Table 1: Performances

	Baseline	Variant 1	Variant 2	Variant 3	Variant 4	Variant 5
# of clusters for each topic	10	given	1	3	5	7
Preprocessing	tokenization	tokenization				
Language model	sense2vec for words	sense2vec for words				
Compositional model	Vector mixture model (BOW (bag-of-words) representation with summarization for each snippet)	Vector mixture model (BOW (bag-of-words) representation with summarization for each snippet)				
Clustering	Kmeans	Kmeans				
Parameters	n_clusters=10	n_clusters=number of subtopics	n_clusters=1	n_clusters=3	n_clusters=5	n_clusters=7
Other						
Average F1	0,6852	0,6803	0,6610	0,6610	0,6656	0,6656
Average Rand Index	0,3732	0,3886	0,6928	0,4726	0,4156	0,3937
Average Adj. Rand Index	0,0284	0,0327	0,0000	0,0504	0,0303	0,0214
Average Jaccard Index	0,1431	0,1944	0,6928	0,3317	0,2347	0,1854
Average *number of created clusters*	10	7,5	1	3	5	7
Average *cluster sizes*	10,0000	17,3611	100,0000	33,3333	20,0000	14,2857

	Variant 6	Variant 7	Variant 8	Variant 9	Variant 10	Variant 11
# of clusters for each topic	20	7	7	7	7	7
Preprocessing	tokenization	tokenization + capitalization removal	tokenization + punctuation removal	tokenization +punctuation removal + en stopwords removal	tokenization + punctuation removal + stemming	tokenization + punctuation removal
Language model	sense2vec for words					word2vec with Brown Corpus
Compositional model	Vector mixture model (BOW (bag-of-words) representation with summarization for each snippet)					
Clustering	Kmeans					
Parameters	n_clusters=20	clusters=7				
Other						
Average F1	0,7001	0,6705	0,6803	0,6656	0,6730	0,6750
Average Rand Index	0,0398	0,3937	0,4036	0,3910	0,3937	0,4023
Average Adj. Rand Index	0,0195	0,0212	0,0384	0,0224	0,0322	0,0006
Average Jaccard Index	0,0891	0,1876	0,1941	0,1876	0,1877	0,2059
Average *number of created clusters*	20	7	7	7	7	7
Average *cluster sizes*	5,0000	14,2857	14,2857	14,2857	14,2857	14,2857

	Variant 12	Variant 13	Variant 14	Variant 15	Variant 16	Variant 17 ***"Chertoy"***
# of clusters for each topic	20	7	7	7	7	7
Preprocessing	tokenization + punctuation removal	tokenization + punctuation removal + POS	tokenization + punctuation removal + POS or without POS if no match with sense2vec	tokenization + punctuation removal + capitalisation removal + en stopwords removal + POS	tokenization + punctuation removal	
Language model	sense2vec for words					
Compositional model	Tuned vector mixture model (BOW (bag-of-words) representation with summarization for each snippet with alfa = 0.5 for non-topic words and alfa = 1.0 for topic words.)	vector mixture model (BOW (bag-of-words) representation with summarization for each snippet)				
Clustering	Kmeans				Affinity Propagation	MeanShift
Parameters	clusters=7				default param	
Other						
Average F1	0,6852	0,6779	0,6663	0,6754	0,6974	0,6832
Average Rand Index	0,3958	0,4383	0,4344	0,3842	0,3691	0,6823
Average Adj. Rand Index	0,0223	0,0617	0,0693	0,0138	0,0299	0,1538
Average Jaccard Index	0,1818	0,2320	0,2407	0,1789	0,1313	0,6249
Average *number of created clusters*	7	7	7	7	14,25	7
Average *cluster sizes*	14,2857	14,2857	14,2857	14,2857	7,0753	14,8214

	Variant 18	Variant 19	Variant 20	Variant 21	Variant 22	Variant 23
# of clusters for each topic	default param	7	7	given	given	given
Preprocessing	tokenization + punctuation removal					
Language model	sense2vec for words		sent2vec for sentences with sent2vec_toronto books_unigrams model	sense2vec for words		
Compositional model	vector mixture model (BOW (bag-of-words) representation with summarization for each snippet)		vector mixture model (summarization of sentences for each snippet)	vector mixture model (BOW (bag-of-words) representation with summarization for each snippet)		
Clustering	Spectral Clustering	Aglomerative Clustering	Kmeans	cosine similarity	cosine similarity with min=0.4	cosine similarity with min=0.3
Parameters	default param	n_clusters=7		n_clusters=given		
Other						
Average F1	0,6656	0,6710	0,6683	0,6610	0,6315	0,6614
Average Rand Index	0,6064	0,4192	0,4791	0,5333	0,4848	0,5309
Average Adj. Rand Index	-0,0391	0,0310	0,0727	-0,0052	-0,0006	-0,0086
Average Jaccard Index	0,5846	0,2170	0,3164	0,4891	0,4188	0,4861
Average *number of created clusters*	8	7	7	3,25	3	3,25
Average *cluster sizes*	12,5000	14,2857	14,2857	33,3333	31,3125	33,2083

	Variant 24	Variant 25	Variant 26	Variant 27	Variant 28	Variant 29
# of clusters for each topic	20	7	7	7	7	7
Preprocessing	tokenization + punctuation removal	tokenization				tokenization + punctuation removal
Language model	sent2vec for sentences with sent2vec_toronto books_unigrams model	sent2vec (sent2vec, plain-text_unigramm - trained Model (Wikipedia 2017))	sent2vec (sent2vec, plain-texts_bigramm - trained Model (Wikipedia 2017))	sent2vec (sent2vec, sent2vec_ wiki_bigrams)	sent2vec (sent2vec, sent2vec_ wiki_unigrams)	sent2vec (sent2vec, plain-text_unigramm - trained Model (Wikipedia 2017))
Compositional model	vector mixture model (summarization of sentences for each snippet)	vector mixture model (BOW (bag-of-words) representation with summarization for each snippet)				
Clustering	MeanShift	Kmeans				
Parameters	default param	n_clusters=given				n_clusters=7
Other		Variant 1 + sent2vec (sent2vec, plain-text_unigramm - trained Model (Wikipedia 2017))	Variant 1 + sentvec, plain-texts_bigramm - trained Model (Wikipedia 2017)	Variant 1 + sent2vec_ wiki_bigrams	Variant 1 + sent2vec_ wiki_unigrams	Variant 8 + sent2vec (sent2vec, plain-text_unigramm - trained Model (Wikipedia 2017))
Average F1	0,6610	0,6823	0,6750	0,6610	0,6750	0,6896
Average Rand Index	0,6129	0,4222	0,4216	0,3913	0,3886	0,4714
Average Adj. Rand Index	0.0775	0,0477	0,0400	0,0191	0,0211	0,1003
Average Jaccard Index	0,6004	0,2343	0,2496	0,2221	0,2083	0,2815
Average *number of created clusters*	7	7,5	7,5	7,5	7,5	7
Average *cluster sizes*	14,4345	17,3611	17,3611	17,3611	17,3611	14,2857

	Variant 30	Variant 31	Variant 32	Variant 33	Variant 34	Variant 35
# of clusters for each topic	7	7	7	given	given	given
Preprocessing	tokenization + punctuation removal					
Language model	sent2vec (sent2vec, plaintexts_bigramm - trained Model (Wikipedia 2017))	sent2vec (sent2vec, sent2vec_wiki_bigrams)	sent2vec (sent2vec, sent2vec_wiki_unigrams)	sent2vec (sent2vec, plaintext_unigramm - trained Model (Wikipedia 2017))	sent2vec (sent2vec, plaintexts_bigramm - trained Model (Wikipedia 2017))	sent2vec (sent2vec, sent2vec_wiki_bigrams)
Compositional model	vector mixture model (BOW (bag-of-words) representation with summarization for each snippet))					
Clustering	Kmeans			cosine similarity		
Parameters	n_clusters=7			n_clusters=given		
Other	Variant 8 + sentvec, plaintexts_bigramm - trained Model (Wikipedia 2017)	Variant 8 + sent2vec_wiki_bigrams	Variant 8 + sent2vec_wiki_unigrams	Variant 21 without POS + sim factor >0 + our sent2vec (sent2vec, plaintext_unigramm - trained Model (Wikipedia 2017))	Variant 21 without POS + sim factor >0, + sentvec, plaintexts_bigramm - trained Model (Wikipedia 2017)	Variant 21 without POS + sim factor >0,+ sent2vec_wiki_bigrams)
Average F1	0,6857	0,6703	0,6610	0,6610	0,6656	0,6679
Average Rand Index	0,4412	0,4102	0,4167	0,4531	0,5479	0,4071
Average Adj. Rand Index	0,0782	0,0057	0,0238	-0.0231	-0,0301	-0,0040
Average Jaccard Index	0,2598	0,2166	0,2458	0,3900	0,5312	0,2428
Average *number of created clusters*	7	7	7	3,75	2,25	6,75
Average *cluster sizes*	14,2857	14,2857	14,2857	30.0000	54,1667	18,3694

	Variant 36	Variant 37	Variant 38	Variant 39	Variant 40
# of clusters for each topic	given	default param	default param	default param	default param
Preprocessing	tokenization + punctuation removal				
Language model	sent2vec (sent2vec, sent2vec_wiki_unigrams)	sent2vec (sent2vec, plaintext_unigramm - trained Model (Wikipedia 2017))	sent2vec (sent2vec, plaintexts_bigramm - trained Model (Wikipedia 2017))	sent2vec (sent2vec, sent2vec_wiki_bigrams)	sent2vec (sent2vec, sent2vec_wiki_unigrams)
Compositional model	vector mixture model (BOW (bag-of-words) representation with summarization for each snippet))				
Clustering	cosine similarity	MeanShift			
Parameters	n_clusters=given	default param			
Other	Variant 21 without POS + sim factor ¿ 0, + sent2vec_wiki_unigrams	Variant 17 + our sent2vec (sent2vec, plaintext_unigramm - trained Model (Wikipedia 2017))	Variant 17 + sentvec, plaintexts_bigramm - trained Model (Wikipedia 2017)	Variant 17 + sent2vec_wiki_bigrams	Variant 17 + sent2vec_wiki_unigrams
Average F1	0,6533	0,6663	0,6739	0,6636	0,6636
Average Rand Index	0,4306	0,6586	0,6582	0,6135	0,6093
Average Adj. Rand Index	0,0390	0,0064	0,0287	-0,0329	-0,0473
Average Jaccard Index	0,2816	0,6496	0,6421	0,5963	0,5901
Average *number of created clusters*	7	5	6	7,5	7,5
Average *cluster sizes*	18,0106	21,9048	17,5595	13,5417	13,5417