Identifying scientific sub-literatures

NSF Specialists Meeting
August 2, 2016

Rob Malouf
San Diego State University

"You shall know a word by the company it keeps."

(J.R. Firth, 1957)

Meaning

In Standard American English (COCA), *reversible* has many senses, but *irreversible* is fairly restricted

There are 2,398 different words in your collocation database for "[word="irreversible"%c]". (Your query "irreversible" returned 1,019 matches in 842 different texts) [0.414 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Mutual information
1	reversible	628	0.007	6	6	9.743
2	coma	1,757	0.02	12	12	9.258
3	damage	25,784	0.288	105	103	8.509
4	degradation	2,569	0.029	10	9	8.445
5	neurological	1,595	0.018	5	5	8.134
6	catastrophic	2,756	0.031	5	5	7.343
7	impairment	3,484	0.039	5	5	7.002
8	Large-scale	4,054	0.045	5	3	6.786
9	commitments	4,073	0.046	5	5	6.78
10	decline	15,654	0.175	16	16	6.515

Meaning

A bottle of tezgüino is on the table.
Everyone likes tezgüino.
Tezgüino makes you drunk.
We make tezgüino out of corn.

Meaning

A bottle of tezgüino is on the table.
Everyone likes tezgüino.
Tezgüino makes you drunk.
We make tezgüino out of corn.

A bottle of _____ is on the table.
Everyone likes _____.
_____ makes you drunk.
We make _____ out of corn.

Meaning

A bottle of tezgüino is on the table.
Everyone likes tezgüino.
Tezgüino makes you drunk.
We make tezgüino out of corn.

A bottle of _____ is on the table.
Everyone likes _____.
_____ makes you drunk.
We make _____ out of corn.

Language variation

Language variation is a constant

Variation in space & time

Social variation

Communities of practice

“A community of practice is a collection of people who engage on an ongoing basis in some common endeavor: a bowling team, a book club, a friendship group, a crack house, a nuclear family, a church congregation”

(Lave & Wenger 1991; Eckert & McConnell-Ginet 1992; Eckert & McConnell-Ginet 1992)

Language variation

reversible vs. *irreversible*

Biomedical contexts:

> *Following a careful screening by a physician for concurrent medical illnesses and* ***reversible*** *causes of pain, persons with CBP undergo evaluation by physical and exercise therapists .*

> *In 1993 , the gene that determines the occurrence of Huntington's disease, an* ***irreversible*** *degeneration of the nervous system , was discovered .*

Language variation

Specialized corpus: 7,317 papers (38,314,863 words) on *non-small cell lung cancer*

There are 1,511 different words in your collocation database for "[word="irreversible"%c]". (Your query "irreversible" returned 665 matches in 338 different texts) [0.157 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood
1	inhibitor	14,683	1.529	103	61	667.649
2	inhibitors	14,363	1.496	92	32	579.564
3	an	92,781	9.662	117	89	371.961
4	EGFR	28,705	2.989	68	32	296.12
5	EGFR-TKIs	616	0.064	25	9	249.617
6	TKIs	1,617	0.168	29	9	241.688
7	TKI	1,875	0.195	24	14	183.778
8	ErbB	892	0.093	19	10	164.865
9	reversible	945	0.098	19	17	162.667
10	arrest	4,366	0.455	24	21	143.552

*Afatinib is an **irreversible** EGFR/HER2 inhibitor developed by Boehringer Ingelheim [11] currently being clinically evaluated in NSCLC .*

Lexicography

Corpus of articles on prescription opioid abuse

Closest synonyms of *recovery*

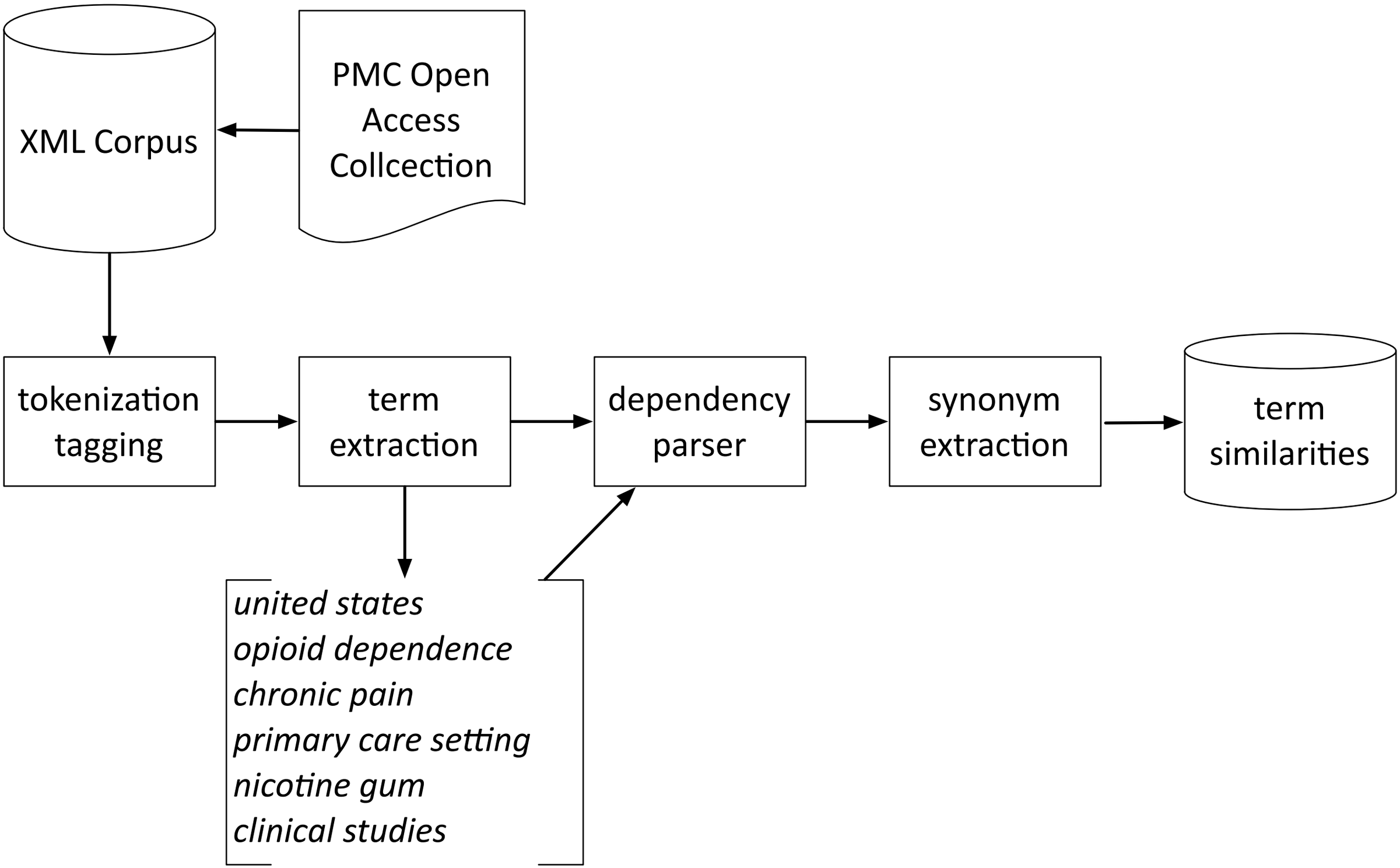
detoxification (0.203)
maintenance treatment (0.188)
therapy (0.174)
induction (0.166)
methadone (0.164)
withdrawal (0.161)
naltrexone (0.160)
treatment (0.156)
medication (0.152)
methadone maintenance (0.151)
mmt (0.150)
dose (0.150)
buprenorphine (0.149)
intervention (0.148)
treatments (0.147)
management (0.145)
administration (0.143)
pharmacotherapy (0.143)
placebo (0.141)
abstinence (0.141)

Lexicography

Corpus of articles on an experimental antidepressant

Closest synonyms of *recovery*

remission (0.127)
relapse (0.096)
inhibition (0.076)
decline (0.073)
improvement (0.073)
recurrence (0.069)
tolerance (0.067)
success (0.065)
examination (0.065)
investigation (0.063)
preoccupation (0.060)
complication (0.060)
card (0.060)
assessment (0.059)
increase (0.059)
episode (0.056)
eect (0.055)
antidepressant action (0.055)
conversation (0.055)
reduction (0.054)


XML Corpus
PMC Open Access Collcection
tokenization tagging
term extraction
dependency parser
synonym extraction
term similarities
united states
opioid dependence
chronic pain
primary care setting
nicotine gum
clinical studies

Building a corpus

Pubmed Central Open Access Collection

A search for *depression* yields a lot of marginally relevant material

MeSH categories

Mental Disorders [F03]
Mood Disorders [F03.600]
▶ Depressive Disorder [F03.600.300]
Depression, Postpartum [F03.600.300.350]
Depressive Disorder, Major [F03.600.300.375]
Depressive Disorder, Treatment-Resistant [F03.600.300.388]
Dysthymic Disorder [F03.600.300.400]
Premenstrual Dysphoric Disorder [F03.600.300.550]
Seasonal Affective Disorder [F03.600.300.775]
Cyclothymic Disorder [F03.600.500]

Focused search turns up 2,487 articles with 9,744,106 words

Building a corpus

All of the articles are ‘about’ depressive disorders, but . . .

1. Relationship between G1287A of the NET Gene Polymorphisms and Brain Volume in Major **Depressive Disorder**: A Voxel-Based MRI Study
   Issei Ueda, Shingo Kakeda, Keita Watanabe, Reiji Yoshimura, Taro Kishi, Osamu Abe, Satoru Ide, Junji Moriya, Asuka Katsuki, Hikaru Hori, Nakao Iwata, Jun Nakamura, Yukunori Korogi
   PLoS One. 2016; 11(3): e0150712. Published online 2016 March 9. doi: 10.1371/journal.pone.0150712
   PMCID: PMC4784887
   Article PubReader PDF–826K Citation

2. The White Matter Microintegrity Alterations of Neocortical and Limbic Association Fibers in Major **Depressive Disorder** and Panic Disorder: The Comparison
   Chien-Han Lai, Yu-Te Wu
   Medicine (Baltimore) 2016 March; 95(9): e2982. Published online 2016 March 7. doi: 10.1097/MD.0000000000002982
   PMCID: PMC4782901
   Article PubReader PDF–361K Citation

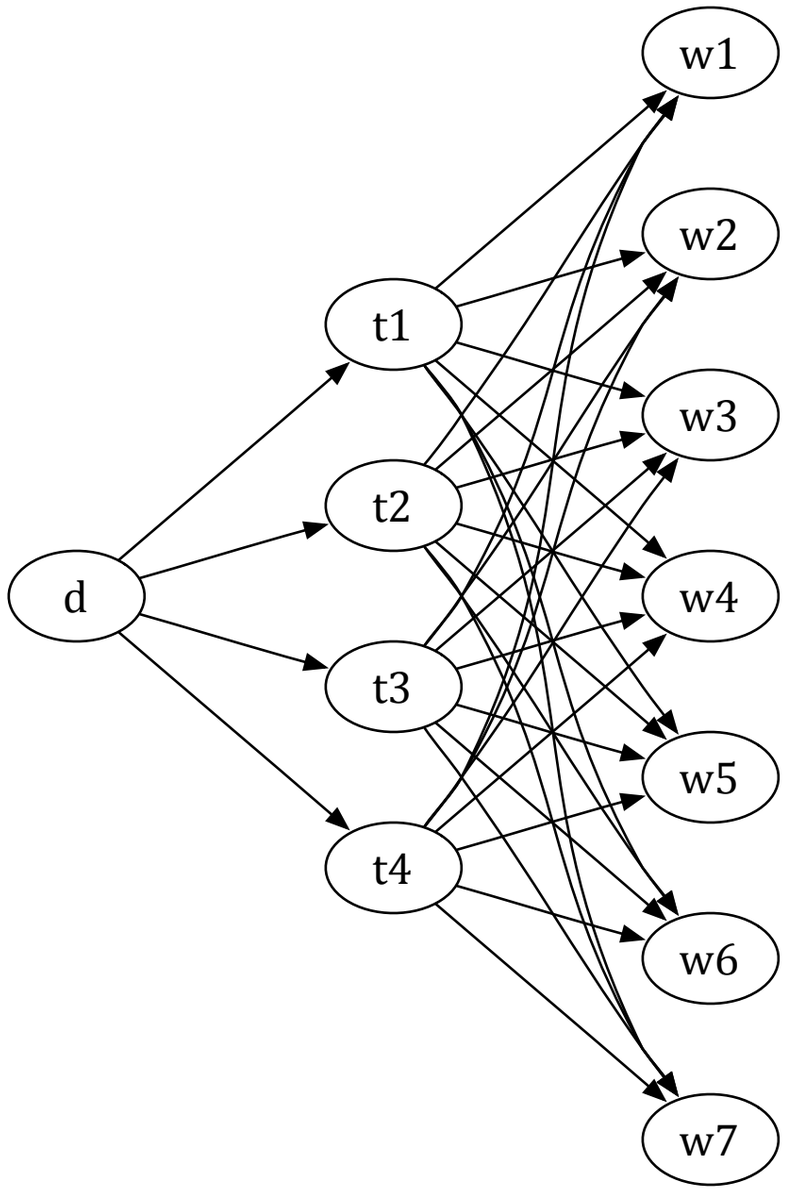
3. Construct Validity of the SF-12v2 for the Homeless Population with Mental Illness: An Instrument to Measure Self-Reported Mental and Physical Health
   Antony Chum, Anna Skosireva, Juliana Tobon, Stephen Hwang
   PLoS One. 2016; 11(3): e0148856. Published online 2016 March 3. doi: 10.1371/journal.pone.0148856
   PMCID: PMC4777288
   Article PubReader PDF–441K Citation

4. German Translation and Validation of the Cognitive Style Questionnaire Short Form (CSQ-SF-D)
   Quentin J. M. Huys, Daniel Renz, Frederike Petzschner, Isabel Berwian, Christian Stoppel, Helene Haker
   PLoS One. 2016; 11(3): e0149530. Published online 2016 March 2. doi: 10.1371/journal.pone.0149530
   PMCID: PMC4774974
   Article PubReader PDF–284K Citation

5. Prevalence and Predisposing Factors for Depressive Status in Chinese Patients with Obstructive Sleep Apnoea: A Large-Sample Survey
   Yaozhang Dai, Xuewu Li, Xin Zhang, Sihua Wang, Jianzhong Sang, Xiufen Tian, Hua Cao
   PLoS One. 2016; 11(3): e0149939. Published online 2016 March 2. doi: 10.1371/journal.pone.0149939
   PMCID: PMC4774961
   Article PubReader PDF–215K Citation

Building a corpus

Topic models (e.g., LDA) use word co-occurrence statistics to infer the latent topic structure of a corpus



Building a corpus

Topic models (e.g., LDA) use word co-occurrence statistics to infer the latent topic structure of a corpus

Top articles in topic #23:

> "Extracting labeled topological patterns from samples of networks"
>
> "Friendship context matters: Examining the domain specificity of alcohol and depression socialization among adolescents"
>
> "An analytical approach to network motif detection in samples of networks with pairwise different vertex labels"
>
> "Explaining ecological clusters of maternal depression in south Western Sydney"

Building a corpus

Topic models perform poorly when faced with a mix of very closely related document types

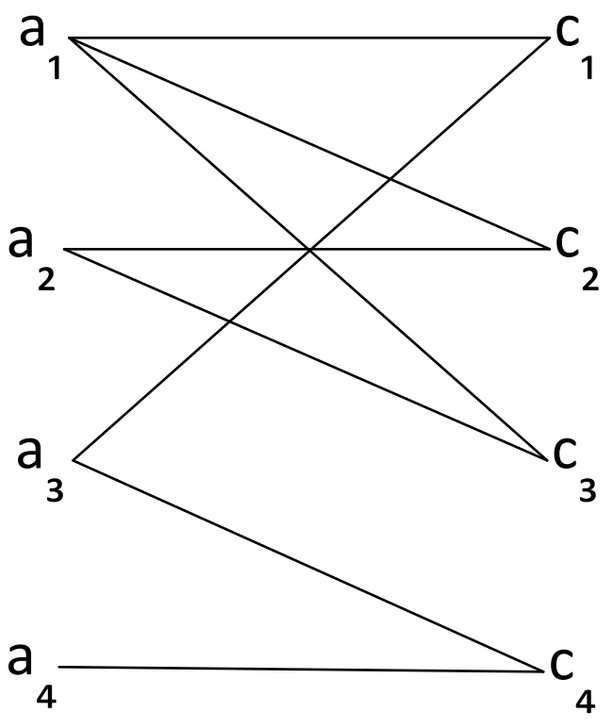
To be valid, sub-corpora must be based on external, non-linguistic criteria

Communities of practice are defined by what they **do**, not how they use language

Citation patterns

Academics use citation to (among other things) situate themselves in a community of practice

Bipartite graph linking articles and their citations



Citation patterns

When do two citations refer to the same source?

Corpus includes 23,967 references with DOIs with 3,659 cited more than once

Exact string match on the entire reference:

99.74% precision, 12.18% recall

Exact string match on the title:

96.67% precision, 56.63% recall

Normalized title match:

96.95% precision, 87.58% recall

Normalized de-stopped title match:

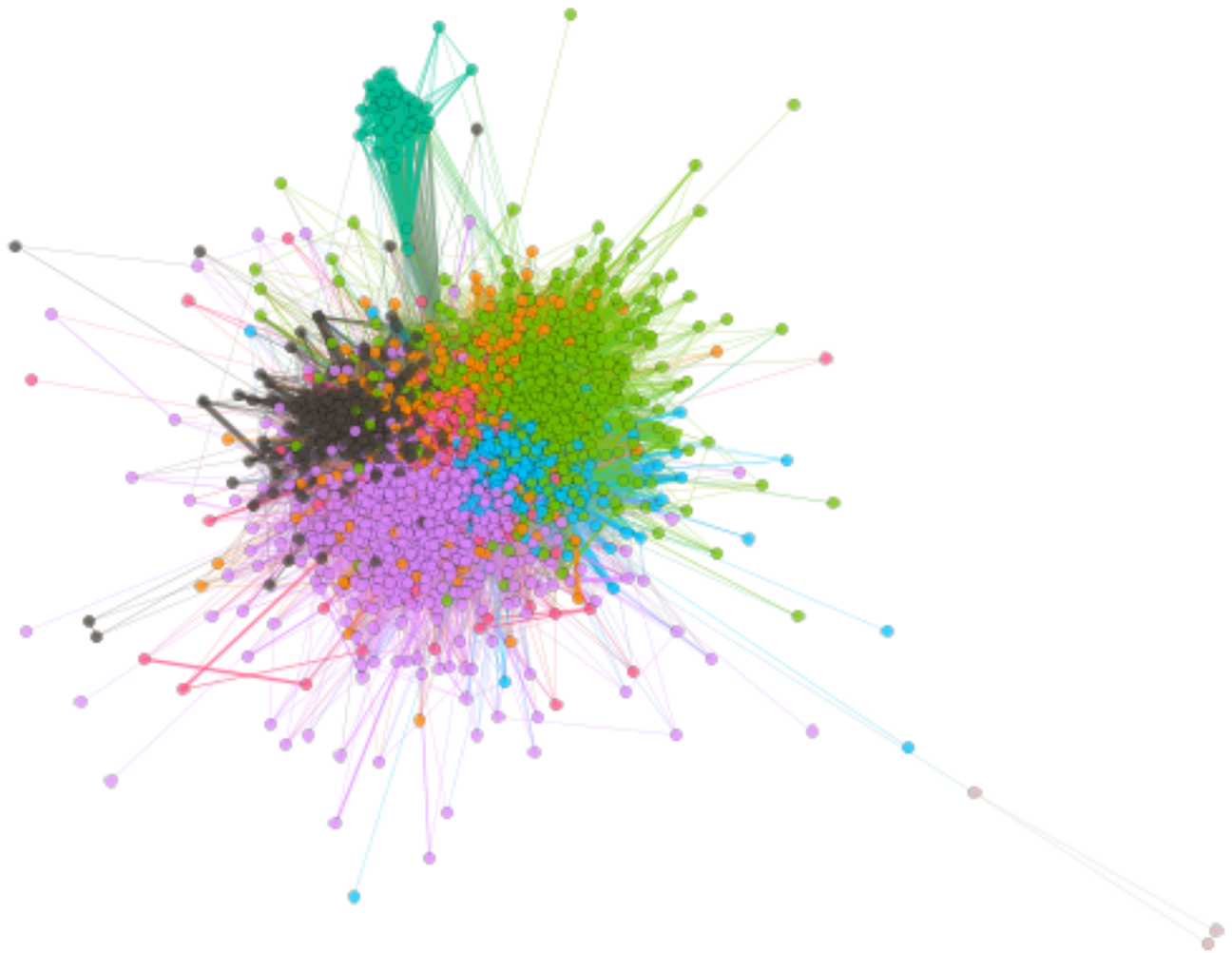
97.03% precision, 88.62% recall

Citation patterns

Academics use citation to (among other things) situate themselves in a community of practice

Bipartite graph linking articles and their citations

Louvain community detection on weighted article projection



Citation patterns

Keywords from the six meaningful communities

mice, brain, mdd, expression, health, hippocampus, rats, intervention, mental_health, protein

mdd, phq-9, health, diabetes, migraine, brain, care, intervention, postpartum, primary_care

placebo, patients, treatment, madrs, escitalopram, women, vortioxetine, trials, remission, baseline

women, postpartum, epds, mothers, maternal, patients, pregnancy, postnatal, ppd, during_pregnancy

hwa-byung, ptsd, rumination, ogm, ruminative_self-focus, imagery, melancholia, amt, memory_specificity, pe

snps, gwas, snp, genetic, disorder, pclo, mdd, disorders, adolescents, association

Prospects

Open source scalable tools

> Python, nltk, networkx, gensim, dask, spark, theano, spark, mongodb

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

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

Theory helps

It’s not always obvious what’s going to be hard

> Multi-word terms, citation matching, corpus partitioning