu locl c, e ×se ee∈LLEEccuu e∈LEcu locl c, e ×se e∈ LEcu locl c, e ×se llooccll c, e cc, ee c, e ×ssee e∈LEcu locl c, e ×se*
- Compute an aggregate score for each city whose local entities are found in a user's tweets

where LE_{cu} is the set of local entities of c found in the profile of user u , $locl(e, c)$ is the localness measure of the entity e with respect to city c

LOCATION PREDICTION

- *cc u,c in descending order to predict the top-k locations of a user*
- *ee,cc) is the localness measure of the entity ee with respect to city cc*
- *cc u,c = e∈LEcu locl c, e ×se ee∈LLEEccuu e∈LEcu locl c, e ×se e∈ LEcu locl c, e ×se llooccll c, e cc, ee c, e ×ssee e∈LEcu locl c, e ×se*
- Compute an aggregate score for each city whose local entities are found in a user's tweets

  where LE_{cu} is the set of local entities of c found in the profile of user u , $locl(e, c)$ is the localness measure of the entity e with respect to city c
- Rank $locScore\ u, c$ in descending order to predict the top-k locations of a user

LOCATION PREDICTION

LOCATION PREDICTION

User Profile

San Francisco International Airport (6), San Francisco (4), Nob Hill (3), San Francisco Museum of Modern Art (1), Beach Blanket Babylon (2), San Francisco Municipal Railway (4), Golden Gate Park (1), San Francisco Bay Area (1), SF Weekly (1), Fox Oakland Theatre (2), Berkley (1), Green Day (1), Oakland (9), San Francisco Bay Area (1), The White Stripes (1), Detroit Metropolitan Wayne County Airport (1), Detroit Historical Museum (1), Detroit Red Wings (4), General Motors (1), Palo Alto (6), SAP AG (8), Facebook (3), PARC (company) (2), Dell (1), Google (1), …

LOCATION PREDICTION

User Profile

San Francisco International Airport (6), San Francisco (4), Nob Hill (3), San Francisco Museum of Modern Art (1), Beach Blanket Babylon (2), San Francisco Municipal Railway (4), Golden Gate Park (1), San Francisco Bay Area (1), SF Weekly (1), Fox Oakland Theatre (2), Berkley (1), Green Day (1), Oakland (9), San Francisco Bay Area (1), The White Stripes (1), Detroit Metropolitan Wayne County Airport (1), Detroit Historical Museum (1), Detroit Red Wings (4), General Motors (1), Palo Alto (6), SAP AG (8), Facebook (3), PARC (company) (2), Dell (1), Google (1), …

Knowledgebase

San Francisco

Nob Hill	0.48214
SF Weekly	0.1875
Golden Gate Park	0.16783
San Francisco International Airport	0.06818
…	

Oakland, CA

Fox Oakland Theatre	0.09375
SF Bay Area	0.12972
Green Day	0.02066
…	

Detroit, MI

Detroit Historical Museum	0.4838
General Motors	0.05538
Detroit Red Wings	0.0232
…	

Palo Alto, CA

PARC (company)	0.03726
Google	0.04678
Facebook	0.05810

LOCATION PREDICTION

User Profile

San Francisco International Airport (6), San Francisco (4), Nob Hill (3), San Francisco Museum of Modern Art (1), Beach Blanket Babylon (2), San Francisco Municipal Railway (4), Golden Gate Park (1), San Francisco Bay Area (1), SF Weekly (1)

Fox Oakland Theatre (2), Berkley (1), Green Day (1), Oakland (9), San Francisco Bay Area (1)

The White Stripes (1), Detroit Metropolitan Wayne County Airport (1), Detroit Historical Museum (1), Detroit Red Wings (4), General Motors (1)

Palo Alto (6), SAP AG (8), Facebook (3), PARC (company) (2), Dell (1), Google (1)

Knowledgebase

San Francisco

Nob Hill	0.48214
SF Weekly	0.1875
Golden Gate Park	0.16783
San Francisco International Airport	0.06818
...	

Oakland, CA

Fox Oakland Theatre	0.09375
SF Bay Area	0.12972
Green Day	0.02066
...	

Detroit, MI

Detroit Historical Museum	0.4838
General Motors	0.05538
Detroit Red Wings	0.0232
...	

Palo Alto, CA

PARC (company)	0.03726
Google	0.04678
Facebook	0.05810

LOCATION PREDICTION

User Profile	Knowledgebase	Location Prediction
San Francisco International Airport (6), San Francisco (4), Nob Hill (3), San Francisco Museum of Modern Art (1), Beach Blanket Babylon (2), San Francisco Municipal Railway (4), Golden Gate Park (1), San Francisco Bay Area (1), SF Weekly (1)	**San Francisco** Nob Hill 0.48214 SF Weekly 0.1875 Golden Gate Park 0.16783 San Francisco International Airport 0.06818 ...	14.5531
Fox Oakland Theatre (2), Berkley (1), Green Day (1), Oakland (9), San Francisco Bay Area (1)	**Oakland, CA** Fox Oakland Theatre 0.09375 SF Bay Area 0.12972 Green Day 0.02066 ...	10.7584
The White Stripes (1), Detroit Metropolitan Wayne County Airport (1), Detroit Historical Museum (1), Detroit Red Wings (4), General Motors (1)	**Detroit, MI** Detroit Historical Museum 0.4838 General Motors 0.05538 Detroit Red Wings 0.0232 ...	8.0600
Palo Alto (6), SAP AG (8), Facebook (3), PARC (company) (2), Dell (1), Google (1)	**Palo Alto, CA** PARC (company) 0.03726 Google 0.04678 Facebook 0.05810	6.9175

IMPLEMENTATION

IMPLEMENTATION

Knowledge base

- All cities of United States with population > 5000 as published in census estimates of 2012
- 4,661 cities and 500714 local entities

IMPLEMENTATION

Knowledge base

- All cities of United States with population > 5000 as published in census estimates of 2012
- 4,661 cities and 500714 local entities

Baseline

- Considers all local entities to be equally local to the city
- Location prediction based only on frequency of entities

EVALUATION

Test Dataset

EVALUATION

Test Dataset

- Published by Cheng et al.
- Collected from September 2009 to January 2010.
- Contains 5119 active users from continental United States with approximately 1000 tweets per user.
- User's location listed in the form of latitude and longitude.

EVALUATION

Evaluation Metrics

EVALUATION

Evaluation Metrics

- Error Distance

  Distance between actual location of the user and the estimated location

- Average Error Distance

  Average of error distance of all users in the test dataset

EVALUATION

Evaluation Metrics

- Error Distance

  Distance between actual location of the user and the estimated location

- Average Error Distance

  Average of error distance of all users in the test dataset

- Accuracy

  Percentage of users predicted within 100 miles of their actual location

EVALUATION

Location Prediction Results

Localness Measure	ACC (%)	AED (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	38.48	599.40	49.85	56.06	64.15
BC	47.91	478.14	57.39	62.18	66.98
Jaccard Index	53.21	433.62	67.41	73.56	78.84
Tversky Index	54.48	429.00	68.72	74.68	79.99

EVALUATION

Localness Measure	ACC (%)	AED (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	**38.48**	**599.40**	**49.85**	**56.06**	**64.15**
BC	47.91	478.14	57.39	62.18	66.98
Jaccard Index	53.21	433.62	67.41	73.56	78.84
Tversky Index	54.48	429.00	68.72	74.68	79.99

EVALUATION

Localness Measure	ACC (%)	AED (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	**38.48**	**599.40**	**49.85**	**56.06**	**64.15**
BC	47.91	478.14	57.39	62.18	66.98
Jaccard Index	53.21	433.62	67.41	73.56	78.84
Tversky Index	54.48	429.00	68.72	74.68	79.99

- PMI is not normalized hence sensitive to the count of the occurrences of local entities in the Wikipedia corpus

EVALUATION

Localness Measure	ACC (%)	AED (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	**38.48**	**599.40**	**49.85**	**56.06**	**64.15**
BC	47.91	478.14	57.39	62.18	66.98
Jaccard Index	53.21	433.62	67.41	73.56	78.84
Tversky Index	54.48	429.00	68.72	74.68	79.99

- PMI is not normalized hence sensitive to the count of the occurrences of local entities in the Wikipedia corpus
  - E.g. PMI of local entities of *Glenn Rock, New Jersey* is higher than those of *San Francisco*

EVALUATION

Localness Measure	ACC (%)	AED (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	38.48	599.40	49.85	56.06	64.15
BC	**47.91**	**478.14**	**57.39**	**62.18**	**66.98**
Jaccard Index	53.21	433.62	67.41	73.56	78.84
Tversky Index	54.48	429.00	68.72	74.68	79.99

EVALUATION

Localness Measure	ACC (%)	AED (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	38.48	599.40	49.85	56.06	64.15
BC	**47.91**	**478.14**	**57.39**	**62.18**	**66.98**
Jaccard Index	53.21	433.62	67.41	73.56	78.84
Tversky Index	54.48	429.00	68.72	74.68	79.99

- Does a good job of assigning low scores to common entities.
  - E.g. *community college, National Weather Service, start up company etc.*

EVALUATION

Localness Measure	ACC (%)	AED (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	38.48	599.40	49.85	56.06	64.15
BC	**47.91**	**478.14**	**57.39**	**62.18**	**66.98**
Jaccard Index	53.21	433.62	67.41	73.56	78.84
Tversky Index	54.48	429.00	68.72	74.68	79.99

- Does a good job of assigning low scores to common entities.
  - E.g. *community college, National Weather Service, start up company etc.*

- Fails for entities with some relevance to the city but no distinguishing factor
  - E.g. *IBM* with respect to *Endicott, New York*

LOCALNESS MEASURE OF ENTITIES

Endicott, New York

From Wikipedia, the free encyclopedia

Endicott is a village in Broome County, New York, United States. The population was 13,392 at the 2010 census.[1] It is part of the Binghamton Metropolitan Statistical Area. The village is named after Henry B. Endicott, a founding member of the Endicott Johnson Corporation shoe manufacturing founded the community as the "Home of the Square Deal".

The village of Endicott is in the town of Union and is west of the city of Binghamton. The community is served by the Greater Binghamton Airport/Edwin A. Link Field. It is part of the "Triple Cities", along with Binghamton and Johnson City.

Contents [hide]

1 History
1.1 The Endicott Johnson Corporation
1.2 IBM
2 Geography

IBM [edit]

Endicott is best known as the "Birthplace of IBM".[2][3] The Computing Tabulating Recording Company (CTR) was founded in Endicott on June 16, 1911, via the merger of the International Time Recording Company (ITR), Tabulating Machine Company, Computing Scale Company, and Bundy Time Recording. These companies used a technology invented by the genius Herman Hollerith whereby stiff paper cards with holes in a systematic pattern, called punched cards could be "read" by machines via electrical contact.

The Computing Tabulating Recording Company changed its name to International Business Machines Corporation (IBM) in 1924. The formation of what soon became IBM consolidated some of the major companies in the industrial time-keeping business, but its new chief executive, Thomas J. Watson, a businessman who turned out to be just as brilliant as George F. Johnson, realized that data processing had far greater potential than just workers punching a time clock. A great motivator of salesmen, Watson sent them

Endicott grew and flourished due to massive numbers of immigrants who came to the area to work for "EJ", predominantly from southern and eastern Europe. "Which way EJ?" was said to be what they asked immigration officials at Ellis Island in New York City, but it is far more likely that they ha memorized the addresses of relatives or friends living in Endicott. The company also maintained recruiting sites in Italy and the Balkans in the early part of the 20th century. Endicott-Johnson's employment in the region reached a peak of about 20,000 in the early 1920s.

In an innovative and far-sighted policy, George F. Johnson made sections of the company's land holdings outside the factory district available to workers to build homes on, with financing provided by the company, and title reverting to the worker when the loan was paid off. Along with extensive co (unheard of at the time and decades before government took over these responsibilities), this "Square Deal" of the early 20th century is commemorated by stone arches erected by the workers in 1920 across Route 17C (Main Street) at the entrances to Endicott and Johnson City.

Endicott-Johnson was hurt by the Depression of the 1930s, but since shoes were a necessity, did better than other manufacturing sectors of the economy. Orders for shoes from the military in World War II in the 1940s propelled employment over the peak attained in the early 1920s. Unfortunately George F. Johnson in 1948 couldn't cope with a more affluent era in the 1950s and 1960s when footwear became mainly a fashion business in the United States. More importantly, little if any money was invested in improving the original 1900 manufacturing technology, which meant that foreign co market share resulted in the closing and sale of the Endicott factories

IBM [edit]

Endicott is best known as the "Birthplace of IBM".[2][3] The Computing Tabulating Recording Company (CTR) was founded in Endicott on June 16, 1911, via the merger of the International Time Recording Company (ITR), Tabulating Machine Company, Computing Scale Company, and Bundy Time

LOCALNESS MEASURE OF ENTITIES

Endicott, New York

From Wikipedia, the free encyclopedia

Endicott is a village in Broome County, New York, United States. The population was 13,392 at the 2010 census.[1] It is part of the Binghamton Metropolitan Statistical Area. The village is named after Henry B. Endicott, a founding member of the Endicott Johnson Corporation shoe manufacturing founded the community as the "Home of the Square Deal".

The village of Endicott is in the town of Union and is west of the city of Binghamton. The community is served by the Greater Binghamton Airport/Edwin A. Link Field. It is part of the "Triple Cities", along with Binghamton and Johnson City.

Contents [hide]

1 History
1.1 The Endicott Johnson Corporation
1.2 IBM
2 Geography

IBM [edit]

Endicott is best known as the "Birthplace of IBM".[2][3] The Computing Tabulating Recording Company (CTR) was founded in Endicott on June 16, 1911, via the merger of the International Time Recording Company (ITR), Tabulating Machine Company, Computing Scale Company, and Bundy Time Recording. These companies used a technology invented by the genius Herman Hollerith whereby stiff paper cards with holes in a systematic pattern, called punched cards could be "read" by machines via electrical contact.

The Computing Tabulating Recording Company changed its name to International Business Machines Corporation (IBM) in 1924. The formation of what soon became IBM consolidated some of the major companies in the industrial time-keeping business, but its new chief executive, Thomas J. Watson, a businessman who turned out to be just as brilliant as George F. Johnson, realized that data processing had far greater potential than just workers punching a time clock. A great motivator of salesmen, Watson sent them

Endicott grew and flourished due to massive numbers of immigrants who came to the area to work for "EJ", predominantly from southern and eastern Europe. "Which way EJ?" was said to be what they asked immigration officials at Ellis Island in New York City, but it is far more likely that they hac memorized the addresses of relatives or friends living in Endicott. The company also maintained recruiting sites in Italy and the Balkans in the early part of the 20th century. Endicott-Johnson's employment in the region reached a peak of about 20,000 in the early 1920s.

In an innovative and far-sighted policy, George F. Johnson made sections of the company's land holdings outside the factory district available to workers to build homes on, with financing provided by the company, and title reverting to the worker when the loan was paid off. Along with extensive cc (unheard of at the time and decades before government took over these responsibilities), this "Square Deal" of the early 20th century is commemorated by stone arches erected by the workers in 1920 across Route 17C (Main Street) at the entrances to Endicott and Johnson City.

Endicott-Johnson was hurt by the Depression of the 1930s, but since shoes were a necessity, did better than other manufacturing sectors of the economy. Orders for shoes from the military in World War II in the 1940s propelled employment over the peak attained in the early 1920s. Unfortunately George F. Johnson in 1948 couldn't cope with a more affluent era in the 1950s and 1960s when footwear became mainly a fashion business in the United States. More importantly, little if any money was invested in improving the original 1900 manufacturing technology, which meant that foreign co market share resulted in the closing and sale of the Endicott factories

IBM [edit]

Endicott is best known as the "Birthplace of IBM".[2][3] The Computing Tabulating Recording Company (CTR) was founded in Endicott on June 16, 1911, via the merger of the International Time Recording Company (ITR), Tabulating Machine Company, Computing Scale Company, and Bundy Time

EVALUATION

Localness Measure	ACC (%)	AED (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	38.48	599.40	49.85	56.06	64.15
BC	47.91	478.14	57.39	62.18	66.98
Jaccard Index	**53.21**	**433.62**	**67.41**	**73.56**	**78.84**
Tversky Index	54.48	429.00	68.72	74.68	79.99

EVALUATION

Localness Measure	ACC (%)	AED (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	38.48	599.40	49.85	56.06	64.15
BC	47.91	478.14	57.39	62.18	66.98
Jaccard Index	**53.21**	**433.62**	**67.41**	**73.56**	**78.84**
Tversky Index	54.48	429.00	68.72	74.68	79.99

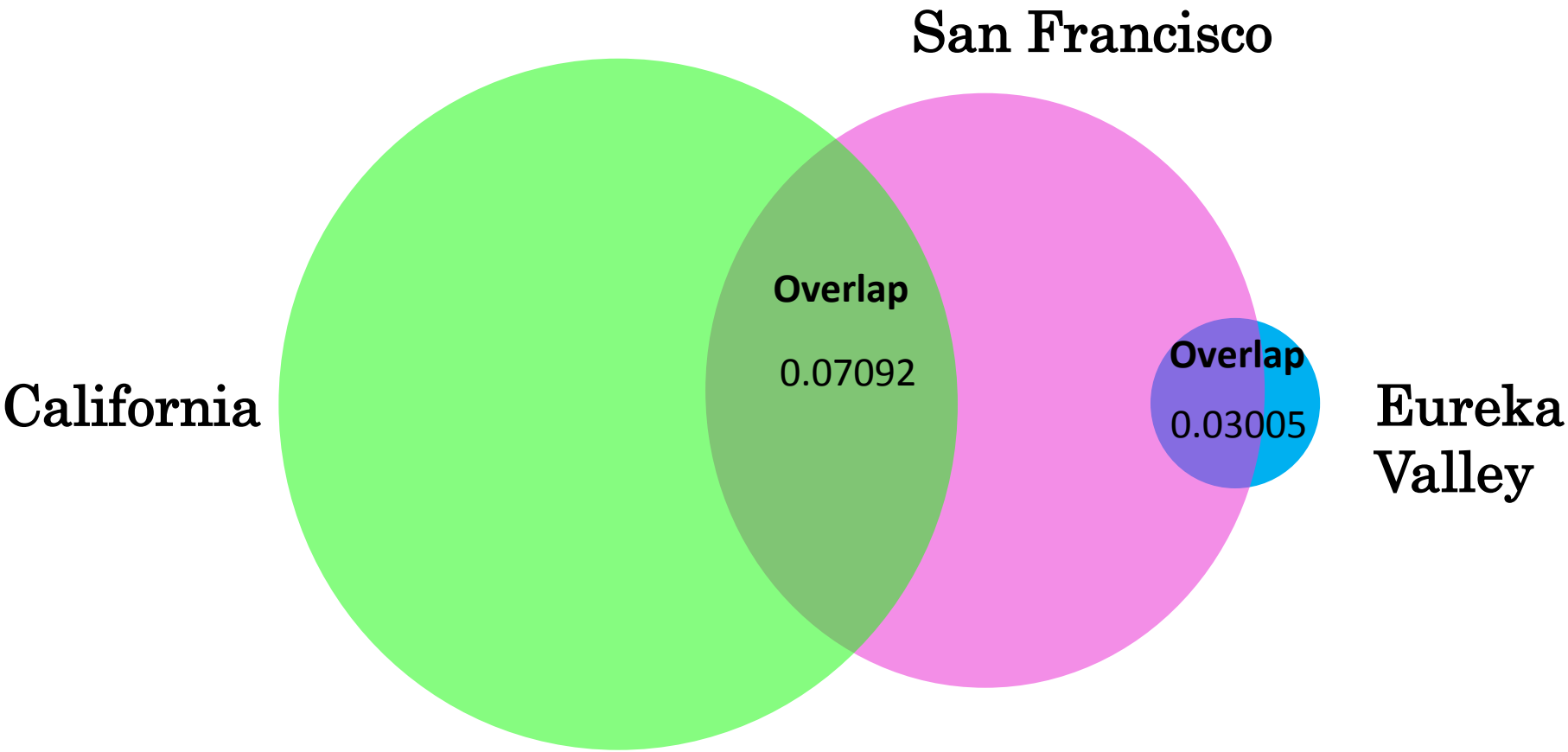
- Underperforms for local entities with fewer entities than the city

EVALUATION

Localness Measure	ACC (%)	AED (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	38.48	599.40	49.85	56.06	64.15
BC	47.91	478.14	57.39	62.18	66.98
Jaccard Index	**53.21**	**433.62**	**67.41**	**73.56**	**78.84**
Tversky Index	54.48	429.00	68.72	74.68	79.99

- Underperforms for local entities with fewer entities than the city
  - E.g. *Eureka Valley* and *California* with respect to *San Francisco.*

EVALUATION


San Francisco
California
Overlap
0.07092
Overlap
0.03005
Eureka Valley

EVALUATION

Localness Measure	ACC (%)	AED (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	38.48	599.40	49.85	56.06	64.15
BC	47.91	478.14	57.39	62.18	66.98
Jaccard Index	53.21	433.62	67.41	73.56	78.84
Tversky Index	**54.48**	**429.00**	**68.72**	**74.68**	**79.99**

EVALUATION

Localness Measure	ACC (%)	AED (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	38.48	599.40	49.85	56.06	64.15
BC	47.91	478.14	57.39	62.18	66.98
Jaccard Index	53.21	433.62	67.41	73.56	78.84
Tversky Index	**54.48**	**429.00**	**68.72**	**74.68**	**79.99**

- Best performing localness measure

EVALUATION

Localness Measure	ACC (%)	AED (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	38.48	599.40	49.85	56.06	64.15
BC	47.91	478.14	57.39	62.18	66.98
Jaccard Index	53.21	433.62	67.41	73.56	78.84
Tversky Index	**54.48**	**429.00**	**68.72**	**74.68**	**79.99**

- Best performing localness measure
- Overcomes the disadvantage of Jaccard Index.

EVALUATION

Localness Measure	ACC (%)	AED (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	38.48	599.40	49.85	56.06	64.15
BC	47.91	478.14	57.39	62.18	66.98
Jaccard Index	53.21	433.62	67.41	73.56	78.84
Tversky Index	**54.48**	**429.00**	**68.72**	**74.68**	**79.99**

- Best performing localness measure
- Overcomes the disadvantage of Jaccard Index.
  - For example: We are able to assign higher localness to *Eureka Valley* (0.7096) than *California* (0.1270) with respect to *San Francisco*

EVALUATION

Comparison with Existing Approaches

Method	ACC (%)	AED (in miles)
Cheng, Caverlee, and Lee, 2010	51.00	535.56
Chang, Lee, Eltaher, and Lee, 2012	49.9	509.3
Wikipedia based Approach	**54.48**	**429.00**

CONCLUSION

CONCLUSION

- **Presented a crowd sourced knowledge based approach, that does not require geo-tagged tweets as a training dataset, to predict the location of a user**

CONCLUSION

- **Presented a crowd sourced knowledge based approach, that does not require geo-tagged tweets as a training dataset, to predict the location of a user**

- Introduced the concept of ***Local Entities*** and preprocessed Wikipedia Hyperlink Graph to extract local entities for each city

CONCLUSION

- **Presented a crowd sourced knowledge based approach, that does not require geo-tagged tweets as a training dataset, to predict the location of a user**

- Introduced the concept of ***Local Entities*** and preprocessed Wikipedia Hyperlink Graph to extract local entities for each city

- Investigated **relatedness measures** to establish the degree of association between a local entity and a city

CONCLUSION

- **Presented a crowd sourced knowledge based approach, that does not require geo-tagged tweets as a training dataset, to predict the location of a user**

- Introduced the concept of ***Local Entities*** and preprocessed Wikipedia Hyperlink Graph to extract local entities for each city

- Investigated **relatedness measures** to establish the degree of association between a local entity and a city

- Evaluated the proposed approach against a benchmark dataset published by Cheng et al. For 5119 users, we are able to predict the location of 55% of users within 100 miles with an average error distance of 429 miles

FUTURE WORK

FUTURE WORK

- Compute the **confidence score** of the prediction based on top-k cities and count of local entities in tweets

FUTURE WORK

- Compute the **confidence score** of the prediction based on top-k cities and count of local entities in tweets
- Investigate other localness measures for score local entities

FUTURE WORK

- Compute the **confidence score** of the prediction based on top-k cities and count of local entities in tweets
- Investigate other localness measures for score local entities
- **Consider semantic types**, categories of local entities and weight the contribution based on types

FUTURE WORK

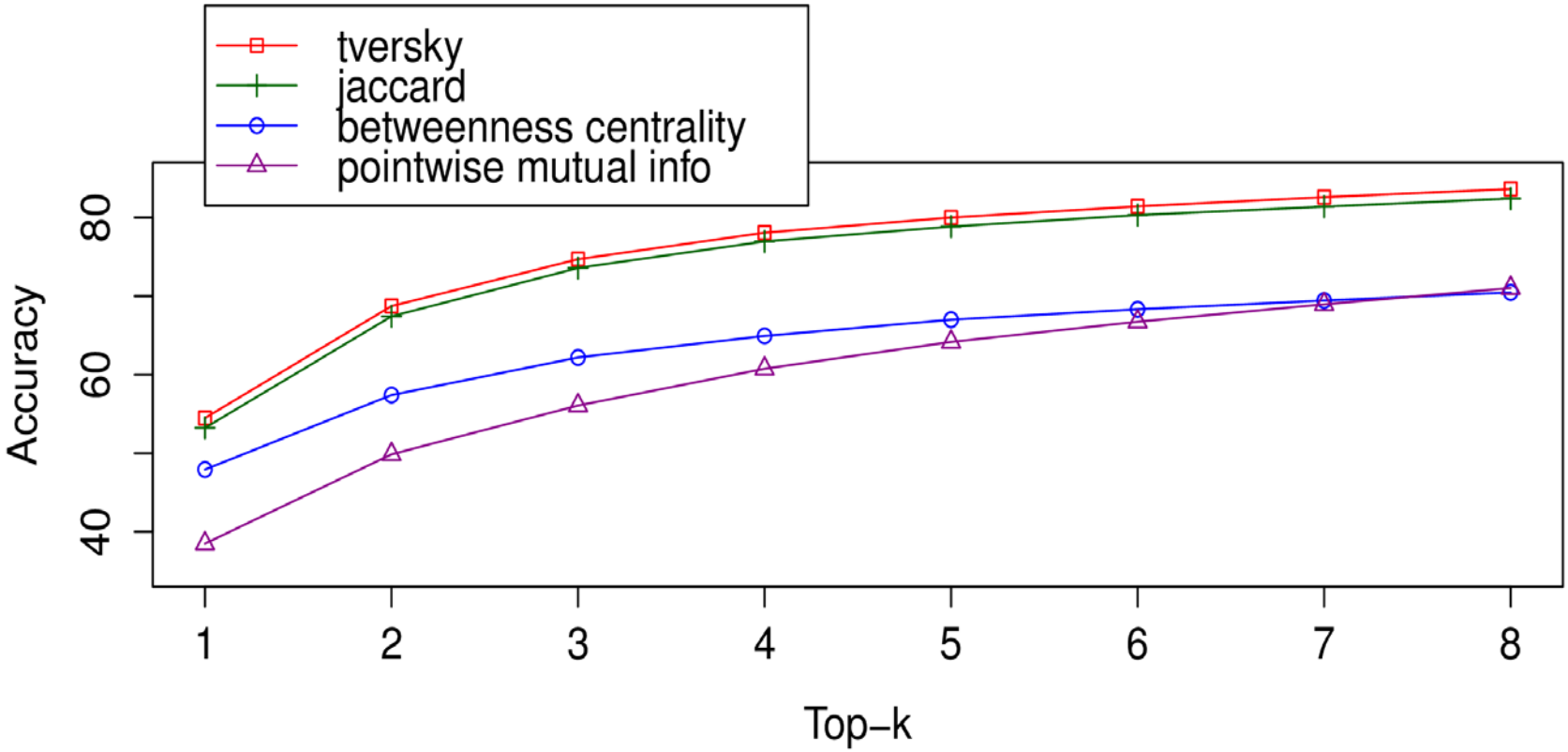
- Compute the **confidence score** of the prediction based on top-k cities and count of local entities in tweets
- Investigate other localness measures for score local entities
- **Consider semantic types**, categories of local entities and weight the contribution based on types
- **Explore other knowledge bases** such as Wikitravel and GeoNames

Thank you!

Contact:
pavan@knoesis.org
@pavankaps

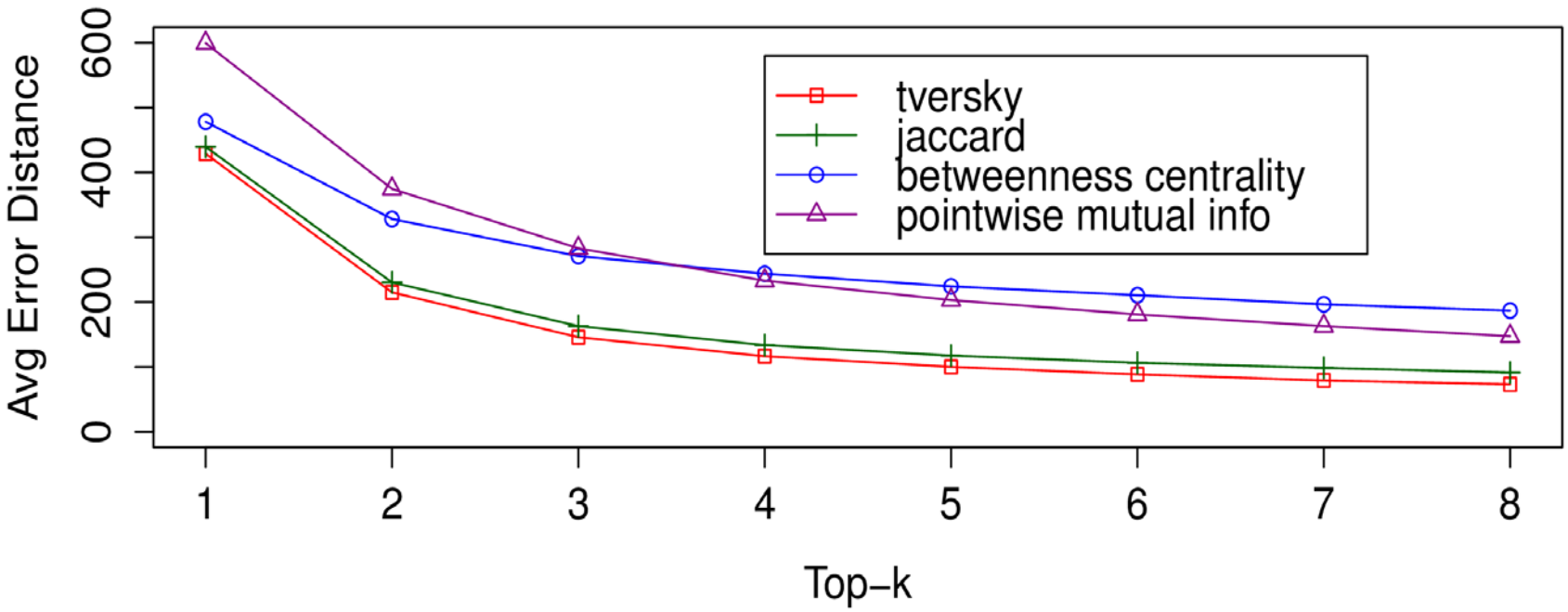
EVALUATION

Top-k Accuracy

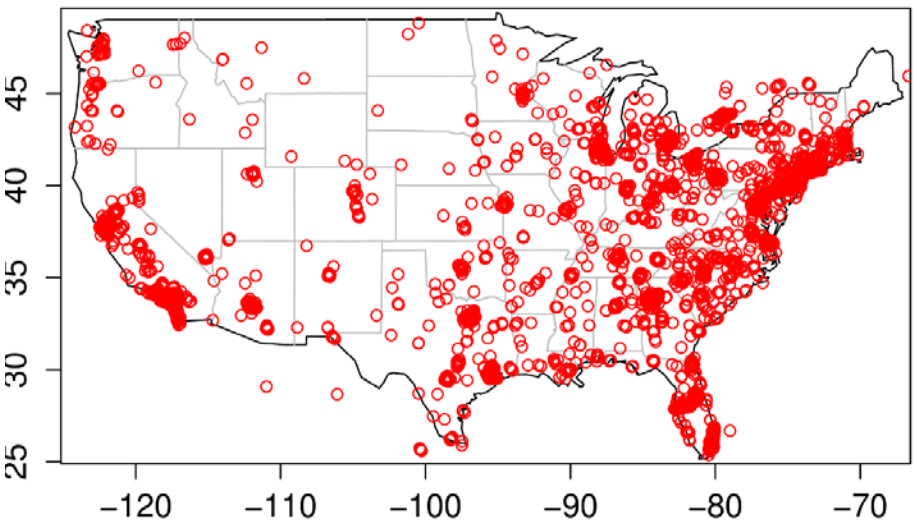


EVALUATION

Top-k Average Error Distance

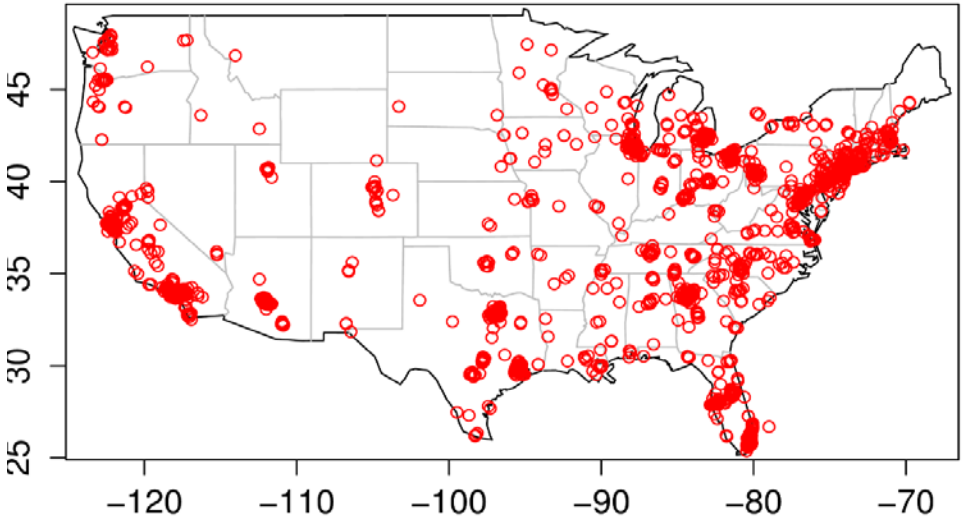


Distribution of users



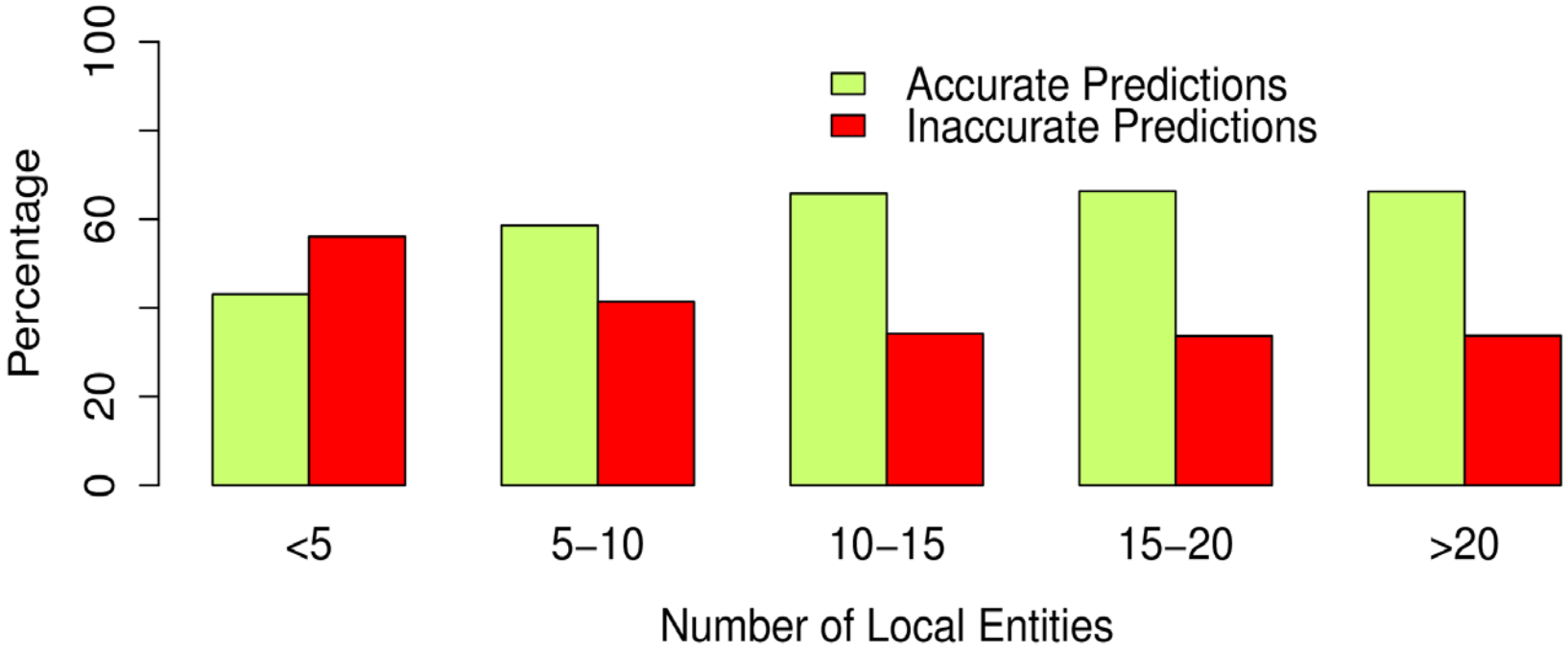
Distribution of all users in the dataset

Distribution of accurately predicted users



EVALUATION

Impact of Local Entities



EVALUATION

Top 100 Cities

EVALUATION

Top 100 Cities

- 2172 users from the dataset are from the top-100 most populated cities of United States
- 60% users predicted within 100 miles of their actual location
- 54% users predicted exactly at the city level