Academic Recommender

*A set of supplementary materials**

*ALL EXAMPLES ARE TAKEN FROM GOOGLE SCHOLAR DATASET

Model Size Comparison

Model Name	Size
glove-wiki-gigaword-50	66 MB
glove-wiki-gigaword-100	128 MB
average_word_embeddings_glove.6B.300d	420 MB
fasttext-wiki-news-subwords-300	958 MB
word2vec-google-news-300	1.7 GB
all-MiniLM-L12-v2	120 MB
bert-base-uncased	420 MB

Edge Case: different words, different semantics

student: 'machine learning, artificial intelligence, deep learning’

prof1 = ‘statistical modeling, machine learning, neural networks, data science'

prof2 = ‘political philosophy, political theory, philosophy’

prof3 = ‘motor control, neuroscience, robotics’

	Method	Student-prof1	Student-prof2	Student-prof3
Token-based methods (handling single word only)	Avg word embedding (Euclidian dist) Model: Model: glove-wiki-gigaword-50	0.5275	0.5308	0.6307
	Max word embedding (Euclidian dist) Model: Model: glove-wiki-gigaword-50	0.7283	0.6252	0.8023
	Avg word embedding (cosine sim) Model: Model: glove-wiki-gigaword-50	0.4725	0.4692	0.3693
	WMD dist Model: Model: glove-wiki-gigaword-50	0.7236	0.9101	1.0

Edge Case: different words, different semantics

student: 'machine learning, artificial intelligence, deep learning’

prof1 = ‘statistical modeling, machine learning, neural networks, data science'

prof2 = ‘political philosophy, political theory, philosophy’

prof3 = ‘motor control, neuroscience, robotics’

	Method (cosine sim)	Student-prof1	Student-prof2	Student-prof3
Phrase-based methods (handling multiple words doc)	Mean-vector (one vector per student/prof) Model: glove-wiki-gigaword-50	0.9951	0.9682	0.9720
	Mean-vector embedding map (one vector per research interest) Model: glove-wiki-gigaword-50	0.6603	0.6146	0.4511
	Sentence transformers (one vector per student/prof) Model: all-MiniLM-L12-v2	0.6896	0.2151	0.3501
	Sentence transformers (one vector per research interest) Model: all-MiniLM-L12-v2	0.5924	0.1790	0.2994
	BERT (one vector per student/prof)	0.8978	0.6298	0.8190

Edge Case: different words, different semantics

We expect **the best matching professor to be prof1**, since research interests are very close and even one is the same (machine learning)

Among prof2 and prof3, while prof2 has some tokens (“theory”) which might be not-too-far from some tokens in student (“learning”, “intelligence”), but prof3 is semantically more relevant. While none of the words are very close, but “machine learning” and “artificial intelligence” can be used in the “motor control” and “robotics” domains. Hence, **prof3 should be closer than prof2**.

Among token-based methods, all are performing very bad. With minor neglection, word mover’s distance (WMD) could be the best option.

Among phrase-based methods, “one vector per student/prof” methods are generally better than “one vector per research interest” methods. Since student and prof1 should be very semantically close, both **sentence transformers (single vector) and BERT** can be good options.

Edge Case: Overqualified/Underqualified

student: 'machine learning, artificial intelligence, deep learning’

prof1 = 'machine learning, deep learning, artificial intelligence'

prof2 = 'machine learning, artificial intelligence, deep learning, psychology, developmental psychology'

prof3 = 'machine learning'

	Method	Student-prof1	Student-prof2	Student-prof3
Token-based methods (handling single word only)	Avg word embedding (Euclidian dist) Model: Model: glove-wiki-gigaword-50	0.4457	0.4882	0.42
	Max word embedding (Euclidian dist) Model: Model: glove-wiki-gigaword-50	0.6848	0.7598	0.5663
	Avg word embedding (cosine sim) Model: Model: glove-wiki-gigaword-50	0.5542	0.5118	0.5797
	WMD dist Model: Model: glove-wiki-gigaword-50	0.0	0.3391	0.5359

Edge Case: Overqualified/Underqualified

student: 'machine learning, artificial intelligence, deep learning’

prof1 = 'machine learning, deep learning, artificial intelligence'

prof2 = 'machine learning, artificial intelligence, deep learning, psychology, developmental psychology'

prof3 = 'machine learning'

	Method (cosine sim)	Student-prof1	Student-prof2	Student-prof3
Phrase-based methods (handling multiple words doc)	Mean-vector (one vector per student/prof) Model: glove-wiki-gigaword-50	0.9999	0.9912	0.9934
	Mean-vector embedding map (one vector per research interest) Model: glove-wiki-gigaword-50	0.8144	0.71	0.8112
	Sentence transformers (one vector per student/prof) Model: all-MiniLM-L12-v2	0.9938	0.7626	0.6106
	Sentence transformers (one vector per research interest) Model: all-MiniLM-L12-v2	0.7664	0.5687	0.8084
	BERT (one vector per student/prof)	0.9897	0.9208	0.8424

Edge Case: Overqualified/Underqualified

We expect **the best matching professor to be prof1**, since research interests are exactly the same as student.

Regarding prof2 and prof3, prof2 is considered an overqualified sample and prof3 is an underqualified one. While prof2 has only one research interest which is the same as student, prof3 has all research interests of student, plus another interests that are far from student's domain. Overall, we expect **prof2 to be much closer to prof1, compared to prof3**.

Is prof1 the best match or prof2? **In my opinion, prof1**. His research interests are the same as student, and not as broad as prof2.

Among token-based methods, word mover's distance (WMD) is the only one that works as expected. Others are mistakenly resulting in prof3 to be better that prof1 and prof2.

Among phrase-based methods, "one vector per student/prof" methods are better considering prof2 as a close (but not best) match: **sentence transformers (single vector) and BERT**

Edge Case: Close words, different semantics

student: 'machine learning, artificial intelligence, deep learning’

prof1 = ‘cloud engineering, deep learning, ML, computer vision'

prof2 = 'psychology, developmental psychology, cognitive learning, receptive learning'

	Method	Student-prof1	Student-prof2
Token-based methods (handling single word only)	Avg word embedding (Euclidian dist) Model: Model: glove-wiki-gigaword-50	0.5379	0.5337
	Max word embedding (Euclidian dist) Model: Model: glove-wiki-gigaword-50	0.9109	0.7493
	Avg word embedding (cosine sim) Model: Model: glove-wiki-gigaword-50	0.4621	0.4663
	WMD dist Model: Model: glove-wiki-gigaword-50	0.6835	0.7946

Edge Case: Close words, different semantics

student: 'machine learning, artificial intelligence, deep learning’

prof1 = ‘cloud engineering, deep learning, ML, computer vision'

prof2 = 'psychology, developmental psychology, cognitive learning, receptive learning'

	Method (cosine sim)	Student-prof1	Student-prof2
Phrase-based methods (handling multiple words doc)	Mean-vector (one vector per student/prof) Model: glove-wiki-gigaword-50	0.9923	0.9822
	Mean-vector embedding map (one vector per research interest) Model: glove-wiki-gigaword-50	0.5974	0.6173
	Sentence transformers (one vector per student/prof) Model: all-MiniLM-L12-v2	0.6784	0.4716
	Sentence transformers (one vector per research interest) Model: all-MiniLM-L12-v2	0.4591	0.3936
	BERT (one vector per student/prof)	0.3396	0.3157

Edge Case: Close words, different semantics

While **prof1** is obviously a much closer match to student, in comparison with prof2, prof2 has several terms (“learning”) that is the same as those in student. Prof1 is semantically close to student, but prof2 might make the model mistakenly consider it a better option, due to similar wording.

Among token-based methods, they all fall into the trap, except **word mover’s distance (WMD)** method. It can successfully plot the difference to some extent.

Among phrase-based methods, “one vector per student/prof” methods are better in distinguishing the huge difference. **Sentence transformers (single vector)** works better than BERT, even with smaller size.

Inference Time Comparison

Example:

student = 'machine learning, artificial intelligence, deep learning, data science'

prof1 = 'statistical modeling, machine learning, neural networks, data science'
prof2 = 'political philosophy, political theory, philosophy'
prof3 = 'motor control, neuroscience, robotics'
prof4 = 'ecology, evolution, population genetics, botany, conservation biology'
prof5 = 'Information Systems, Virtual Work, Virtual Teams, Knowledge Sharing, Project Management'
prof6 = 'Evolutionary Computing, Machine Learning, Bioinformatics, Genetics'
prof7 = 'Soft Matter, Inverse Design, Systems Physics'
prof8 = 'machine learning'

Inference Time Comparison

	Method	Avg ETA in msec (1000 runs)
Token-based methods (handling single word only)	Avg word embedding, distance Model: Model: glove-wiki-gigaword-100	5.25
	Max word embedding, distance Model: Model: glove-wiki-gigaword-100	4.50
	Avg word embedding, cosine sim Model: Model: glove-wiki-gigaword-100	(One-time mapping: 0.12) 3.63
	WMD distance Model: Model: glove-wiki-gigaword-100	4.26
Phrase-based methods (handling multiple words doc)	Mean-vector, single-vector Model: glove-wiki-gigaword-100	(One-time mapping: 3.19) 0.62
	Mean-vector embedding map, multi-vector Model: glove-wiki-gigaword-100	(One-time mapping: 0.75) 1.33
	Sentence transformers, single-vector Model: all-MiniLM-L12-v2	(One-time encoding: 329) 0.64
	Sentence transformers, multi-vector Model: all-MiniLM-L12-v2	(One-time encoding: 932) 103.33
	BERT (base), single-vector	(One-time encoding: 880) 0.30