cases	doc_1		doc_2		decision	id
	authors	• Aslanyan, Grigor • Wetherbee, Ian	**authors**	• Grigor Aslanyan • Ian Wetherbee	DUPLICATES	151
	title	Patents Phrase to Phrase Semantic Matching Dataset	**title**	Patents Phrase to Phrase Semantic Matching Dataset		
	publication_date	2022-08-01 00:00:00	**publication_date**	2022-08-01 00:00:00		
	source	SupportedSources.CORE	**source**	SupportedSources.INTERNET_ARCHIVE		
	journal		**journal**			
	volume		**volume**			
	doi	None	**doi**			
	urls	• http://arxiv.org/abs/2208.01171	**urls**	• https://web.archive.org/web/20220803121109/https://arxiv.org/pdf/2208.01171v1.pdf		
	id	id164172776692272301	**id**	id5076817944455998384		
	abstract	There are many general purpose benchmark datasets for Semantic Textual Similarity but none of them are focused on technical concepts found in patents and scientific publications. This work aims to fill this gap by presenting a new human rated contextual phrase to phrase matching dataset. The entire dataset contains close to $50,000$ rated phrase pairs, each with a CPC (Cooperative Patent Classification) class as a context. This paper describes the dataset and some baseline models.Comment: Presented at the SIGIR PatentSemTech 2022 Workshop. The dataset can be accessed at https://www.kaggle.com/datasets/google/google-patent-phrase-similarity-datase	**abstract**	There are many general purpose benchmark datasets for Semantic Textual Similarity but none of them are focused on technical concepts found in patents and scientific publications. This work aims to fill this gap by presenting a new human rated contextual phrase to phrase matching dataset. The entire dataset contains close to 50,000 rated phrase pairs, each with a CPC (Cooperative Patent Classification) class as a context. This paper describes the dataset and some baseline models.		
	versions		**versions**			