Western

UNIVERSITY · CANADA

Exploratory Search with Archetype-based Language Models

By Brent D. Davis (Now PhD)
(Former) Member of the Insight & Phi Labs @ Western University
Co-Supervised by Drs. Lizotte & Sedig
Committee: Drs. Bauer & Mercer

Western Science

Table of Contents

Content Warning / Disclaimer

- This work contains results from analysis of social media on the subjects of opioid drugs.
- Results include various drug names and other offensive slang
- Depression and associated symptoms are commented on

Overview & Motivation

- Finding information related to a topic is a common task
- Existing information retrieval systems are mostly based around asking the searcher for specialized keywords
- This can be difficult; often the optimal vocabulary is unknown.
- Is there helpful structure in written content that natural language processing and machine learning can discover to help complex search tasks?

Overview & Motivation

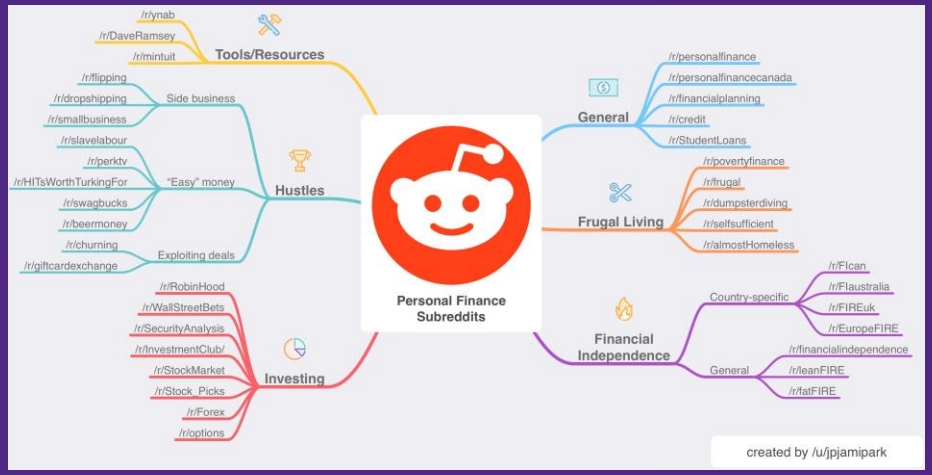
- The output of this work is 3 algorithms which are suitable for exploratory search of documents.
- While generic, specific instantiations of each algorithm are used to prove their applied utility

Definitions

- Unknown Vocabulary Problem
- Exploratory vs Look-up Search
- Archetype
- Representation

Definitions

- Reddit, Subreddit
- Classifier
- Associated Language
- Temporal Representation

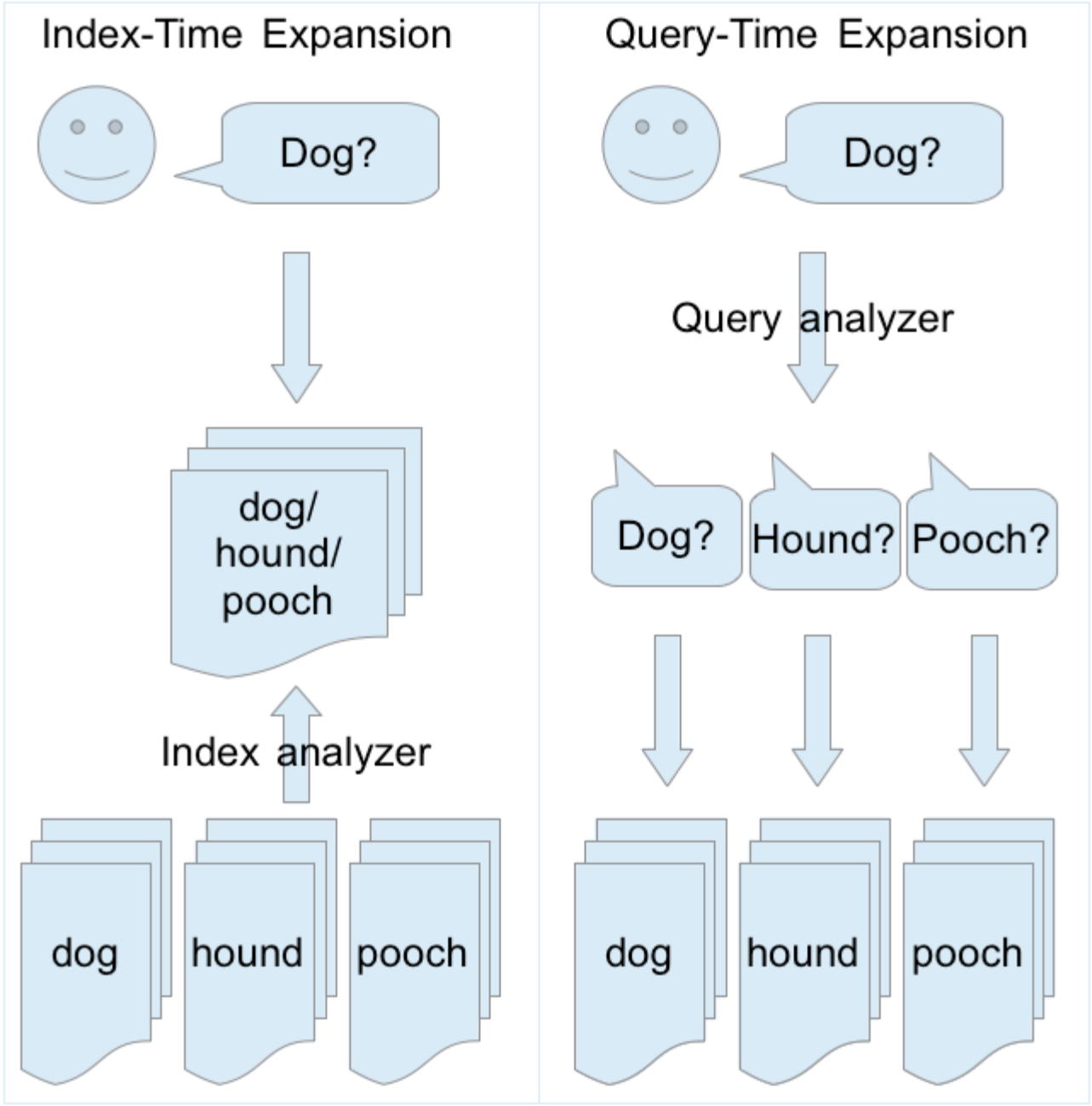


Background

- Related technique: Query Expansion
- Word representations via GloVe
- Author representations via usr2vec

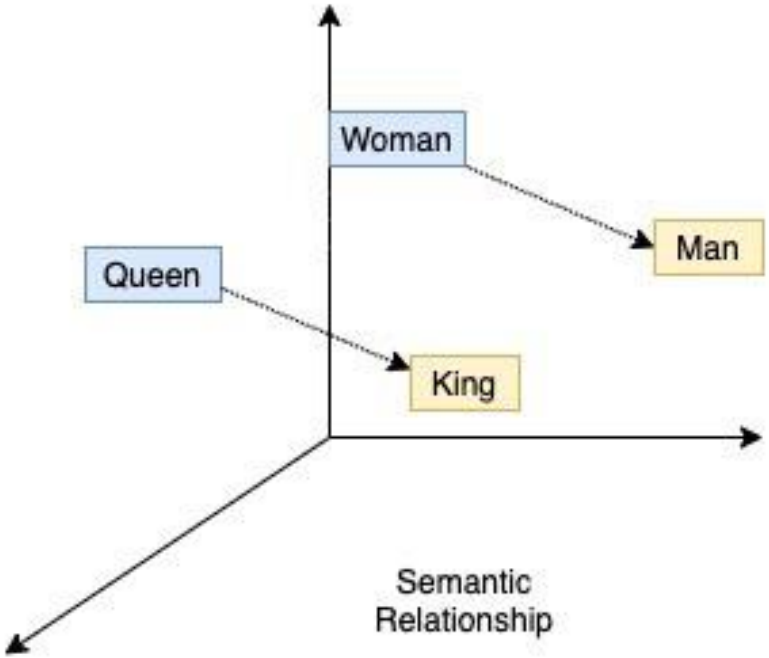
Query Expansion

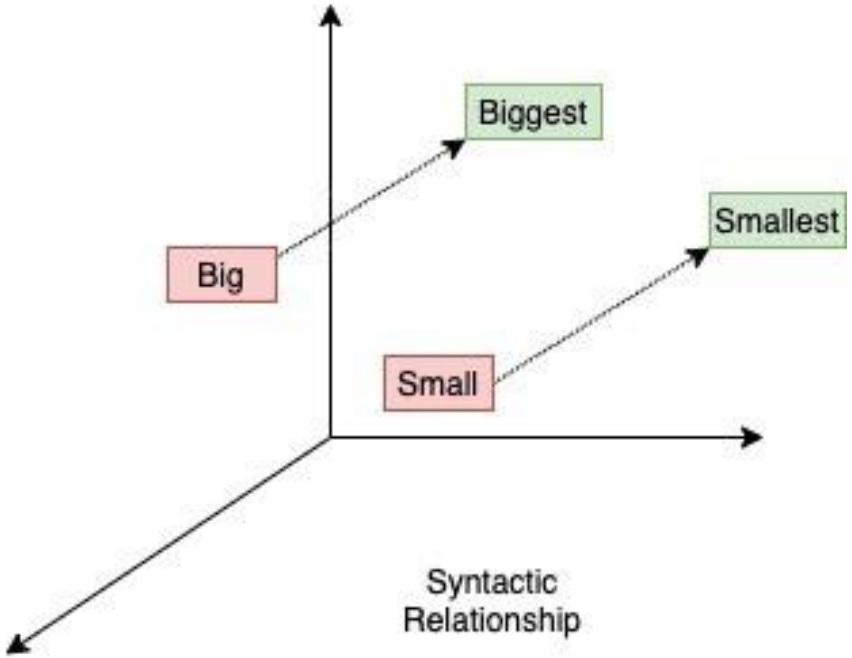
- Big idea of query expansion is to take a pre-formed query and add additional related terms that enhance the search results
- Vector representations of words are well-suited since they can have numeric similarity or distance
- Can expand a query with word vectors by finding *k* nearest neighbours in vector space


Index-Time Expansion
Dog?
dog/
hound/
pooch
Index analyzer
dog
hound
pooch
Query-Time Expansion
Dog?
Query analyzer
Dog?
Hound?
Pooch?
dog
hound
pooch

Training Representations: Word & Author

- Word embeddings are learning associations between words and contexts they are used
- We want ‘happy’ and ‘joy’ to be closer to each other than ‘small’ and ‘large’ would be
- Abstractly, this idea of similarity transfers into author representations
- We want authors who talk about similar topics to be near or similar to each other


Woman
Man
Queen
King
Semantic
Relationship


Biggest
Smallest
Big
Small
Syntactic
Relationship

Training Word Representations

- Global Vectors (GloVe) by Pennington, Socher & Manning (2014) is an unsupervised learning algorithm that is essentially a log-bilinear model with a weighted least-squares objective. From:

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

To:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \hat{w}_j + b_i + \hat{b}_j - \log X_{ij})^2$$

Where f(X) is a weighting function to scale rare and very frequent co-occurrences, *w* is the word vector, *w*-hat is the context vector, *V* is the vocabulary size, and *b's* are bias terms to ensure symmetry in the word co-occurrence matrix ex: (cat,dog) = (dog,cat)

Training Author Representations

- Usr2vec, the original technique for learning author representations, is a 2017 paper by Amir *et al.,*
- Focus is on aggregating an author's linguistic habits as represented by word choices
- This multi-stage representation is intended to allow complex representation of linguistic similarities that allow for an archetypal label to have meaning.

Training Author Representations

- Want to estimate:

$$P(C_j|u_j) \propto \sum_{SC_j} \sum_{w_j \in S} logP(w_i|u_j)$$

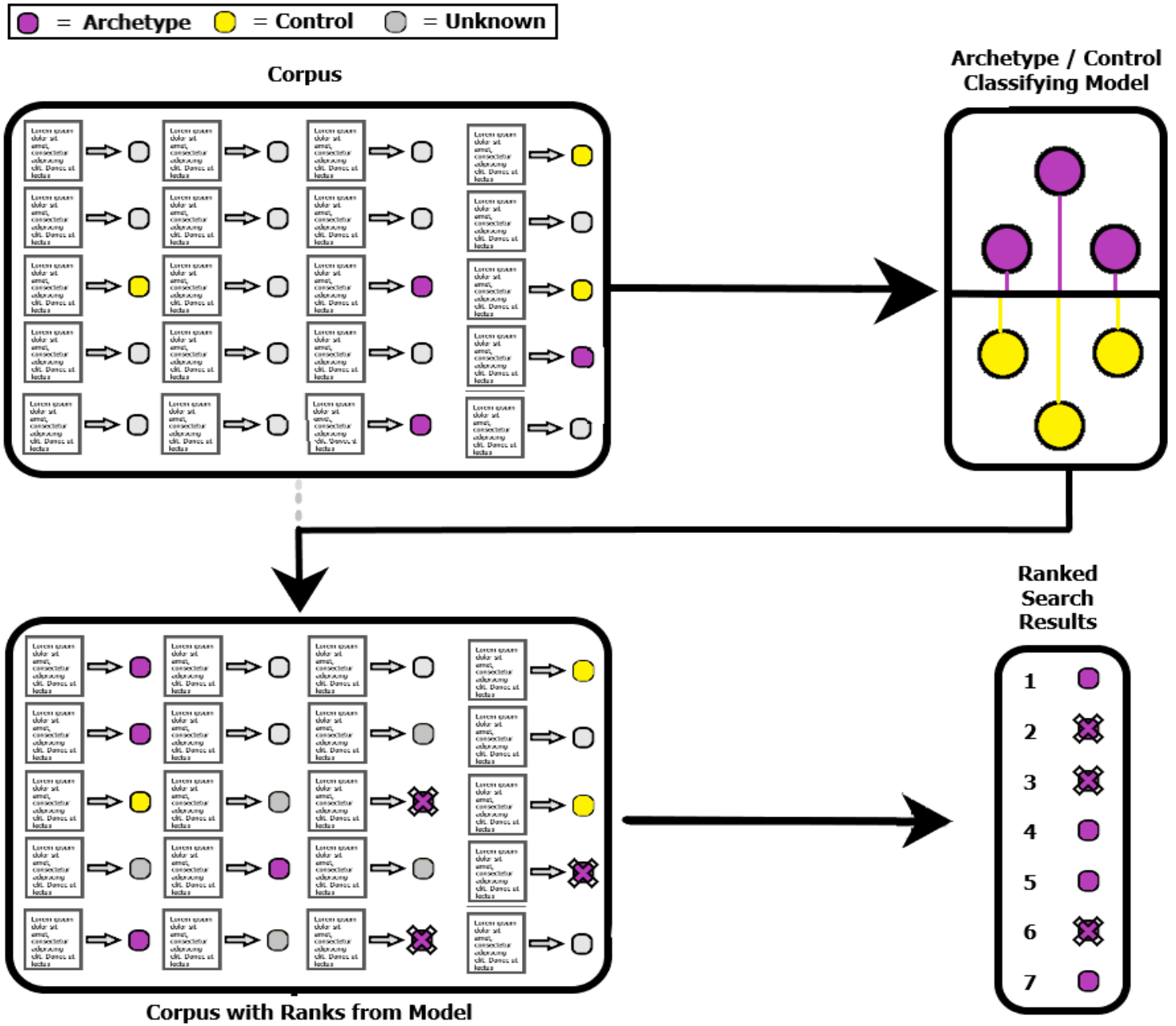
- This is computationally expensive to directly estimate so instead we approximate the probability with:

$$L(w_i, u_j) = \sum_{\widetilde{w}_k \in V} \max(0, 1 - w_i \cdot u_j + \widetilde{w}_k \cdot u_j)$$

Archetype-based Modeling and Search (ABMS)

Motivation:

- Exploratory search needs to go beyond the way people typically search. Utility from finding things that people would miss otherwise
- Vector representations of authors exist in high dimensional space, and are trained to capture similarities and put them near each other.
- How can we use this to explore?


= Archetype
= Control
= Unknown
Corpus
Archetype / Control Classifying Model
Ranked Search Results
1
2
3
4
5
6
7
Corpus with Ranks from Model

Archetype-based Modeling and Search (ABMS)

Algorithm 1: Archetype-Based Modeling and Search

- Input D, a set of documents, which contains the following subsets: A, a subset of archetypes (documents of interest), C a subset of controls (documents not of interest), and U, an unlabeled subset of D, which will be searched to identify additional documents of interest.
- Use the words in all documents in A and develop a word representation W that maps words to vectors.
- Use W and develop a document representation, V, that maps all documents in D to vectors.
- Train a classifier to distinguish V(A) from V(C). The classifier must be able to rank inputs according to their likelihood of belonging to A versus C.
- Apply the classifier to V(U) and rank the unknown documents.

Return as top-ranked documents those most likely to be of interest.

Archetype-based Modeling and Search (ABMS)

- Published in Big Data & Cognitive Computing as “Archetype-based Modeling and Search of Social Media”
- Training of author representations required 18 CPU years, distributing 47,000 neural representations of authors across Compute Canada & sharcnet infrastructure
- https://www.mdpi.com/2504-2289/3/3/44

Western Science

ABMS Limitations

- ABMS requires direct computation of a representation for every author in the corpus of interest.
- Many document types or platforms are not well suited to the focus on author representations
- Some predictions by the model can be difficult to explain.
- Strong correlations may or may not be related to the archetype we wished to pursue

Archetype-based Information Retrieval (AIR)

Motivation:

- ABMS offers exploratory utility, but is computationally expensive and difficult to interpret

- How can we reduce or remove the need to perform computation on target corpuses of large size?

- How can we understand how the model is making decisions?

Searching with AIR

Insight:

- Word representations and author representations exist in the same high dimensional space
- Further, author representations are aggregations of word representations.
- Support Vector Machines (SVMs) produce a separating hyperplane and an associated ‘decision direction’. What words align with this?

Western Science

AIR – BM25 and Elasticsearch

- Information retrieval using keywords is very common – consider the search function on many popular sites (Google, Facebook, Twitter, etc)
- Retrieval systems use these keywords with one or more metrics to produce a ranked score.
- Common ones include *tf-idf* and BM25

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

$$IDF(q_i) = ln\left(\frac{N - n(q_i) + 0.5}{n(q_{i)} + 0.5} + 1\right)$$

Archetype-based Information Retrieval (AIR)

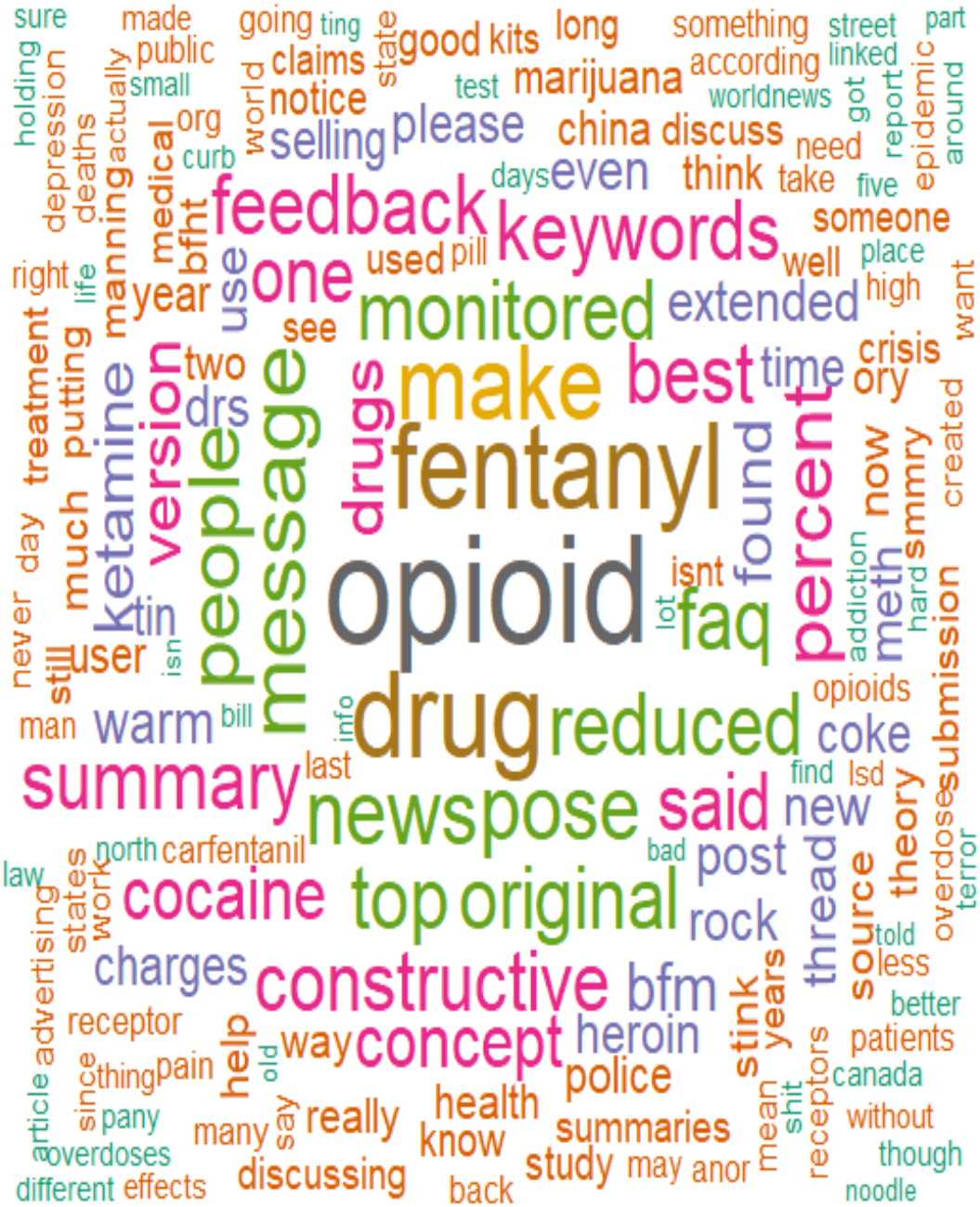
Algorithm 2: Archetype-Based Information Retrieval

- Input D, a set of documents, which contains the following subsets: A, a subset of archetypes (documents of interest), C a subset of controls (documents not of interest), U, an unknown corpus from which to extract relevant documents and q, the number of words to use in the query on U.
- Use the words in all documents in A and develop a word representation W that maps words to vectors, with these words { $w_1, w_2, \ldots, w_i$ } $\in$ W
- Use W and develop a document representation, V, that maps all documents in D to vectors.
- Train a classifier to distinguish V(A) from V(C). The classifier must be able to rank inputs according to their likelihood of belonging to A versus C and output a separating hyperplane H in the form of an q dimensional vector of the same dimensionality as all w $\in$ W.
- Apply a similarity metric (cosine or dot product) between all values w $\in$ W & H
- Retrieve the top q most similar w from sim(w, H) and collect into query Q.
- Index all documents in U with an information retrieval system
- Run a query with keywords Q using similarity metric sim (Ex: BM25). Return the ranked set of results as output.

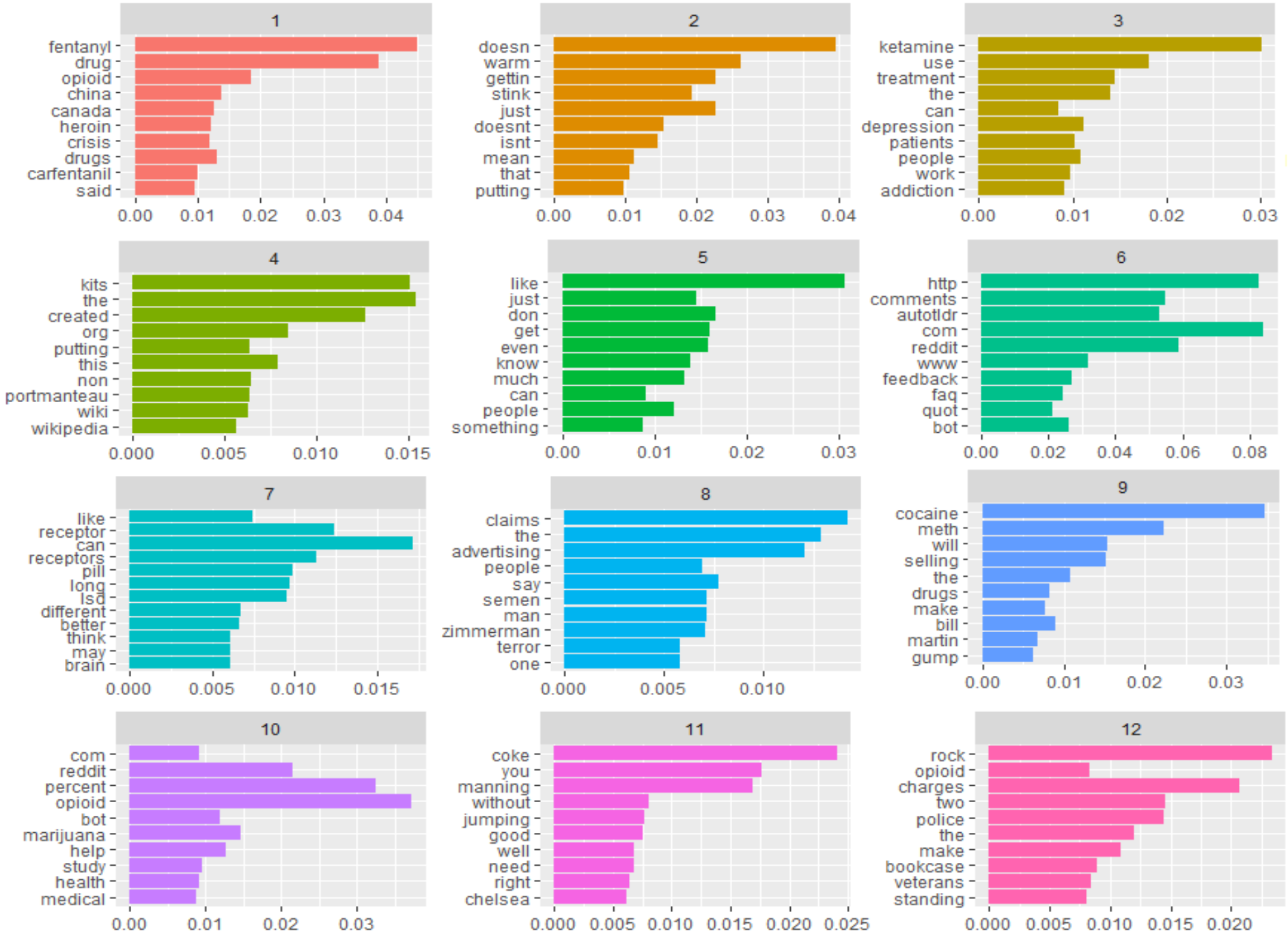
Searching with AIR

- Indexing and querying documents is very fast compared to training author representations
- Keywords extracted can be manually reviewed; this facilitates human-in-the-loop oversight
- Question: How do we know we're finding related content? Reviewing every post is infeasible at scale.

Western Science


opioid
fentanyl
drug
message
make
people
reduced
newspose
monitored
feedback
keywords
best
original
top
constructive
summary
cocaine
drugs
percent
concept
said
faq
found
ketamine
version
one
extended
bfm
heroin
please
even
selling
thread
police
coke
post
rock
new
charges
warm
time
meth
marijuana
china
discuss
think
claims
notice
good
kits
long
something
according
street
part
linked
worldnews
test
sure
made
going
ting
state
public
small
org
curb
world
actually
holding
depression
deaths
manning
medical
bfht
use
year
used
pill
days
need
take
five
got
report
epidemic
around
someone
well
place
high
want
right
life
see
two
drs
crisis
ory
now
smmry
created
treatment
putting
much
tin
user
still
never
day
isn
info
isnt
lot
addiction
hard
submission
opioids
find
lsd
theory
overdose
terror
man
bill
last
north
carfentanil
bad
law
states
work
advertising
article
since
thing
pain
pany
overdoses
different
effects
receptor
help
old
way
say
many
really
health
know
back
discussing
summaries
study
may
anor
mean
shit
receptors
stink
years
source
told
less
better
patients
canada
without
though
noodle


1
two
year
years
police
last
said
day
man
city
old
0.00
0.01
0.02
0.03
0.04
2
think
know
something
point
isn
even
way
actually
didn
say
0.00
0.02
0.04
0.06
3
use
never
game
good
enough
user
pretty
great
still
sub
0.00
0.01
0.02
0.03
0.04
4
new
make
may
money
pany
change
system
used
pay
million
0.00
0.02
0.04
0.06
5
trump
government
president
state
said
country
white
russia
house
politics
0.00
0.02
0.04
0.06
0.08
6
now
time
back
right
since
well
one
around
really
bad
0.00
0.02
0.04
0.06
7
news
concept
found
post
source
notice
discuss
please
submission
user
0.00
0.02
0.04
0.06
8
people
one
many
even
women
life
canada
say
health
men
0.00
0.05
0.10
9
watch
youtube
context
videos
video
mentioned
playlist
subtletv
hard
play
0.00
0.03
0.06
0.09
10
want
work
going
need
good
take
look
better
let
left
0.00
0.01
0.02
0.03
0.04
0.05
11
version
best
bfm
original
faq
top
reduced
keywords
monitored
feedback
0.00
0.02
0.04
0.06
12
message
pose
anor
someone
place
law
case
court
vote
rules
0.00
0.02
0.04
0.06
0.08


1
fentanyl
drug
opioid
china
canada
heroin
crisis
drugs
carfentanil
said
0.00
0.01
0.02
0.03
0.04
2
doesn
warm
gettin
stink
just
doesnt
isnt
mean
that
putting
0.00
0.01
0.02
0.03
0.04
3
ketamine
use
treatment
the
can
depression
patients
people
work
addiction
0.00
0.01
0.02
0.03
4
kits
the
created
org
putting
this
non
portmanteau
wiki
wikipedia
0.000
0.005
0.010
0.015
5
like
just
don
get
even
know
much
can
people
something
0.00
0.01
0.02
0.03
6
http
comments
autotldr
com
reddit
www
feedback
faq
quot
bot
0.00
0.02
0.04
0.06
0.08
7
like
receptor
can
receptors
pill
long
lsd
different
better
think
may
brain
0.000
0.005
0.010
0.015
8
claims
the
advertising
people
say
semen
man
zimmerman
terror
one
0.000
0.005
0.010
9
cocaine
meth
will
selling
the
drugs
make
bill
martin
gump
0.00
0.01
0.02
0.03
10
com
reddit
percent
opioid
bot
marijuana
help
study
health
medical
0.00
0.01
0.02
0.03
11
coke
you
manning
without
jumping
good
well
need
right
chelsea
0.000
0.005
0.010
0.015
0.020
0.025
12
rock
opioid
charges
two
police
the
make
bookcase
veterans
standing
0.000
0.005
0.010
0.015
0.020
Beta Score

AIR & Query Expansion

- We also compared the results of AIR with /r/Opiates against a query expansion from the same GloVe word embedding used to train the model.

- Two observations:
  - There is overlap in AIR and query expansion results.
  - AIR *may* find more that describe personal experience
  - Top 10 results of AIR vs Query Expansion on "Opioid" contain more describing personal experience.

AIR Limitations

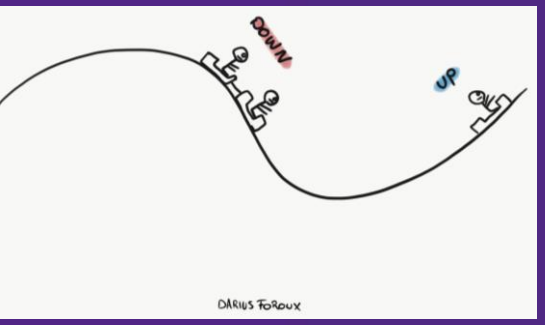
- The words that AIR produces may not be interpretable as to why they are relevant
  - AIR as an algorithm does not depend on the unit of analysis being individual words; bi, tri,n-grams or full sentences are suitable for future work.

- Some of the use case applications for AIR have a temporal component (i.e., binge drinking episodes).
  - AIR uses a single representation that does not include any adjustment for the time period the representation spans

AIR Application

- Recently published in International Journal of Population Data Science (IJPDS): https://ijpds.org/article/view/1716
- Focuses on applications of monitoring population-wide frequency of terms related to depression.
- Proof of concept of more general applicability of approach beyond the usage examining opioid usage across geographies.

Archetype-based Temporal Language Adaptive Stratification (ATLAS)



Motivation:

- ATLAS provides a way to look for change in a single author's archetypal behaviours
- Some behaviours of interest do display periodic or transitive behaviour.
  - Depression can come in cycles; drug use can come and go.
- Language is part of how we build our identity online; can we better explore how identity changes?

ATLAS – Incorporating, Building on ABMS & AIR

- Both ABMS and AIR produce a score; ATLAS weighs each score equally to produce a hybrid weighting.
  - The weighting is an adjustable hyperparameter.
  - Excluding one or the other still allows ATLAS to function.

ATLAS

Algorithm 3.1: Archetype-based Temporal Language Adaptive Stratification - Explore

Train word embedding W on documents $AD \cup BD$.

Train set of author representations Q for each $a_i \in AD$ and $b_i \in BD$

Train set of time separated author representations for each $aq_it_m \in QT$

Train a classifier C to separate aq from bq and retain all classification scores for aq.

Extract decision direction from C, calculate similarity of all $w_i \in W$ and sort

Index all AD using some information retrieval system i.e., Elasticsearch.

Take top x keywords and perform a query on AD with similarity score, i.e., BM25.

Identify all $aq_it_m \in QT$ with $\gamma s(C)+((1-\gamma)s(E)) > archetypal\_threshold$

With a list l that has as elements pairs of aq_it_m, collect items into the list as follows:
For $i = 1$; i < length(faq); i++ do
    for each representation j with the same author as i where $t_j > t_i$ & j has
    $(0.5s(C)+0.5s(e)) < nonarchetypal\_threshold$, append pair i,j to l.

Output list l, containing every pair of archetypal-to-nonarchetypal representations

ATLAS

Algorithm 3.2: Archetype-based Temporal Language Adaptive Stratification - Map

For every pair aq_it_m, aq_jt_n, $\in l$:

Train a classifier to distinguish between the two representations.

With the decision direction from the classifier, extract the top *y* keywords By similarity metric, i.e., cosine or dot product with words $w_i \in W$

Generate a word cloud of the top *y* keywords with word size proportional to the magnitude of *y* with the decision direction via the similar metric.

For every author $a_i \in AD$ with at least one archetypal-to-nonarchetypal pair, do:

Collect all $aq_it_m \in QT$ with $(0.5s(C)+0.5s(E)) > archetypal\_threshold$

Collect all $aq_jt_n \in QT$ with $(0.5s(C)+0.5s(E)) < nonarchetypal\_threshold$

Train a classifier to distinguish between the collection of archetypal representation against nonarchetypal representations

Extract the top *y* keywords from author vocabulary using a similarity metric between the decision direction and $w_i \in W$

Generate a word cloud of the top *y* keywords with word size proportional to the magnitude of *y* with the decision direction via the similar metric.

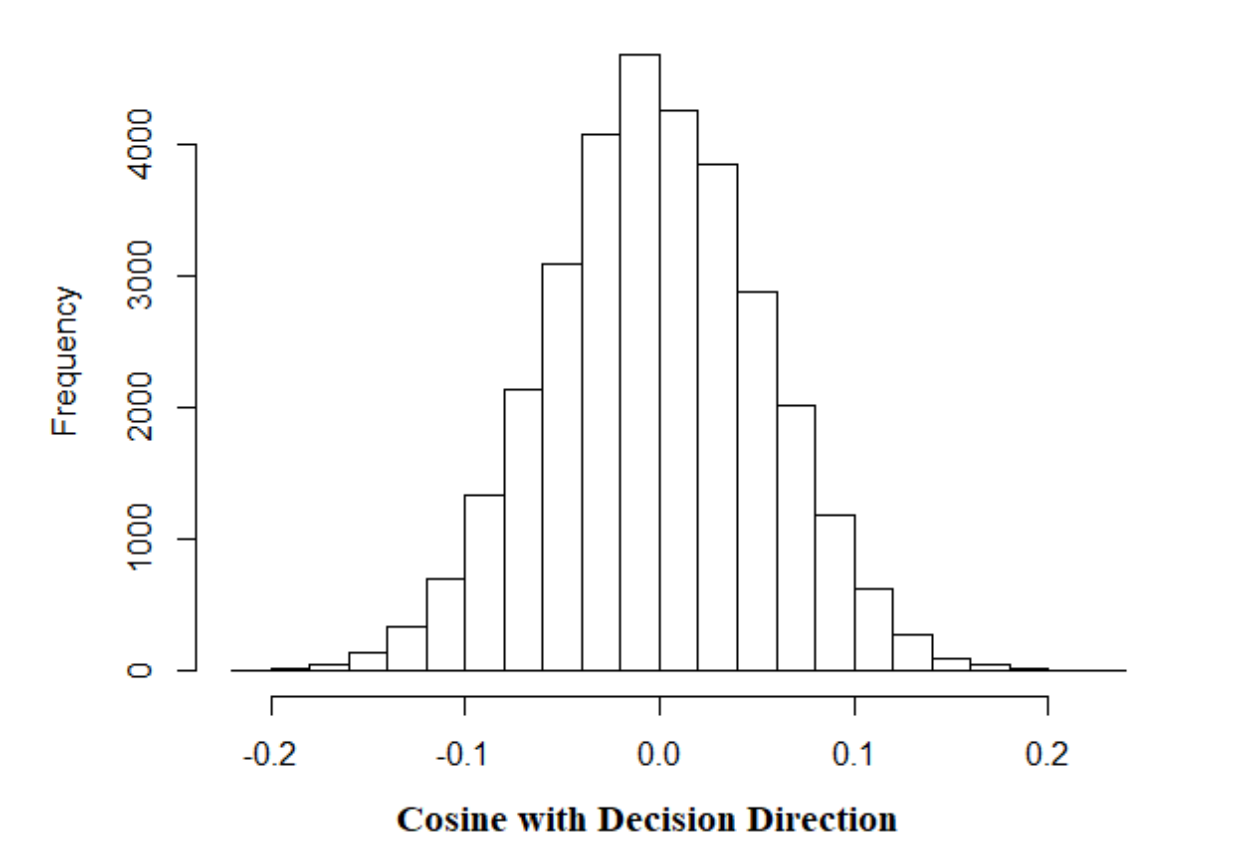
ATLAS – Case Study

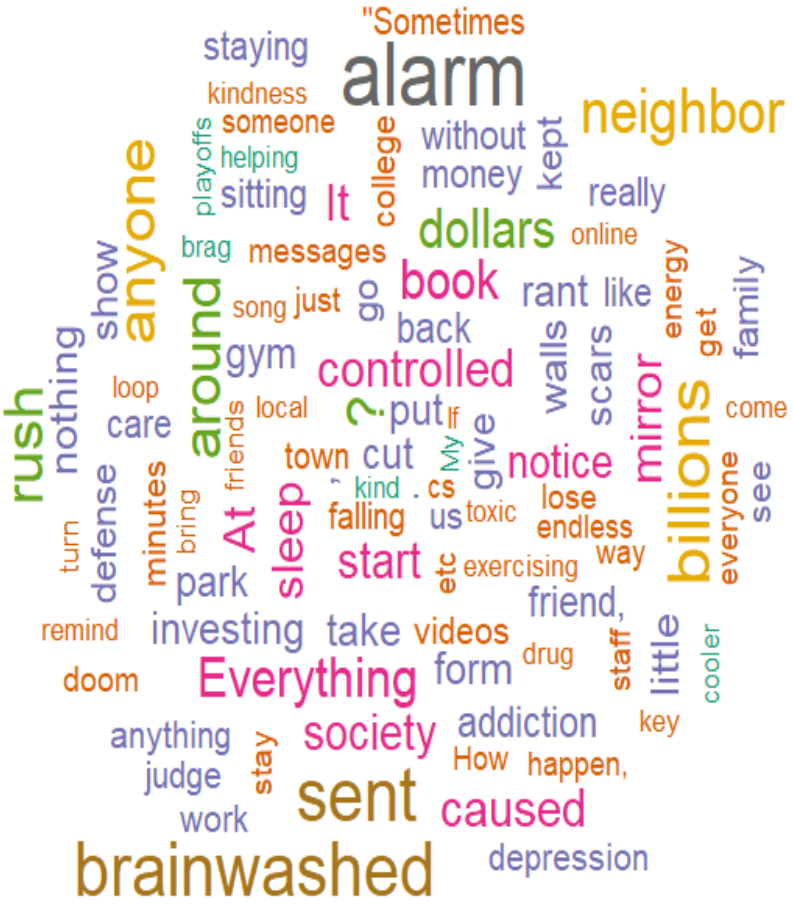
- Depression is a common mental health issue.
- Archetypal authors that self-indicate this condition are active in /r/Depression
  - This is not clinical depression; we are not making a clinical diagnostic tool.
- Is it possible to observe these changes over time? Can we observe resilience behaviours?

ATLAS – Ethics & Case Study

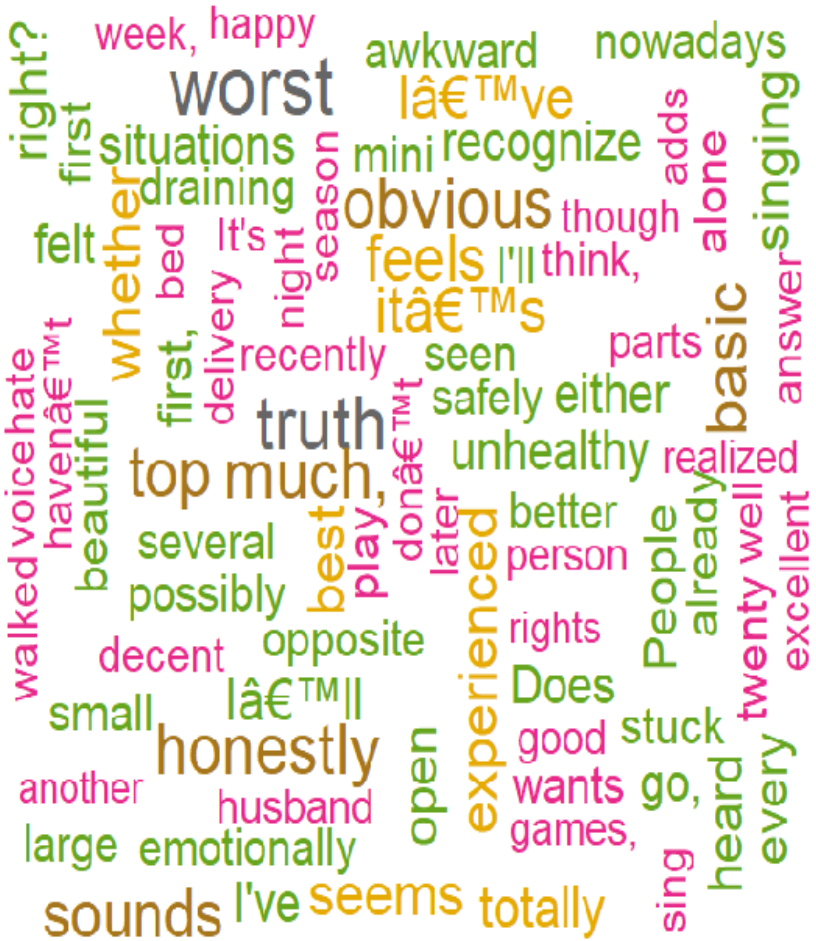
- Previously, ABMS and AIR have been used to assess large scale indicators – population size, topics on social media.
  - They can be used to analyze individuals, but we do not do this except internally to verify it works.
- ATLAS is specialized to analyze individual changes over time. Verification of function in this case requires analyzing an author's posts and then making a judgement on whether they were depressed at point A and recovered at point B
  - I have no qualifications to make that assessment; others do and this may help their work.
- Thus, we use real data to train the model, but evaluate it on synthetic data

ATLAS – Depression Associated Language





Depressive Scoring



Non-Depressive Scoring

Western Science

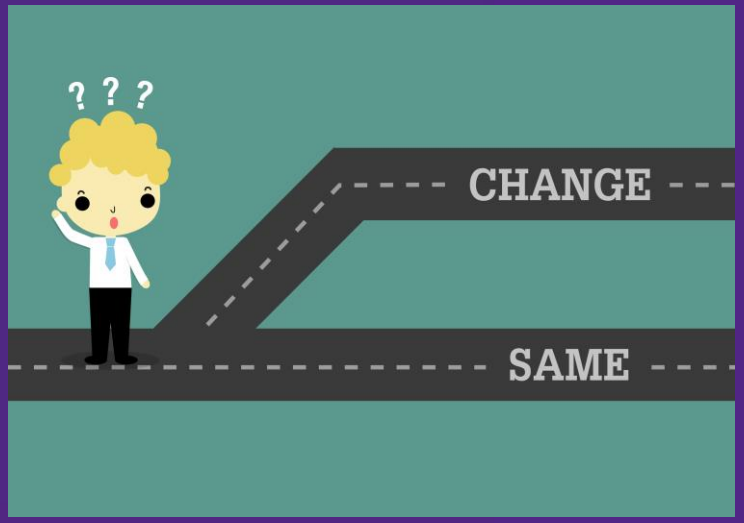
Depressive Scoring

Non-Depressive Scoring

Western Science

ATLAS – Case Study Conclusions

- ATLAS is able to, even in the low-data environment with our synthetic author documents, identify words associated with the change in archetypal behaviour.
- ATLAS shows promise in being able to monitor changes online and at scale.



Conclusions

- ABMS, AIR, and ATLAS are a trio of algorithms accepting generic document input which allow for exploratory search of specialized topics by modeling archetypal behaviour.

- These algorithms have utility in searching social media and in aiding public health causes

Limitations

- All of the algorithms described find correlations, not causations.

- People lie on the internet.
- Complexity of language precludes complete accuracy (i.e., coin collectors and cryptocurrency)

Research Directions

- Can this archetypal framework be adapted to adjust how natural language generation occurs?
  - Give chatbots a ‘persona’
  - Prevent chatbots from discussing or engaging in certain behaviours
- Can causality be applied in the case of ATLAS to try and find causes between archetypal and other behaviours?
- Can heterogeneity in Archetypes be automatically discovered and incorporated?

Acknowledgements

I would like to take this time to thank the following people and organizations for their contributions, in no particular order:

- My supervisors
- My lab members
- My committee
- Computer Science & Western Admin Staff (Prior & Current)
- Ottawa Public Health & Cameron McDermaid
- Geomatics and Cartographic Research Center @ Carleton University
- Canadian Institute For Advanced Research (CIFAR)
- Alberta Machine Intelligence Institute (AMII) & Dr. Fyshe + Students
- MITACS
- Operational Stress & Injury Research Centre @ Parkwood Hospital
- IBM
- Summer Institute on AI & Society
- My beloved wife
- My parents
- The Ineffable

“You are always searching for answers to your questions.

That is because you believe they mean something to you.

As long as you keep desiring answers,
your life will remain a meaningful one.

You are constantly renewing yourself
by thinking and feeling things.”

From Legend of Mana, a Square Enix Game.

Western

UNIVERSITY · CANADA