Better Together: Text + Context

Kenneth Church

DOI: 10.3115/981623.981633 • Corpus ID: 9558665

Word Association Norms, Mutual Information, and Lexicography

Kenneth Ward Church, Patrick Hanks • Published in Annual Meeting of the... 26 June 1989 • Linguistics, Computer Science

TLDR The proposed measure, the association ratio, estimates word association norms directly from computer readable corpora, making it possible to estimate norms for tens of thousands of words.

Abstract The term word association is used in a very particular sense in the psycholinguistic literature. (Generally speaking, subjects respond quicker than normal to the word nurse if it follows a highly associated word such as doctor.) We will extend the term to provide the basis for a statistical description of a variety of interesting linguistic phenomena, ranging from semantic relations of the doctor/nurse type (content word/content word) to lexico-syntactic co-occurrence constraints between verbs and prepositions (content word/function word). This paper will propose an objective measure based on the information theoretic notion of mutual information, for estimating word association norms from computer readable corpora. (The

``You shall know a word by the company it keeps''

1 Meaning and Association

It is common practice in linguistics to classify words not only on the basis of their meanings but also on the basis of their co-occurrence with other words. Running through the whole Firthian tradition, for example, is the theme that "You shall know a word by the company it keeps" (Firth, 1957).

> On the one hand, *bank* co-occurs with words and expression such as *money, notes, loan, account, investment, clerk, official, manager, robbery, vaults, working in a, its actions, First National, of England,* and so forth. On the other hand, we find *bank* co-occurring with *river, swim, boat, east* (and of course *West* and *South*, which have acquired special meanings of their own), *on top of the,* and *of the Rhine.* (Hanks 1987, p. 127)

The search for increasingly delicate word classes is not new. In lexicography, for example, it goes back at least to the "verb patterns" described in Hornby's *Advanced Learner's Dictionary* (first edition 1948). What is new is that facilities for the computational storage and analysis of large bodies of natural language have developed significantly in recent years, so that it is now becoming possible to test and apply informal assertions of this kind in a more rigorous way, and to see what company our words do keep.

Priming & Word Associations

Task: Subject is given two strings and responds “yes” if both are words

Journal of Experimental Psychology
1971, Vol. 90, No. 2, 227–234

FACILITATION IN RECOGNIZING PAIRS OF WORDS: EVIDENCE OF A DEPENDENCE BETWEEN RETRIEVAL OPERATIONS[1]

DAVID E. MEYER[2] AND ROGER W. SCHVANEVELDT
Bell Telephone Laboratories, Murray Hill, New Jersey — *University of Colorado*

EXPERIMENT I

Method

Subjects.—The *S*s were 12 high school students who served as paid volunteers.

Stimuli.—The following test stimuli were used: 48 pairs of associated words, e.g., BREAD-BUTTER and NURSE-DOCTOR, selected from the Connecticut Free Associational Norms (Bousfield, Cohen, & Whitmarsh, 1961); 48 pairs of unassociated words, e.g., BREAD-DOCTOR and NURSE-BUTTER, formed by randomly interchanging the response terms between the 48 pairs of associated words so that there were no obvious associations within the resulting pairs; 48 pairs of nonwords; and 96 pairs involving a word and a nonword. Within each pair of associated words, the second member was either the first or second most frequent free associate given in response to the first member. Within each pair of unassociated words, the second member was never the first or second most frequent free associate of the first member. The median length of strings in the pairs of associated words and pairs of unassociated words was 5 letters and ranged from 3 to 7 letters;

TABLE 1

MEAN REACTION TIMES (RTs) OF CORRECT RESPONSES AND MEAN PERCENT ERRORS IN THE YES–NO TASK

Type of stimulus pair		Correct response	Proportion of trials	Mean RT (msec.)	Mean % errors
Top string	Bottom string				
word	associated word	yes	.25	855	6.3
word	unassociated word	yes	.25	940	8.7
word	nonword	no	.167	1,087	27.6
nonword	word	no	.167	904	7.8
nonword	nonword	no	.167	884	2.6

Title | Add | More | 1–20 | Cited by

Word association norms, mutual information, and lexicography — 4075
KW Church, P Hanks
Computational linguistics 16 (1), 22-29

1555

1454

761

654

602

496

9/15/17

Table 3. Some interesting Associations with "Doctor" in the 1987 AP Corpus (N = 15 million)

I(x, y)	f(x, y)	f(x)	x	f(y)	y
11.3	12	111	*honorary*	621	*doctor*
11.3	8	1105	*doctors*	44	*dentists*
10.7	30	1105	*doctors*	241	*nurses*
9.4	8	1105	*doctors*	154	*treating*
9.0	6	275	*examined*	621	*doctor*
8.9	11	1105	*doctors*	317	*treat*
8.7	25	621	*doctor*	1407	*bills*
8.7	6	621	*doctor*	350	*visits*
8.6	19	1105	*doctors*	676	*hospitals*
8.4	6	241	*nurses*	1105	*doctors*
Some Uninteresting Associations with "Doctor"					
0.96	6	621	*doctor*	73785	*with*
0.95	41	284690	*a*	1105	*doctors*
0.93	12	84716	*is*	1105	*doctors*

There is no data like more data

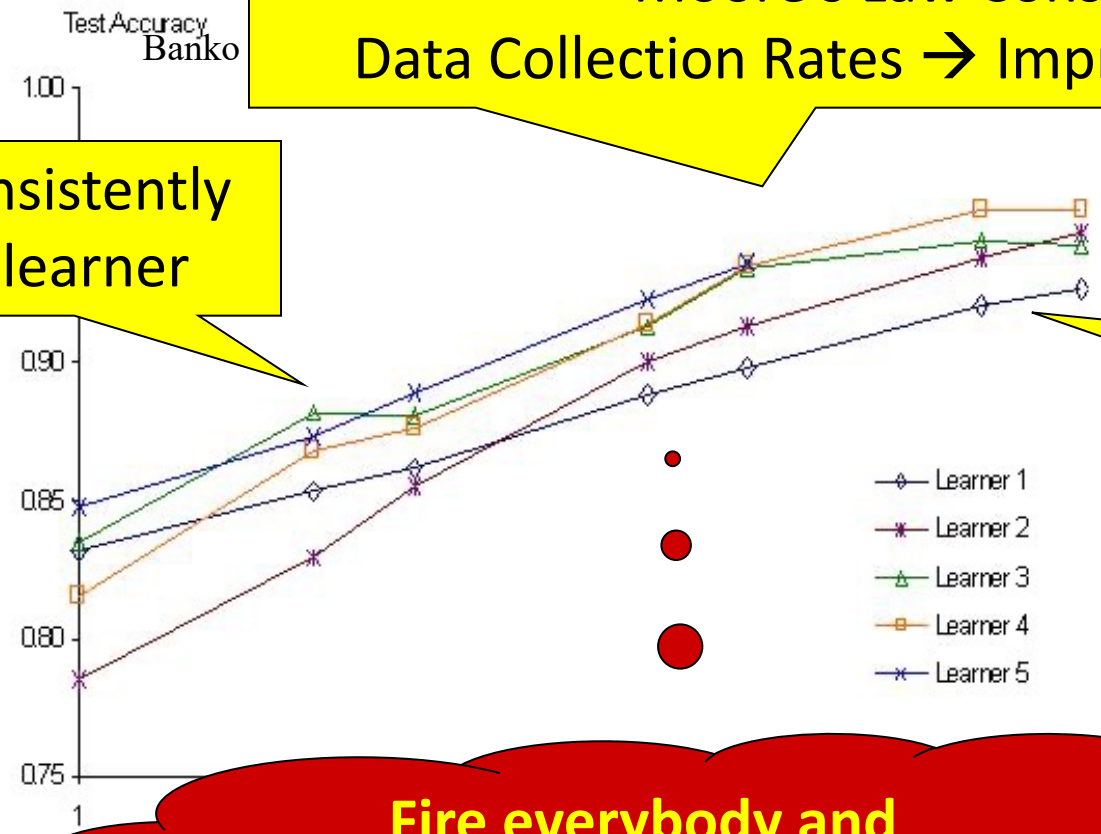
Table 3. Some interesting Associations with "Doctor" in the 1987 AP Corpus (N = 15 million)

I(x, y)	f(x, y)	f(x)	x	f(y)	y
11.3	12	111	*honorary*	621	*doctor*
11.3	8	1105	*doctors*	44	*dentists*
10.7	30	1105	*doctors*	241	*nurses*
9.4	8	1105	*doctors*	154	*treating*
9.0	6	275	*examined*	621	*doctor*
8.9	11	1105	*doctors*	317	*treat*
8.7	25	621	*doctor*	1407	*bills*
8.7	6	621	*doctor*	350	*visits*
8.6	19	1105	*doctors*	676	*hospitals*
8.4	6	241	*nurses*	1105	*doctors*
Some Uninteresting Associations with "Doctor"					
0.96	6	621	*doctor*	73785	*with*
0.95	41	284690	*a*	1105	*doctors*
0.93	12	84716	*is*	1105	*doctors*

"It never pays to think until you've run out of data" – Eric Brill

Moore's Law Constant:
Data Collection Rates → Improvement Rates

No consistently best learner



There is no data like more data

Fire everybody and spend the money on data

Quoted out of context

PMI → Word2vec → BERT → Chatbots

- Historical Precedents
  - Word Associations: doctor primes nurse
  - Perplexity (Shannon), Cloze task, Fill mask
  - ``You shall know a word by the company it keeps''
  - Collocations in Lexicography

Levy & Goldberg (NIPS-2014)

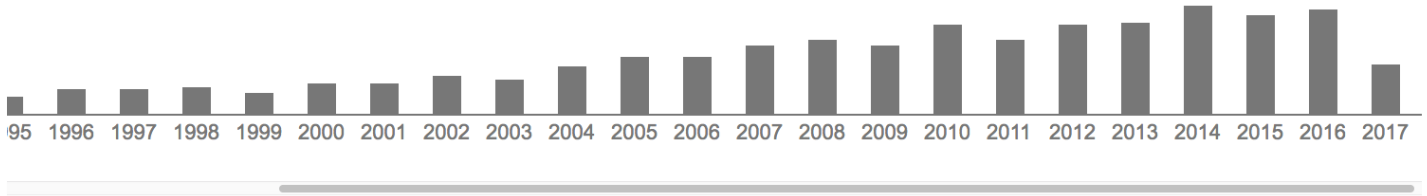
Word2Vec ≈ PMI (Pointwise Mutual Info)

$$sim(x, y) = \cos\big(vec(x), vec(y)\big) \approx PMI(x, y)$$

Word association norms, mutual information, and lexicography [PDF] from aclweb.org

Authors	Kenneth Ward Church, Patrick Hanks
Publication date	1990/3/1
Journal	Computational linguistics
Volume	16
Issue	1
Pages	22-29
Publisher	MIT Press
Description	Abstract The term word association is used in a very particular sense in the psycholinguistic literature.(Generally speaking, subjects respond quicker than normal to the word nurse if it follows a highly associated word such as doctor.) We will extend the term to provide the basis for a statistical description of a variety of interesting linguistic phenomena, ranging from semantic relations of the doctor/nurse type (content word/content word) to lexico-syntactic co-occurrence constraints between verbs and prepositions (content word/ ...
Total citations	Cited by 4269

What happened in 2014?



Scholar articles
Word association norms, mutual information, and lexicography
KW Church, P Hanks - Computational linguistics, 1990
Cited by 4269 - Related articles - All 42 versions

Neural Word Embedding as Implicit Matrix Factorization

Omer Levy
Department of Computer Science
Bar-Ilan University

Yoav Goldberg
Department of Computer Science
Bar-Ilan University

Fill Mask (Cloze Task)

3s

```
pipe = pipeline("fill-mask", model='distilroberta-base')
pipe("She likes <mask>, but he likes that.")
```

```
Some weights of the model checkpoint at distilroberta-base
- This IS expected if you are initializing RobertaForMaske
- This IS NOT expected if you are initializing RobertaForM
[{'score': 0.024591688066720963,
  'token': 123,
  'token_str': ' him',
  'sequence': 'She likes him, but he likes that.'},
 {'score': 0.022750644013285637,
  'token': 9366,
  'token_str': ' pizza',
  'sequence': 'She likes pizza, but he likes that.'},
 {'score': 0.022747714072465897,
  'token': 10017,
  'token_str': ' cats',
  'sequence': 'She likes cats, but he likes that.'},
 {'score': 0.012954547069966793,
  'token': 69,
  'token_str': ' her',
  'sequence': 'She likes her, but he likes that.'},
 {'score': 0.011491155251860619,
  'token': 162,
  'token_str': ' me',
  'sequence': 'She likes me, but he likes that.'}]
```

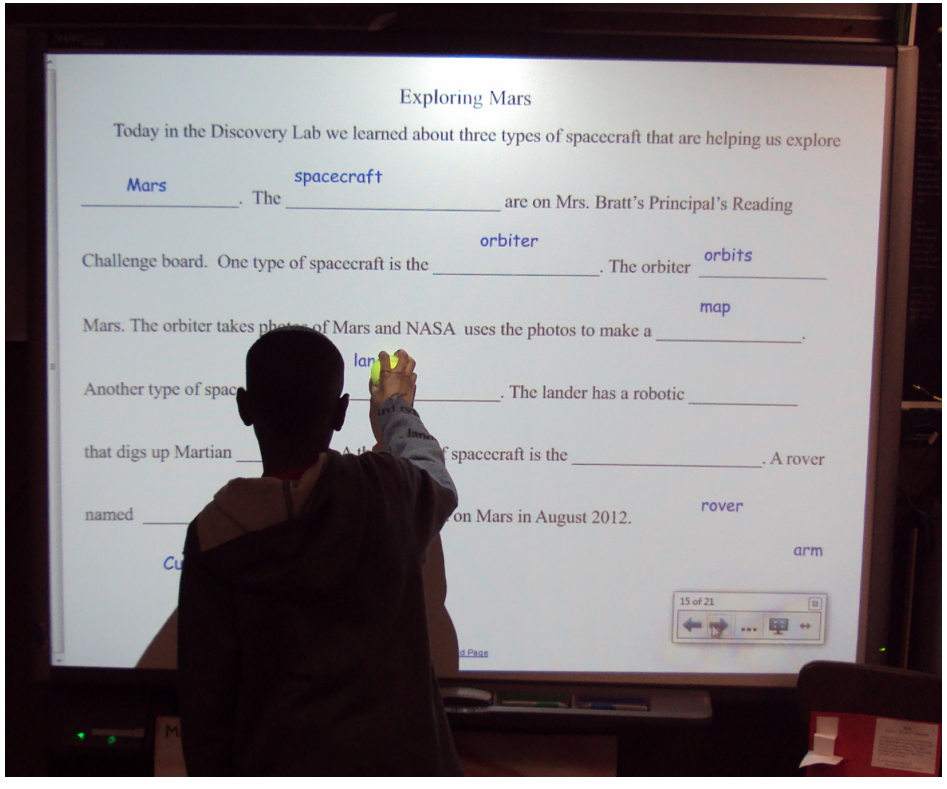
More Inference Tasks: Classification, Fill Mask, …

`echo … | gft_predict -task fill-mask`

BERT: Fill-Mask (Delvin et al, 2019)

Input	Choice 1		Choice 2	
I <mask> you	miss	30%	love	17%
I <mask> him	love	14%	miss	13%
I <mask> her	love	16%	miss	12%
I love <mask>	him	2%	cats	2%
I <mask> New York	love	33%	miss	17%
I love the <mask> guy	pizza	2%	funny	2%
I hate the <mask> guy	fucking	3%	pizza	2%

Cloze Task (1953)



By Laurie Sullivan - DSC01784, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=62727253

Lots of History:

Fill-mask, Cloze Task, Entropy

Prediction and Entropy of Printed English

By C. E. SHANNON

(*Manuscript Received Sept. 15, 1950*)

A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of an ideal predictor are developed.

first line is the original text and the numbers in the second line indicate the guess at which the correct letter was obtained.

(1) THERE IS NO REVERSE ON A MOTORCYCLE A
(2) 1 1 1 5 1 1 2 1 1 2 1 1 1 5 1 1 7 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1 1 1 3 1
(1) FRIEND OF MINE FOUND THIS OUT
(2) 8 6 1 3 1 1 1 1 1 1 1 1 1 1 1 6 2 1 1 1 1 1 1 2 1 1 1 1 1 1
(1) RATHER DRAMATICALLY THE OTHER DAY
(2) 4 1 1 1 1 1 1 1 1 5 1 1 1 1 1 1 1 1 1 1 1 1 6 1 1 1 1 1 1 1 1 1 1 1 1 (9)

Out of 102 symbols the subject guessed right on the first guess 79 times, on the second guess 8 times, on the third guess 3 times, the fourth and fifth guesses 2 each and only eight times required more than five guesses. Results of this order are typical of prediction by a good subject with ordinary literary English. Newspaper writing, scientific work and poetry generally lead to somewhat poorer scores.

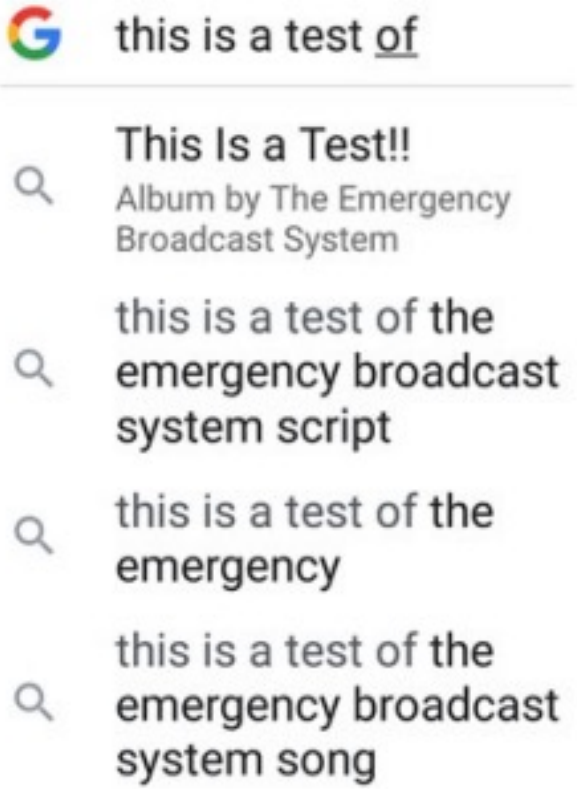
Guess the missing word in a cliché
(from “Emerging trends: Deep nets for poets”)

Fill Mask

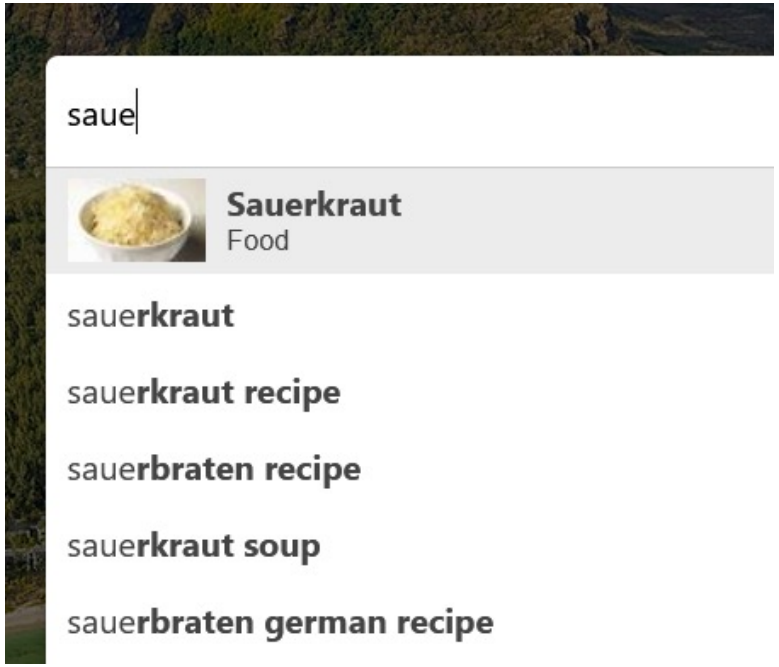
Table 3. Unmasking: replace each input word with [MASK] and predict fillers. Red is added to highlight differences between the top prediction and the original input

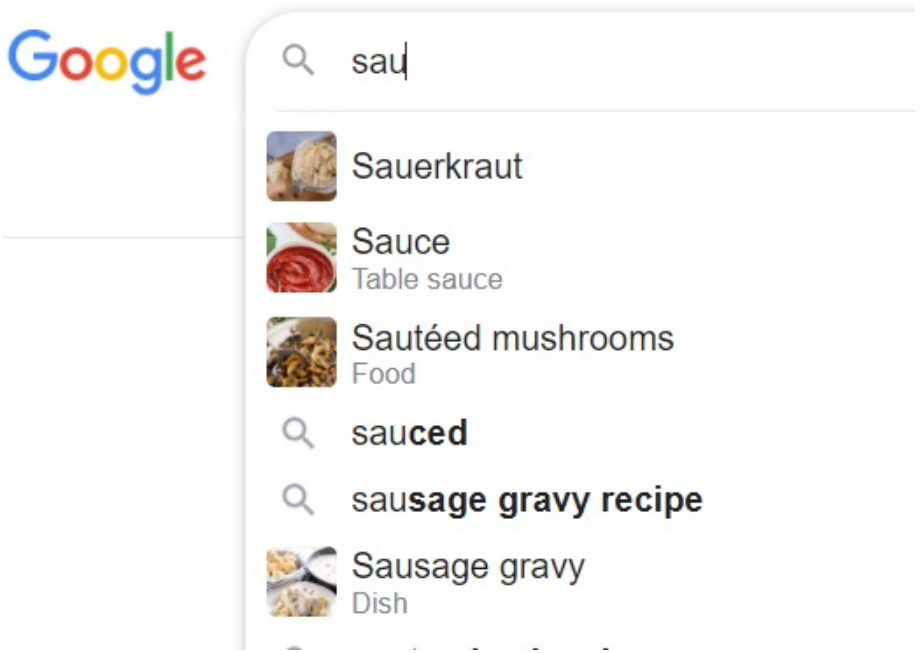
Word	Rank 0	Rank 1	Rank 2
This	this:0.595	it:0.375	there:0.004
is	was:0.628	is:0.334	included:0.007
a	a:0.935	another:0.029	the:0.023
test	part:0.253	subset:0.125	variation:0.082
of	of:0.886	for:0.070	on:0.017
the	the:0.883	an:0.088	our:0.005
emergency	ratio:0.257	television:0.048	satellite:0.031
broadcast	braking:0.620	response:0.070	management:0.024
system	system:0.244	technique:0.077	capability:0.059
.	.:0.969	;:0.029	!:0.001

Autocomplete in Search



Autocomplete ≠ Fill-Mask





How to succeed?
(get beyond publishing)

How to get citations

- First? Last? Best?
- Probably not
- Better:
  - Accessible
  - Helpful to audience
    - Data
    - Tools
    - Survey
    - Accessible

Paradigm Shift (Kuhn)

- Initial promising results
  - promising >> convincing
- Leave enough undone for students to contribute
  - and benefit by doing so

Word2vec is popular (massively cited)

- Word2vec is not first, last or best to discuss
  - Vector spaces, embeddings, analogies, similarity metrics, etc.
- But word2vec is simple and accessible
  - Anyone can download the code and use it in their next paper.
  - Any many do (for better and for worse)
- Available downloads
  - Pre-computed vectors (no training required)
  - Code for training your own vectors on your own corpora

Embeddings:
Similarity of docs/words ≈ cos (dot product)

Applications [edit]

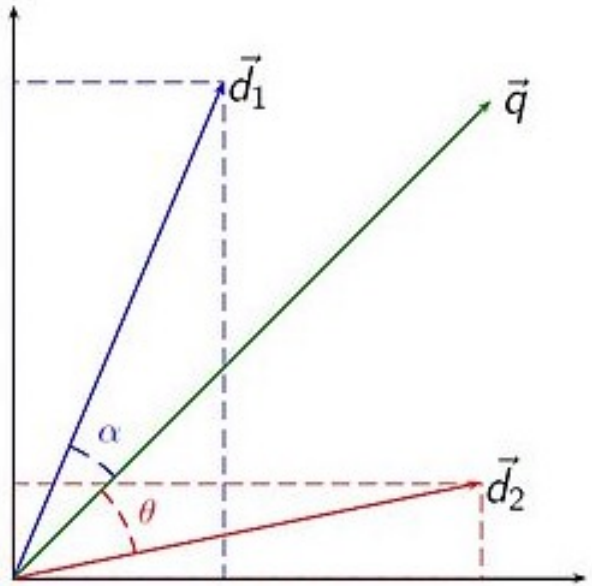
Relevance rankings of documents in a keyword search can be calculated, using the assumptions of document similarities theory, by comparing the deviation of angles between each document vector and the original query vector where the query is represented as the same kind of vector as the documents.

In practice, it is easier to calculate the cosine of the angle between the vectors, instead of the angle itself:

$$\cos\theta = \frac{\mathbf{d_2} \cdot \mathbf{q}}{\|\mathbf{d_2}\| \, \|\mathbf{q}\|}$$

Where $\mathbf{d_2} \cdot \mathbf{q}$ is the intersection (i.e. the dot product) of the document (d_2 in the figure to the right) and the query (q in the figure) vectors, $\|\mathbf{d_2}\|$ is the norm of vector d_2, and $\|\mathbf{q}\|$ is the norm of vector q. The norm of a vector is calculated as such:

$$\|\mathbf{q}\| = \sqrt{\sum_{i=1}^{n} q_i^2}$$



As all vectors under consideration by this model are elementwise nonnegative, a cosine value of zero means that the query and document vector are orthogonal and have no match (i.e. the query term does not exist in the document being considered). See cosine similarity for further information.

man : woman :: king : ***x***

- Vector addition & subtraction
  - $vec(king + woman\ - man) = vec(king) + vec(woman) - vec(man)$
- Analogies
  - $\hat{x} = \underset{x \in V}{\text{ARGMAX}}\, sim(x', king + woman\ - man)$

x	Gender	Number
Queen	f	sg
Monarch	m	sg
Princess	f	sg
Crown˙prince	m	sg
Prince	m	sg
Kings	m	pl
Queen˙Consort	m	sg
Queens	f	pl
Sultan	m	sg
Monarchy	m	sg

Some analogies are easier than others

- Tweets
  - RT @tallinzen: sure, king:queen etc, but did you know word2vec gets real SAT analogies right just 1% of the time?
  - 15 copies of this tweet
    - Some by NLP experts
- Resources Debate
  - WordNet &
  - British National Corpora

Table 2. *Some types of analogies are easier than others, as indicated by accuracies for top choice (A_1), as well as top 2 (A_2), top 10 (A_{10}) and top 20 (A_{20}). The rows are sorted by A_1. These analogies and the type classification come from the questions-words test set, except for the last row, SAT questions. SAT questions are harder than questions-words*

A_1	A_2	A_{10}	A_{20}	N	Analogy type	Example
0.91	0.95	0.98	0.99	1,332	Comparative	young : younger :: wide : wider
0.90	0.94	0.97	0.98	1,599	Nationality-adjective	Ukraine : Ukrainian :: Switzerland : Swiss
0.90	0.93	0.97	0.98	1,332	Plural	woman : women :: snake : snakes
0.87	0.94	1.00	1.00	1,122	Superlative	young : youngest :: wide : widest
0.85	0.90	0.97	1.00	506	Family	uncle : aunt :: stepson : stepdaughter
0.83	0.89	0.97	0.98	335	Capital-countries	Tokyo : Japan :: Tehran : Iran
0.79	0.86	0.94	0.96	4,695	Capital-world	Zagreb : Croatia :: Dublin : Ireland
0.78	0.84	0.98	0.99	1,056	Present-participle	write : writing :: walk : walking
0.71	0.79	0.90	0.92	2,467	City-in-state	Worcester : Massachusetts :: Cincinnati : Ohio
0.68	0.78	0.93	0.95	870	Plural-verbs	write : writes :: work : works
0.66	0.82	0.97	0.98	1,560	Past-tense	writing : wrote :: walking : walked
0.43	0.48	0.64	0.69	812	Opposite	tasteful : distasteful :: sure : unsure
0.35	0.42	0.57	0.62	866	Currency	Vietnam : dong :: USA : dollar
0.29	0.37	0.63	0.73	992	Adjective-to-adverb	usual : usually :: unfortunate : unfortunately
0.01	0.02	0.08	0.10	190	SAT questions	audacious : boldness :: sanctimonious : hypocrisy

Need to Re-organize Courses (Interdisciplinary Re-Org)

Current Courses (and Conferences)

- Machine Learning (ML)
- Vision
- Natural Language (NLP)
- Machine Translation (MT)
- Information Retrieval (IR)
- Speech

Future

- Machine Learning (ML)
- ML Applications
  - Vision/NLP/MT/IR/Speech

Soap Box: My Views

- Interdisciplinary Re-Org: ML/Vision/NLP/MT/IR/Speech
  - Current: People working on one app don't talk to people working on other apps
  - Future: We should work together because we are all using the same ML techniques
- Responsible AI
  - Current: Bias is Bad
  - Future: Embrace Diversity
- Authors View vs. Audience Response
  - Current: Focus on Unambiguous Objective Labels (independent of context)
  - Future: More Room for Subjectivity (and richer semantics)
    - Authors' Position ≠ Audience Response
- End-to-End Optimization vs. Modularity
  - Current: End-to-End → Better Performance
  - Future: Multiple Perspectives are Better Together

Vision: Facts → Opinions

Facts

(b) COCO: *A man and a woman holding a little kid while sitting at a table outside*

Opinions

(a) ArtEmis: *I love everything about this painting of a mother and her two children lovingly interacting with the family pet cat.*

Ambiguity: No Single Correct Gold Label

- A difference of opinion ≠ Error
  - Common Challenge in Language Apps
    - Machine Translation
    - Web Search
    - Even Part of Speech Tagging!
- Standard Solution
  - Score over multiple references (many gold standards)
    - Machine Translation: BLEU
    - Web Search: NDCG

Candidate labels: baseball cap, cap, green hat, hat, head.
Can you guess which one is in the gold standard?

Classic Challenges in Philosophy of Language

Compare & Contrast Language in Visual Genome (VG) with NLP Corpora

- Bounding Box Semantics: Too Limiting
  - <NP> <relation> <NP>
  - NPs in VG are usually rigid designators that mean the same thing in all contexts
    - Few abstract nouns (*ideas*), predicates (verbs, adj), variables (pronouns)
    - Most VG nouns are visual (and less about other senses)
  - More entropy in nouns than relations
    - 8 boring relations cover bulk of the cases
    - Linguists are more interested in predicates than arguments
  - Modifiers are limited to a single box
    - Relative Relations: *girl with green hat* vs *girl on defense*
    - Aggregations over boxes
      - Count vs. Mass: *Cloudy Sky*, *Sandy Beach*
      - Definite vs. Indefinite: *girls playing frisbee* (as opposed to many other people in picture)
      - A horse with two legs?

Verbs, Subjunctive, Focus, Perspective, etc…

- Pictures vs. Videos
  - Captions on pictures have more nouns:
    - *girl with **green hat***
  - Captions on videos have more verbs:
    - *girl **throwing** frisbee*
- Possible Worlds: Subjunctive, Hedges
  - *The girl on defense **might** block the throw,*
  - *but **probably** won't*
- Focus: ***girl on offense*** vs. ***girl on defense***
- Perspective: photographer vs. audience

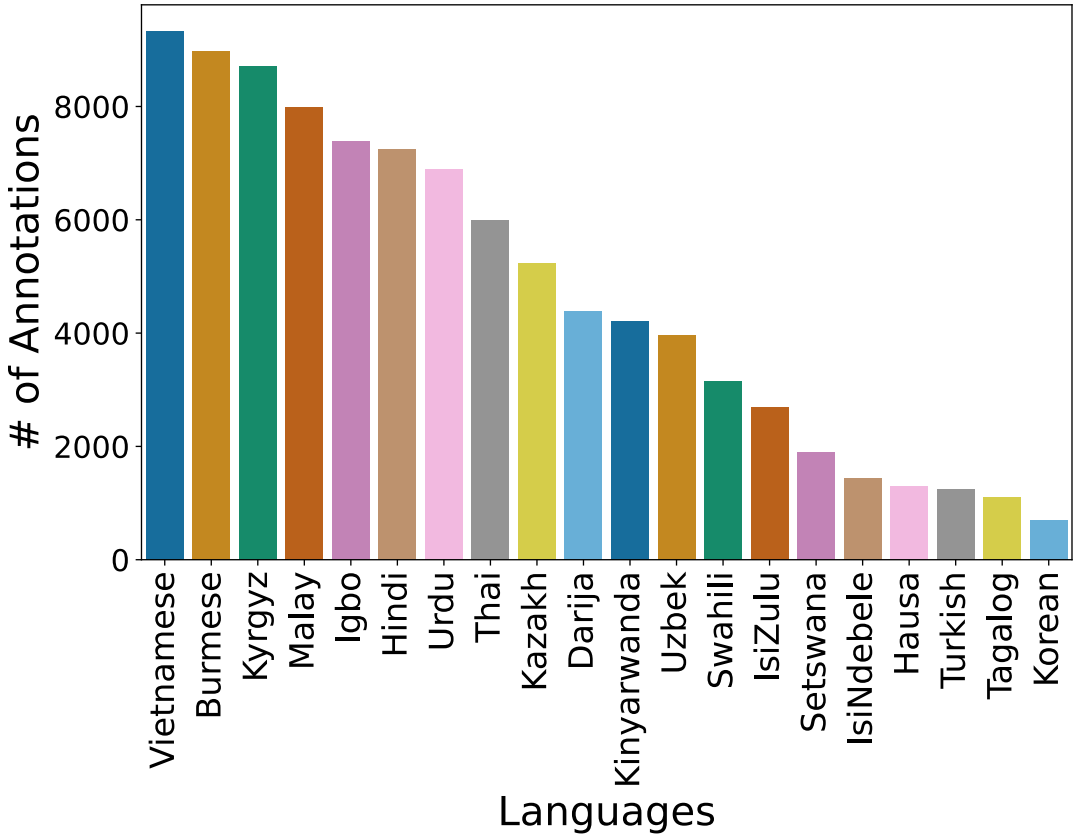
ArtELingo: A Million Emotion Annotations of WikiArt with Emphasis on Diversity over Language and Culture

Youssef Mohamed[1*] **Mohamed Abdelfattah** [1] **Shyma Alhuwaider**[1] **Feifan Li** [1]
Xiangliang Zhang [2] **Kenneth Ward Church** [3] **Mohamed Elhoseiny**[1*]
[1]KAUST [2] University of Notre Dame [3] Northeastern University
{youssef.mohamed,mohamed.abdelfattah,shyma.alhuwaider,feifan.li}@kaust.edu.sa
xzhang33@nd.edu, k.church@northeastern.edu, mohamed.elhoseiny@kaust.edu.sa

a)		شلال طبيعي جميل. مشاعر النمو والحيوية والطاقة موجودة. Translation: Beautiful natural waterfall. Feelings of growth, vitality and energy. **Excitement** **Arabic**	The water that's rushing downward looks like a bride's wedding veil. **Awe** **English**	瀑布就像四蹄生风的白马如潮水涌来，非常的壮观 Translation: The waterfall is like a white horse and wind, it is spectacular. **Contentment** **Chinese**
b)		Translation: Girls sitting with their mother outside the house, exchanging love and affection, pigeons flying over a tree. **Contentment** **Arabic**	The women relaxing while birds are flying about makes me feel relaxed and calm as well. **Contentment** **English**	Translation: Three sisters lying on a bench and watching the birds fly comfortably. **Contentment** **Chinese**
c)		Translation: The use of black and white for painting the forests with all its details brings out a feeling of satisfaction. **Contentment** **Arabic**	The trees are dead and exposing their roots due to erosion and lack of water. **Sadness** **English**	Translation: After the snow in winter, there is snow everywhere, and the dead trees look very depressed. **Sadness** **Chinese**

ArteLingo-23: Embrace Diversity over Language and Culture

- **Africa:** Kinyarwanda, Swahili, IsiZulu, Setswana, Yoruba, Hausa, Igbo.
- **Southeast Asia:** Vietnamese, Indonesian, Thai, Burmese, Malay.
- **Sub-Indian continent:** Tagalog, Tamil, Hindi, Urdu.
- **East Asia:** Korean, *Chinese*.
- **Middle-East:** Turkish, Darija, *Arabic*.
- **Central Asia:** Uzbek
- **Europe and North America:** *English*.



Why study more languages?

- Vision:
  - Depends on both stimulus as well as context
    - Stimulus: Picture
    - Context: Language, Culture, Religion, Politics, Education, Background
- Cliché:
  - *Beauty is in the eye of the beholder*

Few Shot & Zero Shot

Seen Language

خزانة صغيرة على طاولة خشبية تحمل مزهرية بها أزهار حمراء وبيضاء وزرقاء.
Translation: A small dresser on a wooden table holds a vase with red, white, and blue flowers.
Arabic

There are a bunch of flowers in a yellow vase on a table. It looks like something from a restaurant. The table has a yellow cloth on it.
English

花瓶里白色的百合和绿色的小花搭配着，让人感到很美。
Translation: The combination of white lilies and small green flowers in the vase looks nice.
Chinese

Unseen Language

Fleur blanche et rose par un vase très mince
Translation: White and pink flower by a very thin vase.
French

El cálido tono verde de las flores hace que la imagen sea reconfortante y agradable de contemplar.
Translation: The warm green hue of the flowers makes the image look comforting and pleasing.
Spanish

花園中花瓣色的花朵散滿在白色的花瓶上，作者用畢生之力將花卉之美帶給世人。
Translation: The petal-colored flowers in the garden are scattered on the white vase. The author devotes his life to bringing the beauty of flowers to the world.
Chinese(Traditional)

Seen Language

تصور اللوحة امرأة وطفل في مشهد عن قرب. المرأة تمسك ذراع الطفل في ذراعها.
Translation: The painting depicts a woman and child in a close-up view. The woman is holding the child's arm in hers.
Arabic

The baby in the picture looks so calm with the mother closing her eyes and feeling peaceful and content.
English

图片描绘了母亲抚摸着她的孩子，他穿着白色短裤。母亲看起来很温柔。
Translation: The picture depicts a mother stroking her baby, who is wearing white shorts. Mother looks very gentle.
Chinese

Unseen Language

Ce tableau représente une femme vêtue d'une robe orange, assise dans le dos d'un enfant.
Translation: This painting shows a woman in an orange dress, seated behind the back of a child.
French

Esta pintura es una fotografía de un niño durmiendo con su madre.
Translation: This painting is a photograph of a boy sleeping with his mother.
Spanish

इस पेंटिंग में एक महिला मैनीक्योर किया हुआ चश्मा पहने हुए और गले में कपड़ा लपेटे हुए अपने बच्चे को देख रही है
Translation: In this painting, a woman is looking down at her child wearing manicured glasses, holding a cloth wrapped around her neck
Hindi

Anglo-Centered Baseline

- Use English as Pivot Language
- Anglo-Centered Vision Task:
  - Input: Picture
  - Output: English Caption
- If you want a caption in another language:
  - Just translate English caption to that language

- Challenge:
  - Can we do better than that?

Agreement (A) on Emotion Labels:
More A: Landscapes → Less A: Sketches

Figure 5: 8 artworks with genre. Green indicates high agreement in Table 4; red indicates high disagreement.

Genre (G)	$Pr(G\|U)$	$Pr(G\|D)$	A
landscape	0.206	0.097	-1.08
cityscape	0.071	0.036	-0.98
still life	0.043	0.042	-0.03
illustration	0.029	0.029	-0.01
misc	0.167	0.177	0.08
portrait	0.217	0.233	0.10
nude	0.030	0.032	0.11
religious	0.101	0.133	0.40
abstract	0.076	0.112	0.55
sketch	0.061	0.109	0.85

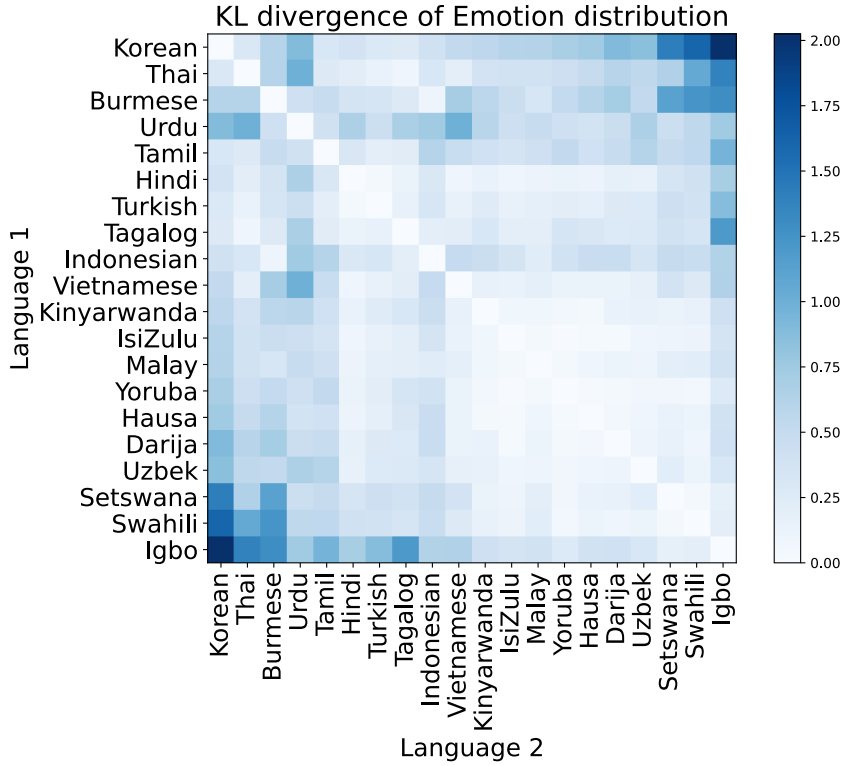
Table 4: Genre sorted by agreement (A). Most agreement: landscapes; Most disagreement: sketches.

Agreement on Emotion Labels

By Genre

Figure 5: 8 artworks with genre. Green indicates high agreement in Table 4; red indicates high disagreement.

By Language (or perhaps Education?)



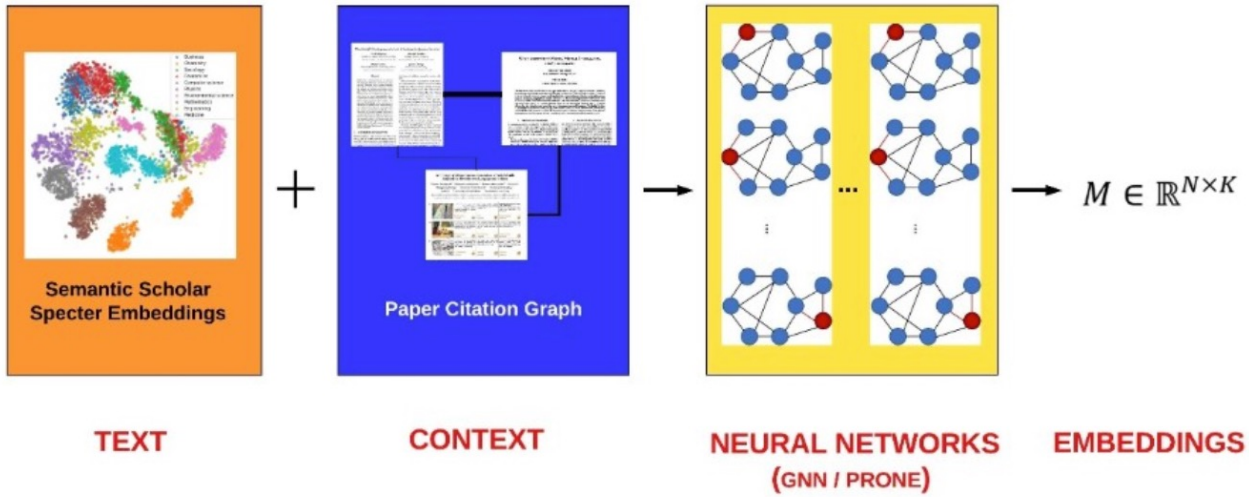
Academic Search

2023 Jelinek Summer Workshop on Speech and Language Technology
https://www.clsp.jhu.edu/2023-jelinek-summer-workshop/

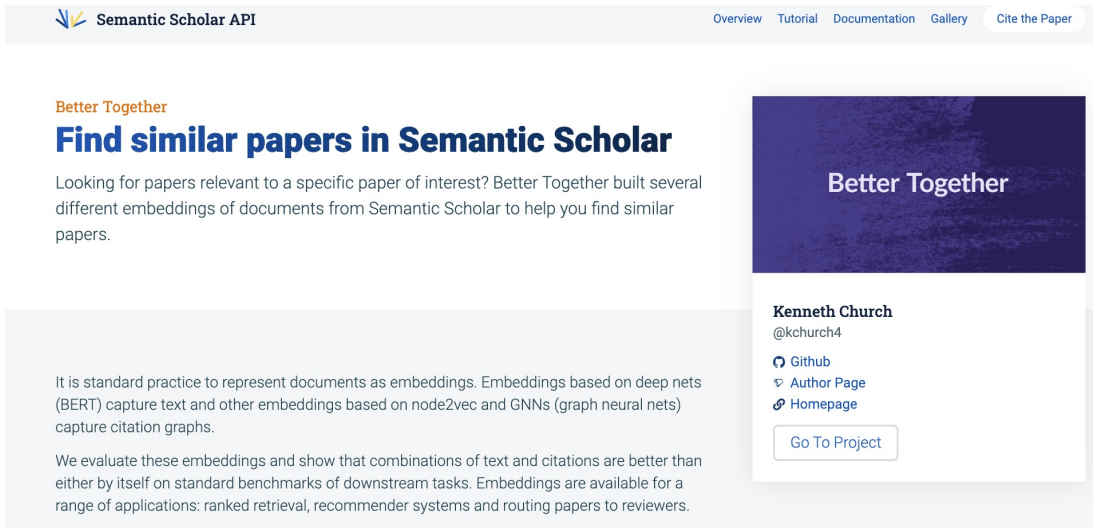
Home > Better Together: Text + Context

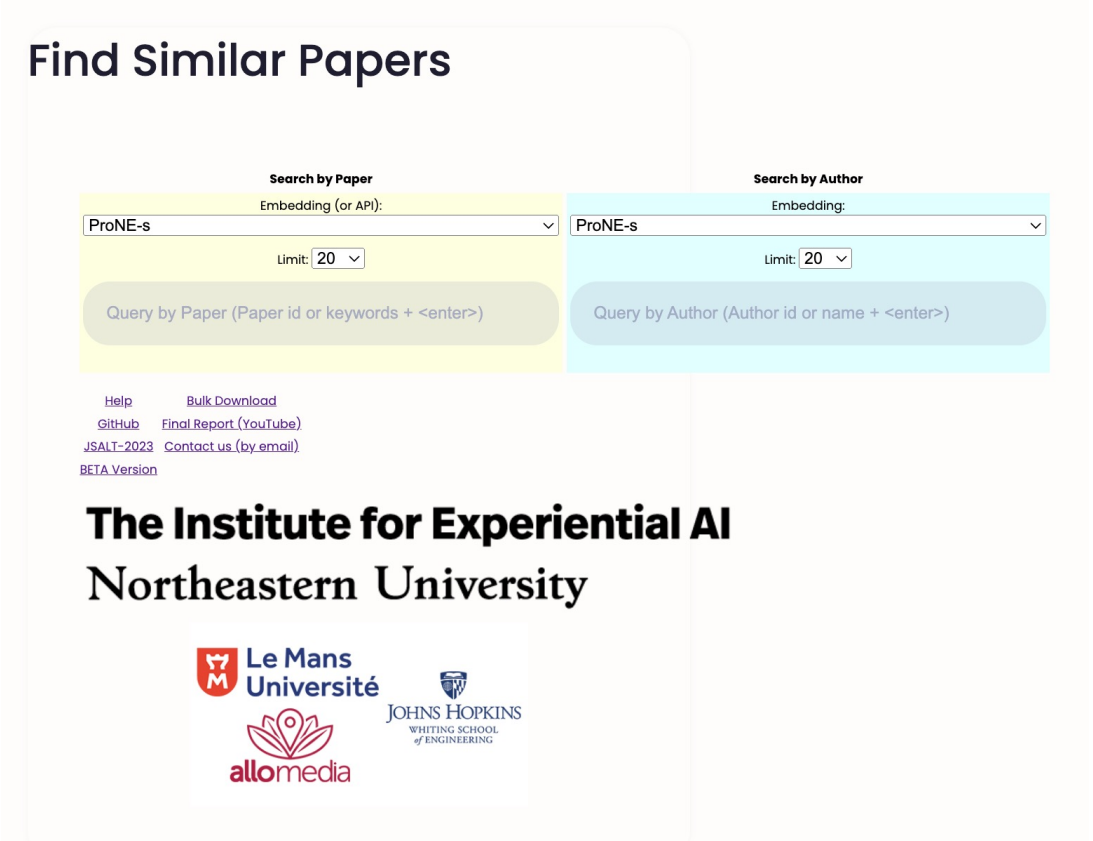
Better Together: Text + Context

Abstract



Google Search: Semantic Scholar Gallery





Google Search: Gallery Semantic Scholar

Find Similar Papers

Search by Paper

Embedding (or API):

ProNE-s

Limit: 20

Query by Paper (Paper id or keywords + <enter>)

Help Bulk Download
GitHub Final Report (YouTube)
JSALT-2023 Contact us (by email)
BETA Version

The Institute for Experienti

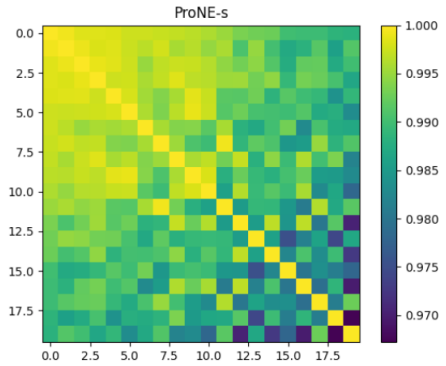
Northeastern University

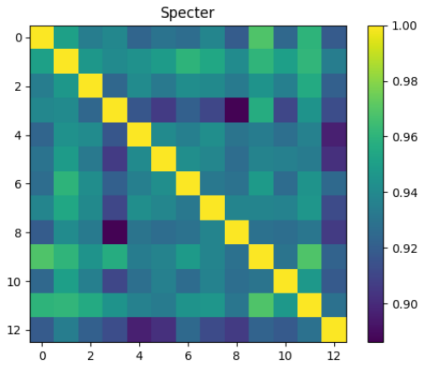
Le Mans Université

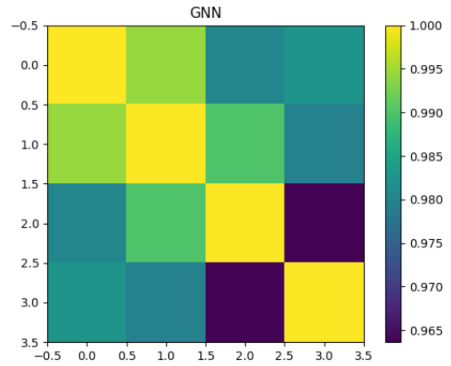
JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

allomedia

Paper: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning







Top

score	citationCount	Paper	Authors	year	More like this	Compare & Contrast	ProNE
	662	InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning	Wenliang Dai, Junnan Li, ..., Steven C. H. Hoi	**2023**	Similar to this	Compare & Contrast	**1.0**
0.999	431	mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality	Qinghao Ye, Haiyang Xu, ..., Feiyan Huang	2023	Similar to this	Compare & Contrast	0.999
0.998	152	MultiModal-GPT: A Vision and Language Model for Dialogue with Humans	T. Gong, Chengqi Lyu, ..., Kai Chen	2023	Similar to this	Compare & Contrast	0.998
0.998	300	Otter: A Multi-Modal Model with In-Context Instruction Tuning	Bo Li, Yuanhan Zhang, ..., Ziwei Liu	2023	Similar to this	Compare & Contrast	0.998
0.998	225	MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models	Chaoyou Fu, Peixian Chen, ..., Rongrong Ji	2023	Similar to this	Compare & Contrast	0.998
0.998	142	SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension	Bohao Li, Rui Wang, ..., Ying Shan	2023	Similar to this	Compare & Contrast	0.998
0.997	546	Improved Baselines with Visual Instruction Tuning	Haotian Liu, Chunyuan Li, ..., Yong Jae Lee	2023	Similar to this	Compare & Contrast	0.997
0.997	818	MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models	Deyao Zhu, Jun Chen, ..., Mohamed Elhoseiny	2023	Similar to this	Compare & Contrast	0.997
0.997	154	OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models	Anas Awadalla, Irena Gao, ..., Ludwig Schmidt	2023	Similar to this	Compare & Contrast	0.997
0.997	187	MMBench: Is Your Multi-modal Model an All-around Player?	Yuanzhan Liu, Haodong Duan, ..., Dahua Lin	2023	Similar to this	Compare & Contrast	0.997

APIs

API	Examples	Arguments	Description
Paper Search	example	help, query, fields	• Find papers matching input query (a string); output fields from S2 for each paper. • See documentation on fields for more information on fields in S2. • A common use case is to request paper ids from titles of papers since many of the APIs below are based on ids in Semantic Scholar (and other sources).
Author Search	example	help, query, fields	• Find authors matching input query (a string); output fields from S2 for each author. • See documentation on fields for more information on fields in S2. • Note: author fields are different from paper fields.
Lookup Paper	simple example, more challenging example	help, id, fields, embeddings	• Input one or more comma separated paper id and output fields from S2, as well as embeddings. • If embeddings argument is specified, then output embedding vectors for each input paper (missing values will have vectors of 0). • See documentation on embeddings for details on how to specify combinations of different embeddings to return.
Lookup Author	example	help, id, fields	• Input author id and output author fields from S2. • Note: author ids are different from paper ids and author fields are different from paper fields.
Lookup Citations	example	help, offset (defaults to 0), limit (defaults to 100; max is 1000), id, fields	• Lookup Citations for paper id and output fields from S2 for each citation. • A useful field to request is contexts; that field returns citing sentences, sentences from other papers that cite the input paper id. • For papers with more than 1000 citations, call this API multiple times with different offsets.
Coauthors	example	help, query, after_year	• Input query (a string); for each matching author ids, returns a list of coauthors filtered by after_year (a 4 digit number). • Note: since Semantic Scholar may have multiple author ids for the same author, the json object contains a list of coauthors for each author matching the input query
Recommend Papers	example	help, id, limit, method, fields, sort_by, score1, score2	• Recommend papers similar to paper id using method. • See documentation on method for choices of methods that are currently supported. • Output fields from S2 for each recommended paper. • The optional arguments, score1 and score2, score recommendations one at a time (for score1) and pairwise (for score2), using one or more of four embeddings.
Recommend Authors	example	help, id, limit, method, fields, sort_by, score1, score2	• Recommend authors near paper id using method • Output fields from S2 for each recommended author.
Compare and Contrast	example1, example2 example2	help, ids (two or more ids, separated by commas)	• Use RAG to compare and contrast the first id with the rest.
Compare and Contrast Texts	example	help, text1, text2	• Use RAG to compare and contrast text1 with text2, where both texts are strings.

APIs

- Paper Search:
  - find paper ids matching input query
- Author Search:
  - find author ids matching input query
- Lookup Paper:
  - lookup fields and embeddings by id
- Lookup Author:
  - lookup fields by id
- Recommend Papers:
  - input paper id and output more paper ids
- Recommend Authors:
  - input paper id and output author ids
- Compare and Contrast:
  - use RAG to compare and contrast one paper id with more paper id(s)

- paper ids:
  - Includes ids from
    - Semantic Scholar (S2)
    - PubMed
    - ACL
    - arXiv
    - MAG (Microsoft Academic Graph)
- embeddings: Specter, ProNE, etc
- fields (from S2): properties of ids
  - title, authors, tldr, abstract, bibtex, references, citations

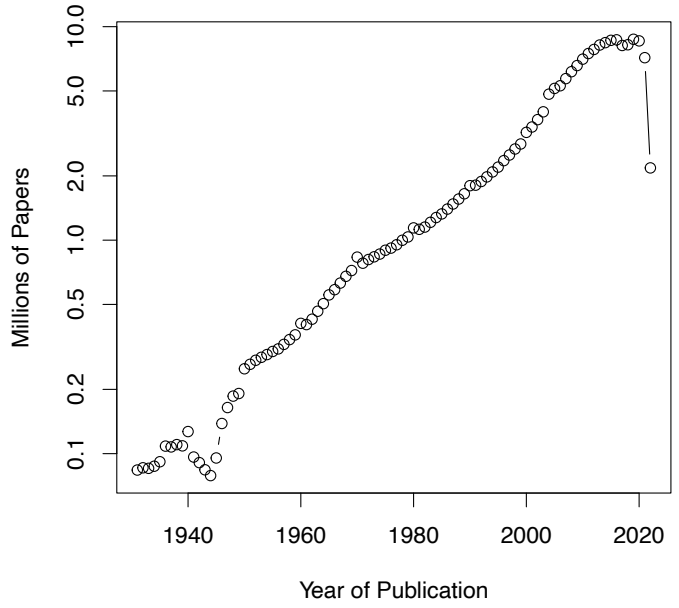
Surveys on Academic Search

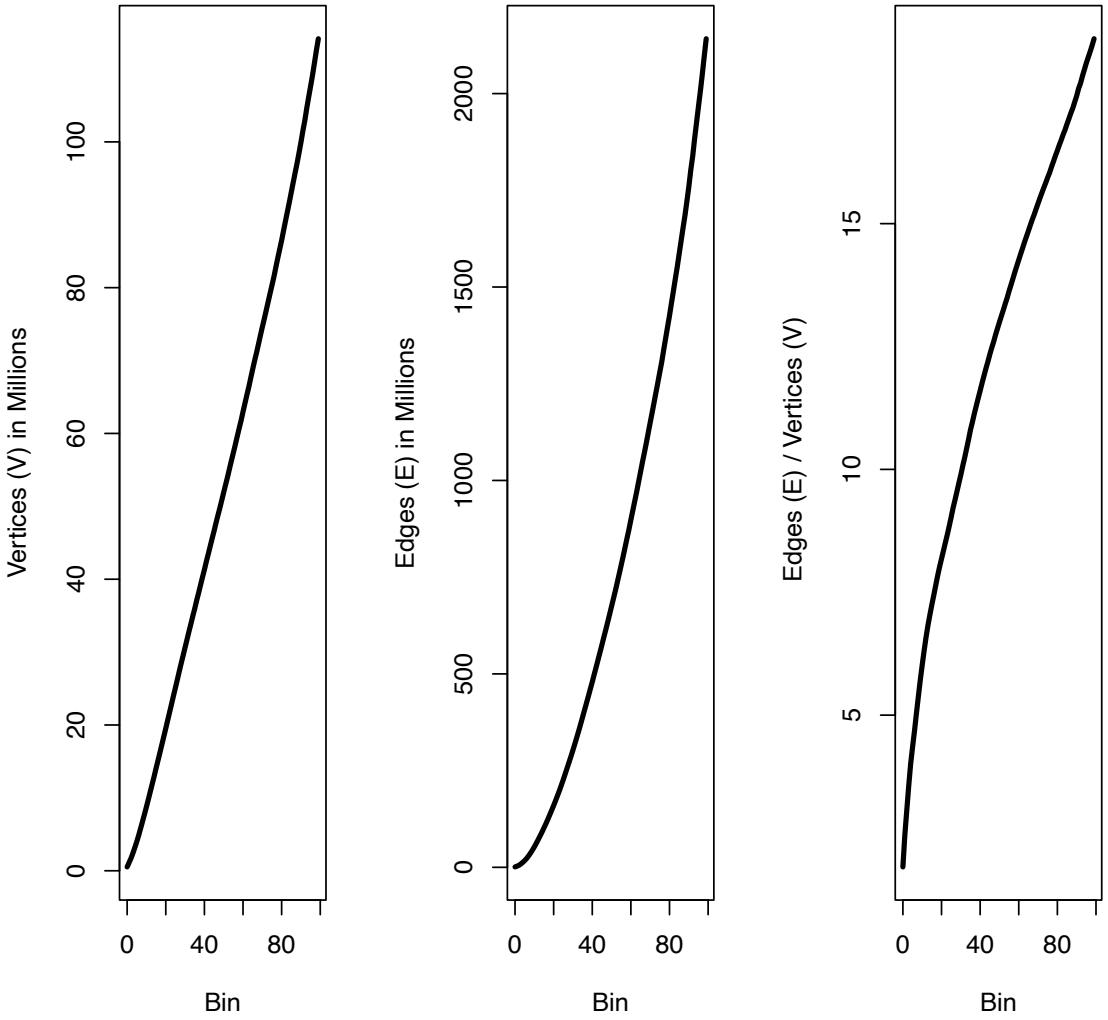
- Content-Based Filtering (CBF): Abstracts (Specter)
- Graph-Based Methods (GB): Citations (ProNE)
- Collaborative Filtering (CF): Clicks
- Better Together: Hybrids/Ensembles of above

- Why study academic search?
  - Academic search is like many important recommendation tasks
    - eCommerce (Product Recommendation), Traffic Analysis (Defense)
  - But data is less sensitive and available
- We will have little to say about CF
  - Because behavioral signals (clicks) are sensitive

Materials

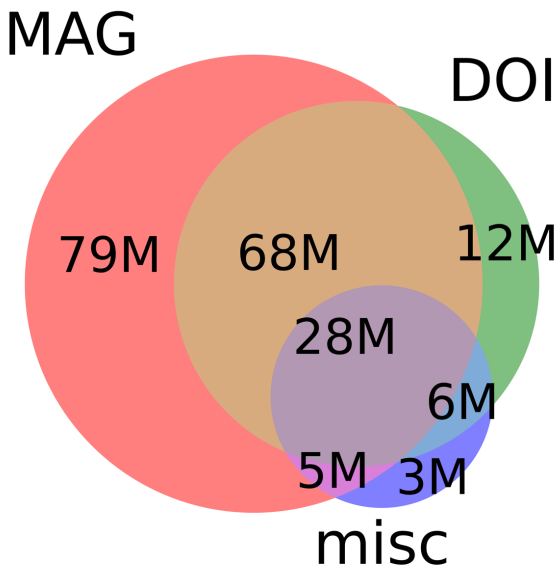
- Large and Growing
  - Semantic Scholar (S2)
    - 200+ Million Papers (nodes)
    - and 2+ Billion Citations (edges)
  - Literature doubles every 9 years!





More on Materials

- Seven Sources:
  - Two Big Sources: MAG, DOI
  - Five More (*misc*): PubMed, PubMedCentral, DBLP, arXiv, ACL
    - arXiv and ACL are tiny
- Many fields of study:
  - Medicine (45M), Chemistry (13M), CS (13M), Biology (13M), Materials Science (10M), Engineering (8M), Physics (7M), Psychology (7M), Mathematics (5M), Political Science (4M), Business (4M), Sociology (3M), Geography (3M), Economics (3M), Environmental Science (3M), Geology (3M), History (2M), Art (2M), Philosophy (1M)
  - Not just CS (Computer Science)



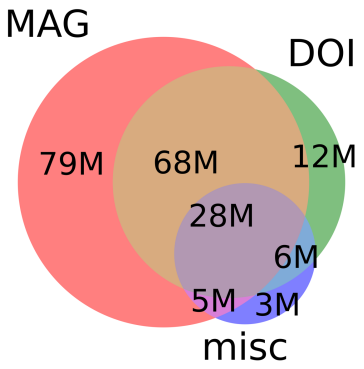
Semantic Scholar: Significant Effort

(source: Dan Weld)

50 person team; 7 year project

207M+ papers; 2B Citations

8M+ monthly active users

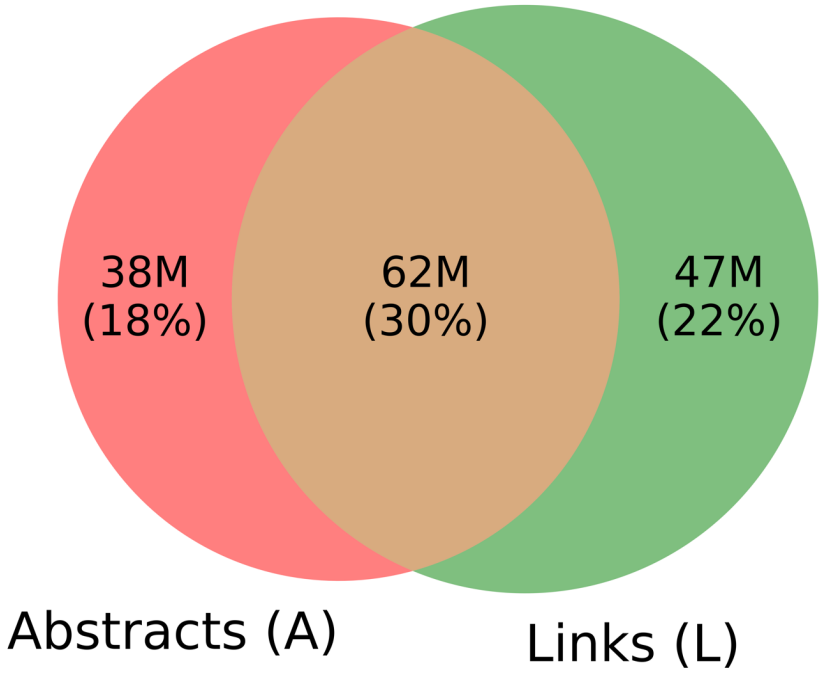


SCALE

	Source	Papers (millions)
1	CorpusId	207.80
2	MAG (Microsoft Academic Graph)	182.18
3	DOI	113.54
4	PubMed	35.03
5	DBLP	6.06
6	PubMedCentral	4.86
7	ArXiv	2.15
8	ACL	0.08

Why Better Together?

- Answer: Coverage
- GNNs assume papers have both
  - Abstracts (A), and
  - Links (L) in Citation Graph
- But not in Semantic Scholar
  - where $|A \cap L| \approx 30\%$
- Too many unrealistic benchmarks:
  - where $|A \cap L| \gg 30\%$



Motivating Scenario: Help Authors Write Sections on Related Work

- Four subtasks
  - **candidate list generation**: list papers to discuss,
  - **organization**: organize list by topic, time, etc.,
  - **summarization**: summarize papers, and
  - **connecting the dots**: explain how papers are relevant to the present discussion

Task: Find papers on ...	Query	CBF	GB
Recommender systems	[9]	[10–14]	[15–19]
Who should review what?	[20]	[21–26]	[27–31]
Citation Recommendation	[32]	[33–37]	[38–41]
RAG	[42]	[43–47]	[48–52]

Table 1: Complementary Recommendations.

Similarities and Differences: CBF & GB

Similarities of CBF & GB

- Embeddings
  - Both CBF and GB
    - represent papers as vectors
  - $sim(a, b) \approx \cos(vec(a), vec(b))$
  - Recommendation ≈ ANN
    - Approx nearest Neighbors
      - Query is a paper (or vector)
      - Output nearby papers in S2

Differences between CBF & GB

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Differences

- Inputs:
  - Titles and abstracts for CBF;
  - citations for GB.
- Interpretations:
  - For CBF, large cosines indicate similar abstracts
  - for GB, large cosines indicate similarity in terms of random walks on citation graph.
- History:
  - Deep networks evolved from NLP
  - whereas spectral clustering has roots in Linear Algebra
    - and was inspired by use cases such as traffic analysis in Applied Math.

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Differences

- Implementation Details:
  - We use Specter (a deep network) for CBF and ProNE (spectral clustering) for GB.
- Computational Bottlenecks:
  - Deep networks are limited by computational cycles,
    - whereas spectral clustering is limited by memory.
  - We use GPUs for deep networks,
    - and terabytes of RAM to compute SVDs for spectral clustering

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

More Differences

- Scale:
  - Larger graphs favor GB because of network effects.
- Time Invariance:
  - CBF embeddings are time invariant because abstracts do not change after publication;
  - GB embeddings improve as papers accumulate citations over time.
- Priors:
  - GB recommendations have more citations,
  - but are less recent
    - (because it takes time to accumulate citations).
- Corner Cases and Missing Values:
  - Multiple perspectives create opportunities to improve robustness and coverage with error detection and imputing missing values.

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Differences in More Detail

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Inputs and Perspectives

- Inputs
  - CBF is based on abstracts
  - and GB is based on citations
- Many of the differences
  - are consequences of inputs
- Perspectives
  - Abstracts:
    - authors' perspective
  - Citations:
    - responses from audience

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

History

- CBF and GB come from
  - different disciplines
  - with different motivations
- Disciplines
  - CBF: Deep Nets & LLMs from CS
  - GB: Eigenvectors (like Page Rank)
    - Linear Algebra
    - Spectral Clustering
    - Applied Math
- Motivations (Use Cases)
  - GB: Traffic Analysis
    - Know who is talking to who
    - But not what they are saying
  - CBF: NLP use cases
    - Situation is reversed
      - Know the content, but not the context of
      - how documents are connected to one another

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Interpretation

- Large cosines suggest papers are similar to one another
  - But for different reasons
- CBF: large cosines →
  - abstracts use similar words
  - in terms of LLMs
- GB: large cosines →
  - papers are near one another
  - in terms of random walks

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Implementation Details

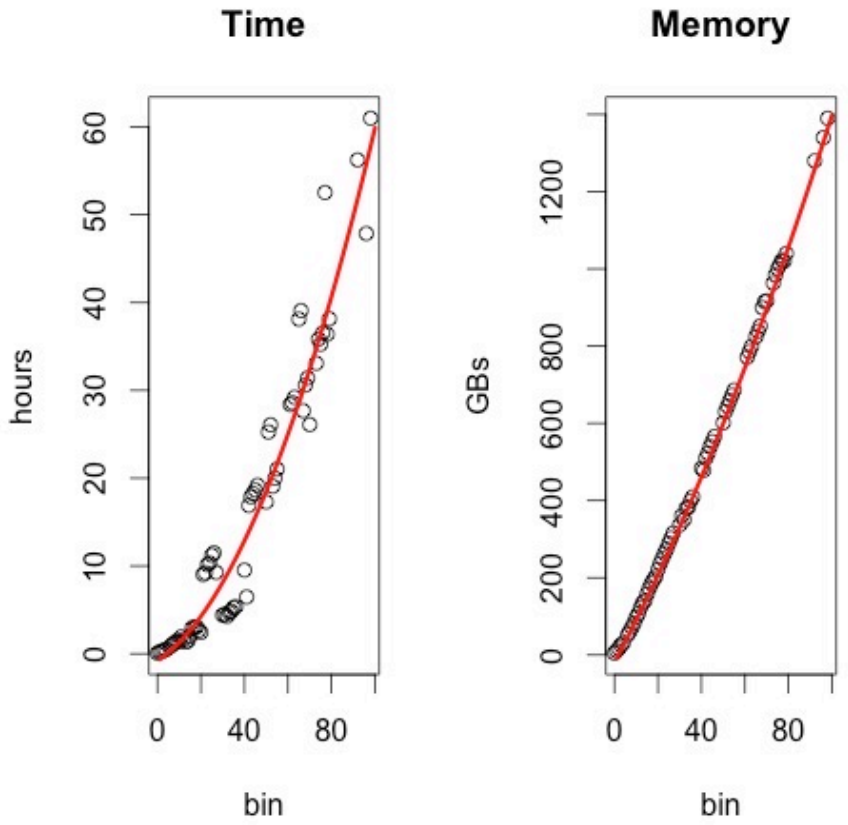
- Embeddings
  - CBF: Specter, a BERT-like deep net
  - GB: ProNE, spectral clustering
- For Specter,
  - no need for training or inference
    - because S2 distributes models and vectors
  - citations are not used for inference,
    - but are used for fine-tuning
- For ProNE, we had to compute them ourselves (heavy lifting)

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

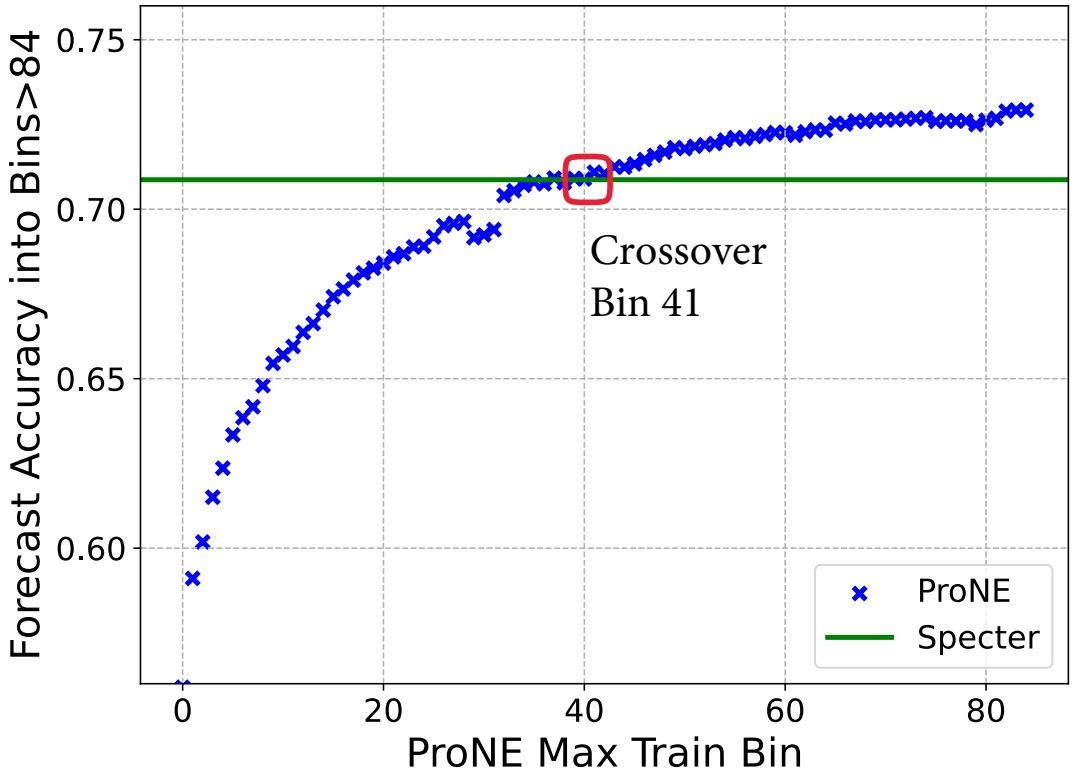
More Implementation Details

- ProNE heavy lifting
  - It takes a few days and a few TBs
  - to compute SVD for larger graphs
    - 100 bins $\approx 100\%$ of S2
- Specter fine-tuning
  - Start with SciBERT
  - fine-tune with a few million triples
    - $< query, pos, neg >$
- ProNE trains on billions
  - as opposed to millions



Scale

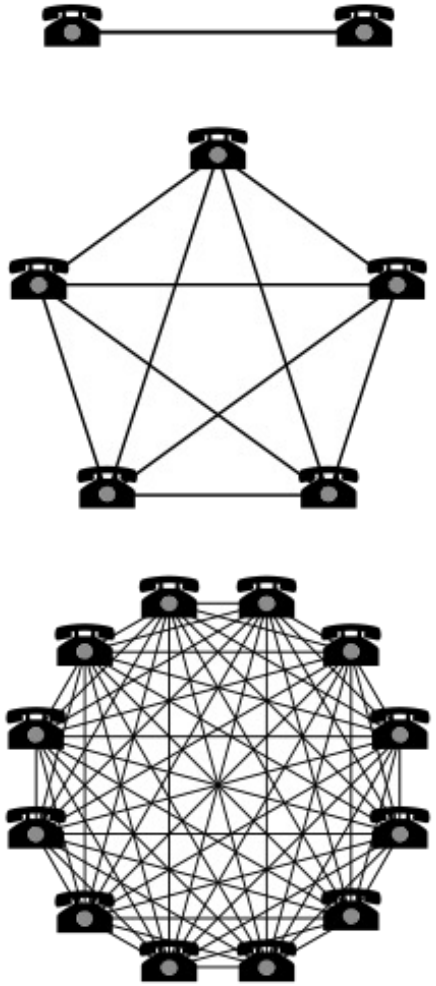
- Scale favors ProNE (GB)
  - because of network effects
- Network effects (Metcalfe's Law)
  - nodes: n
  - edges: n^2
  - paths: 2^n
- Citation Prediction Task
  - Does paper a cite paper b?
  - Given pairs: a, b that are 1-4 hops apart
  - Classification: is this pair 1 hop apart?
- Binning:
  - Train 100 ProNE models



	Vertices (Papers)	Edges (Links)
ogbl-citation2	3M	31M
crossover (bin 41)	42M	499M

Table 2: $|\text{OGB}| \ll$ crossover

Metcalfe's Law (Network Effects)



- History: 3Com was selling small networks
  - 3 = 1 printer + 2 computers
  - Metcalfe argued they should sell bigger networks
    - (and more 3Com products)
    - because of economies of scale
- Economy of Scale:
  - Benefits scale faster than costs
    - Benefits: $\sim n^2$
    - Costs: $\sim n$
- Law has been good for AT&T, Google, Social Media
- Hypo: also good for Academic Search
  - Consequently, we should experiment with large graphs

Time Invariance and Priors

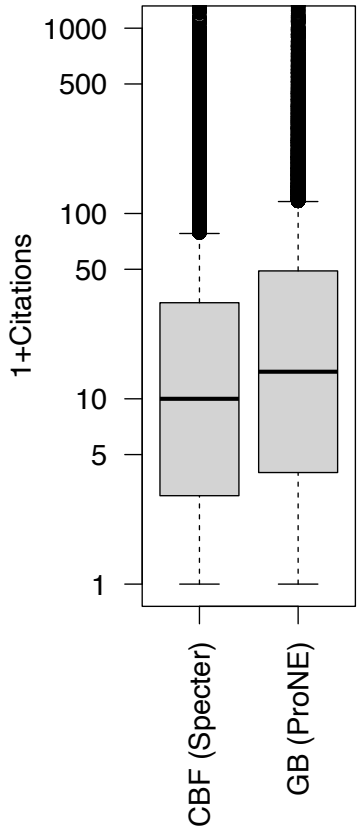
- Time Invariance
  - Authors cannot change abstracts after publication
    - But audience perspective evolves over time
    - Citations accumulate years after publication
- Priors
  - GB returns papers that have more citations but less recent
    - Because it can take time for papers to accumulate citations

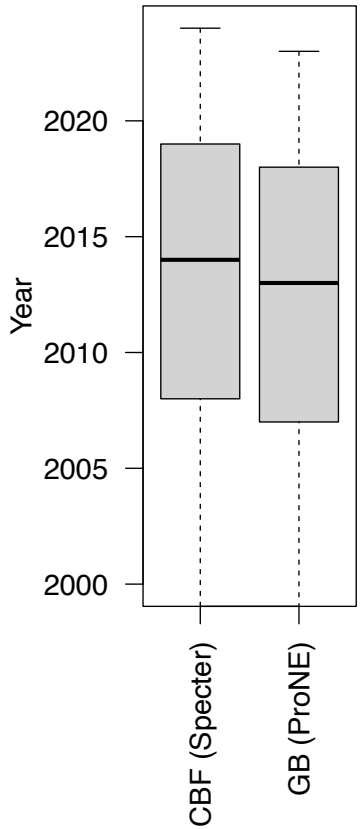
Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Time Invariance and Priors

- Time Invariance
  - Authors cannot change abstracts after publication
    - But audience perspective evolves over time
    - Citations accumulate years after publication
- Priors
  - GB returns papers that have more citations but less recent
    - Because it can take time for papers to accumulate citations





Corner Cases

- Multiple perspectives create opportunities for robustness
  - Detect corner cases by looking for large differences in cosines
    - Duplicate docs:
      - large CBF cosines,
      - but small GB cosines
  - Common corner cases for CBF
    - Duplicate documents
    - Missing/bogus abstracts
    - Non-English (Chinese)
  - GB does not suffer from these corner cases
    - But GB has other corner cases
    - Few (if any) links in citation graph

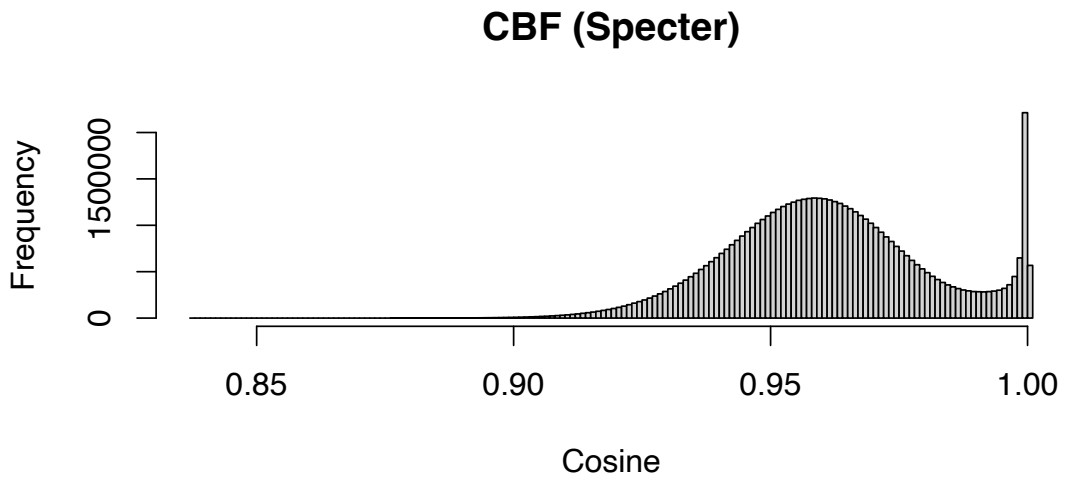
Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

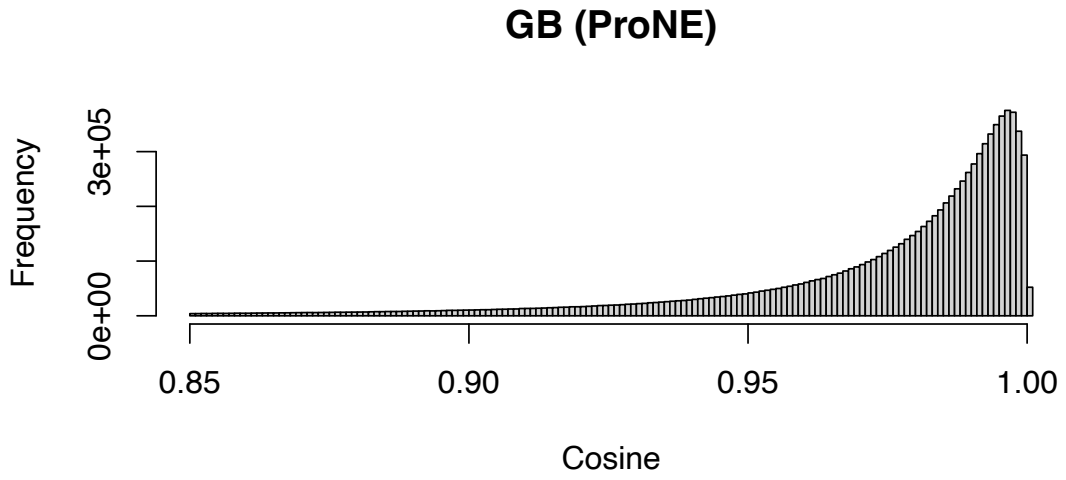
Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.

Corner Cases

- Multiple perspectives create opportunities for robustness
  - Detect corner cases by looking for large differences in cosines
    - Duplicate docs:
      - large CBF cosines,
      - but small GB cosines
  - Common corner cases for CBF
    - Duplicate documents
    - Missing/bogus abstracts
    - Non-English (Chinese)
  - GB does not suffer from these corner cases
    - But GB has other corner cases
    - Few (if any) links in citation graph

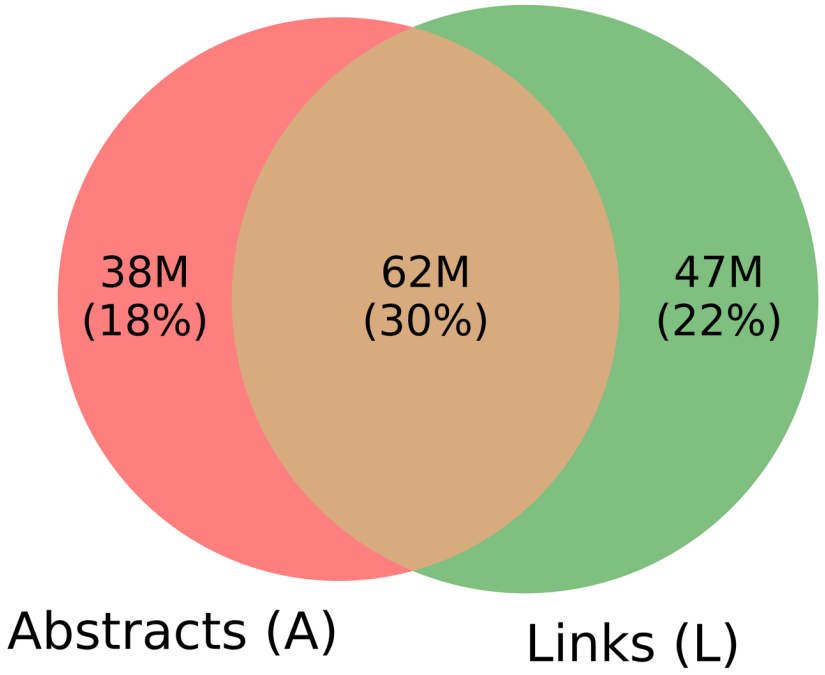
$$cos(q, cand_1)$$





Corner Cases: Imputing Missing Values

- Centroid Approximation
  - Infer a missing vector from
  - the average of its references
- Better Together Approximation
  - Infer a missing vector from
  - the average of papers
  - nearby in another embedding
- Synergies between CBF and GB



Conclusions: Better Together

- Soap Box:
  - Interdisciplinary Re-Org: ML/Vision/NLP/MT/IR/Speech
  - Responsible AI: Embrace Diversity
  - Multiple Perspectives:
    - Authors vs. Audience Response
    - Better Together
- Vision: More perspectives
  - subjectivity, diversity, semantics
- Academic Search Deliverables:
  - APIs, Website, Embeddings
- CBF & GB: Better Together
  - Complementary
  - Synergistic

Feature	CBF	GB
Inputs	Titles and abstracts	Citation graph
Perspective	Authors' position	Audience response
Technology	Deep Nets & LLMs	Spectral Clustering
Discipline	Computer Science	Linear Algebra (Math)
Motivation	Use cases in NLP	Traffic Analysis
Embedding	Specter [6]	ProNE [53]
Interpretation of large cosines	Similar abstracts	Nearby in terms of random walks
Bottleneck	Cycles	Memory
Hardware	GPUs	Terabytes of RAM
Scale	Favor smaller graphs	Favor larger graphs
Invariance	Abstracts are invariant after publication	Citations accumulate after publication
Priors	More recent	More impact (cites)
Corner Cases	Non-English abstracts	Few links in graph

Table 3: Feature table for comparing CBF and GB methods for academic paper recommendation.