Understanding the perception of cities in society

Redaelli Tommaso, 830442, t.redaelli7@campus.unimib.it
Merlo Riccardo, 829805, r.merlo2@campus.unimib.it
Stoffa Giacomo, 830159, g.stoffa1@campus.unimib.it
Riva Leonardo, 830647, l.riva37@campus.unimib.it

Abstract

Every city has its own special features: often people associates words that defines them. These **perceptions** can belong to different fields and can have positive or negative impact. This phenomenon often results in not-always-true **stereotypes**.

The goal is to analyze how the most **meaningful cities** can be **described** by society, with a focus on the links between them, highlighting what they have in common.

WORKFLOW

0. Introduction

- 0.1. Corpus description and techniques
- 0.2. Corpus exploration and preprocessing
- 0.3. Model building and comparison

1. Extract characterizing words for cities

- 1.1. Topic extraction
- 1.2. Method 1 - *"Supervised"*
- 1.3. Method 2 - *"Unsupervised"*
- 1.4. Results comparison

2. Result representation

- 2.1. Wordcloud
- 2.2. Network graph
- 2.3. Viz's comparison between the two methods

3. Result interpretation and disambiguation

- 3.1. WEAT
- 3.2. Method 1 - Topic
- 3.3. Method 2 - Analogy
- 3.4. Results comparison

4. Conclusion

- 4.1. Final results and stereotypes analysis
- 4.2. Limitations and future research

COCA

It is a large genre-balanced corpus of **contemporary american english**.

It is composed of 950 million words, evenly divided into categories, between 1990 and 2019.

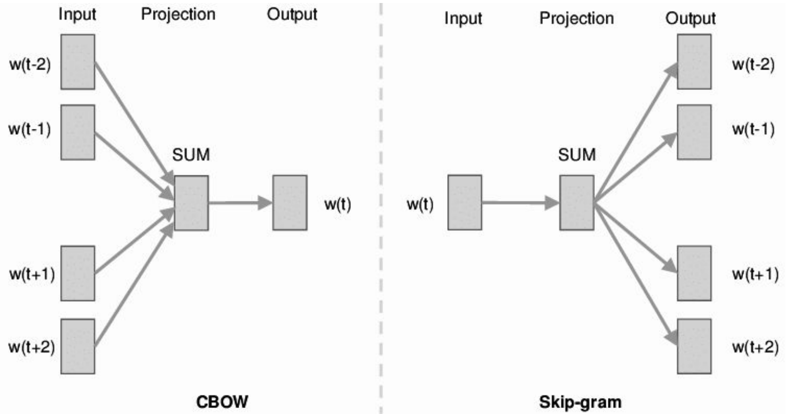
- *spoken*
- *fiction*
- *magazine*
- *newspaper*
- *academic*
- *blogs & web pages*
- *tv & movies subtitles*

Word2Vec

It is an algorithm to learn word associations from a corpus of text, in which each distinct word is represented with a vector.

CBOW: Predicts the **current word from a window** of surrounding context words no matter the order (bag-of-words assumption).

Skip Gram: Uses the **current word to predict** the surrounding window of context words; close context words are weighted heavily than far context words.



WEAT

Word-Embedding Association Test

It measures the **comparative polarization** between a pair of target words (X, Y) and a pair of polarization attributes (A, B) in a single-corpus distributional model

$$s(X,Y,A,B) = \sum_{x \in X} s(x,A,B) - \sum_{y \in Y} s(y,A,B)$$

$$s(w,A,B) = \mathrm{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \mathrm{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

The effect size is

$$\frac{\mathrm{mean}_{x \in X} s(x,A,B) - \mathrm{mean}_{y \in Y} s(y,A,B)}{\mathrm{std\_dev}_{w \in X \cup Y} s(w,A,B)}$$

Corpus preprocessing

1. Use of the standard form, i.e. **WLP**. Verbal tenses and plurals are not useful.

2. Reduction of composed **tag PoS** into their main form, e.g. *nn_pp* → *nn*.

3. **Lemmatization**: only nouns [N], adjectives [J], adverbs [R] e verbs [V] are kept.

4. **Bi-gram** research, to fix composed words, according to the frequency at which two words appear together:

$$\frac{\ln\left(P(word_a, word_b)\right) / \left(P(word_a) \times P(word_b)\right)}{-\ln\left(P(word_a, word_b)\right)}$$

$$\text{where } P(word) = \frac{word\_count}{corpus\_word\_count}$$

↓

"new york"	→	*"new_york"*
"rio de janeiro"	→	*"rio_janeiro"*

Model building and comparison

1 **Word2Vec** models → **CBOW** / **Skip Gram**

Built different models with both techniques, tuning parameters by following a trial-and-error pattern

2 **Evaluation** of the obtained models → For each model, we tested it on a predetermined list of analogies →

Skip Gram performs better:

SG_mc50_npmi → acc. 0.602, std. 0.233
CBOW_mc50_npmi → acc. 0.543, std. 0.213
...

Model building and comparison - *Analogies*

Famous world capitals
Athens : Greece = Baghdad : Iraq

World capitals
Abuja : Nigeria = Accra : Ghana

Currencies
Algeria : dinar = Angola : Kwanza

Cities in state
Chicago : Illinois = Houston : Texas

Adjectives to adverbs
Amazing : Amazingly = Apparent : Apparently

Opposites
Acceptable : Unacceptable = Aware : Unaware

Family
Boy : Girl = Brother : Sister

Nationality adjectives
Albania : Albanian = Argentina : Argentinean

WORKFLOW

0. Introduction

0.1. Corpus description and techniques
0.2. Corpus exploration and preprocessing
0.3. Model building and comparison

1. Extract characterizing words for cities

1.1. Topic extraction
1.2. Method 1 - *"Supervised"*
1.3. Method 2 - *"Unsupervised"*
1.4. Results comparison

2. Result representation

2.1. Wordcloud
2.2. Network graph
2.3. Viz's comparison between the two methods

3. Result interpretation and disambiguation

3.1. WEAT
3.2. Method 1 - Topic
3.3. Method 2 - Analogy
3.4. Results comparison

4. Conclusion

4.1. Final results and stereotypes analysis
4.2. Limitations and future research

Step 1 - Extract characterizing words for cities

1. How to represent the **collective imagination** of a city?

2. Subdivision of the entire image into different **thematic areas** (topic)
   [e.g. food, crime, environment, ...]

3. Given a city C and an area A, words are collected in order to describe C from the point of view of A.
   This process is then developed in two different methods, one "supervised" and one "unsupervised".

Method 1: Supervised - *Formulation*

Idea: topics and their keywords are defined beforehand

crime *= [abduction, terrorism, ...]*
education *= [high_school, learning, ...]*
food *= [restaurant, chef, ...]*
work *= [business, office, ...]*
nightlife *= [disco, theater, ...]*
cold *= [snow, winter, ...]*
hot *= [sun, summer, ...]*
transport *= [taxi, plane, ...]*
environment *= [ecology, sustainability, ...]*
pollution *= [waste, contamination, ...]*

history *= [king, castle, ...]*
gambling *= [roulette, blackjack, ...]*
alcohol *= [liquor, beer, ...]*
music *= [dance, k_pop, ...]*
fitness *= [healthy, workout, ...]*
vegetation *= [tree, florist, ...]*
animals *= [safari, zoo,...]*
sport *= [soccer, football, ...]*
fashion *= [hairdresser, stylist, ...]*
tech *= [technology, smartphone, ...]*

Method 1: Supervised - *Algorithm & Results*

To obtain words that best describe a city:

1. From the list of all words, find the closest ones in terms of cosine distance

2. Set upper limit on count & threshold on similarity, to filter meaningful words

Examples

- Nairobi → [*bombing*: 0.369,
  taxi: 0.338,
  safari: 0.291, ...]
- Athens → [*ancient*: 0.324,
  antiquity: 0.314,
  amphitheater: 0.293, ...]
- Sydney → [*swimming*: 0.383,
  rugby: 0.343,
  labour: 0.316, ...]

Method 2: Unsupervised - *Formulation*

1

Basic idea: given a city and a topic, analogies can be formulated to obtain words related to that specific city in that specific topic.

Example 1: *city : crime = sao_paulo : x*

→ *violent crime, homicide, carjacking, drug trafficking kidnapping, offender, commit crime, robbery, …*

Example 2: *city : pollution = sao_paulo : x*

→ *air_pollution, carbon emission, ozone depletion, emission, ghg emission, deforestation, sulfur dioxide, ...*

2

Setting: definition of the topics of interest, and, for each of them, some keywords that represent them

crime:	*[crime]* ,
education:	*[education, school]* ,
food:	*[food, recipe, alcohol]* ,
sociality:	*[politic, economy]*,
tourism:	*[tourism, vacation]* ,
environment:	*[environment, pollution]* ,
history:	*[history, folklore]* ,
culture:	*[music, dance, art]* ,
geo:	*[vegetation, flora_fauna]*

Method 2: Unsupervised - *Algorithm*

3

Word collection: For each city, for each keyword of each topic, the words that should represent that characteristic in that city are obtained by analogy.

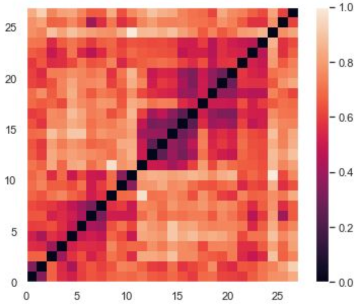
"city" : keyword = city_name : x

- min similarity threshold is set to 0.3, and limits to top 40 words
- due to noise, results that have no direct similarity with 'keyword' and 'city'> 0.2 are discarded

4

Word clustering: To observe both the internal cohesion of a city and the degree of similarity between two of them, it may be useful to aggregate the list of terms obtained.

A DB-SCAN clustering is applied by placing the matrix of the distances between the words collected as input.



"***Crime***" in **Warsaw**:

1. *atrocity, tribunal_hague, radovan_karadzic*
2. *nazi, genocide, kristallnacht, auschwitz*
3. *katyn, nuremberg, espionage, nazism*
4. *ratko_mladic, treblinka, mihailovic, pro_nazi, gestapo, milosevic*

Method 2: Unsupervised - *Results*

5.a

city : food = paris : *x*

escargot, coq_au, provencal, steak_tartare, gourmet, cordon_bleu, wine, hor, s'il_plait, five_course, gefilte_fish, canape, bonbon

5.b

city : crime = kinshasa : *x*

rwanda, habyarimana, hutu, omar_al_bashir, tribunal_hague, rwandan, violence, rwandan_genocide, genocide, extrajudicial_killing, war_crimes, darfur

5.c

city : history = athens : *x*

thucydides, historical, historian, herodotus, egyptology, plato, peloponnesian, greek, aristotle, pericles, greco_roman, epictetus

Supervised vs. Unsupervised - *Pro & cons*

1. The unsupervised method shows more specific and characteristic elements about cities, with a level of detail that the supervised method could not reach, unless the preset lists have not been enlarged enough (expensive). On the other hand, supervised results are more interpretable and reliable, as a result of a hand-made approach

2. However, there's the problem of gathering too much general terms, not specific to cities, but common to several of them [e.g. *environment* → *climate change*], or even terms that are not entirely relevant to the area [e.g. *food* → *<kitchen_book_author>*]

3. Clusterization favors a future elaboration of unsupervised results but, on the other hand, being a partial clustering, it can prevent the identification of singular terms (i.e. unique words that do not have linked terms within those collected).

WORKFLOW

0. Introduction

0.1. Corpus description and techniques
0.2. Corpus exploration and preprocessing
0.3. Model building and comparison

1. Extract characterizing words for cities

1.1. Topic extraction
1.2. Method 1 - *"Supervised"*
1.3. Method 2 - *"Unsupervised"*
1.4. Results comparison

2. Result representation

2.1. Wordcloud
2.2. Network graph
2.3. Viz's comparison between the two methods

3. Result interpretation and disambiguation

3.1. WEAT
3.2. Method 1 - Topic
3.3. Method 2 - Analogy
3.4. Results comparison

4. Conclusion

4.1. Final results and stereotypes analysis
4.2. Limitations and future research

Step 2 - *Results representation*

Two types of representations that show relationships between contents of different cities and distribution of the topics between cities.

1

WordCloud
The purpose is to show which are the most fit words to describe a city, emphasizing topic prevalence using colors.

2

Connection graph
The purpose is to observe how cities are positioned relatively to a specific area of interest, allowing to understand similarity relationships between them, as well as differences in predominance and cohesion of topics.

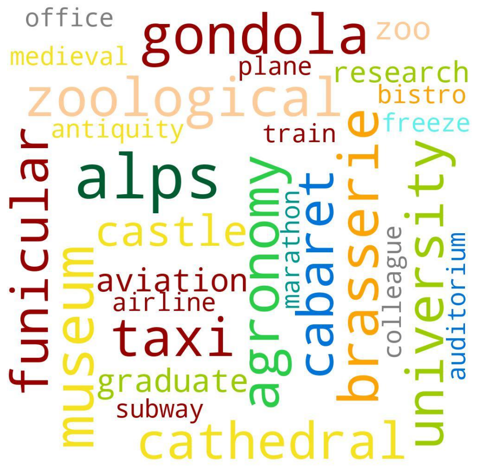
Wordcloud

As in [5], cities are represented in word clouds, with a score:

- The **bigger** the word, the **more similar** to the city (cosine similarity to the city, with a threshold)
- The **color** represents the topic
- The **score** determines how much the topics are meaningful in representing a city
  - Inversely proportional to the number of topic
  - Directly proportional to the number of words of each topic

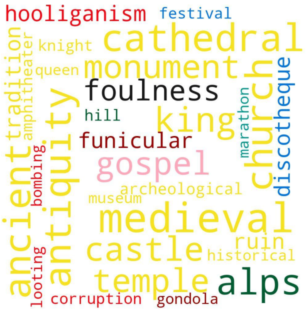
  It is calculated as a product of absolute and relative frequencies of words and topics

Zurich *(0.317)*



Wordcloud (supervised) - *Relevant examples*

Rome
(0.963)

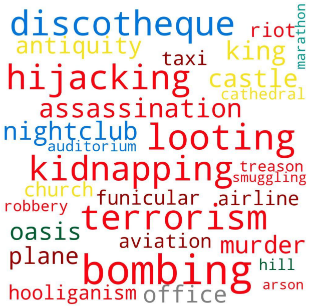


Nashville
(0.816)

Las Vegas
(0.816)

Bologna
(0.730)

Tripoli
(0.566)



Luxembourg
(0.170)

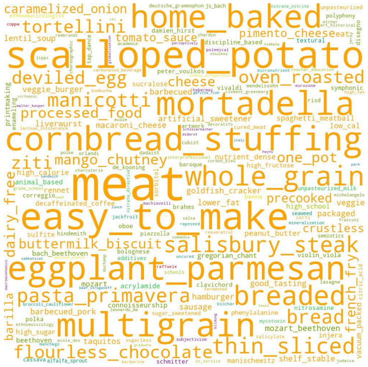
Wordcloud (unsupervised) - *Relevant examples*

Rome
(0.855)

Nashville
(0.915)

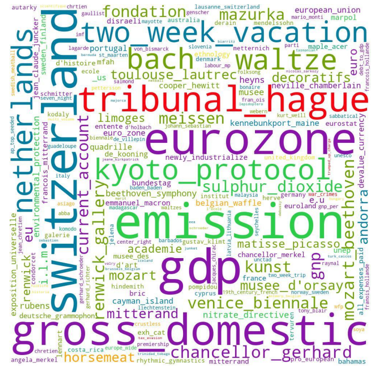
Las Vegas
(0.732)

Bologna
(0.954)



Tripoli
(0.678)

Luxembourg
(0.552)



Network graph

Idea: represent the data for each topic through a graph.

→ **nodes** = cities

→ **edges** = connections between the content of their topics

Cities can be seen under many shades: for topic predominance, for internal cohesion or for different level of connection.

1. **Node size:** how much that city is close to the topic:

$$Dim = \frac{n^\circ\ of\ topic\ words}{n^\circ\ of\ total\ words}$$

2. **Node color:** different for each topic; its intensity is proportional to the cohesion between the words of that city

$$Color = \sum_{c=0}^{\dim(CL)} Cohesion(CL_c) \cdot Freq(CL_c)$$

Network graph

3

Level of connection between two cities is given by the distance between the two mean vectors of the two cities.

Connection with *Bogotà* in *Crime*

source	target	topic	score
bogota	rio_janeiro	crime	0.947666
bogota	madrid	crime	0.902262
bogota	lima	crime	0.884888
bogota	marseille	crime	0.859953
bogota	buenos_aires	crime	0.854765
bogota	caracas	crime	0.850371

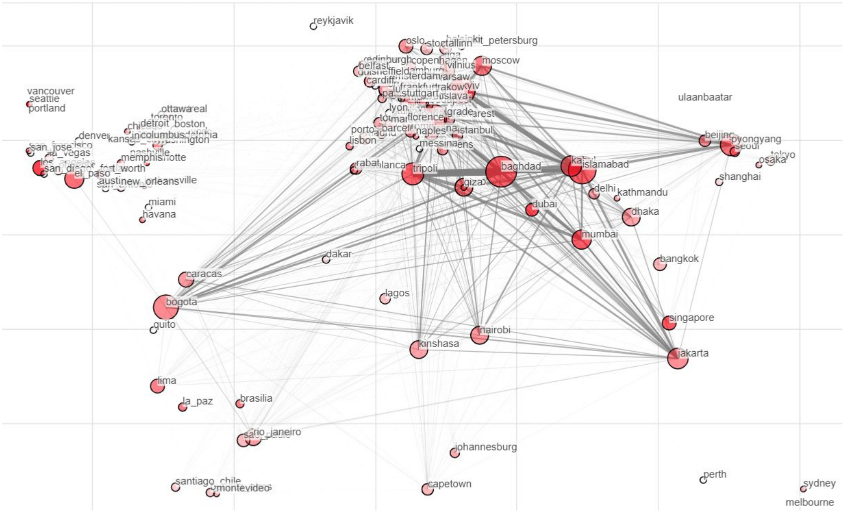
Connection with *Dhaka* in *Environment*

source	target	topic	score
dhaka	jakarta	environment	0.875943
dhaka	ulaanbaatar	environment	0.852511
dhaka	sao_paulo	environment	0.838674
dhaka	santiago_chile	environment	0.831388
dhaka	delhi	environment	0.831137
dhaka	portland	environment	0.821504

Connection with *Warsaw* in *Sociality*

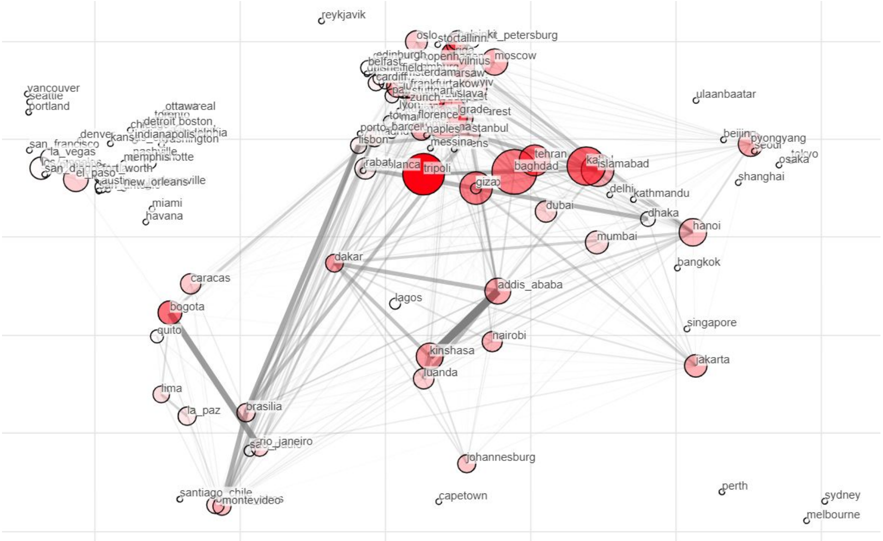
source	target	topic	score
warsaw	prague	sociality	0.914993
warsaw	vilnius	sociality	0.912064
warsaw	riga	sociality	0.909956
warsaw	tallinn	sociality	0.907720
warsaw	bucharest	sociality	0.903156
warsaw	tirana	sociality	0.899807

Network graph - Results



2

Graph obtained by unsupervised method in topic *Crime*



1

Graph obtained by supervised method in topic *Crime*

WORKFLOW

0. Introduction

- 0.1. Corpus description and techniques
- 0.2. Corpus exploration and preprocessing
- 0.3. Model building and comparison

1. Extract characterizing words for cities

- 1.1. Topic extraction
- 1.2. Method 1 - *"Supervised"*
- 1.3. Method 2 - *"Unsupervised"*
- 1.4. Results comparison

2. Result representation

- 2.1. Wordcloud
- 2.2. Network graph
- 2.3. Viz's comparison between the two methods

3. Result interpretation and disambiguation

- 3.1. WEAT
- 3.2. Method 1 - Topic
- 3.3. Method 2 - Analogy
- 3.4. Results comparison

4. Conclusion

- 4.1. Final results and stereotypes analysis
- 4.2. Limitations and future research

Step 3 - *Results interpretation*

The two representations already allow a meaningful exploration of the thematic areas of the cities and the connections between them. Nonetheless, it is possible to spot some instances in which the interpretation is ambiguous, weak or unexpected.
In order to limit this problem, WEAT and SWEAT have been tested.

WEAT
vs
SWEAT

After an attempt of creating some models from city-based sub-corpora, the SWEAT result were not significant, due to the size of the texts, not enough to generate a decent working model.

The final choice fell on **WEAT**

Step 3 - *WEAT usage*

Two different procedures have been developed. For both of them, the problem is represented by defining the **context** through the set of terms A and B on the basis of the topic that wanted to be explored.

1

Topic
Custom version of WEAT called **TWEAT** (T stands for Topic). It works by combining a list of topics to the two lists of attributes. Its goal is to distinguish between negative and positive perception on something

2

Analogies
It works by exploiting analogies, in order to obtain sets of terms related to the same topic but from two opposite points of view, which don't necessarily are "negative" and "positive".

WEAT Disambiguation - *Topic*

a

Definition of the most polarizing attributes, starting from a larger list

POS = [*useful, important, benefit, solution, best, better*]

NEG = [*worse, terrible, horrible, awful, disgusting, dreadful*]

b

Definition of X, Y, A, B

- X → 1st set of cities
- Y → 2nd set of cities
- A → combination of POS with W_T
- B → combination of NEG with W_T

where W_T is a list of Words about a topic T

c

Run classical WEAT, with X, Y, A, B

WEAT Disambiguation - *Topic results*

Music (validation group)

X = [*kabul, baghdad, cairo, damascus, beirut*]
Y = [*nashville, new_york, los_angeles, chicago, denver*]

A = [*useful, important, benefit, solution, best, better*]
B = [*worse, terrible, horrible, awful, disgusting, dreadful*]

T = [*music, song, jazz, lyrics, guitar*]

Effect size = -1.417

Topic impact over WEAT= -0.249

TWEAT p-value = 0.0375

Nightlife (UK vs South America)

X = [*belfast, glasgow, edinburgh, dublin, london, cardiff*]
Y = [*montevideo, buenos_aires, santiago_chile, rio_janeiro, sao_paulo, caracas*]

A = [*useful, important, benefit, solution, best, better*]
B = [*worse, terrible, horrible, awful, disgusting, dreadful*]

T = [*pub, discotheque, nightlife, nightclub, festival, nightspot*]

Effect size = -0.343

Topic impact over WEAT = -0.104

TWEAT p-value = 0.2635

WEAT Disambiguation - *Topic results*

Education (US vs UK)

X = [*toronto, boston, new_york, philadelphia, washington*]
Y = [*edinburgh, leeds, sheffield, dublin, belfast*]

A = [*useful, important, benefit, solution, best, better*]
B = [*worse, terrible, horrible, awful, disgusting, dreadful*]

T = [*university, college, education, school, diploma*]

Effect size with topic = -0.390

Topic impact over WEAT= -0.102

TWEAT p-value = 0.356

Environment (Northern EU vs Brazil)

X = [*oslo, bergen, trondheim, stockholm, gothenburg, uppsala, helsinki, tampere, turku, copenhagen, aarhus, odense*]
Y = [*sao_paulo, rio_janeiro, belo_horizonte, brasilia, salvador, fortaleza, curitiba, manaus, recife, belem, porto_alegre, natal*]

A = [*useful, important, benefit, solution, best, better*]
B = [*worse, terrible, horrible, awful, disgusting, dreadful*]

T = [*environment, renewable, green, sustainability, organic, biological*]

Effect size with topic = 0.010

Topic impact over WEAT= 0.023

TWEAT p-value = 0.4605

WEAT Disambiguation - *Analogies*

a. **Definition of X and Y** (two sets of cities)
e.g. ***X*** = *[oslo, copenhagen, stockholm]* vs. ***Y***= *[brasilia, beijing, pyongyang]*

b. **Definition of disambiguation terms** → **Topic** e.g. *environment* → **Keyword in topic** e.g. *pollution* → **positive POV** e.g. *protection* / **negative POV** e.g. *cause*

c. **Definition of analogies:**
e.g. **positive**: *"environment : protection = pollution : x"*
e.g. **negative**: *"environment : cause = pollution : x"*

→ **POS:** *[protect, safeguard, environmental protection, prevention, …]*
→ **NEG:** *[air pollution, traffic related, smog, headache dizziness, …]*

d. **Run classical WEAT**, with X, Y, A, B

WEAT Disambiguation - *Analogies results*

Environment (Northern EU vs "Polluted Cities")

Pollution → Cause vs Protection

X = [*oslo, copenhagen, stockholm*]
Y = [*brasilia, beijing, pyongyang*]

A = [*protect, enforcement, air_pollution, safeguard, environmental_protection, prevention, abatement, anti_pollution, nonpoint_source, naaqs*]

B = [*air_pollution, nonpoint_source, culprit, pollutant, sulphur_dioxide, particulate, traffic_related, sulfur_dioxide, smog, headache_dizziness*]

Effect size = 1.808

P-value = 0.0

Food (Italy vs Asia)

Ingredients → Fish vs Meat

X = [*rome, bologna, milan*]
Y = [*tokyo, shanghai, singapore*]

A = [*mussels, tuna, sturgeon, salmon, redfish, halibut, shrimp , albacore_tuna, rainbow_trout, tartare, mahi mahi*]

B = [*roast, pork_tenderloin, smoked_sausage, pork_loin, sauce, olive_oil, garlic, andouille_sausage, beef_brisket, parm*]

Effect size = -1.836

P-Value = 0.0

WEAT Disambiguation - *Analogies results*

Culture (America Latina vs European)

Music → Tribal vs Noble

X = [*brasilia, sao_paulo, rio_janeiro, buenos_aires, santiago_chile*]
Y = [*rome, paris, vienna, munich, saint_petersburg*]

A = [*band, dance, drumming, percussion, song, orchestra, jazz, drum, piano, guitar*]

B = [*harmonium, piano, violin, singing, oboe, orchestra, pianoforte, sing, clavichord, bagpipe*]

Effect size = 1.812

P-value = 0.0

History (Colonizers vs Colonies)

Colonialism → Invader vs Oppressed

X = [*lisbon, porto, madrid*]
Y = [*sao_paulo, rio_janeiro, santiago*]

A = [*colonizer, invading, colonization, imperialist, colonialist, invasion, imperialism, colonized, neo_colonialism, conqueror*]

B = [*oppression, oppressor, oppress, marginalized, neo_colonialism, colonized, patriarchy, apartheid, injustice, subjugate*]

Effect size = 1.258

P-value = 0.0

WEAT Disambiguation - *Comparison*

Method 1 - *With Topic*

a. As before, supervised methods are more specific, but this can be restrictive

b. Results are not as good as expected: topics don’t disambiguate that much, even though they manage to polarize more than when they are not used

Method 2 - *With Analogies*

a. As before, unsupervised methods are more powerful but less controllable. Some results may not be fully correlated to the sets A and B under survey

b. Consequently, the measures provided can’t be taken as accurate and truthful but can be interpreted with a more stereotyped key

WORKFLOW

0. Introduction

0.1. Corpus description and techniques
0.2. Corpus exploration and preprocessing
0.3. Model building and comparison

1. Extract characterizing words for cities

1.1. Topic extraction
1.2. Method 1 - *"Supervised"*
1.3. Method 2 - *"Unsupervised"*
1.4. Results comparison

2. Result representation

2.1. Wordcloud
2.2. Network graph
2.3. Viz's comparison between the two methods

3. Result interpretation and disambiguation

3.1. WEAT
3.2. Method 1 - Topic
3.3. Method 2 - Analogy
3.4. Results comparison

4. Conclusion

4.1. Final results and stereotypes analysis
4.2. Limitations and future research

Conclusions

1

How are the different cities in the world perceived? What are the terms/topics most associated with them?

The gathered words manage to picture the major characteristics about cities, in terms of different topics.
Wordclouds show that:

→ some cities are strongly related to one or few topic in their words:

- *Food* in **Bologna:** *[scalopped_potato, mortadella, eggplant_parmesan, home_baked, manicotti, tortellini …]*
- *Culture* in **Vienna**: *[beethoven, mozart, brahms, schubert, waltz, deutsche_grammophon]*
- *Environment* in **Copenhagen:** *[kyoto, carbon_emission, ipcc, ghg_emission, unep, marpol]*
- *Politics and Economy* in **Moscow:** *[kremlin, gorbachev, putin, soviet_union, kgb, espionage, primakov]*

→ on the other hand, other cities whom topics are heterogeneous but still well correlated:

- *Sociality* and *Environment* in **Beijing:** *[economic_growth, export_led, air_pollution, emissions, devalue currency …]*
- *Education*, *Culture* and *Sociality* in **New York:** *[college, 20th_century_art, metropolitan_museum, paul_krugman ..]*
- *Tourism* and *Sociality* in **Dubai:** *[stock_market, emerging_market, mini_vacation, golfing, undiplomatic …]*
- *Crime* and *Drugs* in **Bogotà:** *[cocaine, illicit_drug, alcohol, kidnapping, drug_trafficking, psychoactive_substance …]*

Conclusions

2

How is the world described in collective imagination, in terms of topics?

The network graph allows to make some general considerations:

→ In the ***Crime*** topic, **Middle East**, **Central Africa** and **Latin America** stand out.

→ The ***History*** topic creates strong interconnections in **Europe**, we note the influence of the **UK** in **North America** and **Australia** and the influence of **Spain** and **Portugal** in **Latin America** (the latter is also visible in the ***Culture*** and ***Food*** topics).

→ The ***Education*** topic denotes a large gap between the **Northern** and **Southern** hemisphere of the world, with the exception of **Australia**.

→ The topic ***Environment*** is present and strongly connects **Northern Europe**, **Latin America**, **China** and **India**.

→ From the ***Culture*** topic emerges once again a strong homogeneity in **Europe** linked to everything (classical music and different artistic currents) and a strong connection between the cities of **South America**.

Conclusions

3

Are there any polarizations (positive/negative) regarding to these characteristics? Are common stereotypes confirmed?

There are obviously different points of view in interpreting the topics related to cities. With the final step we tried to disambiguate them, even where cities look strongly connected.

→ *Difference in meaning regarding to the Environment topic in Northern Europe where words are commonly related to danger and climate protection activities, compared to South America or Indian cities where are associated to words more aimed at the causes of environment.*

Additionally is possible to use WEAT to highlight any stereotype of the collective imaginary

→ *Comparing food between meat and fish recipes, we can observe that meat ones are more associated to italian places instead of seafood being more present in oriental cities.*

→ *Opera and classical music is best connected to european countries rather than brazilian cities, in which drums and tribal rhythms are majorly present.*

Considerations, limits and future developments

1 **Model built on the entire COCA corpus** → Ignored any partition in type and year. → Models in different years (evolution of terms and associated topics) → Models in specific corpus (representation of cities in films or news)

2 **Choice of initial words in supervised method** → Strong bias on the specificity of the selected words → Possibility to exploit the model to extend the word list starting from an initial set

Considerations, limits and future developments

3 **Unsupervised analogies results management**

- Less bias, more noise → Research for more targeted analogies or compare them in a pre-compiled set
- DB-SCAN not optimized → Dedicated clustering to highlight specific subtopics

4 **Use of WEAT for disambiguation and stereotypes**

- Topic WEAT ineffective → Possible use of SWEAT on corpus from different cities
- Non-intuitive analogies → Composition of sets A and B supervised with domain-specific terms.

Bibliography

[1] Tomas Mikolov et al. “Distributed representations of words and phrases and their compositionality”. In: Advances in neural information processing systems. 2013, pp. 3111–3119. Available at: https://arxiv.org/abs/1310.4546

[2] Hamilton W., Clark K., Leskovec J. and Jurafsky D., 2020. "Inducing Domain-Specific Sentiment Lexicons From Unlabeled Corpora". Available at: https://arxiv.org/abs/1606.02820

[3] “Understanding Tourism Image of Cities through Social Semantic Network Analysis: A Case Study of Shanghai”, https://gala.gre.ac.uk/id/eprint/19600/7/19600%20MACLEOD_Tourism_in_Contemporary_Cities_2017.pdf (pag 16-38)

[4] “Extracting and analyzing semantic relatedness between cities using news articles”, https://doi.org/10.1080/13658816.2017.1367797

[5] “A Semantic and Sentiment Analysis on Online Neighborhood Reviews for Understanding the Perceptions of People toward Their Living Environments”, https://doi.org/10.1080/24694452.2018.1535886