ARC Welding

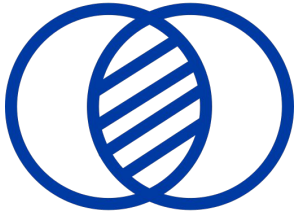
Analysis of Reddit Communities Welding

Jan Sawicki

Executive summary

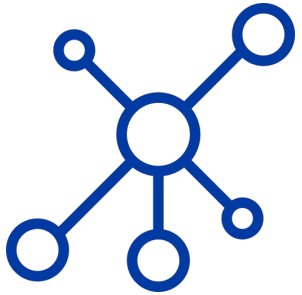
What?

Analyze Reddit...



Why?

...to find community similarities ...



How?

...with NER networks.

ARC Welding: Analysis of Reddit Communities Welding

supervisor: Prof. Maria Ganzha
author: Jan Sawicki

Executive summary
ARC Welding is a project about fusing online communities to find common interest of users

CCS tags
Data extraction and integration; Social advertising; Data mining; Natural language processing; Networks

Reddit

multithreaded download script

Storage

posts with metadata and content

Posts

clean **dataframes** by subreddit

NLP

"**T**he **U**ltimate **D**ataset **F**or **E**verything" categorized subfora are key to the project

hundreds GB from **most popular** subreddits

post processing: score-based filtering content verification

Named entity recognition, sentiment, embeddings, summarization, ...

Similarities

shared **community interests**

"Welding"

processed subreddit **networks**

Networks

subreddit **networks**

Interest mining

Application: community-targeted marketing using reconnoitered conquered sectors

Goal: find similarities
Evaluation: cosine similarity of text embedded findings & community text

Nodes: user interests
Edges: in-post occurrence
Properties: score, awards, sentiment, ...

Mining objects of interest keyword extraction, NER, transformers, sentiment

research engineering

`from python import jupyter, pandas, numpy, pandarallel, psaw, praw, plotly, pyvis, multiprocessing, transformers (pytroch, tensorflow), fasttext, spacy, networkx, ...`

Why Reddit?

Subreddits

Twitter, Instagram don’t have it Facebook has, but private + no API

Insight

Advanced and continually increased user-side expert knowledge of the platform

API

Pushshift!

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. "**The pushshift reddit dataset.**" In Proceedings of the international AAAI conference on web and social media, vol. 14, pp. 830-839. 2020.

?

TBA

Why NOT Reddit

Images and videos

Most of the posts have media content
(text titles are still obligatory)

Very “internet” content

unstructured noise text data, shortcuts, acronym, lack of grammar, slang, subreddit specific phrases (e.g. r/therewasanattempt), a lot of deleted posts

The data

75 **gigabytes**

of raw data scraped with **PRAW** (whole year **2020**)

500K **named entities**

detected with **transformers** and **flair**

200 **subreddits**

selected from top **700** subreddits

13M **edges**

co-ocurrences of named entities in posts (**network edges**)

NER models

dslim/bert-large-NER

*Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "**Bert: Pre-training of deep bidirectional transformers for language understanding.**" arXiv preprint arXiv:1810.04805 (2018).*

*Sang, Erik F., and Fien De Meulder. "**Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition.**" arXiv preprint cs/0306050 (2003).*

flair/ner-english-large

*Schweter, Stefan, and Alan Akbik. "**Flert: Document-level features for named entity recognition.**" arXiv preprint arXiv:2011.06993 (2020).*

Models - comments

ReddiBert?

There is no NER model pretrained on Reddit datasets.

“Special” posts

ASCII art, non-empty empty and other wonder in r/teenagers

Merging NE

There is a bug in transformers: entities are not merged if tokens are of different types. e.g. "I took Ritalin (Methylphenidate) yestarday."

This. Is. SLOW.

For some subreddits it takes **several hours** (per posts from single year)

Network creation

Weighted undirected graph

(no selfloops, no multiedges)

Nodes

Named entities

Edges

Post co-occurence

Weights

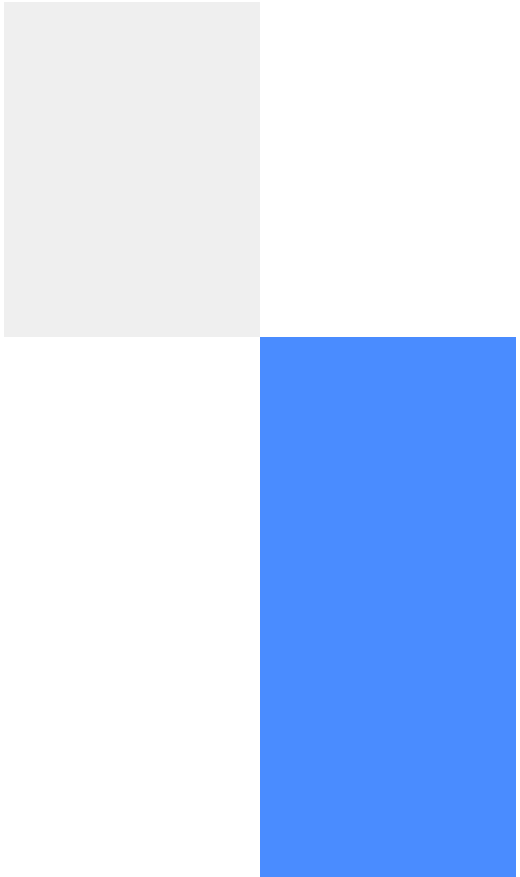
Node weight and edge weights are based on combined score of related posts

The charts

networks

Are they similar?

Question



Network similarity?

Source #1

*Tantardini, Mattia, Francesca Ieva, Lucia Tajoli, and Carlo Piccardi. "**Comparing methods for comparing networks.**" Scientific reports 9, no. 1 (2019): 1-19.*

Source #2

dr inż. Anna **Chmiel**, Faculty of Physics, WUT

Network type	Known Node-Correspondence (KNC)	Unknown Node-Correspondence (UNC)
Undirected Unweighted	–Euclidean (EUC), Manhattan (MAN), Canberra (CAN), Jaccard (JAC) distances –DeltaCon (DCON)	–Global statistics –Spectral Adjacency (EIG-ADJ), Laplacian (EIG-LAP), SNL (EIG-SNL) distances –GCD-11 –MI-GRAAL –NetLSD –Portrait Divergence (PDIV)
Directed Unweighted	–Euclidean, Manhattan, Canberra, Jaccard distances –DeltaCon	–Global statistics –DGCD-129 –MI-GRAAL –Portrait Divergence
Undirected Weighted	–Euclidean, Manhattan, Canberra distances –Weighted Jaccard distance (WJAC)	–Global statistics –Spectral Adjacency, Laplacian, SNL distances –MI-GRAAL –NetLSD –Portrait Divergence
Directed Weighted	–Euclidean, Manhattan, Canberra distances –Weighted Jaccard distance	–Global statistics –MI-GRAAL –Portrait Divergence

Table 1. Classification of network distances.

Metrics

degree
node_weight
edge_weight
node_edge_weight_consistency
degree_pearson_correlation_coeff
exponent
fitting curve: a*(t**b) into node degree

bridges_count, **node_connectivity**

number_of_isolates
dominating_set

average_shortest_path_length
eccentricity of a node v - the maximum distance from v to any other node
radius - min eccentricity
diameter - max eccentricity
center_count
“The center is the set of nodes with eccentricity equal to radius.”
periphery_count
“The periphery is the set of nodes with eccentricity equal to the diameter.”

k_core_size
“A k-core is a maximal subgraph that contains nodes of degree k or more.”

voterank_count
pagerank
based on webpage ranking and designed for directed graphs
largest_clique_fraction, **clique_above_5_fraction**
“For each node v, a maximal clique for v is a largest complete subgraph containing v. The largest maximal clique is sometimes called the maximum clique.“
clustering
“For unweighted graphs, the clustering of a node
u is the fraction of possible triangles through that node that exist,”
global_efficiency
“The efficiency of a pair of nodes in a graph is the multiplicative inverse of the shortest path distance between the nodes.”

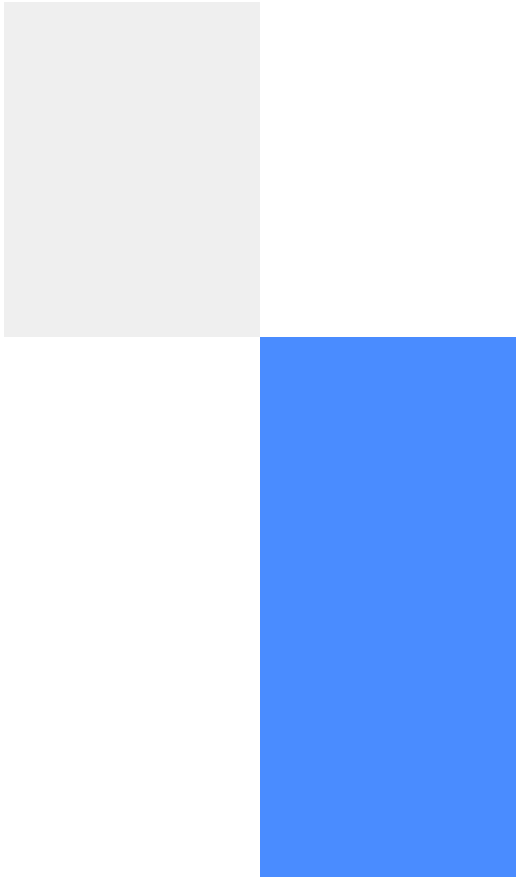
degree_centrality
“The degree centrality for a node v is the fraction of nodes it is connected to.”
betweenness_centrality
“Betweenness centrality of a node v is the sum of the fraction of all-pairs shortest paths that pass through v.”
closeness_centrality
“The closeness of a node is the distance to all other nodes in the graph or in the case that the graph is not connected to all other nodes in the connected component containing that node."
current_flow_closeness_centrality
“Current-flow closeness centrality is (...) based on effective resistance between nodes in a network.”
current_flow_betweenness_centrality
“Current-flow betweenness centrality uses an electrical current model for information spreading”

The charts

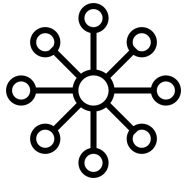
network metrics

No, they are NOT similar

Are they similar?

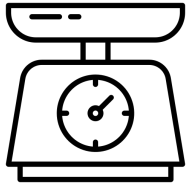


The welding



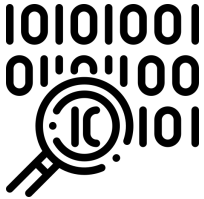
Node degree

Finding nodes with highest degree



Node weight

Finding nodes with highest weight (combined score)



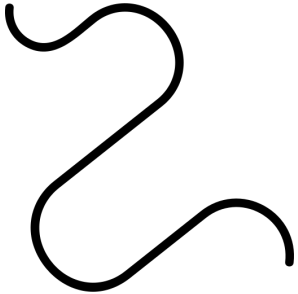
Node embedding

Convert nodes to vectors and find most similar

(selected) Graph network embedding

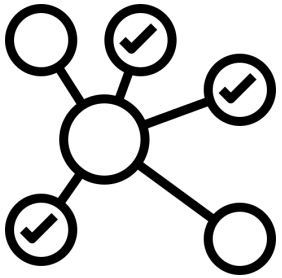
DeepWalk

Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "**Deepwalk: Online learning of social representations.**" In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 701-710. 2014.



LINE

Tang, Jian, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. "**Line: Large-scale information network embedding.**" In Proceedings of the 24th international conference on world wide web, pp. 1067-1077. 2015.



node2vec

Grover, Aditya, and Jure Leskovec. "**node2vec: Scalable feature learning for networks.**" In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 855-864. 2016.

Which one to choose?

Method

All algorithms are based on random walks

Social networks

All networks were tested on social network datasets on YouTube, Facebook, but not Reddit

Time

DeepWalk is super slow

NJIT Data Science Seminar: Steven Skiena, Stony Brook University

Accuracy

node2vec achieves similar results as DeepWalk, but faster

Arsov, Nino, and Georgina Mirceva. "**Network embedding: An overview.**" arXiv preprint arXiv:1911.11726 (2019).

node2vec - a few comments

- Performs **random walks**
- Can be considered an “extension” of
  - **DeepWalk**, which performs **un**
  - **LINE**, which focused on represe (**BFS**)
- 2 crucial parameters
  - Return parameter, **p**
    controls the likelihood of i
  - In-out parameter, **q**
    q > 1 causes bias for BFS, c
- Other parameters:
  - **number of walks** (higher = better), **walk length** (higher = better)
- Directly compares with DeepWalk and LINE in original paper
- **Scalability** tested on Erdos-Renyi graphs with sizes from 100 to 1,000,000
- Tested on real networks (**Facebook**, **PPI** - Protein-Protein Interactions, **arXiv** ASTRO-PH)

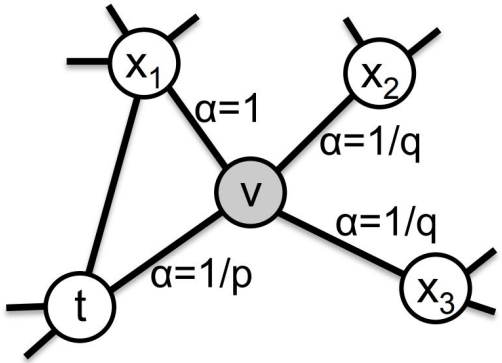
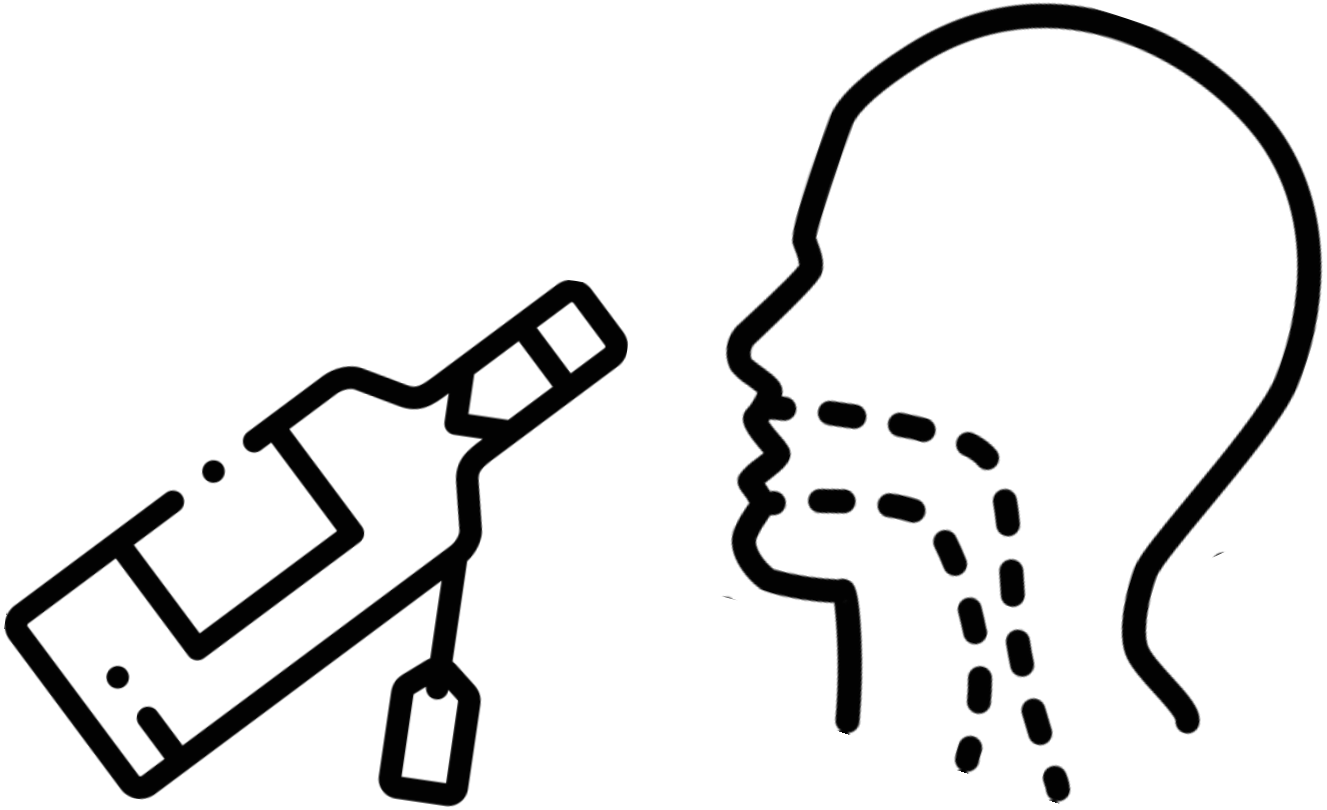
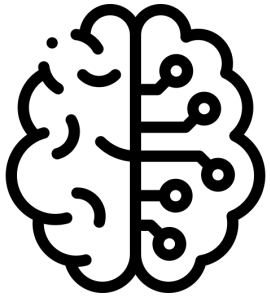


Figure 2: Illustration of the random walk procedure in *node2vec*. The walk just transitioned from t to v and is now evaluating its next step out of node v. Edge labels indicate search biases α.

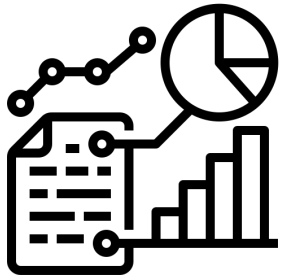


Bottlenecks

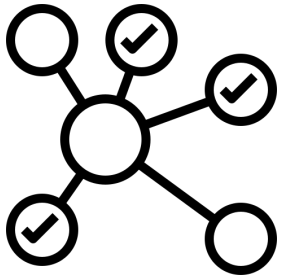
Posts download



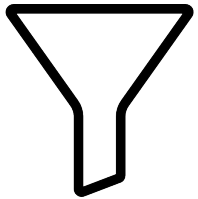
NER



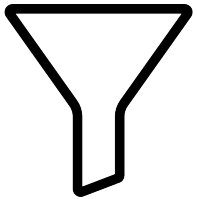
network **metrics**



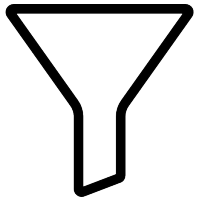
node2vec
cosine similarity



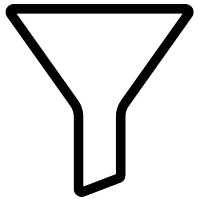
subreddit selection
(by **subscribers**)



posts filtering
(by **score**)



network reduction
(by **score**)



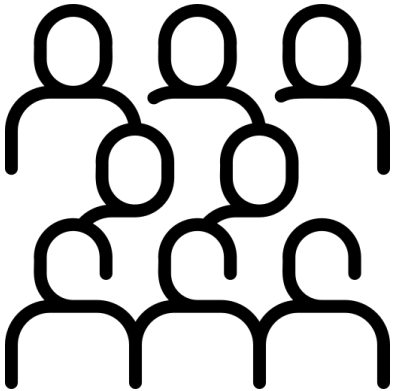
subreddit pairs
selection
(by **crossposts**)

Cool, cool.

How to evaluate this?

- Classification metrics - accuracy, recall, F1, ...
- NLP metrics - GLUE, BLUE, METEOR, ...
- Evaluate oneself
- Create a new metric
- Network metrics (PageRank)
- Semantic similarity using text embeddings

How to evaluate this?



Manual annotation

Ask annotators to
evaluate the results

Why Reddit?

Subreddits

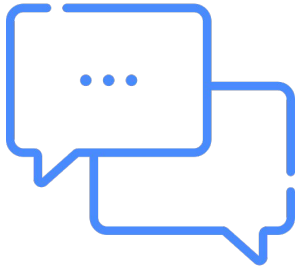
Twitter, Instagram don't have it Facebook has, but private + no API

Insight

Advanced and continually increased user-side expert knowledge of the platform

API

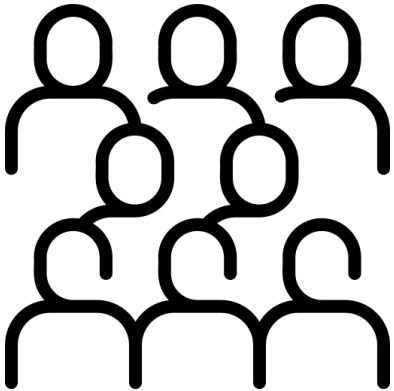
Pushshift is love, Pushshift is life.



?

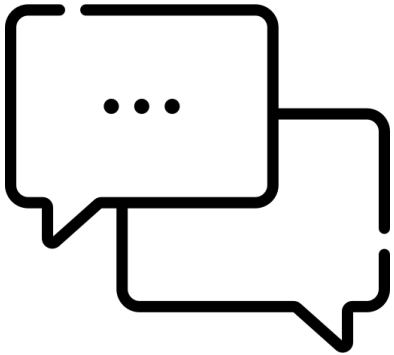
TBA

How to evaluate this?



Manual annotation

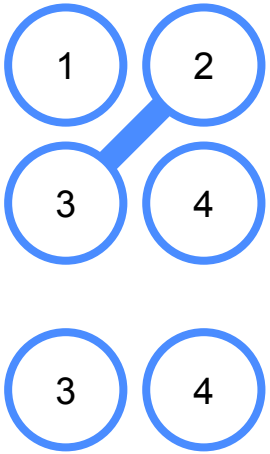
Ask annotators to evaluate the results



Crossposts

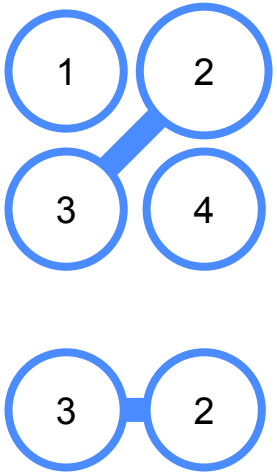
Posts can be “reposted” from one subreddit to another subreddit

Methodology shift



Old approach

1. Choose subreddits
2. Find crossposts



New approach

1. Find crossposts
2. Filter subreddits

Are crossposts a good estimator of interests between subreddit?

Well...

Crossposts?

Cons

- There is not many of them
- They not always have named entities at all
- Is is debatable whether they are the proper representation of cross-subreddit interests

Pros

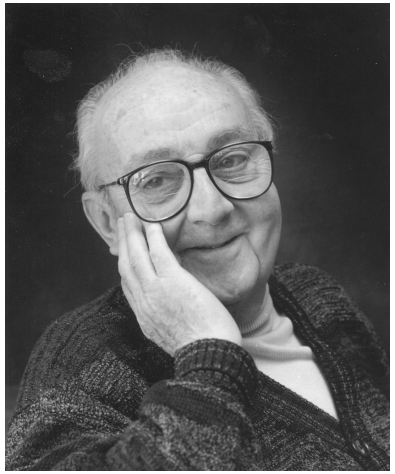
- They achieve high scores
- They are well spread in time (over the year)

*"An **approximate answer** to the **right question** is worth far more than a **precise answer** to the **wrong one**."*

Tukey, John W. "**The future of data analysis.**" The annals of mathematical statistics 33, no. 1 (1962): 1-67.

*"All **models are wrong**, but some are **useful**"*

George E. P. Box. "**Science and Statistics.**" *Journal of the American Statistical Association* 71, no. 356 (1976): 791–99. https://doi.org/10.2307/2286841.



Finis

Jan Sawicki

j.sawicki@mini.pw.edu.pl